Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **58**(1), 2016, 99–119 doi: 10.1111/anzs.12144

EVALUATING THE CONTRIBUTIONS OF INDIVIDUAL VARIABLES TO A QUADRATIC FORM

PAUL H. GARTHWAITE[1,*] AND INGE KOCH[2]

The Open University and University of Adelaide

Summary

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 criterion used to select it. An intuitively reasonable criterion is proposed and the partition to which it leads is determined. The partition is based on a transformation that maximises the sum of the correlations between individual variables and the variables to which they transform under a constraint. Properties of the partition, including optimality properties, are examined. 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 such rotation, whereby the assessed contributions are unchanged for variables that the rotation does not affect directly. Application of the partition to Hotelling's one- and two-sample test statistics, Mahalanobis distance and discriminant analysis is described and illustrated through examples. It is shown that bootstrap confidence intervals for the contributions of individual variables to a partition are readily obtained.

Key words: Corr-max transformation; collinearity; discriminant analysis; Hotelling; Mahalanobis distance; rotation.

1. Introduction

Quadratic forms feature as statistics in various multivariate contexts. Well-known examples include Hotelling's T^2 statistic and the 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 $(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, is large. It might be helpful to have a measure of the contribution of individual variables to the size of this quadratic form.

*Author to whom correspondence should be addressed.
[1]Department of Mathematics and Statistics, The Open University, MK7 6AA, UK
e-mail: paul.garthwaite@open.ac.uk
[2]School of Mathematical Sciences, University of Adelaide, SA 5005, Australia.

Acknowledgements. The authors thank an associate editor and two referees for helpful comments that led to substantial improvements in this paper. We also thank the University of Adelaide and the Open University for facilitating visits and for 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.

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 certain optimality criteria.

Our method of forming a partition is based on a transformation that we call the corr-max transformation. Garthwaite, Critchley, Anaya-Izquierdo & Mubwandarikwa (2012) focussed on a transformation referred to as the cos-square transformation, but also proposed a second transformation called the cos-max transformation. The latter is closely related to the corr-max transformation that we introduce here. However, while the cos-max transformation was designed to transform a data matrix, the intended use of the corr-max transformation is the transformation of 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 maximise 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). The coordinate axes corresponding to these variables can then be rotated to yield a set of variables with little collinearity. We adapt the corr-max transformation so that contributions to the quadratic form, as measured by the partition, will only change for those variables that are affected by the rotation. We refer to this feature as the rotation invariance property.

The task of determining which variables have most influence on a Mahalanobis distance has attracted some attention in the literature. The Mahalanobis-Taguchi system, which features in statistical process control, estimates the covariance matrix $\mathbf{\Sigma}$ from 'normal' data and computes the Mahalanobis distances for a set of 'abnormal' data points, in order to determine signal-to-noise ratios for individual variables and hence identify variables that are useful diagnostics of abnormality (Taguchi & Jugulum 2002; Das & Datta 2007). In ecology, the Mahalanobis distance has been used in the construction of maps that show suitable habitat areas for a particular species. Pixels on the map are equated to points in multidimensional space on the basis of environmental variables and the Mahalanobis distance is used to give a measure of the distance from a point to the mean of the ecological niche. To identify the minimum set of basic habitat requirements for a species, Rotenberry, Preston & Knick (2006) proposed a decomposition of the Mahalanobis distance that exploits the eigenvectors of $\mathbf{\Sigma}$. Based on work in Rotenberry, Knick & Dunn (2002), they argued that the variables that load heavily on the eigenvectors corresponding to the smallest eigenvalues are the most influential in determining habitat suitability. Calenge, Darmon, Basille & Jullien (2008) added a step to the method of Rotenberry *et al.* (2006), forming a further eigenvector decomposition with the aim of producing biologically meaningful axes. The decomposition we propose here is simpler to implement and has a straightforward interpretation, making it

more likely to be put into practice. Rogers (2015) adopted it as a tool for identifying the key climate variables in determining future changes in the distribution of vector-borne diseases, illustrating its use through application to dengue, an important tropical disease. He referred to the contributions of variables being measured on the *Garthwaite–Koch scale*, citing a technical report (Garthwaite & Koch 2013) that forms the basis of the present paper.

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 this 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. Bootstrap confidence intervals for the contributions of individual variables can be constructed to quantify uncertainty in these contributions and to increase our insight into the relative importance of these contributions. These ideas are illustrated in Section 6. Concluding comments are given in Section 7.

2. Rationale for a partition

Let Q be the quadratic form

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

where $\boldsymbol{X}=(X_1,\ldots,X_m)^\top$ is an $m\times 1$ random vector whose variance is proportional to $\boldsymbol{\Sigma}$ and where $\boldsymbol{\mu}$ is a given $m\times 1$ vector that is not necessarily the mean of $\boldsymbol{X}$. This type of quadratic form arises in various applications. For example, in Hotelling's one-sample T^2 statistic, $\boldsymbol{X}$ would take the value of a sample mean, $\boldsymbol{\Sigma}$ would be the population variance, and $\boldsymbol{\mu}$ would be the hypothesised 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 $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ is diagonal is obvious. However, if $\boldsymbol{\Sigma}$ 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

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

and, to aid explanation, suppose the three components of $\mathbf{x}=(x_1,x_2,x_3)^\top$ correspond to standardised variables, *age* (x_1), *height* (x_2), and *weight* (x_3). In this example 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 $\boldsymbol{\Sigma}^{-1}$ 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 to be 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. Usually 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 some other quantity that is based on Q. The transformation that underlies the partition is defined in the next section.

Before ending this section we introduce further notation that will be used in the remainder of the paper. Bold-face italic capital letters $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{W}^{\diamond}$, $\widehat{\boldsymbol{W}}$, etc. are $m \times 1$ random vectors. Subscripts are added to denote components of the vector: $\boldsymbol{X} = (X_1, \ldots, X_m)^\top$, $\widehat{\boldsymbol{W}} = (\widehat{W}_1, \ldots, \widehat{W}_m)^\top$, etc. The notation $\widehat{\boldsymbol{\Sigma}}$ is used to denote a generic estimate of the $m \times m$ population variance matrix, $\boldsymbol{\Sigma}$. Likewise $\widehat{\boldsymbol{\Sigma}}_1$ is used to denote the standard unbiased estimate of $\boldsymbol{\Sigma}$ given by one sample and $\widehat{\boldsymbol{\Sigma}}_p$ is used to denote the standard pooled estimate of $\boldsymbol{\Sigma}$ based on independent samples from two populations that both have variance $\boldsymbol{\Sigma}$. The symbols σ_i^2, $\widehat{\sigma}_i^2$ are used to denote the ith diagonal entries of $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$, respectively. The symbols $\mathbf{D}$ and $\widehat{\mathbf{D}}$ are used to denote $m \times m$ diagonal matrices with ith diagonal entries equal to $\sigma_i^{-1}$and $\widehat{\sigma}_i^{-1}$, respectively $(i = 1, \ldots, m)$. Thus $\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}$ and $\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}}$ have diagonal entries equal to 1. The symbol $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is used to denote the $n \times m$ data matrix whose rows are the n observations and $\tilde{\mathbf{x}}_i$ is used to denote the ith column of $\mathbf{X}$. The symbols $\mathbf{A}$, $\widehat{\mathbf{A}}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{H}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Psi}$ are used to denote $m \times m$ matrices and $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_d$ are used to denote $m \times m$ and $d \times d$ orthogonal matrices, respectively.

3. The corr-max transformation

To form our partition, we consider transformations of the form

$$\boldsymbol{X} \mapsto \boldsymbol{W} = \mathbf{A}(\boldsymbol{X} - \boldsymbol{\mu}), \tag{2}$$

where $\boldsymbol{W}$ is an $m \times 1$ vector and

$$\boldsymbol{W}^\top \boldsymbol{W} = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{3}$$

for any value of $\boldsymbol{X}$. Then

$$Q = \sum_{i=1}^{m} W_i^2, \tag{4}$$

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

(a) the components of $\boldsymbol{W}$ are uncorrelated and have identical variances, 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.

The following theorem gives the transformation that maximises $\Sigma_{i=1}^{m}\text{cor}(X_i, W_i)$ under the constraints that (2) and (3) hold, where $\text{cor}(\cdot,\cdot)$ denotes correlation. Proofs of theorems are given in Appendix A.

Theorem 1. *Suppose* $\boldsymbol{W}=\mathbf{A}(\boldsymbol{X}-\boldsymbol{\mu})$ *and* $\text{var}(\boldsymbol{X})\propto\boldsymbol{\Sigma}$. *If (3) holds for all* $\boldsymbol{X}$, *then the components of* $\boldsymbol{W}$ *are uncorrelated with identical variances. If, in addition,* $\mathbf{A}$ *is chosen to maximise* $\Sigma_{i=1}^{m}\text{cor}(X_i, W_i)$, *then* $\mathbf{A}=(\mathbf{D\Sigma D})^{-1/2}\mathbf{D}$, *where* $\mathbf{D}$ *is a diagonal matrix such that* $\mathbf{D\Sigma D}$ *has diagonal entries equal to 1.*

We define the *corr-max transformation* to be the transformation given by (2) with $\mathbf{A}=(\mathbf{D\Sigma D})^{-1/2}\mathbf{D}$. From Theorem 1, this transformation yields a $\boldsymbol{W}$ that satisfies requirement (a). For (b), we first note that it is always possible to scale and translate X_i so that it has the same variance and the same mean as W_i, whence the degree to which X_i equates to W_i would primarily be determined by its correlation with W_i. (Perfect correlation would imply that they were identical.) Moreover, scaling and translation do not change the nature of a variable. Otherwise, for example, temperature measurements on the Celsius and Fahrenheit scales would not be equivalent. Hence, the degree to which W_i equates to the ith x-variable is largely determined by $\text{cor}(X_i, W_i)$. Consequently under a sensible criterion the corr-max transformation satisfies (b) as fully as possible, since it maximises $\Sigma_{i=1}^{m}\text{cor}(X_i, W_i)$ when the constraint equations (2) and (3) hold. The extent to which the corr-max transformation satisfies (b) is discussed further in Section 7.

Theorem 1 completes the specification of our partition for the case where $\boldsymbol{\Sigma}$ is known. To summarise, if $\boldsymbol{X}$ takes the value $\mathbf{x}$ and $\text{var}(\boldsymbol{X})\propto\boldsymbol{\Sigma}$, the corr-max transformation yields the new vector $\mathbf{w}=(\mathbf{D\Sigma D})^{-1/2}\mathbf{D}(\mathbf{x}-\boldsymbol{\mu})$ and the contribution of the ith x-variable to $Q(\mathbf{x})=(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^{\top}$ is defined to be w_i^2 $(i=1,\ldots,m)$.

When $\boldsymbol{\Sigma}$ is unknown, we replace it in the foregoing method with an estimate, $\widehat{\boldsymbol{\Sigma}}$ say. In some contexts this type of substitution can have drawbacks but here it seems appropriate, since it yields properties similar to Theorem 1, but in terms of maximising sample correlations, which we denote by $\text{cor}_s(\cdot,\cdot)$, rather than population correlations. This result is given in Theorem 2. Its proof is similar to that of Theorem 1 and is omitted.

Theorem 2. *Suppose that the sample variance of* $\boldsymbol{X}$ *is proportional to* $\widehat{\boldsymbol{\Sigma}}$ *and* $\Sigma_{j=1}^{m}\text{cor}_s(X_j, \widehat{W}_j)$ *is to be maximised, subject to* $\widehat{\boldsymbol{W}}=\widehat{\mathbf{A}}(\boldsymbol{X}-\boldsymbol{\mu})$ *and* $\widehat{\boldsymbol{W}}^{\top}\widehat{\boldsymbol{W}}=(\boldsymbol{X}-\boldsymbol{\mu})^{\top}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})^{\top}$ *for any* $\boldsymbol{X}$. *Then* $\widehat{\mathbf{A}}=(\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}})^{-1/2}\widehat{\mathbf{D}}$ *and*

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

where $\widehat{\mathbf{D}}$ *is diagonal and* $\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}}$ *has diagonal entries equal to 1.*

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 the values of these correlations and, more generally, the correlations $\text{cor}(X_i, W_j)$ and $\text{cor}_s(X_i, \widehat{W}_j)$ for $i=1,\ldots,m; j=1,\ldots,m$. It has the

interesting implications that $\mathrm{cor}(X_i, W_j) = \mathrm{cor}(X_j, W_i)$ and $\mathrm{cor}_s(X_i, \widehat{W}_j) = \mathrm{cor}_s(X_j, \widehat{W}_i)$ for all i and j, since $(\mathbf{D\Sigma D})^{1/2}$ and $(\widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{D}})^{1/2}$ are both symmetric matrices.

Theorem 3. *Suppose* $\boldsymbol{W} = (\mathbf{D\Sigma D})^{-1/2}\mathbf{D}(\boldsymbol{X} - \boldsymbol{\mu})$ *and* $\mathrm{var}(\boldsymbol{X}) \propto \mathbf{\Sigma}$. *Then* $\mathrm{cor}(X_i, W_j)$ *equals the* (i,j)*th entry of* $(\mathbf{D\Sigma D})^{1/2}$. *Similarly, if* $\widehat{\boldsymbol{W}} = (\widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{D}})^{-1/2}\widehat{\mathbf{D}}(\boldsymbol{X} - \boldsymbol{\mu})$ *and the sample variance of* $\boldsymbol{X}$ *is proportional to* $\widehat{\mathbf{\Sigma}}$, *then* $\mathrm{cor}_s(X_i, \widehat{W}_j)$ *equals the* (i,j)*th entry of* $(\widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{D}})^{1/2}$.

So far we have only considered the partition of a quadratic form, but the corr-max transformation also gives a useful partition of the bilinear form $\mathbf{U}^\top\mathbf{\Sigma}^{-1}\mathbf{V}$, provided $\mathrm{var}(\mathbf{U}) \propto \mathbf{\Sigma}$ and $\mathrm{var}(\mathbf{V}) \propto \mathbf{\Sigma}$. Theorem 4 gives the relevant result. Its proof follows from the proof of Theorem 1.

Theorem 4. *Suppose* $\mathrm{var}(\boldsymbol{U}) \propto \mathbf{\Sigma}$ *and* $\mathrm{var}(\boldsymbol{V}) \propto \mathbf{\Sigma}$ *where* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *are* $m \times 1$ *random vectors. Let* $\boldsymbol{W}^* = \mathbf{A}\boldsymbol{U}$ *and* $\boldsymbol{W}^\circ = \mathbf{A}\boldsymbol{V}$ *where* $\mathbf{A}$ *is a square matrix. Under the constraint* $(\boldsymbol{W}^*)^\top\boldsymbol{W}^\circ = \boldsymbol{U}^\top\mathbf{\Sigma}^{-1}\boldsymbol{V}$, *both* $\Sigma_{i=1}^m \mathrm{cor}(U_i, W_i^*)$ *and* $\Sigma_{i=1}^m \mathrm{cor}(V_i, W_i^\circ)$ *are maximised when* $\mathbf{A} = (\mathbf{D\Sigma D})^{-1/2}\mathbf{D}$.

Both $\boldsymbol{U} \mapsto \boldsymbol{W}^*$ and $\boldsymbol{V} \mapsto \boldsymbol{W}^\circ$ are corr-max transformations, since $\mathbf{A} = (\mathbf{D\Sigma D})^{-1/2}\mathbf{D}$. From this, and from the theorem, it is reasonable to identify the ith components of $\boldsymbol{W}^*$ and $\boldsymbol{W}^\circ$ with the ith components of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Then $(\boldsymbol{W}^*)^\top\boldsymbol{W}^\circ$ is our partition of $\boldsymbol{U}^\top\mathbf{\Sigma}^{-1}\boldsymbol{V}$, giving $W_i^*W_i^\circ$ as the contribution of the ith x-variable to $\boldsymbol{U}^\top\mathbf{\Sigma}^{-1}\boldsymbol{V}$. In Section 5 we use the theorem to form a partition of Fisher's linear discriminant function. When $\mathbf{\Sigma}$ is estimated from data, results corresponding to Theorem 4 hold with $\widehat{\mathbf{A}} = (\widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{D}})^{-1/2}\widehat{\mathbf{D}}$.

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) showed that the x-variables involved in a collinearity can be identified using the cos-max transformation. Example 2 in Section 5.2 illustrates this.

Collinear variables can be replaced by non-collinear variables via orthogonal rotation of coordinate axes. This can clarify the relationship between x-variables and a quadratic form, as examples will illustrate. Only axes corresponding to collinear variables need be rotated. The results of a rotation are sensitive to scale, 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 $\mathrm{var}(\boldsymbol{X}) \propto \mathbf{\Sigma}$ and $\mathbf{D\Sigma D}$ has diagonal entries all equal to 1, so the components of $\mathbf{D}\boldsymbol{X}$ have identical variances. Let $\mathbf{\Gamma}$ be an $m \times m$ orthogonal matrix and put $\boldsymbol{Y} = \mathbf{\Gamma D}(\boldsymbol{X} - \boldsymbol{\mu})$, so that $\boldsymbol{Y}$ is obtained by a re-scaling of $\boldsymbol{X} - \boldsymbol{\mu}$, followed by an orthogonal rotation. Suppose that we want to apply a transform

$$\mathbf{Y} \mapsto \mathbf{W}^\diamond = \mathbf{CY}, \tag{6}$$

in such a manner that there are large correlations between Y_i and $W_i^\diamond$ for $i = 1,\ldots,m$. The components of $\mathbf{Y}$ are not all equally important – after rotation some components will have a smaller variance than others and those with smaller variances are deemed to be less important, as in principal components analysis. The corr-max transformation would choose $\mathbf{C}$ to maximise $\Sigma_{i=1}^m \text{cor}(Y_i, W_i^\diamond)$ but now, to reflect the differing importance of some variables, we choose $\mathbf{C}$ to maximise $\Sigma_{i=1}^m \{\text{var}(Y_i)\}^{1/2}\text{cor}(Y_i, W_i^\diamond)$. This gives greater weight to the Y_i with greater variance. Theorem 5 gives the resulting matrix $\mathbf{C}$ and properties of the transformation.

Theorem 5. *Let $\boldsymbol{Y} = \boldsymbol{\Gamma}\mathbf{D}(\boldsymbol{X} - \boldsymbol{\mu})$ where $\boldsymbol{\Gamma}$ is a given orthogonal matrix and var$(\boldsymbol{X}) \propto \boldsymbol{\Sigma}$. Suppose $\Sigma_{i=1}^m \{\text{var}(Y_i)\}^{1/2}\text{cor}(Y_i, W_i^\diamond)\}$ is to be maximised, subject to $\boldsymbol{W}^\diamond = \boldsymbol{CY}$ and $(\boldsymbol{W}^\diamond)^\top \boldsymbol{W}^\diamond = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$, where $\mathbf{C}$ is a square matrix. Then:*

(i) the components of $\boldsymbol{W}^\diamond$ are uncorrelated and have identical variances;

(ii)

$$\mathbf{C} = (\boldsymbol{\Gamma}\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}\boldsymbol{\Gamma}^\top)^{-1/2} = \boldsymbol{\Gamma}(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}\boldsymbol{\Gamma}^\top;$$

(iii)

$$\boldsymbol{W}^\diamond = \boldsymbol{\Gamma}(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}\mathbf{D}(\boldsymbol{X} - \boldsymbol{\mu}); \tag{7}$$

(iv) $[\{\text{var}(Y_i)\}^{1/2}\text{cor}(Y_i, W_j^\diamond]$ is equal to the (i, j)th entry of $(\boldsymbol{\Gamma}\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}\boldsymbol{\Gamma}^\top)^{1/2}$;

(v) with $\boldsymbol{W}^\diamond$ written as $\mathbf{W}^\diamond(\boldsymbol{\Gamma})$ to highlight that it is a function of $\boldsymbol{\Gamma}$,

$$\mathbf{W}^\diamond(\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}\mathbf{W}^\diamond(\mathbf{I}) \tag{8}$$

where $\boldsymbol{W}^\diamond(\mathbf{I}) = \mathbf{I}(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}\mathbf{D}(\mathbf{X} - \boldsymbol{\mu})$.

The transformation from $\boldsymbol{X} - \boldsymbol{\mu}$ to $\boldsymbol{W}^\diamond$ will be referred to as the *adapted* corr-max transformation. It is identical to the ordinary corr-max transformation if there is no rotation, that is when $\boldsymbol{\Gamma} = \mathbf{I}$. If $\boldsymbol{\Sigma}$ is unknown, we replace it with an estimate, $\widehat{\boldsymbol{\Sigma}}$, and put

$$\widehat{\boldsymbol{W}}^\diamond = \boldsymbol{\Gamma}(\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}})^{-1/2}\widehat{\mathbf{D}}(\boldsymbol{X} - \boldsymbol{\mu}). \tag{9}$$

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 $\boldsymbol{W}^\diamond$ or $\widehat{\boldsymbol{W}}^\diamond$.

From equation (8) we obtain the same result whether (a) we multiply $\mathbf{D}(\boldsymbol{X} - \boldsymbol{\mu})$ by the rotation matrix $\boldsymbol{\Gamma}$ and transform the result, or (b) we transform $\mathbf{D}(\boldsymbol{X} - \boldsymbol{\mu})$ and multiply the result by $\boldsymbol{\Gamma}$. That is, with the adapted corr-max transformation, the operations of rotation and transformation are commutative.

This property allows us to rotate the coordinate axes corresponding to x-variables involved in a collinearity while neither affecting the identity of other x-variables, nor altering assessments of the latter variables' contributions to the quadratic form. To elucidate, suppose that we want to rotate the first d of the m axes. Then the rotation matrix $\boldsymbol{\Gamma}$ has the following block-diagonal form:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Gamma}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-d} \end{bmatrix}, \tag{10}$$

where $\boldsymbol{\Gamma}_d$ is a $d \times d$ orthogonal matrix. Multiplying $\boldsymbol{X}$ by $\boldsymbol{\Gamma}$ only changes the first d components of $\boldsymbol{X}$ and leaves its other components unchanged, so the latter components are the original

variables. Moreover, under the transformation in equation (7), the last $m-d$ components of $\widehat{\boldsymbol{W}}^{\diamond}$ are unaffected by $\boldsymbol{\Gamma}_d$; the rotation only changes its first m components. Thus, under the adapted corr-max transformation, the rotation of selected axes will leave some variables unchanged (those corresponding to unrotated axes) and the contributions of those variables to the quadratic form, as measured by the partition, will also be unchanged. We refer to this as the *rotation invariance property*.

Ideally, a partition yields 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 might rotate the axes corresponding to variables involved in the collinearity, and then apply the transformation. 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.

A rotation is attractive if it yields meaningful quantities. If, say, the only collinearity was between the first two variables, X_1 and X_2, a sensible rotation might be

$$\boldsymbol{\Gamma}_2 = \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{bmatrix} \text{ and } \boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Gamma}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-2} \end{bmatrix},$$

which constructs 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 and the new variables will also have a low correlation if the variance of X_1 is similar to the variance of X_2. If rotation is used to counteract more than one distinct collinearity between the x variables, then Γ_m should have a block diagonal form, with a separate block for each collinearity. An example is given in Section 5.2. When a collinearity involves more than two variables, constructing meaningful variables that have low correlations can be a challenging task. An approach based on orthogonal contrasts that might sometimes be useful is described in Appendix B.

While rotation can be helpful when collinearities are present, we should stress that rotation is never essential. The standard corr-max transformation of Section 3 can be applied whenever $\boldsymbol{\Sigma}$ is a positive-definite matrix, even if $\boldsymbol{\Sigma}$ contains high correlations, and it will yield a sensible partition of a quadratic form, as we discuss further in Section 7. Hence axes should only be rotated when the new variables that are constructed have an understandable interpretation.

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 collinearity is present and some x-axes are rotated 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

$$\Theta = \delta(\boldsymbol{X}-\boldsymbol{\mu})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}), \tag{11}$$

with $\widehat{\boldsymbol{\Sigma}}$ an estimate of $\boldsymbol{\Sigma}$, $\text{var}(\boldsymbol{X}) \propto \boldsymbol{\Sigma}$ and δ a known positive scalar. From equation (5), the corr-max transformation yields $\widehat{\boldsymbol{W}} = (\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}})^{-1/2}\widehat{\mathbf{D}}(\boldsymbol{X}-\boldsymbol{\mu})$, and the contribution of the ith x-variable to Θ is evaluated as δw_i^2, where $(w_1,\ldots,w_m)^\top$ is the value of $\widehat{\boldsymbol{W}}$ given by data.

Before the partition can be applied, $\boldsymbol{X}$, $\widehat{\boldsymbol{\Sigma}}$, δ, and $\boldsymbol{\mu}$ must be identified and it must be checked that $\text{var}(\boldsymbol{X}) \propto \boldsymbol{\Sigma}$. (The matrix $\widehat{\mathbf{D}}$ 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 analyst should examine the correlations between components of $\widehat{\boldsymbol{W}}$ and the corresponding components of $\mathbf{X}$; rotation of x-axes might be considered if some correlations are low. (In our examples we consider rotating axes when correlations are 0.8 or lower.)

In the following four applications, the first three have precisely the form given in (11), 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 $\text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, giving a sample mean $\overline{\boldsymbol{X}}$ 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,

$$T_1^2 = n(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_0)^\top \widehat{\boldsymbol{\Sigma}}_1^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_0). \tag{12}$$

Let the role of $\boldsymbol{X}$ in (11) be played by $\overline{\boldsymbol{X}}$, so that $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}/n$. 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, $\text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\text{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, 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,

$$T_2^2 = \{n_1 n_2/(n_1+n_2)\}(\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_2)^\top \widehat{\boldsymbol{\Sigma}}_p^{-1}(\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_2), \tag{13}$$

where $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_2$ are the sample means and $\widehat{\boldsymbol{\Sigma}}_p$ is the pooled estimate of $\boldsymbol{\Sigma}$ derived from the two samples. Let the role of $\boldsymbol{X}$ in (11) be played by $\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2$, so $\text{var}(\boldsymbol{X}) \propto \boldsymbol{\Sigma}$. 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 $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ are two $m \times 1$ vectors, then the Mahalanobis distance between them is

$$(\boldsymbol{X}_{(1)}-\boldsymbol{X}_{(2)})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{(1)}-\boldsymbol{X}_{(2)}). \tag{14}$$

Here $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ 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 $\text{var}(\boldsymbol{X}_{(i)}) = k_i\boldsymbol{\Sigma}$ $(i = 1, 2)$ where k_1 or k_2 (but not both) may equal 0. We also suppose that $\text{E}(\widehat{\boldsymbol{\Sigma}}) \propto \boldsymbol{\Sigma}$ so, for example, $\widehat{\boldsymbol{\Sigma}}$ might be the maximum likelihood estimate or an unbiased estimate of $\boldsymbol{\Sigma}$. Let $\boldsymbol{X} = \boldsymbol{X}_{(1)} - \boldsymbol{X}_{(2)}$, so $\text{var}(\boldsymbol{X}) \propto \boldsymbol{\Sigma}$. 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.* Suppose an observation needs to be classified as belonging to one of two classes that are characterised by the multivariate normal distributions $\text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\text{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, with sample means $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_2$ and common estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_p$. A new observation $\boldsymbol{X}^*$ is classified as belonging

to class 1 on the basis of Fisher's linear discriminant function if

$$r(\boldsymbol{X}^*) = \left\{ \boldsymbol{X}^* - \frac{1}{2}(\overline{\boldsymbol{X}}_1 + \overline{\boldsymbol{X}}_2) \right\}^\top \widehat{\boldsymbol{\Sigma}}_p^{-1} (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) > 0. \tag{15}$$

Consider the transformations

$$\widehat{\mathbf{W}}^\circ = (\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}_p\widehat{\mathbf{D}})^{-1/2}\widehat{\mathbf{D}}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \tag{16}$$

and

$$\widehat{\mathbf{W}}^* = (\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}_p\widehat{\mathbf{D}})^{-1/2}\widehat{\mathbf{D}}\{\mathbf{X}^* - \frac{1}{2}(\overline{\boldsymbol{X}}_1 + \overline{\boldsymbol{X}}_2)\}. \tag{17}$$

Since $\text{var}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \propto \boldsymbol{\Sigma}$ and $\text{var}[\boldsymbol{X}^* - \frac{1}{2}(\overline{\boldsymbol{X}}_1 + \overline{\boldsymbol{X}}_2)] \propto \boldsymbol{\Sigma}$, Theorem 4 applies. Hence the ith components of both $\widehat{\mathbf{W}}^\circ$ and $\widehat{\mathbf{W}}^*$ can be identified with the ith x-variable. Let W_i° and W_i^* denote these components. Because $r(\mathbf{X}^*) = \Sigma_{i=1}^m W_i^\circ W_i^*$, the contribution of the ith x-variable to $r(\mathbf{X}^*)$ is given by the observed value of $W_i^\circ W_i^*$.

We use two examples to explore how the transformation and partition work in practice. In the first example we apply the transformation without rotation of variables and consider applications (a), (c) and (d). In the second example, given in the next subsection, we illustrate application (b) and apply the transformation to both rotated and un-rotated variables.

Example 1: Swiss bank notes Flury & Riedwyl (1988) present data on 100 genuine Swiss 1000-franc bank notes. Six measurements were made on each note: *length* (length), *left-ht* (height measured on the left), *right-ht* (height measured on the right), *lower* (distance from the inner frame to the lower border), *upper* (distance from the inner frame to the upper border), and *diagonal* (length of the diagonal). These measurements are the data values of $\boldsymbol{X} = (X_1, \ldots, X_6)^\top$. Their sample standard deviations are (0.388, 0.364, 0.355, 0.643, 0.649, 0.447) and the reciprocals of these standard deviations form the diagonal entries of $\widehat{\mathbf{D}}$. The sample correlation matrix of $\mathbf{X}$ is:

$$\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}} = \begin{pmatrix} 1.000 & 0.411 & 0.416 & 0.229 & 0.057 & 0.032 \\ 0.411 & 1.000 & 0.664 & 0.242 & 0.208 & -0.265 \\ 0.416 & 0.664 & 1.000 & 0.255 & 0.133 & -0.150 \\ 0.229 & 0.242 & 0.255 & 1.000 & -0.632 & -0.001 \\ 0.057 & 0.208 & 0.133 & -0.632 & 1.000 & -0.260 \\ 0.032 & -0.265 & -0.150 & -0.001 & -0.260 & 1.000 \end{pmatrix} \tag{18}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix. It can be seen that no correlation is larger than 0.664. The mean vector for the banknote measurements is $\overline{\mathbf{x}} = (214.969, 129.943, 129.720, 8.305, 10.168, 141.517)^\top$.

If the corr-max transformation is applied to a vector $\boldsymbol{X}$ to yield a vector $\widehat{\boldsymbol{W}}$, the correlations between components of $\boldsymbol{X}$ and the corresponding components of $\widehat{\boldsymbol{W}}$ are equal to the diagonal entries of $(\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}})^{1/2}$. These diagonal entries 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 $\widehat{\mathbf{W}}$, 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, $\boldsymbol{\mu}_0 = (215.007, 129.979, 129.756, 8.369, 10.233, 141.562)^\top$. These values have been chosen so that, for each variable, the hypothesised population mean exceeds

the sample mean by 0.1 standard deviations. The value of the test statistic, given by equation (12), is $T_1^2 = 8.74$. We have already calculated $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}}$. Setting $\boldsymbol{X}$ and $\boldsymbol{\mu}$ equal to $\bar{\mathbf{x}}$ and $\boldsymbol{\mu}_0$ respectively in equation (5) gives $\widehat{\boldsymbol{W}} = -(0.051, 0.053, 0.055, 0.164,\ 0.182, 0.138)^\top$. 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 0.51^2, 0.53^2, 0.55^2, 1.64^2, 1.82^2 and 1.38^2. These values sum to 8.75, 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 hypothesised population mean by an equivalent amount, the last three x-variables (*lower, upper,* and *diagonal*) make larger contributions to the T_1^2 statistic than the first three x-variables (*length*, *left-ht* and *right-ht*).

As an example involving Mahalanobis distance, suppose the measurements for an additional banknote that might be a forgery are $\mathbf{x}_2 = (215.8, 129.7, 129.0, 6.9, 8.6,\ 143.2)^\top$. The Mahalonobis distance between $\mathbf{x}_2$ and the mean value of $\boldsymbol{X}$ in the sample of 100 genuine banknotes $\bar{\mathbf{x}}$, is given by equation (14) with $\boldsymbol{X}_{(1)} = \bar{\mathbf{x}}$ and $\boldsymbol{X}_{(2)} = \mathbf{x}_2$. The value of this distance is 55.69, which gives clear evidence the note is a forgery ($p < 0.0001$). Our partition can be used to determine which characteristics of the new banknote distinguish it from the genuine banknotes. We put $\boldsymbol{X} = \bar{\boldsymbol{x}} - \boldsymbol{x}_2$ and $\boldsymbol{\mu} = \mathbf{0}$ in equation (5), to obtain $\widehat{\boldsymbol{W}}$. As $\delta = 1$, the contribution of the ith x-variable to the Mahalanobis distance is the square of the ith component of $\widehat{\boldsymbol{W}}$. These squared values are (8.64 0.87 4.54 16.66 15.12 9.86). Hence the measurements that most distinguish the new banknote from genuine banknotes are X_4 (*lower*) and X_5 (*upper*).

The Swiss bank notes dataset given by Flury & 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

$$\widehat{\boldsymbol{\Sigma}}_p = \begin{pmatrix} 0.1371 & 0.0448 & 0.0406 & -0.0217 & 0.0169 & 0.0085 \\ 0.0448 & 0.0988 & 0.0663 & 0.0163 & 0.0186 & -0.0241 \\ 0.0406 & 0.0663 & 0.1076 & 0.0198 & 0.0154 & 0.0052 \\ -0.0217 & 0.0163 & 0.0198 & 0.8473 & -0.3768 & 0.1191 \\ 0.0169 & 0.0186 & 0.0154 & -0.3768 & 0.4128 & -0.0487 \\ 0.0085 & -0.0241 & 0.0052 & 0.1191 & -0.0487 & 0.2555 \end{pmatrix}. \tag{19}$$

Table 1 summarises the analysis. The first two rows, $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_2$, show the sample means of the genuine and faked bank notes, respectively. The note to be classified is $\boldsymbol{X}^*$. Equation (15) gives -20.34 as the value of $r(\boldsymbol{X}^*)$, indicating that the new note should be classified as coming from the same population as the fakes. Applying equations (16) and (17) we obtain $\widehat{\boldsymbol{W}}^\circ$ and $\widehat{\boldsymbol{W}}^*$ (fourth and fifth rows). The contributions of individual x-variables to $r(\boldsymbol{X}^*)$ are evaluated as the diagonal entries of $\widehat{\boldsymbol{W}}^\circ(\widehat{\boldsymbol{W}}^*)^\top$, shown in the last row. 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, *lower*, *upper,* and *diagonal*, underlie the outcome of the discrimination rule, as they make much larger contributions to $r(\boldsymbol{X}^*)$ (in absolute value) than the first three variables.

In Section 1 we noted that Rotenberry *et al.* (2006) examined eigenvectors corresponding to small eigenvalues in order to determine influential variables on a Mahalanobis distance.

TABLE 1

Values from the discriminant analysis for a Swiss bank note

	X_1	X_2	X_3	X_4	X_5	X_6
$\overline{\boldsymbol{X}}_1$	214.969	129.943	129.720	8.305	10.168	141.517
$\overline{\boldsymbol{X}}_2$	214.823	130.300	130.193	10.530	11.133	139.45
$\boldsymbol{X}^*$	214.4	130.1	130.3	9.7	11.7	139.8
$\widehat{\boldsymbol{W}}^\circ$	0.38	−0.001	−1.48	−4.16	−2.80	4.56
$\widehat{\boldsymbol{W}}^*$	−1.44	−0.64	1.49	1.21	2.10	−1.46
$[\widehat{\boldsymbol{W}}^\circ(\widehat{\boldsymbol{W}}^*)^\top]_{ii}$	−0.55	0.001	−2.21	−5.04	−5.89	−6.67

Before leaving this example we illustrate their method by applying it to the case where we have 100 genuine banknotes and an additional banknote that might be a forgery. Their approach is to decompose the quadratic form $Q=(\overline{\mathbf{x}}-\mathbf{x}_2)^\top\widehat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{x}}-\mathbf{x}_2)$ as

$$Q=\sum_{j=1}^{m}u_j^2/\lambda_j$$

where $u_j=(\overline{\mathbf{x}}-\mathbf{x}_2)^\top\boldsymbol{\gamma}_j$, $\lambda_1>\ldots>\lambda_m$ are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_m$ are the corresponding eigenvectors. They focus on small eigenvalues because, if λ_j is small, then $X^\top\boldsymbol{\gamma}_j$ varies little for the X values in the hundred genuine banknotes, so that a large value of $(\overline{\mathbf{x}}-\mathbf{x}_2)^\top\boldsymbol{\gamma}_j$ is more indicative of forgery. The eigenvalues of the sample covariance matrix of the genuine banknotes are 0.69, 0.36, 0.19, 0.087, 0.080 and 0.041, so interest centres on either just the smallest eigenvalue or the smallest three eigenvalues. The following are the eigenvectors for the three smallest eigenvalues:

	length	*left-ht*	*right-ht*	*lower*	*upper*	*diagonal*
Smallest	−0.011	0.737	−0.666	−0.050	−0.062	0.072
2^{nd} smallest	0.113	−0.360	−0.481	0.559	0.548	0.116
3^{rd} smallest	0.786	−0.243	−0.280	−0.243	−0.244	−0.354

Based on just the eigenvector corresponding to the smallest eigenvalue, X_2 and X_3 (*left-ht* and *right-ht*) are clearly the most important variables, since in the case of that eigenvector they have much larger coefficients (in absolute value) than the other variables. However, if the three eigenvectors displayed above are all considered relevant, then deciding which x-variables are important is not clear-cut and requires the analyst to make an intuitive judgment. Moreover, there is no obvious method of determining the relative quantitative importance of different variables, and with any such method the answers are likely to depend on whether the three smallest eigenvalues or only the very smallest are considered "small".

5.2. Collinearities, rotation and quadratic forms

Some advantages of the (un-adapted) corr-max transformation are diminished 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 we apply the cos-max transformation to data that have been standardised to have means of 0 and variances of 1, making the cos-max and corr-max transformations 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 creates meaningful non-collinear variables from the 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 & 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, $\widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}_p\widehat{\mathbf{D}}$, takes the value

$$\begin{pmatrix} 1.00 & 0.68 & 0.05 & 0.10 & 0.06 & 0.15 & 0.06 & 0.63 & 0.65 \\ 0.68 & 1.00 & -0.04 & -0.11 & -0.04 & 0.05 & -0.15 & 0.34 & 0.34 \\ 0.05 & -0.04 & 1.00 & 0.78 & 0.86 & 0.14 & -0.05 & -0.04 & -0.05 \\ 0.10 & -0.11 & 0.78 & 1.00 & 0.90 & 0.13 & 0.01 & -0.04 & -0.06 \\ 0.06 & -0.04 & 0.86 & 0.90 & 1.00 & 0.15 & -0.06 & -0.08 & -0.11 \\ 0.15 & 0.05 & 0.14 & 0.13 & 0.15 & 1.00 & 0.12 & 0.21 & 0.21 \\ 0.06 & -0.15 & -0.05 & 0.01 & -0.06 & 0.12 & 1.00 & 0.16 & 0.16 \\ 0.63 & 0.34 & -0.04 & -0.04 & -0.08 & 0.21 & 0.16 & 1.00 & 0.97 \\ 0.65 & 0.34 & -0.05 & -0.06 & -0.11 & 0.21 & 0.16 & 0.97 & 1.00 \end{pmatrix} \tag{20}$$

Under the cos-max transformation, a data matrix $\mathbf{X}$ is transformed to $(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{X}$. 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 $\mathbf{X}_s^\top\mathbf{X}_s$. If we put $(\mathbf{X}_s^\top\mathbf{X}_s)^{-1/2} = \mathbf{H} = (\boldsymbol{h}_1,\ldots,\boldsymbol{h}_m)^\top$ then, as Garthwaite *et al.* (2012) pointed out, the variance inflation factor for the jth variable (VIF_j) is equal to $\boldsymbol{h}_j^\top\boldsymbol{h}_j$. 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, $\mathbf{X}_s^\top\mathbf{X}_s = \widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}_p\widehat{\mathbf{D}}$, so examining the rows of $(\widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}_p\widehat{\mathbf{D}})^{-1/2}$ identifies variables involved in collinearities. (When $\mathbf{X}_s^\top\mathbf{X}_s = \widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}_p\widehat{\mathbf{D}}$, the corr-max and cos-max transformations are the same.)

We put $(\widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}_p\widehat{\mathbf{D}})^{-1/2} = (\boldsymbol{h}_1,\ldots,\boldsymbol{h}_m)^\top$ and give the values of the $\boldsymbol{h}_j^\top$ in Table 2. Values greater than 0.80 in absolute value 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 $\boldsymbol{h}_5^\top\boldsymbol{h}_5$. A VIF above 10 is often treated as indicative of collinearity (Neter, Wasserman & Kutner 1983 p. 392) On this basis, X_8 (*Bfat*) and X_9 (*SSF*) are involved in collinearites and, from the bold-face numbers in the display of $\boldsymbol{h}_8$ and $\boldsymbol{h}_9$, there is a collinearity between them. Weaker boundaries for flagging a collinearity have also been proposed; Menard (1995, p. 66) suggests a VIF above 5 should raise concern and O'Brien (2007) reports that boundary values as low as 4 have been suggested as rules of thumb. A boundary of 4 or 5 would indicate one further collinearity, between X_4 (*Hg*) and X_5 (*Hc*).

TABLE 2

Rows of $(\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}_p\widehat{\mathbf{D}})^{-1/2}$ *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
$\boldsymbol{h}_1^\top$	**1.66**	−0.61	0.06	−0.23	0.00	−0.01	−0.06	−0.17	−0.41	3.38
$\boldsymbol{h}_2^\top$	−0.61	**1.36**	−0.03	0.23	−0.08	0.00	0.14	−0.06	0.07	2.31
$\boldsymbol{h}_3^\top$	0.06	−0.03	**1.75**	−0.30	−0.80	−0.03	0.02	0.05	−0.10	3.83
$\boldsymbol{h}_4^\top$	−0.23	0.23	−0.30	**2.09**	**−1.12**	0.00	−0.04	0.04	0.00	5.82
$\boldsymbol{h}_5^\top$	0.00	−0.08	−0.80	**−1.12**	**2.48**	−0.08	0.06	−0.09	0.22	8.15
$\boldsymbol{h}_6^\top$	−0.01	0.00	−0.03	0.00	−0.08	**1.04**	−0.05	−0.07	−0.06	1.09
$\boldsymbol{h}_7^\top$	−0.06	0.14	0.02	−0.04	0.06	−0.05	**1.04**	−0.07	−0.02	1.11
$\boldsymbol{h}_8^\top$	−0.17	−0.06	0.05	0.04	−0.09	−0.07	−0.07	**3.27**	**−2.43**	16.64
$\boldsymbol{h}_9^\top$	−0.41	0.07	−0.1	0.00	0.22	−0.06	−0.02	**−2.43**	**3.38**	17.59

If the corr-max transformation is applied to $\mathbf{X}=(X_1,\ldots,X_9)^\top$, then the following are the sample correlations between each x variable and the variable to which 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 the axes associated with these variable pairs we replace the corr-max transformation by the adapted corr-max transformation given by equation (9), with $\boldsymbol{\Gamma}$ set equal to the following block-diagonal orthogonal matrix:

$$\boldsymbol{\Gamma}=\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2^{-1/2} & 2^{-1/2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2^{-1/2} & -2^{-1/2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1/2} & 2^{-1/2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1/2} & -2^{-1/2} \end{pmatrix}. \tag{21}$$

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 new variables and their transformed values while leaving the corresponding correlations of all other variables unchanged. The correlations between the rotated x variables and the variables to which they transform are displayed below. 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 ($\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$), and the pooled standard deviations (S.D.) 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.	11.69	8.07	0.336	0.929	2.60	1.801	43.3	4.45	27.3

The reciprocals of the standard deviations constitute the non-zero (diagonal) entries of the matrix $\widehat{\mathbf{D}}$. When Hotelling's T^2 test is used to compare the means of the two groups we obtain a T_2^2 statistic equal to 1199.1. This value gives, as you might expect, very clear evidence of differences between the two groups ($p \simeq 0.0000$). However, the question of which quantities contribute most to the T_2^2 value is still relevant.

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

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

The partition allows us to evaluate 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 whose axes were rotated, *Hg* and *Hc*, their average makes a substantial contribution (13%) while the contribution from their difference is only 1%.

6. Bootstrap confidence intervals

The corr-max transformation gives point estimates of the contributions of individual variables to a quadratic form. Obtaining theoretical results that give interval estimates of these contributions is difficult, but the bootstrap can be used to obtain approximate confidence intervals. We elucidate the procedure through examples.

6.1. Confidence interval for contributions to a Mahalanobis distance

In Example 1 there were 100 genuine Swiss 1000-franc bank notes and an additional bank note that might be a forgery. The Mahalanobis distance between the potential forgery and the mean of the genuine bank notes was 55.69 and the contributions of the six individual variables were estimated as (8.64 0.87 4.54 16.66 15.12 9.86). To obtain bootstrap confidence intervals for these contributions we generated 100 000 resamples from the 100 genuine bank notes. Each resample was a random sample of size 100 drawn *with replacement* from the 100 genuine notes.

Each resample was used in the same way as the original sample. The Mahalanobis distance between the potential forgery and the mean of the resample was calculated, with the resample being used to estimate the covariance matrix, $\widehat{\boldsymbol{\Sigma}}$. The contributions of individual

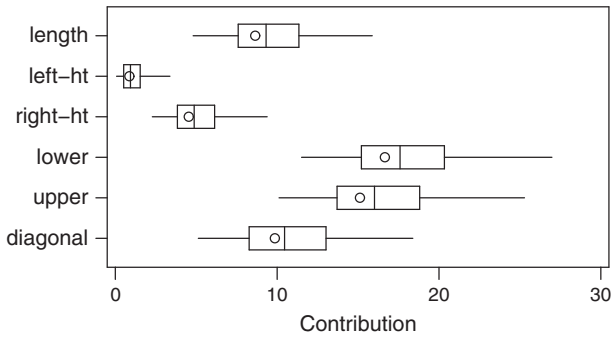

Figure 1. Confidence intervals for the contributions of individual variables to the Mahalanobis distance of a potential banknote forgery.

variables to the Mahalanobis distance were then evaluated using the corr-max transformation. This gave 100 000 estimates of the contribution of each variable and the kth smallest of these is equated to the $(k/1000)$th percentile of the bootstrap distribution. The median for a variable's contribution is thus the 50 000th smallest value and the endpoints of an approximate 95% confidence interval are the 2500th smallest and 2500th largest values. (This is the simplest method of forming bootstrap confidence intervals. As is well known, it typically works reasonably well but produces some bias, so work is underway to explore its performance in the current context and compare it with other bootstrap methods.)

Figure 1 gives 'pseudo-boxplots' for the contributions of each of the six variables. As in a conventional boxplot, the ends of the box indicate the interquartile range of the data and the line within the box marks the median. However, we used the whiskers to depict the central 95% confidence interval, rather than the trimmed range. The circles show the point estimates $(8.64,\ldots,9.86)$ given by the actual data. The figure indicates that measurements of the height on the left and right sides (left-ht and right-ht) contribute comparatively little to the Mahalanobis distance, while the distances from the inner frame to the lower border (lower) and from the inner frame to the upper border (upper) contribute noticeably more. Other firm conclusions are difficult to make, because there is substantial uncertainty as to the contributions of variables.

6.2. Confidence interval for contributions to a two-sample T^2 statistic

Example 2 involves the study of a group of 100 female athletes and a group of 102 male athletes. Nine variables were measured on each athlete and two pairs of variables were rotated to reduce collinearities. The T_2^2 statistic for comparing the two groups was calculated and gave overwhelming evidence that the groups differed. To form bootstrap confidence intervals for the contributions of individual variables to this statistic, the groups must be resampled separately - a resample consists of the measurements of 100 athletes randomly drawn with replacement from the female athletes and 102 athletes drawn with replacement from the male athletes. The T_2^2 statistic was determined for each of 100 000 resamples and the contribution of individual variables/variable combinations to the statistic in each resample was evaluated using the adapted corr-max transformation.

Pseudo boxplots derived from the results are given in Figure 2. These show that the primary contributions to the T^2 statistic are clearly from $B+S$ and $B-S$, the combination

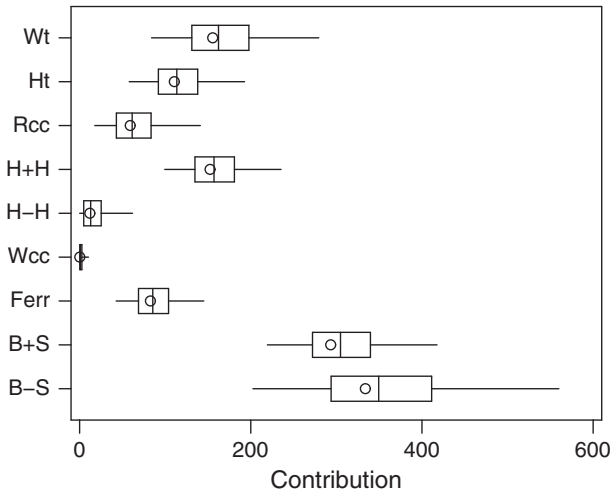

Figure 2. Confidence intervals for the contributions of individual variables to a two-sample T^2 statistic for comparing female and male athletes.

variables that are formed from the sums and differences of *Bfat* (percentage of body fat) and *SSF* (sum of skin folds). Other variables contribute noticeably less, but the only variable that clearly makes almost no contribution is the white cell blood count (*Wcc*). The confidence intervals are skewed to the right and the larger contributions tend to have wider confidence intervals. These appear to be characteristic traits and can also be seen in Figure 1.

7. Concluding comments

The corr-max transformation is straightforward to calculate. The matrices $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{D}}$ are readily determined and a spectral decomposition gives, say, $\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}} = \mathbf{H}\boldsymbol{\Psi}\mathbf{H}^\top$ where $\boldsymbol{\Psi}$ is a diagonal matrix of eigenvalues of $\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}}$ and $\mathbf{H}$ is an orthogonal matrix whose columns are eigenvectors. After $\boldsymbol{\Psi}$ and $\widehat{\boldsymbol{\Sigma}}$ have been determined, $(\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}})^{-1/2}$ is set equal to $\mathbf{H}\boldsymbol{\Psi}^{-1/2}\mathbf{H}^\top$. Hence, the corr-max transformation and the partition it yields are readily implemented in any programming language that offers matrix functions. To facilitate use of the partition in some important applications, programs have been written in R that determine the contributions of individual variables to a quadratic form in the contexts of Hotelling's one-sample and two-sample T^2 tests, Mahalanobis distance, and the classification of an item to one of two populations on the basis of Fisher's linear discriminant function. These programs are available from URL: http://users.mct.open.ac.uk/paul.garthwaite.

The rotation of variables has received much attention in this paper, and further comment is needed to give a balanced perspective on its role in partitioning a quadratic form. As in equation (5), let $\widehat{\boldsymbol{W}} = (\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}})^{-1/2}\widehat{\mathbf{D}}(\boldsymbol{X} - \boldsymbol{\mu})$. When the correlations are high between each component of $\widehat{\boldsymbol{W}}$ and the corresponding component of $\boldsymbol{X}$, then the partition is clearly a sensible way of evaluating the contribution of each x-variable. When some of these correlations are low, they can sometimes be increased dramatically through rotations that yield interpretable variables. This potential benefit of rotation was illustrated in Section 5.2. However, finding suitable rotations that yield interpretable variables is not always possible. Moreover, even when such rotations can be found, there are attractions in the simplicity of forming a

partition that retains the original x-variables. We briefly return to the athletes data to show that low correlations do not preclude a transparent relationship between the x-variables and the contributions allocated to them by the partition.

Let $\boldsymbol{X}^{\#}$ denote the difference between an athlete's measurements and the average for their gender. Put $\boldsymbol{X}^{*} = \widehat{\mathbf{D}}\boldsymbol{X}^{\#}$, so that the components of $\boldsymbol{X}^{*} = (X_1^{*},\ldots,X_9^{*})$ are standardized values of each variable. Let $(\widehat{W}_1,\ldots,\widehat{W}_9)^{\top} = \widehat{\boldsymbol{W}} = (\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}_p\widehat{\mathbf{D}})^{-1/2}\boldsymbol{X}^{*}$. Then $\delta\widehat{W}_i^2$ is the contribution of the ith variable to the quadratic form $\boldsymbol{\Theta}$ in equation (11). 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 2) to determine their contribution to $\boldsymbol{\Theta}$.

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

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 standardised 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,

$$\begin{aligned}\widehat{W}_8^2 &= \{-0.17X_1^{*} - 0.06X_2^{*} + 0.05X_3^{*} + 0.04X_4^{*} - 0.09X_5^{*} - 0.07X_6^{*} \\ &\quad - 0.07X_7^{*} + 0.84X_8^{*} + 2.43(X_8^{*} - X_9^{*})\}^2\end{aligned} \tag{22}$$

$$\begin{aligned}\widehat{W}_9^2 &= \{-0.41X_1^{*} + 0.07X_2^{*} - 0.10X_3^{*} + 0.00X_4^{*} + 0.22X_5^{*} - 0.06X_6^{*} \\ &\quad - 0.02X_7^{*} + 2.43(X_9^{*} - X_8^{*}) + 0.95X_9^{*}\}^2\end{aligned} \tag{23}$$

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, inasmuch as the large terms in (22) both involve X_8^{*} while those in (23) both involve X_9^{*}. We should note though, that while our method gives contributions to $\widehat{W}_8^2$ and $\widehat{W}_9^2$ that seem reasonable, other methods may give different values that also seem reasonable. We should also note that, in this example, the low correlation on which we focused stems from a single collinearity between just two variables. With multiple collinearities involving several variables, the relationship between the x-variables and the contributions allocated to them would be less straightforward.

Experiments are often laborious and costly 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. It can always be applied, provided that $\widehat{\boldsymbol{\Sigma}}$ is positive-definite, and yields a well-defined numerical evaluation of the contributions of individual x-variables to the quadratic farm. With multiple collinearities involving several variables, it can be difficult to judge intuitively whether an evaluation is a sensible reflection of these contributions and then the credibility of an evaluation must stem from the method used to

produce it. Our method is derived from a clear, understandable criterion that gives it a sound basis. In our experience the method has never given an evaluation that seems unreasonable and we recommend its use for the decomposition of a quadratic form for any positive-definite matrix $\widehat{\boldsymbol{\Sigma}}$. In reporting results, the method used to obtain the decomposition should be stated so as to define the evaluated contributions unambiguously.

Appendix A: Proofs of theorems

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

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

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

Proof. Let $\mathrm{var}(\boldsymbol{Y})=k\boldsymbol{\Omega}^{-1}$ where k is a scalar. Observe that $\mathrm{E}(\boldsymbol{Y}^\top\mathbf{B}\boldsymbol{Y})=\mathrm{E}[tr(\boldsymbol{Y}^\top\mathbf{B}\boldsymbol{Y})]=\mathrm{E}[tr(\mathbf{B}\boldsymbol{Y}\boldsymbol{Y}^\top)]=tr[\mathbf{B}\mathrm{E}(\boldsymbol{Y}\boldsymbol{Y}^\top)]=ktr[\mathbf{B}\boldsymbol{\Omega}^{-1}]=ktr[\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}]=ktr[(\mathbf{B}^\top)^{-1}]$. Hence we wish to maximise $tr[(\mathbf{B}^\top)^{-1}]$ under the constraint that $\mathbf{B}^\top\mathbf{B}=\boldsymbol{\Omega}$ or, equivalently, that $\mathbf{B}^{-1}(\mathbf{B}^\top)^{-1}=\boldsymbol{\Omega}^{-1}$. From Lemma 1, $(\mathbf{B}^\top)^{-1}=\boldsymbol{\Omega}^{-1/2}$, so $\mathbf{B}=\boldsymbol{\Omega}^{1/2}$.

Proof of Theorem 1. For any $\boldsymbol{X}$, by assumption $(\boldsymbol{X}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})=\boldsymbol{W}^\top\boldsymbol{W}=(\boldsymbol{X}-\boldsymbol{\mu})^\top\mathbf{A}^\top\mathbf{A}(\boldsymbol{X}-\boldsymbol{\mu})$. Choosing $\boldsymbol{X}$ so that only one entry of $(\boldsymbol{X}-\boldsymbol{\mu})$ is non-zero shows that the diagonal entries of $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{A}^\top\mathbf{A}$ are equal. Choosing $\boldsymbol{X}$ so that only two entries of $(\boldsymbol{X}-\boldsymbol{\mu})$ are non-zero then does the same for off-diagonal entries, so $\boldsymbol{\Sigma}^{-1}=\mathbf{A}^\top\mathbf{A}$. Since $\mathrm{var}(\boldsymbol{X})\propto\boldsymbol{\Sigma}$, it follows that $\mathrm{var}(\mathbf{W})\propto\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top=\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top=\mathbf{I}$. Consequently the components of $\mathbf{W}$ are uncorrelated and have identical variances.

For the next part of the theorem, let $\mathbf{W}^*=\mathbf{A}\{\boldsymbol{X}-\mathrm{E}(\boldsymbol{X})\}$ and $\mathrm{var}(\boldsymbol{X})=k\boldsymbol{\Sigma}$. Then $\mathrm{E}(\boldsymbol{W}^*)=\mathbf{0}$ and $\mathrm{var}(\boldsymbol{W}^*)=k\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top=k\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top=k\mathbf{I}$. Let $\mathbf{Z}=\mathbf{D}\{\boldsymbol{X}-\mathrm{E}(\mathbf{X})\}$ and put $\mathbf{Z}=(Z_1,\ldots,Z_m)^\top$, so that $\mathrm{E}(\mathbf{Z})=0$, $\mathrm{var}(\mathbf{Z})=k\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}$, and $\mathrm{var}(Z_i)=k$ for $i=1,\ldots,m$. We know that $\mathrm{cor}(X_i,W_i^*)=\mathrm{cor}(Z_i,W_i^*)=\mathrm{E}(Z_iW_i^*/k)$. In addition $\mathrm{cor}(X_i,W_i)=\mathrm{cor}(X_i,W_i^*)$ because $\boldsymbol{W}-\boldsymbol{W}^*=\mathbf{A}\{\mathrm{E}(\boldsymbol{X})-\boldsymbol{\mu}\}$. Thus $\Sigma_{i=1}^m\mathrm{cor}(X_i,W_i)=\mathrm{E}(\mathbf{Z}^\top\boldsymbol{W}^*/k)=\mathrm{E}(\mathbf{Z}^\top\mathbf{A}\mathbf{D}^{-1}\mathbf{Z}/k)$. Also the constraint $\mathbf{A}^\top\mathbf{A}=\boldsymbol{\Sigma}^{-1}$ is equivalent to $(\mathbf{A}\mathbf{D}^{-1})^\top(\mathbf{A}\mathbf{D}^{-1})=\mathbf{D}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{D}^{-1}$. Hence $\mathbf{A}$ must be chosen to maximise $\mathrm{E}(\mathbf{Z}^\top\mathbf{A}\mathbf{D}^{-1}\mathbf{Z})$, where $\mathrm{E}(\mathbf{Z})=0$, $\mathrm{var}(\mathbf{Z})\propto\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}$ and $(\mathbf{A}\mathbf{D}^{-1})^\top(\mathbf{A}\mathbf{D}^{-1})=(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1}$. From Lemma 2, $\mathbf{A}\mathbf{D}^{-1}=(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}$. Thus, $\mathbf{A}=(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}\mathbf{D}$ and $\mathbf{W}=(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}\times\mathbf{D}(\boldsymbol{X}-\boldsymbol{\mu})$.

Proof of Theorem 3. Let $\mathrm{var}(X)=k\boldsymbol{\Sigma}$, $\mathbf{Z}=\mathbf{D}\{\mathbf{X}-\mathrm{E}(\mathbf{X})\}$ and put $\mathbf{Z}=(Z_1,\ldots,Z_m)^\top$. It then follows that $\mathrm{cor}(X_i,W_j)=\mathrm{cor}(Z_i,W_j)$. Now $\mathrm{var}(\mathbf{Z})=k\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}$ and $\mathrm{var}(\mathbf{W})=k(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}\times(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}=k\mathbf{I}$. Also, $\mathrm{E}[\mathbf{Z}\{\boldsymbol{W}-\mathrm{E}(\boldsymbol{W})\}^\top]=\mathrm{E}[\boldsymbol{Z}\{(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}\mathbf{Z}\}^\top]=\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2}=$

$\text{var}(\boldsymbol{Z})(\mathbf{D\Sigma D})^{-1/2} = k(\mathbf{D\Sigma D})^{1/2}$. Hence $cov(Z_i, W_j)$ is the (i,j) entry of $[k(\mathbf{D\Sigma D})^{1/2}]$. Since $\text{var}(Z_i) = \text{var}(W_j) = k$, both $\text{cor}(Z_i, W_j)$ and $\text{cor}(X_i, W_j)$ equal the (i,j)th entry of $(\mathbf{D\Sigma D})^{1/2}$. Similar reasoning shows that $\text{cor}_s(X_i, \widehat{W}_j)$ is the (i,j)th entry of $(\widehat{\mathbf{D}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{D}})^{1/2}$.

Proof of Theorem 5. Part (i) follows from reasoning similar to the first steps of the proof of Theorem 1. To prove (ii), let $\mathbf{\Phi} = \mathbf{\Gamma D\Sigma D\Gamma}^\top$ and let $\boldsymbol{V} = \boldsymbol{Y} - \text{E}(\boldsymbol{Y})$, so that $\text{E}(\boldsymbol{V}) = 0$ and $\text{var}(\mathbf{V}) \propto \mathbf{\Phi}$. Put $\boldsymbol{W}^* = \mathbf{C}\boldsymbol{V}$, so that $\text{var}(\boldsymbol{W}^*) = \mathbf{I}$ since $\mathbf{C}^\top\mathbf{C} = \mathbf{\Phi}^{-1}$. It then follows that $\Sigma_{i=1}^m [\{\text{var}(Y_i)\}^{1/2}\text{cor}(Y_i, W_i^\diamond)] = \Sigma_{i=1}^m [\{\text{var}(V_i)\}^{1/2}\text{cor}(V_i, W_i^*)] = \Sigma_{i=1}^m cov(V_i, W_i^*) = \mathbf{V}^\top \boldsymbol{W}^* = \boldsymbol{V}^\top \boldsymbol{C}\boldsymbol{V}$. From Lemma 2, $\mathbf{V}^\top\mathbf{CV}$ is maximised when $\mathbf{C} = \mathbf{\Phi}^{-1/2} = (\mathbf{\Gamma D\Sigma D\Gamma}^\top)^{-1/2}$. Additionally, $(\mathbf{\Gamma D\Sigma D\Gamma}^\top)^{-1/2}$ is the unique symmetric square-root of $(\mathbf{\Gamma D\Sigma D\Gamma}^\top)^{-1} = \mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1}\mathbf{\Gamma}^\top$. As $[\mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1/2}\mathbf{\Gamma}^\top][\mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1/2}\mathbf{\Gamma}^\top] = \mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1}\mathbf{\Gamma}^\top$, it follows that $(\mathbf{\Gamma D\Sigma D\Gamma}^\top)^{-1/2} = \mathbf{\Gamma}(\mathbf{D\Sigma D})^{-1/2}\mathbf{\Gamma}^\top$. Part (iii) follows immediately from (ii) and the definition of $\mathbf{Y}$. The proof of (iv) is analogous to the proof of Theorem 3. Part (v) is immediate from equation (7).

Appendix B: Orthogonal matrices from contrasts

Suppose the first four x-axes are rotated by the transformation $(Y_1, \ldots, Y_4)^\top = \mathbf{\Gamma}_d(X_1^\diamond, \ldots, X_4^\diamond)^\top$, where

$$\mathbf{\Gamma}_d = \begin{pmatrix} 2^{-1} & 2^{-1} & -2^{-1} & -2^{-1} \\ 2^{-1/2} & -2^{-1/2} & 0 & 0 \\ 0 & 0 & 2^{-1/2} & -2^{-1/2} \\ 2^{-1} & 2^{-1} & 2^{-1} & 2^{-1} \end{pmatrix} \tag{24}$$

and $X_j^\diamond = (X_j - \mu_j)/\sigma_j$ for $j = 1\ldots, 4$, so that the $X_j^\diamond$ are standardised x-variables. It is readily checked that $\mathbf{\Gamma}_d$ is an orthogonal matrix. The new variables are

$$Y_1 = (X_1^\diamond + X_2^\diamond)/2 - (X_3^\diamond + X_4^\diamond)/2 \quad Y_2 = (X_1^\diamond - X_2^\diamond)/2^{1/2}$$
$$Y_3 = (X_3^\diamond - X_4^\diamond)/2^{1/2} \quad Y_4 = (X_1^\diamond + X_2^\diamond + X_3^\diamond + X_4^\diamond)/2$$

Thus Y_1 is the difference between the average of the first two $X^\diamond$ variables and the average of the other two $X^\diamond$ variables, Y_2 and Y_3 are each proportional to the difference between a pair of $X^\diamond$ variables, and Y_4 is proportional of the average of the four $X^\diamond$ variables. Consequently $Y_1, \ldots, Y_4$ may all be meaningful combinations of the x-variables, although this is obviously context dependent. The point of this example is that the top three rows of $\mathbf{\Gamma}_d$ form a (non-unique) set of orthogonal contrasts and contrasts can prove useful when meaningful linear combinations of variables are sought. It is certainly the case that in the analysis of experiments, contrasts among factor levels are commonly constructed for that reason.

More generally, one approach to finding rotation matrices that give meaningful new variables is to look for a set of meaningful orthonormal contrasts. Suppose the first d axes are to be rotated, so that linear combinations of $X_1^\diamond, \ldots, X_d^\diamond$ are to be formed. A complete set of orthonormal contrasts consists of $d - 1$ linear combinations

$$Y_i = \sum_{j=1}^{d} \alpha_{ij} X_j^{\diamond} \; i = 1,\ldots,d-1, \tag{25}$$

such that $\Sigma_{j=1}^{d} \alpha_{ij} = 0$, $\Sigma_{j=1}^{d} \alpha_{ij}^2 = 1$ and, for $i \neq k (k = 1,\ldots,d-1)$, $\Sigma_{j=1}^{d} \alpha_{ij}\alpha_{kj} = 0$. The set of orthonormal contrasts is not unique, giving flexibility in constructing the Y_i. To form a rotation matrix from these contrasts, we set the (i,j)th entry of $\mathbf{\Gamma}_d$ equal to α_{ij} $(i = 1,\ldots,d-1; j = 1,\ldots,d)$ and the (d,j)th entry of $\mathbf{\Gamma}_d$ equal to $d^{-1/2}$ $(j = 1,\ldots,d)$.

References

CALENGE, C., DARMON, G., BASILLE, A. & JULLIEN, J.-M. (2008). The factorial decomposition of the Mahalanobis distances in habitat selection studies. *Ecology* **89**, 555–566.

COOK, R. D. & WEISBERG, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.

DAS, P. & DATTA, S. (2007). Exploring the effects of chemical composition in hot rolled steel product using Mahalanobis distance scale under Mahalanobis-Taguchi system. *Comp. Mater. Sci.* **38**, 671–677.

FLURY, B. & RIEDWYL, H. (1988). *Multivariate Statistics: A Practical Approach*. Cambridge: Cambridge University Press.

GARTHWAITE, P. H., CRITCHLEY, F., ANAYA-IZQUIERDO, K. & MUBWANDARIKWA, E. (2012). Orthogonalization of vectors with minimal adjustment. *Biometrika* **99**, 787–798.

GARTHWAITE, P. H. & KOCH, I. (2013). Evaluating the contributions of individual variables to a quadratic form. Technical Report 13/07, Statistics Group, The Open University, UK. Available from URL: http://statistics.open.ac.uk/2013_technical_reports.

MENARD, S. (1995). *Applied Logistic Regression Analysis*. Thousand Oaks, CA: Sage.

NETER, J., WASSERMAN, W. & KUTNER, M. H. (1983). *Applied Linear Regression Models*. Illinois: Irwin.

O'BRIEN, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **41**, 673–690.

ROGERS, D. J. (2015). Dengue: recent past and future threats. *Philos. T. Roy. Soc. B* **370**, 20130562. (doi:10.1098/rstb.2013.0562).

ROTENBERRY, J. T., KNICK, S. T. & DUNN, J. E. (2002). A minimalist approach to mapping species' habitat: Pearson's planes of closest fit. In *Predicting Species Occurrences: Issues of Accuracy and Scale*, J. M. SCOTT, P. J. HEGLUND, M. L. MORRISON, J. B. HAUFLER, M. G. RAPHAELl, W. A. WALL & F. B. SAMSON eds., 281–289. Washington: Island Press.

ROTENBERRY, J. T., PRESTON, K. L. & KNICK, S. T. (2006). GIS-based niche modeling for mapping species habitat. *Ecology* **87**, 1458–1464.

TAGUCHI, G. & JUGULUM, R. (2002). *The Mahalanobis-Taguchi Strategy: A Pattern Technology System*. New York: Wiley.