Evaluating the contributions of individual variables to a quadratic form

Paul H. Garthwaite[a*], **Inge Koch**[b]

[a] *Department of Mathematics and Statistics, The Open University, MK7 6JA, UK*

[b] *School of Mathematical Sciences, University of Adelaide, SA 5005, Australia*

ABSTRACT

Quadratic forms capture multivariate information in a single number, making them useful, for example, in hypothesis testing. When a quadratic form is large and hence interesting, it might be informative to partition the quadratic form into contributions of individual variables. In this paper it is argued that meaningful partitions can be formed, though the precise partition that is determined will depend on the criteria used to select it. An intuitively reasonable criteria is proposed and the partition to which it leads is determined. The partition is based on a transformation that maximizes the sum of the correlations between individual variables and the variables to which they transform under a constraint. Properties of the partition are examined and it is shown that the partition is optimal under two other criteria. The contributions of individual variables to a quadratic form are less clear-cut when variables are collinear, and forming new variables through rotation can lead to greater transparency. The transformation is adapted so that it has an invariance property under rotation. Application of the partition to Hotelling's one- and two-sample T^2 statistics, Mahalanobis distance and discriminant analysis is described.

Keywords: Corr-max transformation; Collinearity; Discriminant analysis; Hotelling; Mahalanobis distance; Rotation.

* Corresponding address: Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes MK7 6JA, UK

E-mail address: paul.garthwaite@open.ac.uk (Paul Garthwaite). *Telephone:* +44 1908 655974

1. Introduction

Quadratic forms feature as a statistic in various multivariate contexts, such as Hotelling's T^2 and Mahalanobis distance. When the value of a quadratic form is large, then an obvious question is:

Which variables cause it to be large?

To illustrate, suppose $\mathbf{x}$ is an observation that should come from a distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. However, it appears to come from a different distribution because the Mahalonobis distance, equal to the quadratic form $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, is large. It might be helpful to have a measure of the contribution of individual variables to the size of this quadratic form.

When variables are correlated, it is not immediately apparent that a sensible answer to this question can be given. However, we shall argue that the question can be answered in a meaningful way and we will propose a method of partitioning a quadratic form into contributions from individual variables. This does not imply that there is a "best" way of forming such a partition, other than in some simple situations where arguments of symmetry can be used. However, although a partition of a quadratic form may be arbitrary to a degree, it can still be useful and informative. We show that the partition we propose meets optimality criteria.

Our method of forming a partition is based on a transformation that we call the corr-max transformation. The transformation is very closely related to the cos-max transformation introduced by Garthwaite *et al.* (2012). However, while the cos-max transformation was designed to transform a data matrix, the motivation for the corr-max transformation is to transform a random vector. The cos-max transformation adjusts a data matrix by a minimal amount while yielding a matrix with orthonormal columns; each of the original variables is associated with exactly one of these columns. The corr-max transformation yields a vector whose covariance matrix is proportional to the identity matrix, while each of the original variables is associated with exactly one component of the transformed vector. The strength of the associations is measured by correlations and the transformation is chosen to maximize the sum of these correlations (hence our name for the transformation).

Collinearities between variables will reduce the strength of some associations. The variables that are involved in a collinearity can be identified using the cos-max transformation (Garthwaite *et al.*, 2012). Then multiplication by an orthogonal matrix will rotate variables and may be used to remove collinearities. The corr-max transformation can be adapted so that rotation will only affect those components of the transformed vector that correspond to variables that are rotated. Those that correspond to variables that are not rotated will not be affected and their contributions to the quadratic form, as measured by the partition, will be unchanged. We refer to this as the rotation invariance property.

In Section 2 we argue that the value of a quadratic form can be meaningfully partitioned into separate contributions of individual variables and give the criteria that determine the corr-max transformation and our proposed partition. In Section 3 we obtain the transformation and the partition. In Section 4 the transformation is adapted to have the rotation invariance property and ways to exploit the property are suggested. In Section 5 we describe use of the partition in contexts where Hotelling's T^2 statistic or Mahalanobis distance arise, and in discriminant problems involving two groups. Concluding comments are given in Section 6.

2. Rationale and partition

Let Q be the quadratic form

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

where $X = (X_1, \ldots, X_m)^T$ is an $m \times 1$ random vector whose variance is proportional to $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ is a given $m \times 1$ vector that is not necessarily the mean of X. This type of quadratic form arises in various applications. For example, in Hotelling's one-sample T^2 statistic, X would take the value of a sample mean, $\mathbf{\Sigma}$ would be the population variance, and $\boldsymbol{\mu}$ would be the hypothesized population mean. The purpose of this paper is to give a method of evaluating the contributions of individual variables to Q. Before doing so, we must first consider whether it is possible, in principle, to meaningfully answer the question, *What are the contributions of individual variables to a quadratic form?*

Clearly a good answer can easily be given when $\mathbf{\Sigma}$ is the identity matrix: the contribution of each variable is then the square of the corresponding component of $\mathbf{x} - \boldsymbol{\mu}$. Extension to the case where $\mathbf{\Sigma}$ is diagonal is obvious. However, if $\mathbf{\Sigma}$ is not diagonal then it is less clear that Q can be partitioned between variables in a meaningful way. To examine this issue, we consider an example.

Specifically, let

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{pmatrix},$$

and, to aid explanation, suppose the three components of $\mathbf{x} = (x_1, x_2, x_3)^T$ correspond to standardised variables, *age* (x_1), *height* (x_2), and *weight* (x_3). Then the contribution of *age* (x_1) to Q is always clear, since

$$Q = (x_1, x_2, x_3) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_1^2 + (x_2, x_3) \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix}.$$

If $x_2 = x_3$, then *height* and *weight* contribute equally to Q, from symmetry. Hence, even though $\mathbf{\Sigma}^{-1}$ is not diagonal, the contributions of each variable to Q can be determined: *age* contributes x_1^2 while *height* and *weight* each contribute $\{Q - x_1^2\}/2$.

To expand this example, suppose x_3 were slightly greater in magnitude than x_2. Then the contribution of *age* to Q would still be x_1^2 while, in dividing the balance of Q between *height* and *weight*, it seems reasonable to give *weight* slightly the greater portion. Other situations are also readily constructed where common sense can indicate, approximately, the contributions of each variable to Q. In most situations though, there will be no partition of Q that is unquestionably better than any alternative. However, it may still be the case that sensible methods of partitioning Q broadly agree on the contributions made by individual variables. We construct a partition that helps interpret the results of some statistical analyses by giving a clearer relationship between the data variables and a test statistic or other quantity that is based on Q.

To form our partition, we consider transformations of the form $W = \mathbf{A}(X - \boldsymbol{\mu})$, such that W is an $m \times 1$ vector and

$$W^T W = (X - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (X - \boldsymbol{\mu}) \tag{2}$$

for any value of X. Then

$$Q = \sum_{i=1}^{m} W_i^2, \tag{3}$$

where $W = (W_1, \ldots, W_m)^T$, so W yields a partition of Q. The partition will be useful and meaningful if

(a) the components of W are uncorrelated, and

(b) it is reasonable to identify W_i with the ith x-variable, as the contribution of that x-variable to Q can then sensibly be defined as W_i^2.

As equation (2) holds for any X, it follows that $\mathbf{A}^T\mathbf{A} = \boldsymbol{\Sigma}^{-1}$. Also, $\text{var}(X) \propto \boldsymbol{\Sigma}$, so we have that $\text{var}(W) \propto \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{I}$, and (a) follows.

For (b), we first note that a linear transformation does not change the nature of a variable. To give a familiar example, standardizing a variable does not change the quantity to which it relates. As another example, a thermometer may make measurements on the centigrade or Fahrenheit scales and the measurements, which are linear transformations of each other, both equate to temperature. It is always possible to apply a linear transformation to W_i such that the transformed variable has the same mean and the same variance as X_i. Then the degree to which the rescaled W_i equates to X_i would primarily be determined by its correlation with X_i. (Perfect correlation would imply that they were identical.) Thus, as linear transformations do not affect correlation or change the nature of a variable, the degree to which W_i equates to the ith x-variable is largely determined by $\mathrm{corr}(X_i, W_i)$, where $\mathrm{corr}(.\,,\,.)$ denotes correlation.

In consequence, to try to meet (b) as fully as possible we will choose the transformation matrix $\mathbf{A}$ so that $\sum_{i=1}^{m} \mathrm{corr}(X_i, W_i)$ is maximized, subject to $\mathbf{A}^T\mathbf{A} = \mathbf{\Sigma}^{-1}$. This criterion for choosing $\mathbf{A}$ defines the corr-max transformation. In Section 6, the extent to which the corr-max transformation meets (b) is discussed further.

Before ending this section we summarize further notation used in several places. Capital letters without a subscript X, Y, $W^{\diamond}$, $\widehat{W}$, etc are $m \times 1$ random vectors. Subscripts are added to denote components of the vector: $X = (X_1, \ldots, X_m)^T$, $\widehat{W} = (\widehat{W}_1, \ldots, \widehat{W}_m)^T$, etc.

$\widehat{\mathbf{\Sigma}}$: generic estimate of $m \times m$ population variance matrix, $\mathbf{\Sigma}$.

$\widehat{\mathbf{\Sigma}}_1$: the standard unbiased estimate of $\mathbf{\Sigma}$ given by one sample.

$\widehat{\mathbf{\Sigma}}_p$: the standard pooled estimate of $\mathbf{\Sigma}$ based on independent samples from two populations that have variance $\mathbf{\Sigma}$.

$\mathbf{D}$: an $m \times m$ diagonal matrix with diagonal elements equal to the reciprocal of the square-root of the corresponding diagonal element of $\mathbf{\Sigma}$, if $\mathbf{\Sigma}$ is known, or $\widehat{\mathbf{\Sigma}}$, if $\mathbf{\Sigma}$ is unknown. Thus $\mathbf{D\Sigma D}$ ($\mathbf{\Sigma}$ known) or $\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D}$ ($\mathbf{\Sigma}$ unknown) has diagonal elements of 1.

$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$: the $n \times m$ data matrix whose rows are the n observations.

$\tilde{\mathbf{x}}_i$: the ith column of $\mathbf{X}$.

$\mathbf{A}$, $\widehat{\mathbf{A}}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{H}$, $\mathbf{\Omega}$, $\mathbf{\Psi}$: $m \times m$ matrices.

$\mathbf{\Gamma}$, $\mathbf{\Gamma}_d$: $m \times m$ and $d \times d$ orthogonal matrices, respectively.

3. The cos-max transformation

We first find the transformation matrix of the corr-max transformation. This is the matrix $\mathbf{A}$ that maximizes $\sum_{i=1}^{m} \text{corr}(X_i, W_i)$, subject to $\mathbf{A}^T\mathbf{A} = \mathbf{\Sigma}^{-1}$. Its formula is derived in Theorem 2. A preparatory result, useful in its own right, is given in Theorem 1. Proofs of theorems are given in the Appendix.

Theorem 1. *Suppose* $\mathbf{B}$ *is a square matrix and* $tr(\mathbf{B})$ *is to be maximized under the condition that* $\mathbf{B}^T\mathbf{B} = \mathbf{\Omega}^{-1}$, *where* $\mathbf{\Omega}$ *is a positive-definite matrix. Then* $\mathbf{B} = \mathbf{\Omega}^{-1/2}$, *the symmetric square-root of* $\mathbf{\Omega}$.

Theorem 2. *Let* $var(X) \propto \mathbf{\Sigma}$ *and let* $\mathbf{D\Sigma D}$ *have diagonal elements of 1 where* $\mathbf{D}$ *is diagonal. Suppose* $\sum_{i=1}^{m} corr(X_i, W_i)$ *is to be maximised, subject to* $W = \mathbf{A}(X - \boldsymbol{\mu})$, *where* $\mathbf{A}$ *is a square matrix,* $\mathbf{A}^T\mathbf{A} = \mathbf{\Sigma}^{-1}$ *and* $\boldsymbol{\mu}$ *is any given* $m \times 1$ *vector. Then* $\mathbf{A} = (\mathbf{D\Sigma D})^{-1/2}\mathbf{D}$ *and* $W = (\mathbf{D\Sigma D})^{-1/2}\mathbf{D}(X - \boldsymbol{\mu})$.

Theorem 2 completes the specification of our partition when $\mathbf{\Sigma}$ is known. To summarise, if $\text{var}(X) \propto \mathbf{\Sigma}$ and $X = \mathbf{x}$, the following method is used to partition the quadratic form $Q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})^T$ and evaluate the contribution of individual variables to $Q(\mathbf{x})$.

1. The $m \times m$ positive-definite diagonal matrix $\mathbf{D}$ is formed for which $\mathbf{D\Sigma D}$ has diagonal elements equal to 1.

2. The corr-max transformation is applied to $\mathbf{x} - \boldsymbol{\mu}$, giving $\mathbf{w} = (\mathbf{D\Sigma D})^{-1/2}\mathbf{D}(\mathbf{x} - \boldsymbol{\mu})$. Then $Q(\mathbf{x}) = \mathbf{w}^T\mathbf{w}$ and w_i^2 is defined to be the contribution of the ith x-variable to $Q(\mathbf{x})$ for $i = 1, \ldots, m$.

When $\mathbf{\Sigma}$ is unknown, we replace it in the above method with an estimate, $\widehat{\mathbf{\Sigma}}$ say. In some contexts, this type of substitution can have drawbacks, but here it seems appropriate, as it can yield properties similar to Theorem 2, but in terms of maximizing sample correlations rather than population correlations.

To illustrate, suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a simple random sample of n independent observations from the distribution of X. Define $\widehat{W} = \widehat{\mathbf{A}}(X - \boldsymbol{\mu})$ and let $\mathrm{corr}_s(X_i, \widehat{W}_i)$ denote the sample correlation between the ith component of X and the ith component of $\widehat{W}$. To obtain $\mathrm{corr}_s(X_j, \widehat{W}_j)$, put $\mathbf{w}_j = \widehat{\mathbf{A}}(\mathbf{x}_j - \boldsymbol{\mu})$, so that $\widehat{W} = \mathbf{w}_j$ when $X = \mathbf{x}_j$, and let w_{ji} and x_{ji} denote the i components of $\mathbf{w}_j$ and $\mathbf{x}_j$ $(j = 1, \ldots, n)$. Then $(x_{1i}, w_{1i}), \ldots, (x_{ni}, w_{ni})$ are a random sample from the joint distribution of $(X_i,\ \widehat{W}_i)$ and standard formula give $\mathrm{corr}_s(X_i, \widehat{W}_i)$. Theorem 3 gives a result about maximizing $\sum_{i=1}^{m} \mathrm{corr}_s(X_i, \widehat{W}_i)$ that corresponds to the result in Theorem 2.

Theorem 3. *Let $\widehat{\boldsymbol{\Sigma}}_1$ be the unbiased estimate of var(X) that is given by a random sample, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, taken from the distribution of X. Suppose also that $\sum_{j=1}^{m} corr_s(X_j, \widehat{W}_j)$ is to be maximized, subject to $\widehat{W} = \widehat{\mathbf{A}}(X - \boldsymbol{\mu})$ and $\widehat{\mathbf{A}}^T\widehat{\mathbf{A}} = \widehat{\boldsymbol{\Sigma}}_1^{-1}$. Let $\mathbf{D}\widehat{\boldsymbol{\Sigma}}_1\mathbf{D}$ have diagonal elements of 1 where $\mathbf{D}$ is diagonal. Then $\widehat{\mathbf{A}} = (\mathbf{D}\widehat{\boldsymbol{\Sigma}}_1\mathbf{D})^{-1/2}\mathbf{D}$ and*

$$\widehat{W} = (\mathbf{D}\widehat{\boldsymbol{\Sigma}}_1\mathbf{D})^{-1/2}\mathbf{D}(X - \boldsymbol{\mu}). \tag{4}$$

Theorems 2 and 3 give two closely related optimality properties held by the corr-max transformation. To motivate another optimality property, suppose that, in the transformation $W = \mathbf{A}(X - \boldsymbol{\mu})$, $\mathbf{A}$ were a diagonal matrix. Then the transformation would rescale components of X to give W, so X and W would be very similar, differing only in the choice of the scales on which they were measured. However, we require $\mathbf{A}^T\mathbf{A} = \boldsymbol{\Sigma}^{-1}$, so setting $\mathbf{A}$ equal to a diagonal matrix is not an option. Instead, we might choose to maximize $\mathrm{tr}(\mathbf{A})$, with the aim of obtaining a transformation that largely rescales the components of X to obtain W. This transformation would be sensitive to the scales of the x-variables. As $\mathbf{D}(X - \boldsymbol{\mu})$ is scale and location invariant, a preferable criteria is to put $W = \mathbf{BD}(X - \boldsymbol{\mu})$ and maximize $\mathrm{tr}(\mathbf{B})$ under the condition $(\mathbf{BD})^T\mathbf{BD} = \boldsymbol{\Sigma}^{-1}$. From Theorem 1, this gives $\widehat{\mathbf{B}} = (\mathbf{D}\widehat{\boldsymbol{\Sigma}}\mathbf{D})^{-1/2}$, so that the corr-max transformation is again optimal.

While the corr-max transformation yields a sensible method of partitioning Q into contributions of individual variables, other reasonable methods may well give a slightly different partition, but differences should be small when there is a close relationship between each W_i variable and the x-variable with which it is paired. Information about the strength of these relationships is provided by the correlations between X_i and W_i ($i = 1, \ldots, m$). The following theorem gives a simple means of finding their values and, more generally, the correlations $\mathrm{corr}_s(X_i, \widehat{W}_j)$ for $i = 1, \ldots, m$; $j = 1, \ldots, m$. It has the interesting implication that $\mathrm{corr}_s(X_i, \widehat{W}_j) = \mathrm{corr}_s(X_j, \widehat{W}_i)$ for all i and j.

Theorem 4. *Suppose that* $\widehat{W} = (\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2}\mathbf{D}(X - \boldsymbol{\mu})$*, where each diagonal element of* $\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D}$ *equals 1 and* $var_s(X) = k\widehat{\mathbf{\Sigma}}$*, where* $var_s(\,.\,)$ *denotes sample variance. Then* $corr_s(X_i, \widehat{W}_j)$ *equals the* (i, j) *element of* $(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{1/2}$.

4. Rotation invariance property

When the correlations between X_i and $\widehat{W}_i$ are weak for some values of i, there will generally be strong collinearities between some of the x-variables. The standard diagnostic for detecting collinearities are variance inflation factors. Suppose the values of $X_1, \ldots, X_m$ are observed on each of n items ($n > m$) and let R_j^2 denote the multiple correlation coefficient when X_j is regressed on the other X variables. Then the variance inflation factor for X_j, VIF_j say, is defined to be $(1 - R_j^2)^{-1}$. This will be large if X_j is involved in a collinearity. Garthwaite *et al.* (2012) show that the x-variables involved in a collinearity can be identified using the cos-max transformation, or another transformation that they also propose, the cos-square transformation. An example given in Section 5.2 uses the cos-max transformation to identify collinearities.

Collinearities can be removed by orthogonal rotation of those variables that are involved in the collinearities. This clarifies the relationship between x-variables and a quadratic form, as will be illustrated in Section 5. The results of a rotation are sensitive to the scales of the x-variables, so before rotation we scale the x-variables. This is the same as in principal component analysis, where variables are frequently scaled to have identical variances before applying the principal component transformation (which is an orthogonal rotation).

Here, $\text{var}(X) \propto \mathbf{\Sigma}$ and $\mathbf{D\Sigma D}$ has diagonal elements of 1, so the components of $\mathbf{D}X$ have identical variances. Let $\mathbf{\Gamma}$ be an $m \times m$ orthogonal matrix and put $Y = \mathbf{\Gamma D}(X - \boldsymbol{\mu})$, so that Y is obtained by scaling $X - \boldsymbol{\mu}$, followed by an orthogonal rotation. Suppose that we want to transform Y to $W^\diamond = \mathbf{C}Y$, with large correlations between Y_i and $W_i^\diamond$ for $i = 1, \ldots, m$. The components of Y are not all equally important – after rotation some components will have a smaller variance than others and those with smaller variances are less important, as in principal components analysis. The corr-max transformation would choose $\mathbf{C}$ to maximize $\sum_{i=1}^m \text{corr}(Y_i, W_i^\diamond)$ but now, to reflect the differing importance of some variables, we choose $\mathbf{C}$ to maximize $\sum_{i=1}^m \{\text{var}(Y_i)\}^{1/2}\text{corr}(Y_i, W_i^\diamond)$. This gives greater weight to the Y_i with greater variance. As $\text{var}(Y) \propto \mathbf{\Gamma D \Sigma D \Gamma}^T$, the constraint we impose on $\mathbf{C}$ is that $\mathbf{C}^T\mathbf{C} = (\mathbf{\Gamma D \Sigma D \Gamma}^T)^{-1}$. Theorem 5 gives $\mathbf{C}$.

Theorem 5. *Let $var(Y) \propto \mathbf{\Gamma D \Sigma D \Gamma}^T$ and suppose that $\sum_{i=1}^m \{var(Y_i)\}^{1/2} corr(Y_i, W_i^\diamond)$ is to be maximized, subject to $W^\diamond = \mathbf{C}Y$, where $\mathbf{C}^T\mathbf{C} = (\mathbf{\Gamma D \Sigma D \Gamma}^T)^{-1}$. Then $\mathbf{C} = (\mathbf{\Gamma D \Sigma D \Gamma}^T)^{-1/2} = \mathbf{\Gamma}(\mathbf{D \Sigma D})^{-1/2}\mathbf{\Gamma}^T$.*

Analogous to Theorem 4 (and obtained through a similar proof), $[\{\text{var}(Y_i)\}^{1/2}\text{corr}(Y_i, W_j^\diamond]$ is equal to the $(i,\, j)$ element of $(\mathbf{\Gamma D \Sigma D \Gamma}^T)^{1/2}$.

As $W^\diamond = \mathbf{\Gamma}(\mathbf{D \Sigma D})^{-1/2}\mathbf{\Gamma}^T Y$, $Y = \mathbf{\Gamma D}(X - \boldsymbol{\mu})$ and $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}$, it is immediate that $W^\diamond = \mathbf{\Gamma}(\mathbf{D \Sigma D})^{-1/2}\mathbf{D}(X - \boldsymbol{\mu})$. If $\mathbf{\Sigma}$ is unknown, we replace it with an estimate, $\widehat{\mathbf{\Sigma}}$, and put

$$\widehat{W}^\diamond = \mathbf{\Gamma}(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2}\mathbf{D}(X - \boldsymbol{\mu}). \tag{5}$$

The transformation from $X - \boldsymbol{\mu}$ to $W^\diamond$ or $\widehat{W}^\diamond$ will be referred to as the *adapted* corr-max transformation. It is identical to the ordinary corr-max transformation if there is no rotation, when $\mathbf{\Gamma} = \mathbf{I}$. In the partition associated with this transformation, the contribution of the ith variable to the quadratic form is evaluated as $(w_i^\diamond)^2$, where $w_i^\diamond$ is the value taken by the ith component of $W^\diamond$ or $\widehat{W}^\diamond$.

The adapted corr-max transformation arose from consideration of the transformation $W^\diamond = \mathbf{C}\{\mathbf{\Gamma}\mathbf{D}(X - \boldsymbol{\mu})\}$, where $\mathbf{C}$ is the transformation matrix and $\mathbf{\Gamma}\mathbf{D}(X - \boldsymbol{\mu})$ is a rotation of $\mathbf{D}(X - \boldsymbol{\mu})$ by $\mathbf{\Gamma}$. To highlight that $W^\diamond$ is a function of $\mathbf{\Gamma}$, write it as $W^\diamond(\mathbf{\Gamma})$. From equation (5),

$$W^\diamond(\mathbf{\Gamma}) = \mathbf{\Gamma} W^\diamond(\mathbf{I}).$$

Thus we obtain the same result whether (a) we rotate $\mathbf{D}(X - \boldsymbol{\mu})$ by $\mathbf{\Gamma}$ and transform the result, or (b) we transform $\mathbf{D}(X - \boldsymbol{\mu})$ and rotate the result by $\mathbf{\Gamma}$. That is, with the adapted corr-max transformation, the operations of rotation and transformation are commutative: they may be performed in either order with the same result.

The property allows us to rotate x-variables that are involved in a collinearity without affecting those components of the transformed vector that correspond to unrotated variables. To illustrate, suppose that we want to rotate the first d of the m variables. Then the rotation matrix $\mathbf{\Gamma}$ has the following block-diagonal form:

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-d} \end{bmatrix}, \tag{6}$$

where $\mathbf{\Gamma}_d$ is a $d \times d$ orthogonal matrix. Multiplying the data matrix $\mathbf{X}$ by $\mathbf{\Gamma}$ only rotates the first d columns of $\mathbf{X}$ and leaves its other columns unchanged. Moreover, under the transformation in equation (5), the last $m-d$ columns of $\widehat{W}^\diamond$ are unaffected by $\mathbf{\Gamma}_d$; the rotation only changes its first m columns. That is, under the adapted corr-max transformation, the rotation of some columns of $\mathbf{X}$ will not affect the transformed values of the unrotated columns. We refer to this as the *rotation invariance property.*

One consequence is that the contributions of un-rotated variables to the quadratic form, as measured by the partition, will be unchanged.

An aim in forming a partition is to obtain orthogonal components that are closely related on a one-to-one basis with meaningful quantities. When these quantities cannot be the original x variables because of a collinearity, the rotation invariance property suggests that we rotate the variables involved in the collinearity so as to reduce the correlations between them, and then apply the transformation. Then there should still be close pairwise relationships between each unrotated variable and the variable to which it transforms, as these relationships are not compromised by the rotation. Also, there should now be close relationships between the quantities obtained through rotation and the variables to which they transform.

The rotation should be chosen so as to yield meaningful quantities. If, say, the only collinearity was between the first two variables, it might be sensible to construct two new variables, one proportional to the sum of X_1 and X_2, and the other proportional to their difference. This will often create variables that have a natural interpretation; the new variables will also have a low correlation if the variance of X_1 is similar to the variance of X_2. The new variables could be created by setting $\mathbf{\Gamma}_m$ equal to the orthogonal matrix

$$\begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ -2^{-1/2} & 2^{-1/2} \end{bmatrix}.$$

If there is more than one distinct collinearity between the x variables, the $\mathbf{\Gamma}_m$ should itself have a block diagonal form, with a separate block for each collinearity. An example is given in Section 5.2.

5. Applications

In Section 5.1 we describe some common applications in which the corr-max transformation yields a partition that quantifies the contributions of individual variables to a test statistic. In Section 5.2 an example is given in which collinearities are removed through rotation while applying the transformation.

5.1 Hotelling T^2, Mahalanobis distance and discriminant analysis

The standard application in which the partition is useful is where a statistic of interest, Θ say, has the form

$$\Theta = \delta(X - \boldsymbol{\mu})^T \widehat{\boldsymbol{\Sigma}}^{-1}(X - \boldsymbol{\mu}), \tag{7}$$

with $\widehat{\boldsymbol{\Sigma}}$ an estimate of $\boldsymbol{\Sigma}$, $\text{var}(X) \propto \boldsymbol{\Sigma}$ and δ a positive scalar. From equation (4), the corr-max transformation yields $\widehat{W} = (\mathbf{D}\widehat{\boldsymbol{\Sigma}}\mathbf{D})^{-1/2}\mathbf{D}(X - \boldsymbol{\mu})$, and the contribution of the ith x-variable to Θ is evaluated as δw_i^2, where $(w_1, \ldots, w_m)^T$ is the value of $\widehat{W}$ given by data.

To apply the partition, X, $\widehat{\boldsymbol{\Sigma}}$, δ, and $\boldsymbol{\mu}$ must be identified and it must be checked that $\text{var}(X) \propto \boldsymbol{\Sigma}$. (The matrix $\mathbf{D}$ is obtained from $\widehat{\boldsymbol{\Sigma}}$.) The individual contributions, δw_i^2 for $i = 1, \ldots, m$, then follow automatically. After using the transformation, the correlations between components of $\widehat{W}$ and the corresponding components of X should be examined; rotation of components of X might be considered if some correlations are low.

In the following four applications, the first three have precisely the form given in (7), while the fourth is closely related to it.

(a) *Hotelling's one-sample T^2 statistic.* A random sample of size n is taken from $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, giving a sample mean $\overline{X}$ and sample covariance $\widehat{\boldsymbol{\Sigma}}_1$. The standard test of the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is based on Hotelling's one-sample T^2 statistic,

$$T_1^2 = n(\overline{X} - \boldsymbol{\mu}_0)^T \widehat{\boldsymbol{\Sigma}}_1^{-1}(\overline{X} - \boldsymbol{\mu}_0). \tag{8}$$

Let $X = \overline{X}$, so $\text{var}(X) = \boldsymbol{\Sigma}/n$. The partition is obtained by putting $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_1$, $\delta = n$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_0$.

(b) *Hotelling's two-sample T^2 statistic.* Two random samples of sizes n_1 and n_2 are drawn from the multivariate normal distributions, $N_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_m(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, that have the same covariance matrix. Then the hypothesis $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ is tested using Hotelling's two-sample T^2 statistic,

$$T_2^2 = \{n_1 n_2/(n_1 + n_2)\}(\overline{X}_1 - \overline{X}_2)^T \widehat{\boldsymbol{\Sigma}}_p^{-1}(\overline{X}_1 - \overline{X}_2), \tag{9}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the sample means and $\widehat{\boldsymbol{\Sigma}}_p$ is the pooled estimate of $\boldsymbol{\Sigma}$ derived from the two samples. Let $X = \overline{X}_1 - \overline{X}_2$, so $\text{var}(X) \propto \boldsymbol{\Sigma}$. Put $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_p$, $\delta = n_1 n_2/(n_1+n_2)$ and $\boldsymbol{\mu} = \mathbf{0}$ to obtain the contributions of individual variables to T_2^2.

(c) *Mahalanobis distance.* If $X_{(1)}$ and $X_{(2)}$ are two $m \times 1$ vectors, then the Mahalanobis distance between them is

$$(X_{(1)} - X_{(2)})^T \widehat{\boldsymbol{\Sigma}}_M^{-1}(X_{(1)} - X_{(2)}). \tag{10}$$

Here $X_{(1)}$ and $X_{(2)}$ must be independent, but either or both of them could be individual observations, or sample means, or one of them could be a vector of known constants. We suppose $\text{var}(X_{(i)}) = k_i \boldsymbol{\Sigma}$ $(i = 1, 2)$ where k_1 or k_2 (but not both) may equal 0. Let $X = X_{(1)} - X_{(2)}$, so $\text{var}(X) \propto \boldsymbol{\Sigma}$. Choose an unbiased estimate of $\boldsymbol{\Sigma}$ as $\widehat{\boldsymbol{\Sigma}}_M$, put $\delta = 1$ and $\boldsymbol{\mu} = \mathbf{0}$. Then the partitioning gives the contributions of individual variables to the Mahalanobis distance.

(d) *Fisher's linear discriminant function.* A quantity with similarities to a quadratic form arises when Fisher's linear discriminant function is used to discriminate between two classes. Suppose an observation needs to be classified as belonging to one of two classes that are characterised by the multivariate normal distributions $N_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_m(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, with sample means $\overline{X}_1$ and $\overline{X}_2$ and common estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_p$. A new observation X^* is classified as belonging to class 1 if

$$\mathfrak{r}(X^*) = \left\{ X^* - \frac{1}{2}(\overline{X}_1 + \overline{X}_2) \right\}^T \widehat{\mathbf{\Sigma}}_p^{-1}(\overline{X}_1 - \overline{X}_2) > 0. \tag{11}$$

Consider the transformations

$$\widehat{W}^\circ = (\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D})^{-1/2}\mathbf{D}(\overline{X}_1 - \overline{X}_2) \tag{12}$$

and

$$\widehat{W}^* = (\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D})^{-1/2}\mathbf{D}\{X^* - \frac{1}{2}(\overline{X}_1 + \overline{X}_2)\}. \tag{13}$$

As $\mathrm{var}(\overline{X}_1 - \overline{X}_2) \propto \mathbf{\Sigma}$ and $\mathrm{var}[X^* - \frac{1}{2}(\overline{X}_1 + \overline{X}_2)] \propto \mathbf{\Sigma}$, the ith components of both $\widehat{W}^\circ$ and $\widehat{W}^*$ can be equated to the ith x-variable. Let W_i° and W_i^* denote these components. As $\mathfrak{r}(X^*) = \sum_{i=1}^m W_i^\circ W_i^*$, the contribution of the ith x-variable to $\mathfrak{r}(X^*)$ is given by the data value of $W_i^\circ W_i^*$.

In Example 1 we explore how the transformation and partition work in practice, examining applications (a), (c) and (d) for illustration.

Example 1: Swiss bank notes. Flury and Riedwyl (1988) give data on 100 genuine Swiss 1000-franc bank notes. Six measurements were made on each note: length, height (measured on the left), height (measured on the right), distance of inner frame to the lower border, distance of inner frame to the upper border, and length of the diagonal. These measurements are the data values of $X = (X_1, \ldots, X_6)^T$. Their sample standard deviations are (0.39, 0.36, 0.36, 0.64, 0.65, 0.45) and the inverses of these standard deviations form the diagonal elements of $\mathbf{D}$. The sample covariance matrix of X is equated to $\widehat{\mathbf{\Sigma}}$ and its correlation matrix is:

$$\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D} = \begin{pmatrix} 1.00 & 0.41 & 0.42 & 0.23 & 0.06 & 0.03 \\ 0.41 & 1.00 & 0.66 & 0.24 & 0.21 & -0.26 \\ 0.42 & 0.66 & 1.00 & 0.25 & 0.13 & -0.15 \\ 0.23 & 0.24 & 0.25 & 1.00 & -0.63 & 0.00 \\ 0.06 & 0.21 & 0.13 & -0.63 & 1.00 & -0.26 \\ 0.03 & -0.26 & -0.15 & 0.00 & -0.26 & 1.00 \end{pmatrix}. \tag{14}$$

It can be seen that there are no high correlations. The mean vector for the banknote measurements, $\overline{\mathbf{x}}$, is (214.97, 129.94, 129.72, 8.31, 10.17, 141.52).

If the corr-max transformation is applied to a vector X to yield a vector $\widehat{W}$, the correlations between components of X and the corresponding components of $\widehat{W}$ are equal to the diagonal elements of $(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{1/2}$. These diagonal elements are 0.96, 0.90, 0.91, 0.91, 0.91 and 0.98. They are all large, indicating close one-to-one relationships between each x-variable and its corresponding component of $\widehat{W}$, so rotation of x-variables is unnecessary.

Hotelling's one-sample T^2 statistic might be used to test the hypothesis that the population mean vector is, say, $\boldsymbol{\mu}_0 = (215.01, 129.98, 129.76, 8.37, 10.23, 141.57)^T$. These values have been chosen so that, for each variable, the hypothesized population mean exceeds the sample mean by 0.1 standard deviations. The value of the test statistic, given by equation (8), is $T_1^2 = 8.71$. We have $\mathbf{D}$ and $\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D}$. Setting X and $\boldsymbol{\mu}$ equal to $\overline{\mathbf{x}}$ and $\boldsymbol{\mu}_0$ in equation (4) gives $\widehat{W} = -(0.054, 0.055, 0.052, 0.164, 0.181, 0.136)^T$. As $\delta = 100$, the contribution of the ith x-variable to T_1^2 is $100w_i^2$, so the contributions of the six x-variables are 0.54^2, 0.55^2, 0.52^2, 1.64^2, 1.81^2 and 1.36^2. These sum to 8.68, which differs slightly from T_1^2 because we have listed all contributions to 2 decimals only, and not given their precise values. The actual sum of the contributions equals T_1^2 as the theory tells us. Although for each component the sample mean differs from the hypothesized population mean by an equivalent amount, the last three x-variables make larger contributions to the T_1^2 statistic than the first three x-variables.

As an example involving Mahalanobis distance, suppose the measurements for an additional banknote that might be a forgery are $\mathbf{x}_2 = (215.8, 129.7, 129.0, 6.9, 8.6, 143.2)^T$. The Mahalonobis distance between $\mathbf{x}_2$ and the mean value of X in the sample of 100 genuine banknotes $\overline{\mathbf{x}}$, is given by equation (10) with $X_{(1)} = \overline{\mathbf{x}}$ and $X_{(2)} = \mathbf{x}_2$. Its value is 51.95, which suggests the note is a forgery. Our partition can be used to determine which characteristics of the new banknote distinguish it from the genuine banknotes. We put $X = \overline{\mathbf{x}} - \mathbf{x}_2$ and $\boldsymbol{\mu} = \mathbf{0}$ in equation (4), to obtain $\widehat{W}$. As $\delta = 1$, the contribution of the ith x-variable to the Mahalanobis distance is the square of the ith component of $\widehat{W}$. These squared values are (5.60 1.02 4.21 16.16 14.82 10.14). Hence the measurements that most distinguish the new banknote from genuine banknotes are the distance from the inner frame to the lower border (X_4), the distance from the inner frame to the upper border (X_5) and, to a lesser extent, the length of the diagonal (X_6).

The dataset of Swiss bank notes given by Flury and Riedwyl (1988) contained 100 faked bank notes in addition to the 100 genuine notes. As an example that involves Fisher's discriminant rule, we consider the task of using these data to classify a note as *genuine* or from the same popultation as the *fakes*. The pooled sample covariance matrix based on all 200 notes is

$$\widehat{\boldsymbol{\Sigma}}_p = \begin{pmatrix} 0.137 & 0.045 & 0.041 & -0.022 & 0.017 & 0.009 \\ 0.045 & 0.099 & 0.066 & 0.016 & 0.019 & -0.024 \\ 0.041 & 0.066 & 0.108 & 0.020 & 0.015 & 0.005 \\ -0.022 & 0.016 & 0.020 & 0.847 & -0.377 & 0.119 \\ 0.017 & 0.019 & 0.015 & -0.377 & 0.413 & -0.049 \\ 0.009 & -0.024 & 0.05 & 0.119 & -0.049 & 0.256 \end{pmatrix}, \tag{15}$$

and the sample mean of the genuine and faked bank notes are, respectively, $\overline{X}_1 = (214.97, 129.94, 129.72, 8.31, 10.17, 141.52)^T$ and $\overline{X}_2 = (214.82, 130.30, 130.19, 10.53, 11.13, 139.45)^T$. The note to be classified is $X^* = (214.4, 130.1, 130.3, 9.7, 11.7, 139.8)^T$. Equation (11) gives -20.34 as the value of $\mathfrak{r}(X^*)$, indicating that the new note should be classified as coming from the same population as the fakes.

Applying equations (12) and (13), $\widehat{W}^\circ = (0.38, -0.001, -1.48, -4.16, -2.80, 4.56)^T$ and $\widehat{W}^* = (-1.44, -0.64, 1.49, 1.21, 2.10, -1.46)^T$. The contributions of individual x-variables to $\mathfrak{r}(X^*)$ are evaluated as the diagonal elements of $\widehat{W}^\circ(\widehat{W}^*)^T$: -0.55, 0.001, -2.21, -5.04, -5.89 and -6.67. Unlike the previous two examples, some of these values are negative; negative values suggest the new note is from the same population as the faked notes. The last three variables $X_4, \ldots X_6$ underlie the outcome of the discrimination rule, as they make much larger contributions to $\mathfrak{r}(X^*)$ (in absolute value) than the first three variables.

5.2 Collinearities, rotation and quadratic forms

Some advantages of the (un-adapted) corr-max transformation are lost when strong collinearities are present: not every X variable will be closely related to the transformed variable with which it is paired. Here we examine a dataset in which collinearities are present and illustrate use of the cos-max transformation matrix to identify the variables that are collinear. To identify collinearities, the cos-max transformation is applied to data that have been standardized to have means of 0 and variances of 1, making the cos-max and corr-max very similar, as will be seen.

The dataset contains two strata whose means will be compared using Hotelling's two-sample T^2 statistic. We partition the test statistic into the contributions of individual variables/variable combinations by applying the adapted corr-max transformation. The rotation matrix ($\mathbf{\Gamma}$) we use in the transformation removes collinearites while creating meaningful variables from those variables that are involved in the collinearities.

Example 2: Female and male athletes. The data relate to the following nine measurements ($X_1, \ldots, X_9$) that were made on $n_1 = 100$ female and $n_2 = 102$ male athletes collected at the Australian Institute of Sport (Cook and Weisberg, 1994): *Wt* (weight), *Ht* (height), *Rcc* (red blood cell count), *Hg* (hemoglobin), *Hc* (hematocrit), *Wcc* (white blood cell count), *Ferr* (plasma ferratin concentration), *Bfat* (% body fat), and *SSF* (sum of skin folds). It is assumed that the two groups (females and males) may have different means, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, but have a common covariance matrix $\mathbf{\Sigma}$. Let $\widehat{\mathbf{\Sigma}}_p$ denote the pooled estimate of $\mathbf{\Sigma}$. The pooled correlation matrix, $\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D}$, takes the value

$$
\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D} = \begin{pmatrix}
1.00 & 0.68 & 0.05 & 0.10 & 0.06 & 0.15 & 0.06 & 0.63 & 0.65 \\
0.68 & 1.00 & -0.04 & -0.11 & -0.04 & 0.05 & -0.15 & 0.34 & 0.34 \\
0.05 & -0.04 & 1.00 & 0.78 & 0.86 & 0.14 & -0.05 & -0.04 & -0.05 \\
0.10 & -0.11 & 0.78 & 1.00 & 0.90 & 0.13 & 0.01 & -0.04 & -0.06 \\
0.06 & -0.04 & 0.86 & 0.90 & 1.00 & 0.15 & -0.06 & -0.08 & -0.11 \\
0.15 & 0.05 & 0.14 & 0.13 & 0.15 & 1.00 & 0.12 & 0.21 & 0.21 \\
0.06 & -0.15 & -0.05 & 0.01 & -0.06 & 0.12 & 1.00 & 0.16 & 0.16 \\
0.63 & 0.34 & -0.04 & -0.04 & -0.08 & 0.21 & 0.16 & 1.00 & 0.97 \\
0.65 & 0.34 & -0.05 & -0.06 & -0.11 & 0.21 & 0.16 & 0.97 & 1.00
\end{pmatrix}. \quad (16)
$$

Under the cos-max transformation, a data matrix $\mathbf{X}$ is transformed to $(\mathbf{X}^T\mathbf{X})^{-1/2}\mathbf{X}$. Let $\mathbf{X}_s$ denote the data matrix after variables have been centred and scaled so that the correlation matrix of $\mathbf{X}_s$ is $\mathbf{X}_s^T\mathbf{X}_s$. Put $(\mathbf{X}_s^T\mathbf{X}_s)^{-1/2} = \mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_m)^T$. Garthwaite *et al.* (2012) point out that the variance inflation factor for the jth variable (VIF$_j$) is then equal to $\mathbf{h}_j^T\mathbf{h}_j$. Moreover, if VIF$_j$ is large, indicating a collinearity, then large components of $\mathbf{h}_j$ correspond to the variables that underlie the collinearity. In the present example, $\mathbf{X}_s^T\mathbf{X}_s = \mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D}$, so examining the rows of $(\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D})^{-1/2}$ identifies variables involved in collinearities. (When $\mathbf{X}_s^T\mathbf{X}_s = \mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D}$, the corr-max and cos-max transformations are the same.)

We put $(\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D})^{-1/2} = (\mathbf{h}_1, \ldots, \mathbf{h}_m)^T$ and give their values in Table 1. Values above 0.75 are given in bold-face type. The last column of the table gives the VIF for each variable, e.g. 8.15 is the VIF for X_5 and equals $\mathbf{h}_5^T\mathbf{h}_5$. A VIF above 10 is often treated as a collinearity (Neter *et al.*, 1983, p. 392). On this basis, X_8 (Bfat) and X_9 (SSF) are involved in collinearites and, from the bold-face numbers in $\mathbf{h}_8$ and $\mathbf{h}_9$, there is a collinearity between them. There also appears to be a weak collinearity between X_4 (Hg) and X_5 (Hc).

If the corr-max transformation is applied to $X = (X_1, \ldots, X_9)^T$, then the following are the sample correlations between each x variable and the variable to which

Table 1

Rows of $(\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D})^{-1/2}$ and variance inflation factors for data on athletes

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	VIF
$\mathbf{h}_1^T$	**1.66**	-0.61	0.06	-0.23	0.00	-0.01	-0.06	-0.17	-0.41	3.38
$\mathbf{h}_2^T$	-0.61	**1.36**	-0.03	0.23	-0.08	0.00	0.14	-0.06	0.07	2.31
$\mathbf{h}_3^T$	0.06	-0.03	**1.75**	-0.30	-0.80	-0.03	0.02	0.05	-0.10	3.83
$\mathbf{h}_4^T$	-0.23	0.23	-0.30	**2.09**	**-1.12**	0.00	-0.04	0.04	0.00	5.82
$\mathbf{h}_5^T$	0.00	-0.08	-0.80	**-1.12**	**2.48**	-0.08	0.06	-0.09	0.22	8.15
$\mathbf{h}_6^T$	-0.01	0.00	-0.03	0.00	-0.08	**1.04**	-0.05	-0.07	-0.06	1.09
$\mathbf{h}_7^T$	-0.06	0.14	0.02	-0.04	0.06	-0.05	**1.04**	-0.07	-0.02	1.11
$\mathbf{h}_8^T$	-0.17	-0.06	0.05	0.04	-0.09	-0.07	-0.07	**3.27**	**-2.43**	16.64
$\mathbf{h}_9^T$	-0.41	0.07	-0.1	0.00	0.22	-0.06	-0.02	**-2.43**	**3.38**	17.59

it transforms:

Variable:	Wt	Ht	Rcc	Hg	Hc	Wcc	Ferr	Bfat	SSF
Correlation:	0.84	0.91	0.83	0.80	0.76	0.99	0.99	0.76	0.75

The correlations for Hg and Hc are a little low, suggesting that remedial action might be taken to offset both the mild collinearity between this pair of variables as well as the stronger collinearity between Bfat and SSF. To rotate both these variable pairs, the corr-max transformation was replaced by the adapted corr-max transformation of equation (5), with $\mathbf{\Gamma}$ set equal to the following block-diagonal orthogonal matrix,

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2^{-1/2} & 2^{-1/2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2^{-1/2} & -2^{-1/2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1/2} & 2^{-1/2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1/2} & -2^{-1/2} \end{pmatrix}. \tag{17}$$

We refer to the variables to which Bfat and SSF transform as B+S and B−S, and those from Hg and Hc as H+H and H−H. Rotation dramatically increased the correlations between the variables involved in rotations and their transformed values while leaving the corresponding correlations of all other variables unchanged. The following are the correlations between the rotated x variables and the variables to which they transform. These show a close one-to-one relationship between the two sets of variables.

Variable:	Wt	Ht	Rcc	H+H	H−H	Wcc	Ferr	B+S	B−S
Correlation:	0.84	0.91	0.83	0.91	0.96	0.99	0.99	0.95	0.99

Before rotation, the sample means for the female and male athletes ($\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$), and the pooled standard deviations (S.D.s) were as follows.

	Wt	Ht	Rcc	Hg	Hc	Wcc	Ferr	Bfat	SSF
Female:	67.34	174.59	4.405	13.560	40.48	6.994	57.0	17.85	87.0
Male:	82.52	185.51	5.027	15.553	45.65	7.221	96.4	9.25	51.4
S. D.s	11.69	8.07	0.336	0.929	2.60	1.801	43.3	4.45	27.3

The inverses of the standard deviations give the non-zero elements of the diagonal matrix $\mathbf{D}$. Hotelling's T^2 test was used to compare the means of the two groups. The T_2^2 statistic equalled 1199.1 giving, as you might expect, very clear evidence of differences between the two groups. The question is still relevant though, of which quantities contribute most to the T_2^2 value.

Putting $\widehat{\mathbf{w}}^{\diamond} = \mathbf{\Gamma}(\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D})^{-1/2}\mathbf{D}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ gives

$$\widehat{\mathbf{w}}^{\diamond} = (-1.75,\ -1.48,\ -1.08,\ -1.74,\ -0.49,\ -0.06,\ -1.28,\ 2.41,\ 2.57).$$

The partition evaluates the contributions of individual x-variables/variable combinations to T_2^2 as proportional to the squares of the components of $\widehat{\mathbf{w}}^{\diamond}$:

$$3.08,\ 2.19,\ 1.17,\ 3.02,\ 0.24,\ 0.00,\ 1.64,\ 5.81,\ 6.61.$$

(When multiplied by δ, which here equals $100(102)/(100 + 102)$, these sum to the value of the T_2^2 statistic, 1199.1, apart from rounding error.) On the scale given by our partition, the largest contributors to the size of T^2 are the average of Bfat and SSF (contributing 24%) and the difference between these same two quantities (contributing 28%). With the other pair of variables that were rotated, Hg and Hc, their average makes a significant contribution (13%) while the contribution from their difference is only 1%.

6. Concluding comments

The corr-max transformation is straightforward to calculate. $\widehat{\mathbf{\Sigma}}$ and $\mathbf{D}$ are readily determined and a spectral decomposition gives, say, $\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D} = \mathbf{H}\mathbf{\Psi}\mathbf{H}^T$ where $\mathbf{\Psi}$ is a diagonal matrix of eigenvalues of $\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D}$ and $\mathbf{H}$ is an orthogonal matrix whose columns are eigenvectors. Then $(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2} = \mathbf{H}\mathbf{\Psi}^{-1/2}\mathbf{H}^T$. It has also been shown that, using the corr-max transformation, the value of a quadratic form can readily be partitioned into contributions of individual x-variables. This has been detailed for important applications that occur frequently in practice. We have only obtained *point* estimates of the contributions of variables to a quadratic form, but bootstrap methods would be a simple way of extending the approach to obtain interval estimates.

A crucial point, which was discussed in Section 2 but which we return to now, is that the partition provides a *meaningful* way of evaluating the contributions of individual x-variables. As in equation (4), let $\widehat{W} = (\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2}\mathbf{D}(X - \boldsymbol{\mu})$. When the correlations are high between each component of $\widehat{W}$ and the corresponding component of X, then the partition is clearly a sensible way of evaluating the contribution of each x-variable. It was also illustrated that rotation may be used to construct interpretable variables to dramatically increase correlations if they are low. However, while rotation can add clarity, there are also attractions in the simplicity of forming a partition that retains the original x-variables. When this is done, there is still a transparent relationship between the x-variables and the contributions allocated to them by the partition. We return to the data on athletes to show this.

Let $X^{\#}$ denote the difference between an athlete's measurements and the average for their gender. Put $X^* = \mathbf{D}X^{\#}$, so that the components of $X^* = (X_1^*, \ldots, X_9^*)$ are standardized values of each variable. Let $(\widehat{W}_1, \ldots, \widehat{W}_9)^T = \widehat{W} = (\mathbf{D}\widehat{\mathbf{\Sigma}}_p\mathbf{D})^{-1/2}X^*$. Then $\delta\widehat{W}_i^2$ is the contribution of the ith variable to the quadratic form $\mathbf{\Theta}$ in equation (7). We focus on the two most highly correlated variables, Bfat (X_8) and SSF (X_9). The partition uses the following equations (obtained from Table 1) to determine their contribution to $\mathbf{\Theta}$.

$$
\begin{aligned}
\widehat{W}_8^2 &= (-0.17X_1^* - 0.06X_2^* + 0.05X_3^* + 0.04X_4^* - 0.09X_5^* - 0.07X_6^* \\
&\qquad - 0.07X_7^* + 3.27X_8^* - 2.43X_9^*)^2 \\
\widehat{W}_9^2 &= (-0.41X_1^* + 0.07X_2^* - 0.10X_3^* + 0.00X_4^* + 0.22X_5^* - 0.06X_6^* \\
&\qquad - 0.02X_7^* - 2.43X_8^* + 3.38X_9^*)^2
\end{aligned}
$$

These formulae show precisely how the contributions of individual variables to Θ are calculated. In particular, the formulae show that the difference between X_8^* and X_9^* has a substantial impact on the assessed contributions of Bfat and SSF. This arises from the high correlation between them (the correlation is 0.96), so that a large difference between their standardized differences is unexpected and so increases Θ. The role of the interaction between Bfat and SSF can be further clarified by writing $\widehat{W}_8^2$ and $\widehat{W}_9^2$ as,

$$\widehat{W}_8^2 = \{-0.17X_1^* - 0.06X_2^* + 0.05X_3^* + 0.04X_4^* - 0.09X_5^* - 0.07X_6^* - 0.07X_7^* + 0.84X_8^* + 2.43(X_8^* - X_9^*)\}^2 \quad (18)$$

$$\widehat{W}_9^2 = \{-0.41X_1^* + 0.07X_2^* - 0.10X_3^* + 0.00X_4^* + 0.22X_5^* - 0.06X_6^* - 0.02X_7^* + 2.43(X_9^* - X_8^*) + 0.95X_9^*\}^2 \quad (19)$$

Written in this way, $\widehat{W}_8^2$ and $\widehat{W}_9^2$ seem a very reasonable reflection of the respective contributions of Bfat and SSF to the quadratic form, as the large terms in (18) both involve X_8^* while those in (19) both involve X_9^*.

Experiments are often laborious to conduct and the scientists who conduct them would like to glean as much as possible from the data they gather. Not infrequently, a quadratic form is central to a multivariate statistical analysis and then the scientists might reasonably expect the quadratic form to yield more than just a p-value from a hypothesis test. The method developed in this paper provides a means of learning more about a quadratic form and hence should prove useful.

Acknowledgements

The authors thank the University of Adelaide and the Open University for facilitating visits and their hospitality. PHG benefitted from a travel award from the University of Adelaide. This work was supported by a project grant from the Medical Research Council.

Appendix. Proofs of theorems

Proof of Theorem 1. As $\mathbf{\Omega}$ is positive-definite, $\mathbf{B}$ is of full rank, so the singular value decomposition theorem gives $\mathbf{B} = \mathbf{F}\mathbf{\Lambda}^{1/2}\mathbf{G}^T$, where $\mathbf{\Lambda}$ is a diagonal matrix and $\mathbf{F}$ and $\mathbf{G}$ are orthogonal matrices. Then $\mathbf{B}^T\mathbf{B} = \mathbf{G}\mathbf{\Lambda}\mathbf{G}^T$, so $\mathbf{G}\mathbf{\Lambda}\mathbf{G}^T$ is the unique spectral decomposition of $\mathbf{\Omega}^{-1}$. Also, $\max\{\text{tr}(\mathbf{B})\} = \max\{\text{tr}(\mathbf{F}\mathbf{\Lambda}^{1/2}\mathbf{G}^T)\} = \max\{\text{tr}(\mathbf{\Lambda}^{1/2}\mathbf{G}^T\mathbf{F})\}$. Now $\mathbf{G}$ and $\mathbf{F}$ are orthogonal matrices, so the maximum value that the (i,i) element of $\mathbf{G}^T\mathbf{F}$ can take is 1, and it can only equal 1 if the ith columns of $\mathbf{G}$ and $\mathbf{F}$ are equal. Hence $\text{tr}(\mathbf{\Lambda}^{1/2}\mathbf{G}^T\mathbf{F})$ is maximized when $\mathbf{G} = \mathbf{F}$. Thus $\mathbf{B} = \mathbf{G}\mathbf{\Lambda}^{1/2}\mathbf{G}^T = \mathbf{\Omega}^{-1/2}$.

Lemma 1. *Suppose that $E(Y) = \mathbf{0}$ and var$(Y) \propto \mathbf{\Omega}$, and that $E(Y^TV)$ is to be maximized, where $V = \mathbf{B}Y$, $\mathbf{B}$ is a square matrix and $\mathbf{B}^T\mathbf{B} = \mathbf{\Omega}^{-1}$. Then $\mathbf{B} = \mathbf{\Omega}^{-1/2}$.*

Proof. Let $\text{var}(Y) = k\mathbf{\Omega}$ where k is a scalar. We have that $\text{E}(Y^TV) = \text{E}(Y^T\mathbf{B}Y) = \text{E}[\text{tr}(Y^T\mathbf{B}Y)] = \text{E}[\text{tr}(\mathbf{B}YY^T)] = \text{tr}[\mathbf{B}\text{E}(YY^T)] = k\text{tr}[\mathbf{B}\mathbf{\Omega}] = k\text{tr}[\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}] = k\text{tr}[(\mathbf{B}^T)^{-1}]$. Hence $\text{tr}[(\mathbf{B}^T)^{-1}]$ must be maximized under the constraint that $\mathbf{B}^T\mathbf{B} = \mathbf{\Omega}^{-1}$ or, equivalently, that $\mathbf{B}^{-1}(\mathbf{B}^T)^{-1} = \mathbf{\Omega}$. From Theorem 1, $(\mathbf{B}^T)^{-1} = \mathbf{\Omega}^{1/2}$.

Proof of Theorem 2. Let $W^* = \mathbf{A}\{X - \text{E}(X)\}$ and $\text{var}(X) = k\mathbf{\Sigma}$. Then $\text{E}(W^*) = 0$ and $\text{var}(W^*) = k\,\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T = k\,\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = k\,\mathbf{I}$. Let $Z = \mathbf{D}\{X - \text{E}(X)\}$ and put $Z = (Z_1, \ldots, Z_m)^T$, so that $\text{E}(Z) = 0$, $\text{var}(Z) = k\,\mathbf{D}\mathbf{\Sigma}\mathbf{D}$, and $\text{var}(Z_i) = k$ for $i = 1, \ldots, m$. We have that $\text{corr}(X_i, W_i^*) = \text{corr}(Z_i, W_i^*) = \text{E}(Z_iW_i^*/k)$. Also, $\text{corr}(X_i, W_i) = \text{corr}(X_i, W_i^*)$ as $W_i - W_i^* = \mathbf{A}\{\text{E}(X) - \boldsymbol{\mu}\}$. Thus, $\sum_{i=1}^m \text{corr}(X_i, W_i) = \text{E}(Z^TW^*/k)$. Also, the constraint $\mathbf{A}^T\mathbf{A} = \mathbf{\Sigma}^{-1}$ is equivalent to $(\mathbf{A}\mathbf{D}^{-1})^T(\mathbf{A}\mathbf{D}^{-1}) = \mathbf{D}^{-1}\mathbf{\Sigma}^{-1}\mathbf{D}^{-1}$. Hence, $\mathbf{A}$ must be chosen to maximize $\text{E}(Z^TW^*)$, where $W^* = \mathbf{A}\mathbf{D}^{-1}Z$ and $(\mathbf{A}\mathbf{D}^{-1})^T(\mathbf{A}\mathbf{D}^{-1}) = (\mathbf{D}\mathbf{\Sigma}\mathbf{D})^{-1}$. From Lemma 1, $\mathbf{A}\mathbf{D}^{-1} = (\mathbf{D}\mathbf{\Sigma}\mathbf{D})^{-1/2}$. Thus, $\mathbf{A} = (\mathbf{D}\mathbf{\Sigma}\mathbf{D})^{-1/2}\mathbf{D}$ and $W = (\mathbf{D}\mathbf{\Sigma}\mathbf{D})^{-1/2}\mathbf{D}(X - \boldsymbol{\mu})$.

Lemma 2. *Given an $n \times m$ matrix $\mathbf{Y} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m)$ of rank m ($m \leq n$), let $\mathbf{B}$ be the square matrix such that $\mathbf{Y}\mathbf{B} = (\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_m)$, $\sum_{i=1}^m \mathbf{y}_i^T\mathbf{u}_i$ is maximized, and $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_m$ are a set of orthonormal vectors. Then $\mathbf{B} = (\mathbf{Y}^T\mathbf{Y})^{-1/2}$.*

Lemma 2 is proved in Garthwaite *et al.* (2012). The transformation from $\mathbf{Y}$ to $\mathbf{Y}\mathbf{B}$

is the cos-max transformation.

Proof of Theorem 3. Put $\overline{\mathbf{x}} = \sum \mathbf{x}_j/n$ and $\overline{\mathbf{w}} = \sum \mathbf{w}_j/n$. Also, put $\mathbf{X}_c = (\mathbf{x}_1 - \overline{\mathbf{x}}, \ldots, \mathbf{x}_n - \overline{\mathbf{x}})^T$ and $\mathbf{W}_c = (\mathbf{w}_1 - \overline{\mathbf{w}}, \ldots, \mathbf{w}_n - \overline{\mathbf{w}})^T$. Thus $\mathbf{X}_c$ and $\mathbf{W}_c$ are centred data matrices; each of their columns has a mean of 0. Let $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{w}}_i$ denote the ith columns of $\mathbf{X}_c$ and $\mathbf{W}_c$, respectively $(i = 1, \ldots, m)$. Then, as $\mathrm{cov}(X_i, W_i) = \mathrm{cov}(X_i{+}a, W_i{+}b)$ for any constants a and b,

$$\sum_{i=1}^{m} \mathrm{corr}_s(X_i, W_i) = \sum_{i=1}^{m} \frac{\tilde{\mathbf{x}}_i^T \tilde{\mathbf{w}}_i}{\{(\tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_i)(\tilde{\mathbf{w}}_i^T \tilde{\mathbf{w}}_i)\}^{1/2}} = \sum_{i=1}^{m} \left\{ \left(\frac{\tilde{\mathbf{x}}_i}{(\tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_i)^{1/2}} \right)^T \left(\frac{\tilde{\mathbf{w}}_i}{(\tilde{\mathbf{w}}_i^T \tilde{\mathbf{w}}_i)^{1/2}} \right) \right\}.$$

For $i = 1, \ldots, m$, let $\tilde{\mathbf{y}}_i = \tilde{\mathbf{x}}_i/(\tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_i)^{1/2}$ and $\tilde{\mathbf{u}}_i = \tilde{\mathbf{w}}_i/(\tilde{\mathbf{w}}_i^T \tilde{\mathbf{w}}_i)^{1/2}$. Put $\mathbf{Y} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m)$ and $\mathbf{U} = (\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_m)$. Now $\widehat{\Sigma} = \mathbf{X}_c^T \mathbf{X}_c/(n-1)$, so the diagonal elements of $\mathbf{D}$ are $\{\tilde{\mathbf{x}}_j^T \tilde{\mathbf{x}}_j/(n-1)\}^{-1/2}$ and $\mathbf{Y} = \mathbf{X}_c \mathbf{D}/(n-1)^{1/2}$. Also, $\mathbf{W}_c = \mathbf{X}_c \widehat{\mathbf{A}}^T$, so $\mathbf{W}_c^T \mathbf{W}_c$ is diagonal as $\widehat{\mathbf{A}}^T \widehat{\mathbf{A}} = \widehat{\mathbf{\Sigma}}^{-1}$. Hence, $\mathbf{U}^T \mathbf{U}$ is also diagonal. Moreover, $\tilde{\mathbf{u}}_i^T \tilde{\mathbf{u}}_i = 1$ for $i = 1, \ldots, m$, so $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_m$ are a set of orthonormal vectors. From Lemma 2, under the constraint that $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_m$ are orthonormal vectors, $\sum_{i=1}^m {\tilde{\mathbf{y}}_i}^T \tilde{\mathbf{u}}_i$ is maximized when $\mathbf{U} = \mathbf{Y}(\mathbf{Y}^T \mathbf{Y})^{-1/2}$. Equivalently, under the constraint that $\mathbf{W}_c^T \mathbf{W}_c$ is diagonal, $\sum_{i=1}^m \mathrm{corr}_s(X_i, W_i)$ is maximized when $\mathbf{W}_c = \mathbf{E}\mathbf{X}_c \mathbf{D}(\mathbf{D}\mathbf{X}_c^T \mathbf{X}_c \mathbf{D})^{-1/2}$, where $\mathbf{E}$ is any positive-definite diagonal matrix. Under the more restrictive constraint that $\mathbf{W}_c = \mathbf{X}_c \widehat{\mathbf{A}}^T$ with $\widehat{\mathbf{A}}^T \widehat{\mathbf{A}} = \widehat{\mathbf{\Sigma}}^{-1}$, it follows that $\sum_{i=1}^m \mathrm{corr}_s(X_i, W_i)$ is maximized when $\widehat{\mathbf{A}}^T = \mathbf{D}(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2}$.

Proof of Theorem 4. Let $Z = \mathbf{D}\{X - \mathrm{E}(X)\}$ and put $Z = (Z_1, \ldots, Z_m)^T$. Then $\mathrm{corr}_s(X_i, \widehat{W}_j) = \mathrm{corr}_s(Z_i, \widehat{W}_j)$. Now $\mathrm{var}_s(Z) = k\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D}$ and $\mathrm{var}_s(\widehat{W}) = k(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2}\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D}(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2} = k\,\mathbf{I}$. Also, $\mathrm{E}[Z\{\widehat{W} - \mathrm{E}(\widehat{W})\}^T] = \mathrm{E}[Z\{(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2}Z\}^T] = \mathrm{E}(ZZ^T)(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2} = \mathrm{var}_s(Z)(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{-1/2} = k(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{1/2}$. Hence $\mathrm{cov}_s(Z_i, \widehat{W}_j)$ is the (i,j) element of $[k(\mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D})^{1/2}]$. The result follows as $\mathrm{var}_s(Z_i) = \mathrm{var}_s(\widehat{W}_j) = k$.

Proof of Theorem 5. Let $\mathbf{\Phi} = \mathbf{\Gamma}\mathbf{D}\mathbf{\Sigma}\mathbf{D}\mathbf{\Gamma}^T$ and let $V = Y - E(Y)$, so

$E(V) = 0$ and $\text{var}(V) \propto \mathbf{\Phi}$. Put $W^* = \mathbf{C}V$, so $\text{var}(W^*) = \mathbf{I}$ as $\mathbf{C}^T\mathbf{C} = \mathbf{\Phi}^{-1}$. Then $\sum_{i=1}^m [\{\text{var}(Y_i)\}^{1/2}\text{corr}(Y_i, W_i^\diamond)] = \sum_{i=1}^m [\{\text{var}(V_i)\}^{1/2}\text{corr}(V_i, W_i^*)] = \sum_{i=1}^m \text{cov}(V_i, W_i^*) = V^T W^*$. From Lemma 2, $V^T W^*$ is maximized when $\mathbf{C} = \mathbf{\Phi}^{-1/2} = (\mathbf{\Gamma D \Sigma D \Gamma}^T)^{-1/2}$. Also, $(\mathbf{\Gamma D \Sigma D \Gamma}^T)^{-1/2}$ is the unique symmetric square-root of $(\mathbf{\Gamma D \Sigma D \Gamma}^T)^{-1} = \mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1}\mathbf{\Gamma}^T$. As $[\mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1/2}\mathbf{\Gamma}^T].[\mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1/2}\mathbf{\Gamma}^T] = \mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1}\mathbf{\Gamma}^T$, it follows that $(\mathbf{\Gamma D \Sigma D \Gamma}^T)^{-1/2} = \mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1/2}\mathbf{\Gamma}^T$.

References

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. Cambridge University Press, Cambridge.

Garthwaite, P. H., Critchley, F., Anaya-Izquierdo, K., and Mubwandarikwa, E. (2012). Orthogonalization of vectors with minimal adjustment. *Biometrika*, **99**, 787–798.

Neter, J., Wasserman, W., and Kutner, M. H. (1983). *Applied Linear Regression Models*. Irwin, Illinois.