ORTHOGONALIZATION OF VECTORS WITH MINIMAL ADJUSTMENT

BY PAUL H. GARTHWAITE*, FRANK CRITCHLEY AND EMMANUEL MUBWANDARIKWA†

The Open University

A new transformation is proposed that gives orthogonal components with a one-to-one correspondence between the original vectors and the components. The aim is that each component should be close to the vector with which it is paired, orthogonality imposing a constraint. If the vectors have been normalized, then the transformation minimizes the sum of the squared correlations between each vector and its associated component. The transformation has a strong dilution property: duplicating a vector, perhaps several times, has no effect on the orthogonal components that correspond to non-duplicated vectors. Properties of the transformation give it varied uses, notably as a diagnostic tool for identifying collinearities and as a method of choosing Bayesian prior weights for model averaging. These applications are described and we also discuss regression and experimental design, broader contexts in which the transformation should prove useful.

1. Introduction. Suppose a set of non-orthogonal vectors are to be adjusted by minimal amounts so as to transform them to a set of orthonormal components. Specifically, let \((x_1, \ldots, x_m)\) be a set of \(n \times 1\) vectors and put \(X = (x_1, \ldots, x_m)\). Assume \(X'X\) is non-singular. We devise a transformation of these vectors to a set of \(n\)-dimensional components \(u_1, \ldots, u_m\) such that

\[
\phi = \sum_{i=1}^{m} (x_i'u_i)^2
\]

is maximized subject to the conditions

1. \(U = XA\) for some \(m \times m\) matrix \(A\), where \(U = (u_1, \ldots, u_m)\);
2. \(U'U = I_m\), the \(m \times m\) identity matrix.
3. \(x_i'u_i > 0\) for \(i = 1, \ldots, m\).

*This work was completed while PHG was a visitor at the University of New South Wales
†Supported by a PhD studentship from the Open University

AMS 2000 subject classifications: Primary 62H99
Keywords and phrases: Collinearity, cos-square, dilution, orthogonal components, prior weights, transformation
Condition 1 implies that $u_1, \ldots, u_m$ are obtained from a linear transformation of $(x_1, \ldots, x_m)$. Condition 2 implies that $u_1, \ldots, u_m$ form a set of standardized orthogonal components.

It is anticipated that the transformation will be most useful when $x_1, \ldots, x_m$ are standardized vectors. That is, for each $i$, $x_i'$ is a unit vector. Then $x_i$ and $u_i$ each denote directional vectors in $n$-space and $x_i' u_i$ is the cosine of the angle between them. The square of this cosine gets larger as $x_i$ and $u_i$ become more co-linear. Hence in maximizing $\phi$ in equation (1), we are seeking a set of orthonormal components $u_1, \ldots, u_m$ such that $u_i$ is close to $x_i$. Note that under this transformation the $x_i$ vectors retain their identity; $x_i$ is associated with $u_i$. This contrasts with principal components or factor analysis, where a component typically relates to a number of $x$-variables and where an $x$-variable may relate to more than one component. If each $x$-vector is normalized so as to have a mean of 0 and a length of 1 (so, for each $i$, $1' x_i = 0$ and $x_i' x_i = 1$, where $1$ is an $n \times 1$ vector of 1's), then $x_i' u_i$ is the correlation between $x_i$ and $u_i$. Thus when the $x$-vectors are normalized, the transformation finds orthogonal components that maximize the squared correlation between each $x$-vector and its associated component.

We will refer to this transformation as the cos-square transformation. It has an unusual and useful property. Suppose the set of $x$-vectors is increased by adding duplicates of one of them. For definiteness, let $x_{m+1}, \ldots, x_{m+k}$ be virtual duplicates of $x_m$. (We assume $x_m, x_{m+1}, \ldots, x_{m+k}$ all differ very slightly so that $(X^*)' X^*$ is positive-definite, where $X^* = (x_1, \ldots, x_{m+k})$.) Then the first $m - 1$ orthogonal components, $u_1, \ldots, u_{m-1}$, are the same when the transformation is applied to the set of vectors $x_1, \ldots, x_m$ as when it is applied to the augmented set of vectors $x_1, \ldots, x_{m+k}$. That is, under the cos-square transformation, duplicating a vector does not change the orthogonal components associated with the other vectors. We refer to this property as the strong dilution property.

In Section 2, an algorithm to perform the transformation is described and properties of the transformation and examined. There are varied applications where the transformation will prove useful and some are described in Section 3. They include diagnostics to detect collinearities, regression on orthogonal components, and the selection of Bayesian prior weights for weighted model averaging.

2. The transformation. In this section an algorithm is given for finding a matrix $A$ that maximizes $\phi$ in equation (1) subject to conditions 1 and 2.

Theorem 1. Suppose $C$ is a diagonal $m \times m$ matrix of positive constants
and let $c_i$ denote its $(i, i)$ element $(i = 1, \ldots, m)$. Let $Q\Lambda Q' = XC'XC$ be the spectral decomposition of $XC'XC$, where $\Lambda$ is the diagonal matrix of eigenvalues and $Q$ is the matrix whose columns are the corresponding normalized eigenvectors. If $U = XA$ satisfies conditions 1–3, then $\sum_{i=1}^{m} c_i x'_i u_i$ is maximized when $A = CQ\Lambda^{-1/2}Q'$.

**Proof.** Put

$$B = \Lambda^{1/2}Q'C^{-1}AQ,$$

where $\Lambda^{1/2}$ denotes a diagonal matrix with non-negative diagonal elements and $\Lambda^{1/2}\Lambda^{1/2} = \Lambda$. Then $B'B = QA'C^{-1}Q\Lambda Q'C^{-1}AQ = QA'X'XAQ = Q'U'UQ$. Now $Q$ is an orthogonal matrix. Hence, $B'B = I_m$ if and only if $U'U = I_m$. Thus condition 2 is equivalent to $B'B = I_m$. We seek to maximize $\sum_{i=1}^{m} c_i x'_i u_i$ under the constraint $B'B = I_m$. Thus, $\text{trace}(\Lambda^{1/2}B)$ is to be maximized under the constraint $B'B = I_m$. This has solution $B = I_m$ so, from equation (2), $A = CQ\Lambda^{-1/2}Q'$.

The following iterative algorithm exploits Theorem 1 to find $u_1, \ldots, u_m$ that maximize $\sum_{i=1}^{m} (x'_i u_i)^2$ under conditions 1–3. The algorithm repeatedly maximizes $\sum_{i=1}^{m} c_i x'_i u_i$ until convergence, at each iteration setting each $c_i$ equal to the most recent estimate of $x'_i u_i$.

**Algorithm**

1. Set $C_1$ equal to the $m \times m$ identity matrix and put $i = 1$.
2. At the $i$th iteration, perform a spectral decomposition of $C_iX'XC_i$, giving

$$C_iX'XC_i = Q_i\Lambda_i Q_i',$$

where, for $j = 1, \ldots, m$, the $j$th column of $Q_i$ is the normalized eigenvector corresponding to the eigenvalue in the $(j, j)$ element of $\Lambda_i$.
3. Put $E_i = C_i^{-1}Q_i\Lambda_i^{1/2}Q_i'$.
4. Set $C_{i+1}$ equal to a diagonal matrix, with diagonal equal to the diagonal of $E_i$.
5. Return to step (2) until convergence, when $C_{i+1} \approx C_i$.
6. Set $(u_1, \ldots, u_m)$ equal to $XC_iQ_i\Lambda_i^{-1/2}Q_i'$.

**Remarks on Algorithm**

1. We will refer to $C$, $Q$ and $\Lambda$ as the cosine matrix, eigenvector matrix and eigenvalue matrix of the transformation, where these are the values

```
of \( C_i, Q_i \) and \( \Lambda_i \) at convergence. We also refer to \( \langle C, Q, \Lambda \rangle \) as the matrix triple of the transformation.

2. If \( A_i = C_i Q_i A_i^{-1/2} Q_i' \) and \( U_i = X A_i \), then \( X' U_i = C_i^{-1} C_i X' X C_i Q_i A_i^{-1/2} Q_i' = C_i^{-1} Q_i A_i Q_i' Q_i A_i^{-1/2} Q_i' = C_i^{-1} Q_i A_i^{1/2} Q_i' \). Hence, step (3) sets \( E_i \) equal to \( X' U_i \), so the diagonal elements of \( C_i+1 \) and \( X' U_i \) are equal.

3. Suppose the diagonal elements of \( C_i \) are positive. Then the diagonal elements of \( C_i^{-1} Q_i A_i^{1/2} Q_i' \), and hence those of \( C_i+1 \) (step 4), are also positive. Since \( C_1 \) has positive diagonal elements (step 1), so do all \( C_i \), which is a condition in Theorem 1.

4. The following theorem gives convergence properties of the algorithm.

**Theorem 2.** If \( X' X \) is a positive-definite matrix, then the algorithm for the cos-square transformation of \( x_1, \ldots, x_m \) converges. At convergence \( \sum_{j=1}^m (x_j' u_j)^2 \) is maximized subject to conditions 1–3 and the matrix triple \( \langle C, Q, \Lambda \rangle \) is unique, apart from the ordering of eigenvectors and eigenvalues in \( Q \) and \( \Lambda \).

**Proof.** At the \( i \)th iteration of the algorithm, \( C_{i+1} \) is chosen to maximise \( \text{trace}(C_i C_{i+1}) \). As the option of setting \( C_{i+1} \) equal to \( C_i \) is available, \( \text{trace}(C_i C_{i+1}) \) is monotonic non-decreasing as \( i \) increases. Also, \( \text{trace}(C_i C_{i+1}) \) is bounded above by \( \sum_{j=1}^m (x_j' w_j)^2 \), where \( w_j = x_j/\|x_j\| \). Hence the algorithm converges.

The optimal components are \( u_1, \ldots, u_m \). If the diagonal elements of \( C_i \) equalled \( x_1' u_1, \ldots, x_m' u_m \), then at the next iteration \( C_{i+1} \) would be set equal to \( C_i \) and the algorithm will have converged. Hence one value of \( C \) for which the algorithm converges is when the true optimum is reached. We must show it is unique.

The conditions for convergence are (i) \( Q \Lambda Q' \) is the spectral decomposition of \( CX' X C \) and (ii) diagonal \( \langle C \rangle = \text{diagonal}(C^{-1} Q \Lambda^{1/2} Q') \). Let \( c_1, \ldots, c_m \) and \( d_1, \ldots, d_m \) denote the diagonal elements of \( C \) and \( Q \Lambda^{1/2} Q' \), respectively. From condition (i), any element of \( Q \Lambda^{1/2} Q' \) has an expansion of the form \( \sum_j f_j(X) c_j^{(j)} \ldots c_j^{(m)} \) where \( \sum_{j=1}^m e_{jk} = 1 \). As \( c_j > 0 \) for all \( i \), there is at most one set of values \( c_1, \ldots, c_m \) that satisfy the set of equations \( c_j^2 = d_j \) for \( j = 1, \ldots, m \). Thus there is at most one matrix \( C \) for which

\[
\text{diagonal}(C C) = \text{diagonal}(Q \Lambda^{1/2} Q').
\]

Condition (ii) implies that (3) must hold at convergence and we know the algorithm converges. In consequence, there is a unique \( C \) that satisfies (3) and, using condition (i), the matrix triple \( \langle C, Q, \Lambda \rangle \) is unique, apart from the ordering of eigenvectors and eigenvalues in \( Q \) and \( \Lambda \). \( \square \)
ORTHOGONALIZATION OF VECTORS

5

REMARKS ON TRANSFORMATION

1. The transformation matrix is

\[
A = C Q \Lambda^{-1/2} Q
\]

and the transformation can be applied to any \(1 \times m\) row vector or \(k \times m\) matrix, simply by post-multiplying the vector/matrix by \(A\).

2. The transformation is determined by \(X'X\), rather than \(X\). It will often be convenient to base the transformation on the covariance matrix obtained from \(X'X\), which results in \(A\) being scaled by \(1/(n-1)\) but does not change \(C, Q\) or \(u_1, \ldots, u_m\). If standardization of \(x_1, \ldots, x_m\) is required, the transformation should be based on the correlation matrix.

3. The strong dilution property. In an appendix we prove the following theorem, which shows that the cos-square transformation has the strong dilution property. That is, when a vector is duplicated (perhaps many times) the components corresponding to the non-duplicated vectors are unchanged. The theorem also gives related invariance properties.

THEOREM 3. Suppose that \(k\) duplicates \(x_{m+1}, \ldots, x_{m+k}\) are added to the set of vectors \(x_1, \ldots, x_m\). Let \(X^* = (x_1, \ldots, x_{m+k})\). Assume the duplicates are virtually identical to \(x_m\) while \((X^*)'X^*\) is positive-definite. Let \(u_1, \ldots, u_m\) and \(u_1^*, \ldots, u_{m+k}^*\) denote the components resulting from the cosine transformation when it is applied to \((x_1, \ldots, x_m)\) and \((x_1, \ldots, x_{m+k})\), respectively. Also, let \(A\) and \(A^*\) be the corresponding transformation matrices. Then, as \(x_{m+j} \to x_m\) for \(j = 1, \ldots, k\),

\[\begin{align*}
(i) & \quad u_i^* \to u_i \text{ for } i = 1, \ldots, m-1, \\
(ii) & \quad \sum_{i=1}^{m+k} (x_i' u_i^*)^2 \to \sum_{i=1}^{m} (x_i' u_i)^2, \text{ and} \\
(iii) & \quad A = \begin{pmatrix} A_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \text{ and } A^* = \begin{pmatrix} A_{11}^* & A_{12}^* \\ A_{21}^* & A_{22}^* \end{pmatrix},
\end{align*}\]

where \(A_{11}\) and \(A_{11}^*\) are \((m-1) \times (m-1)\) matrices, then \(A_{11}^* \to A_{11}\).

Regarding the components that correspond to the duplicates of \(x_m\), the determinant of \(U'U\) equals \(|A'X'XA| = |A'| |X'X| |A|\). Hence if \(X'X\) is singular (so that \(|X'X| = 0\)), \(U'U\) cannot equal the identity matrix. It follows that the components \(u_{m+1}, \ldots, u_{m+k}\) become ill-defined at the limit when \(x_m = x_{m+1} = \ldots = x_{m+k}\).

For \(i = 1, \ldots, m\) let \(p_i = (x_i' u_i)^2/\sum_{j=1}^{m} (x_j' u_j)^2\) and, for \(i = 1, \ldots, m + k\), let \(p_i^* = (x_i' u_i^*)^2/\sum_{j=1}^{m+k} (x_j' u_j^*)^2\). From Theorem 3, as \(x_{m+j} \to x_m\) for \(j = \ldots,
for \( i = 1, \ldots, m - 1 \). Also, \( \mathbf{p} = (p_1, \ldots, p_m)' \) and \( \mathbf{p}^* = (p_1^*, \ldots, p_{m+k}^*)' \) each have positive components that sum to 1, so they could be probability vectors. This standardization can be useful in applications (cf. Section 3.2). A different standardization is obtained by putting 
\[
 v_i = \frac{x_i'}{u_i} / \left( \sum_{j=1}^{m+1} (x_j' u_j)^2 \right)^{1/2}
\]
for \( i = 1, \ldots, m \) and 
\[
 v_i^* = \frac{x_i'u_i^*}{(\sum_{j=1}^{m+k} (x_j'u_j^*)^2)^{1/2}}
\]
for \( i = 1, \ldots, m+k \). Then 
\[
 v_i^* \rightarrow v_i \quad \text{for} \quad i = 1, \ldots, m - 1,
\]
while \( \mathbf{v} = (v_1, \ldots, v_m)' \) and \( \mathbf{v}^* = (v_1^*, \ldots, v_{m+k}^*)' \) are each of unit length.

4. Applications.

4.1. Detection and identification of collinearities. A widely recommended approach for detecting collinearities is to calculate a variance inflation factor for each column of \( \mathbf{X} \) [Farrar and Glauber (1967)]. Suppose \( x_1, \ldots, x_m \) are observations of variables \( X_1, \ldots, X_m \) and let \( R_i^2 \) denote the multiple correlation coefficient when \( X_i \) is regressed on the other \( X \)-variables. The variance inflation factor for \( X_i \), \( VIF_i \) say, is defined to be
\[
 VIF_i = (1 - R_i^2)^{-1}.
\]
\( VIF_i \) will be large if \( X_i \) is involved in a collinearity. Hence, the VIF may be calculated for each variable and used as a diagnostic to identify which variables are involved in collinearities. However, the VIFs do not indicate the number of collinearities or directly identify which variables are associated with each of them [Wetherill (1986, p. 87)]. The most common approach for identifying which variables form a collinearity is to determine the eigenvectors and eigenvalues of \( \mathbf{X}' \mathbf{X} \). Near-zero eigenvalues imply a collinearity and, in principle, the corresponding eigenvectors identify the \( X \)-variables involved; components of the eigenvector that are large in magnitude should correspond to those \( X \)-variables that are most influencing the collinearity.

As in Section 2, let \( \mathbf{U} = \mathbf{X} \mathbf{A} \), where \( \mathbf{A} \) is the transformation matrix for the cos-square transformation of \( \mathbf{X} \) and \( \mathbf{U} = (u_1, \ldots, u_m) \). Then \( \mathbf{A} \) may be used both as a diagnostic for determining the number of collinearities and as a means of identifying the variables that contribute to each collinearity. The transformation aims to find orthogonal vectors \( u_1, \ldots, u_m \), where \( u_i \) links strongly to \( x_i \) but not to the other \( x \)-vectors. Consequently, most of the off-diagonal elements of \( \mathbf{A} \) will be close to 0. However, if there is a collinearity between \( k \) of the \( X \)-variables, then the space these variables comfortably
span will have a smaller dimension than \( k \), making it difficult to transform them to \( k \) orthogonal \( u \)-vectors, so some diagonal elements of \( A \) will be quite far from 0. Thus a collinearity is indicated by large off-diagonal elements.

Suppose now that the \( X \)-variables had been standardized so that \( X'X \) is a correlation matrix. Then the VIFs may be determined from the transformation matrix. It is well known that \( VIF_i \) equals the \((i, i)\) diagonal element of \( (X'X)^{-1} \) when \( X'X \) is a correlation matrix [e.g. Farrar and Glauber (1967)].

Also, \( A = CQ\Lambda^{-1/2}Q \), so \( AA' = C(Q\Lambda^{-1}Q')C = C(C^{-1}(X'X)^{-1}C^{-1})C = (X'X)^{-1} \). Thus if \( a_i' \) is the \( i \)th row of \( A \), then

\[ VIF_i = a_i'a_i. \]

That is, the rows of \( A \) determine the VIFs, with large values of \( a_i'a_i \) indicating a collinearity. Moreover, most components of \( a_i \) will typically be small; the values that have greater magnitude cause \( VIF_i \) to be large (obviously) and correspond to the variables that form the collinearity.

Examining \( A \) provides more information about collinearities than examining the eigenvectors that correspond to small eigenvalues. This is because a collinearity relates to only one eigenvector, whereas each variable forming a collinearity corresponds to a separate row-vector in \( A \).

As an example we examine data on 180 pitprops cut from Corsican pine [Jeffers 1967]. The data have been widely analyzed, most commonly in the context of principal components analysis, where components are formed from 13 physical variables. Table 1 gives the sample correlation matrix for these variables, \( X_1, \ldots, X_{13} \). There are quite strong correlations between \( X_1 \) and \( X_2 \), between \( X_3 \) and \( X_4 \), and between \( X_6 \) and \( X_7 \), as well as a number of moderate correlations. Table 2 gives the transformation matrix when \( X'X \) is set equal to this correlation matrix. Values above 0.5 are given in bold-face type. The last column of the table gives the VIF for each variable; e.g. 13.71 is the VIF for \( X_2 \) and equals \( a_2'a_2 \). A VIF value above 10 is often treated as indicative of a collinearity [Neter (1983, p. 392)] and these values are also given in bold-face. On this basis, there is a collinearity between \( X_1 \) and \( X_2 \), and another between \( X_3 \) and \( X_4 \). From the components of \( a_3, a_4 \) and \( a_5 \), there is a suggestion that the latter collinearity also involves \( X_5 \); from practical considerations that is plausible, as \( X_4 \) and \( X_5 \) are the specific gravity of a pitprop before and after being oven-dried, while \( X_3 \) is the initial moisture content of the prop. \( X_7 \) also has a VIF above 10. It seems to be moderately collinear with \( X_6 \) and, to a lesser extent, \( X_{10} \).

In comparison, eigenvalues and eigenvectors provide more limited information about collinearities. The smallest three eigenvalues for the pitprop data are 0.05, 0.04 and 0.04, and the next
<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
<th>X11</th>
<th>X12</th>
<th>X13</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td>0.954</td>
<td>0.364</td>
<td>0.342</td>
<td>0.313</td>
<td>0.496</td>
<td>0.424</td>
<td>0.592</td>
<td>0.545</td>
<td>0.084</td>
<td>-0.019</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td></td>
<td>0.969</td>
<td>0.297</td>
<td>0.284</td>
<td>0.291</td>
<td>0.503</td>
<td>0.419</td>
<td>0.648</td>
<td>0.569</td>
<td>0.076</td>
<td>-0.036</td>
<td>0.144</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td></td>
<td></td>
<td>0.882</td>
<td>0.875</td>
<td>0.883</td>
<td>0.813</td>
<td>0.411</td>
<td>0.125</td>
<td>0.081</td>
<td>0.162</td>
<td>0.220</td>
<td>0.126</td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td></td>
<td></td>
<td></td>
<td>0.220</td>
<td>0.153</td>
<td>0.174</td>
<td>0.059</td>
<td>0.137</td>
<td>0.014</td>
<td>0.097</td>
<td>0.149</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.381</td>
<td>0.296</td>
<td>0.004</td>
<td>0.211</td>
<td>0.274</td>
<td>0.061</td>
<td>0.024</td>
<td>0.184</td>
<td></td>
</tr>
<tr>
<td>X6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.372</td>
<td>0.090</td>
<td>0.465</td>
<td>0.675</td>
<td>0.036</td>
<td>-0.232</td>
<td></td>
</tr>
<tr>
<td>X7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.372</td>
<td>0.090</td>
<td>0.465</td>
<td>0.675</td>
<td>0.036</td>
<td></td>
</tr>
<tr>
<td>X8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.** Correlation matrix for the physical properties of props.
<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
<th>$X_9$</th>
<th>$X_{10}$</th>
<th>$X_{11}$</th>
<th>$X_{12}$</th>
<th>$X_{13}$</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>3.04</td>
<td>-1.92</td>
<td>-0.23</td>
<td>-0.09</td>
<td>0.10</td>
<td>-0.07</td>
<td>-0.22</td>
<td>-0.17</td>
<td>-0.01</td>
<td>-0.21</td>
<td>-0.07</td>
<td>-0.06</td>
<td>-0.15</td>
<td>13.14</td>
</tr>
<tr>
<td>$a_2$</td>
<td>-1.88</td>
<td>3.14</td>
<td>0.00</td>
<td>-0.05</td>
<td>0.04</td>
<td>0.01</td>
<td>-0.22</td>
<td>-0.00</td>
<td>-0.33</td>
<td>-0.28</td>
<td>-0.08</td>
<td>-0.05</td>
<td>-0.21</td>
<td>13.71</td>
</tr>
<tr>
<td>$a_3$</td>
<td>-0.26</td>
<td>0.01</td>
<td>2.69</td>
<td>-2.00</td>
<td>0.51</td>
<td>0.02</td>
<td>0.28</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.08</td>
<td>-0.06</td>
<td>-0.03</td>
<td>0.01</td>
<td>11.66</td>
</tr>
<tr>
<td>$a_4$</td>
<td>-0.10</td>
<td>-0.06</td>
<td>-1.92</td>
<td>2.88</td>
<td>-0.57</td>
<td>-0.31</td>
<td>0.03</td>
<td>0.10</td>
<td>-0.03</td>
<td>-0.02</td>
<td>0.00</td>
<td>-0.04</td>
<td>0.01</td>
<td>12.42</td>
</tr>
<tr>
<td>$a_5$</td>
<td>0.13</td>
<td>0.06</td>
<td>0.60</td>
<td>-0.70</td>
<td>1.27</td>
<td>-0.13</td>
<td>-0.13</td>
<td>-0.03</td>
<td>0.02</td>
<td>0.07</td>
<td>0.04</td>
<td>0.09</td>
<td>0.03</td>
<td>2.53</td>
</tr>
<tr>
<td>$a_6$</td>
<td>-0.09</td>
<td>0.02</td>
<td>0.02</td>
<td>-0.34</td>
<td>-0.12</td>
<td>1.99</td>
<td>-1.64</td>
<td>0.04</td>
<td>0.08</td>
<td>0.36</td>
<td>0.03</td>
<td>-0.13</td>
<td>0.09</td>
<td>6.93</td>
</tr>
<tr>
<td>$a_7$</td>
<td>-0.21</td>
<td>-0.21</td>
<td>0.23</td>
<td>0.03</td>
<td>-0.09</td>
<td>-1.26</td>
<td>3.12</td>
<td>0.01</td>
<td>-0.18</td>
<td>-0.71</td>
<td>-0.02</td>
<td>0.13</td>
<td>0.23</td>
<td>12.03</td>
</tr>
<tr>
<td>$a_8$</td>
<td>-0.22</td>
<td>-0.00</td>
<td>-0.03</td>
<td>0.12</td>
<td>-0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
<td>1.28</td>
<td>-0.18</td>
<td>-0.27</td>
<td>-0.09</td>
<td>0.15</td>
<td>0.09</td>
</tr>
<tr>
<td>$a_9$</td>
<td>-0.01</td>
<td>-0.42</td>
<td>-0.02</td>
<td>-0.04</td>
<td>0.02</td>
<td>0.08</td>
<td>-0.24</td>
<td>-0.18</td>
<td>1.34</td>
<td>-0.15</td>
<td>-0.07</td>
<td>-0.02</td>
<td>0.01</td>
<td>2.10</td>
</tr>
<tr>
<td>$a_{10}$</td>
<td>-0.23</td>
<td>-0.32</td>
<td>0.08</td>
<td>-0.02</td>
<td>0.06</td>
<td>0.33</td>
<td>-0.85</td>
<td>-0.24</td>
<td>-0.13</td>
<td>1.96</td>
<td>0.36</td>
<td>0.17</td>
<td>0.19</td>
<td>5.12</td>
</tr>
<tr>
<td>$a_{11}$</td>
<td>-0.09</td>
<td>-0.12</td>
<td>-0.06</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.03</td>
<td>-0.09</td>
<td>-0.08</td>
<td>0.44</td>
<td>1.12</td>
<td>0.05</td>
<td>0.09</td>
<td>1.51</td>
</tr>
<tr>
<td>$a_{12}$</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.04</td>
<td>-0.05</td>
<td>0.09</td>
<td>-0.15</td>
<td>0.19</td>
<td>0.16</td>
<td>-0.02</td>
<td>0.21</td>
<td>0.05</td>
<td>1.13</td>
<td>-0.01</td>
<td>1.43</td>
</tr>
<tr>
<td>$a_{13}$</td>
<td>-0.20</td>
<td>-0.29</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.10</td>
<td>0.33</td>
<td>0.10</td>
<td>0.01</td>
<td>0.23</td>
<td>0.09</td>
<td>-0.01</td>
<td>1.21</td>
<td>1.77</td>
</tr>
</tbody>
</table>
smallest is 0.19. Hence there appear to be three collinearities. The
eigenvectors corresponding to the three smallest eigenvalues are:
\((-0.00, -0.05, 0.12, -0.02, 0.01, -0.54, 0.76, 0.03, -0.05, -0.32, -0.05, 0.05, 0.04), (0.39, -0.41, 0.53, -0.59, 0.20, 0.08, -0.04, -0.05, 0.05, 0.06, 0.00, 0.00, 0.01)\) and \((-0.57, 0.58, 0.41, -0.38, 0.12, 0.06, 0.00, 0.02, -0.06, 0.00, -0.01, 0.00, -0.01)\). The first of these eigenvectors suggests a collinearity between
\(X_6, X_7\) and (perhaps) \(X_{10}\). This concurs with one of the VIF findings. The
other two eigenvectors both suggest a collinearity that involves \(X_1, X_2, X_3\)
and \(X_4\) but, in contrast to the VIF analysis, they do not suggest that there
is one collinearity between \(X_1\) and \(X_2\) and another between \(X_3\) and \(X_4\).

The transformation matrix of the cos-square transformation is only one
of many ways of decomposing the VIF. Indeed, suppose the quantity to
be maximized in equation (1) is \(\sum_{i=1}^{m} (x_i^Tu_i)^\nu\) for any \(\nu \geq 1\), rather than
\(\sum_{i=1}^{m} (x_i^Tu_i)^2\). Let \(A^\#\) denote the resulting transformation matrix. In the
notation of Theorem 1, \(A^\# = C^{\nu-1}Q\Lambda^{-1/2}Q'\) where \(Q\Lambda Q' = C^{\nu-1}X'XC^{\nu-1}\).
Thus \(A^\#(A^\#)' = (X'X)^{-1}\), so \(A^\#\) is also a decomposition of the VIF. Consequentially, some additional criterion is needed to choose \(\nu\). One criterion is
to require the transformation to have property (iii) of Theorem 3 so that, if
one \(X\)-variable is duplicated, the relationship between the other \(X\)-variables
and the VIF is unaffected, which seems appropriate. This favours setting \(\nu\)
equal to 2, which results in the cos-square transformation.

4.2. Bayesian prior weights and other weighting schemes. The
application that motivated the present work arises in Bayesian model averaging.
Suppose we have models \(M_1, \ldots, M_m\) and that \(p_i\) is the prior probability
(prior weight) that \(M_i\) is the true model. Suppose also that we have data \(D\)
and \(f_i(D \mid M_i)\) is the likelihood of the data if \(M_i\) is true. Then in forming a
Bayesian weighted model average, the posterior weight given to \(M_i\) is

\[
(8) \quad w_i = \frac{p_i f_i(D \mid M_i)}{\sum_{j=1}^{m} p_j f_j(D \mid M_j)},
\]

for \(i = 1, \ldots, m\). As this formula shows, the influence of the prior prob-
obabilities \(p_1, \ldots, p_m\) do not dissipate as data are gathered, but instead have
a multiplicative effect on the \(w_i\). Hence, these prior probabilities should be
chosen with care.

In practice, the most common choice for the \(p_i\) is to set each of them equal
to \(1/m\). However, this does not seem the best choice if some models are very
similar to each other while other models are quite distinct. In regression, for
example, if a small subset of the explanatory variables are believed impor-
tant, then many models might be constructed by combining that subset of
variables with two or three of the other explanatory variables. Then giving
the same prior probability to each model would strongly favour the subset
of variables believed important. Instead, the probabilities given to models
that are similar should be reduced because of that similarity. Otherwise
too little probability may be placed “...on good, but unique models, as a
consequence of massing excess probability on large sets of bad, but similar
models” [Chipman et al. (2001, p. 79)].

Garthwaite and Mubwandarikwa (2008) suggest using the cos-square trans-
formation to choose prior probabilities (prior weights). Let \( M_1, \ldots, M_m \) form
the set of models to which prior probabilities must be assigned. They sup-
pose an \( m \times m \) correlation matrix \( \mathbf{R} \) is determined that reflects the simi-
larity between models. For example, \( \mathbf{R} \) might be formed from the correla-
tions between predictions given by the models at a range of design points.
The matrix triple \( \langle \mathbf{C}, \mathbf{Q}, \mathbf{\Lambda} \rangle \) is determined for the cos-square transformation
with \( \mathbf{X}'\mathbf{X} \) set equal to \( \mathbf{R} \). If \( c_1, \ldots, c_m \) are the diagonal of \( \mathbf{C} \), then \( M_i \) is
given the prior probability \( p_i = c_i^2 / \sum_{j=1}^{m} c_j^2 \). [In the notation of Theorem 3,
\( p_i \) also equals \( (\mathbf{x}_i'\mathbf{u}_i)^2 / \sum_{j=1}^{m} (\mathbf{x}_j'\mathbf{u}_j)^2 \).] Garthwaite and Mubwandarikwa call
this method the cos-square weighting scheme. It has a dilution property: models
that are more highly correlated with other models receive a smaller
prior probability. As an illustration, suppose five models \( M_1, \ldots, M_5 \) have
the following correlation matrix:

\[
\begin{array}{cccccc}
M_1 & 1 & 0.9 & 0.8 & 0.7 & 0.6 \\
M_2 & 0.9 & 1 & 0.7 & 0.6 & 0.5 \\
M_3 & 0.8 & 0.7 & 1 & 0.5 & 0.4 \\
M_4 & 0.7 & 0.6 & 0.5 & 1 & 0.3 \\
M_5 & 0.6 & 0.5 & 0.4 & 0.3 & 1 \\
\end{array}
\]

Then the prior probabilities given to these models are 0.10, 0.19, 0.22, 0.24
and 0.25, respectively, so that each model receives a larger probability than
any model that has higher correlations than it has, and vice-versa.

A related issue concerns how prior probabilities should be revised if the
set of models is augmented by adding models that are virtually identical to
a model already in the set. George (1999), Clyde (1999), and Chipman et
al (2001) argue that the prior probability given to a model that has been
duplicated should be divided between that model and its duplicates, while
the prior probabilities given to other models should be unchanged. Following
George (1999), this is referred to as “diluting” the prior to compensate for
the presence of similar models and it is termed the strong dilution property
where a regression problem initially involves two uncorrelated predictors $X_1$ and $X_2$. These yield three regression models $\{X_1\}$, $\{X_2\}$ and $\{X_1, X_2\}$. A new predictor variable is introduced, $X_3$, and this variable is very highly correlated with $X_2$, but not with $X_3$. It is argued that the probability allocated to $\{X_1\}$ should be unchanged when $X_3$ is introduced as a potential predictor, whereas the probability allocated to $\{X_2\}$ and $\{X_1, X_2\}$ should be “diluted” across all the new models containing $X_3$. George (1999, p. 176) writes,

“$X_3$ has not really added any new models to the mix. Instead, models containing $X_3$ are merely equivalent substitutes for the corresponding models containing $X_2$. Introducing $X_3$ has essentially resulted in relabelling a set of equivalent models. The probability of such a set should not increase as a result of this relabelling, and it is dilution that prevents this from happening.”

From equation (5), the cos-square weighting scheme has the strong dilution property. For example, suppose a model $M_6$ is added to the models whose correlation matrix is given above. If $M_6$ were virtually identical to $M_5$, then the prior probabilities given to $M_1, \ldots, M_4$ would be unchanged, while the probability of 0.25 previously given to $M_5$ would be divided between $M_5$ and $M_6$.

The cos-square transformation is also useful for forming weighted averages in contexts other than choosing Bayesian prior weights. For example, suppose a variety of different measurements is made on each of a set of objects and a univariate similarity (or dissimilarity) measure is to be constructed from both these measurements and, perhaps, functions of these measures that are thought to be useful, such as ratios, products and linear combinations. There may well be high correlations between some quantities and it seems sensible to take these correlations into account in forming a similarity measure. The cos-square transformation would yield suitable weights based just on the correlation structure, though the weights would generally need to be adjusted to reflect other factors. Conflict can arise between wanting to use all available variables and wanting to avoid the over-representation of some features because of correlations between variables. Using a mechanism that takes due account of correlations avoids the problem.

4.3. Regression and Design. In multiple regression, there is a preference for explanatory variables that are orthogonal. Methods such as principal component regression, latent root regression and partial least squares are advocated not only for their dimension reduction properties, but also because they give orthogonal components. An obvious advantage of orthogonal explanatory variables is that the relationship between each explanatory variable and the dependent variable is not affected by which other explanatory
variables are included in the model. For example, forward stepwise regression and backwards stepwise regression will lead to the same model if predictors are orthogonal.

Strategies have been proposed for using regression on orthogonal components to select variables. For example, Jolliffe (1972, 1973) gives methods of selecting variables following principal component regression. However, methods that give orthogonal components do not, in general, lend themselves naturally to the task of variable selection. An exception is the cos-square transformation, since it retains a one-to-one correspondence between the original variables and the orthogonal components.

The pitprop data reported by Jeffers (1967) contained a dependent variable, the maximum compression stress of a prop ($Y$), as well as the thirteen $X$-variables whose correlation matrix is given in Table 1. Correlations between $Y$ and the $X$-variables are given in Jeffers (1967), Mardia et al. (1979, p. 178) and elsewhere. If $Y$ is regressed against the orthogonal components $u_1, \ldots, u_{13}$ obtained from the cos-square transformation, the proportions of the variance of $Y$ accounted for by each of these components (times 100%) are 7.00, 0.26, 41.15, 7.92, 2.34, 3.12, 1.82, 6.51, 1.35, 0.41, 0.05, 0.67 and 0.49. Jeffers (1967) and Mardia et al. (1979, p.246) both performed a principal components regression for the pitprops data. Jeffers selected five principal components which together accounted for 63.97% of the variation in $Y$ while Mardia et al. selected eight principal components which together accounted for 72.79 of this variation. By comparison, the best five predictors from $u_1, \ldots, u_{13}$ account for 65.70% of the variation in $Y$ (slightly better than best five principal components) and the best eight predictors account for 71.21% (slightly poorer than the best eight principal components). Hence, there is disparity as whether the cos-square orthogonal components or an equal number of principal components make better predictors but, in any case, differences seem small. Regarding variable selection, the variables corresponding to the best five cos-square components are $X_1$, $X_3$, $X_4$, $X_6$ and $X_8$, which form a good set of predictor variables; Mardia et al (1979, p. 178) give $(X_1, X_3, X_6, X_8, X_{11})$ as the optimal set of five predictors. However, differences are greater when the number of predictor variables is increased from five to eight: three of the variables chosen by the cos-square transformation are not in the optimum set of eight predictor variables.

Experimental design is another area where orthogonalization is often sought. The cos-square transformation should prove useful if a design is to be adjusted to make it orthogonal and small adjustment is preferred. As an example, suppose $n$ units are meant to represent a population of $m$-dimensional objects. Then $n$ objects might be sampled from the popula-
tion and taken as representative of it. However, if \( \mathbf{x}_i \) denotes the vector of sample values for the \( i \)th dimension \((i = 1, \ldots, m)\), then \( \mathbf{x}_1, \ldots, \mathbf{x}_m \) are unlikely to be a set of orthogonal vectors. If orthogonality is required, then the cos-square transformation could be applied to \( \mathbf{x}_1, \ldots, \mathbf{x}_m \) to obtain (after rescaling) orthogonal components that are similar to \( \mathbf{x}_1, \ldots, \mathbf{x}_m \).

**APPENDIX A: PROOF OF THEOREM 3**

The vector \( \mathbf{x}_m \) will be duplicated. Partitioning \( \mathbf{X}'\mathbf{X} \) to separate \( \mathbf{x}_m \) from \( \mathbf{x}_1, \ldots, \mathbf{x}_{m-1} \), put

\[
\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{r} \\ \mathbf{r}' & r_m \end{pmatrix},
\]

where \( \mathbf{R}_{11} \) is a \((m-1) \times (m-1)\) matrix. Duplicates \( \mathbf{x}_{m+1}, \ldots, \mathbf{x}_{m+k} \) are added to \( \mathbf{x}_1, \ldots, \mathbf{x}_m \) to form the matrix \( \mathbf{X}^* = (\mathbf{x}_1, \ldots, \mathbf{x}_{m+k}) \). Each duplicate differs from \( \mathbf{x}_m \) by only a small amount and we put \( \mathbf{x}_{m+i} = \mathbf{x}_m + \alpha \xi_i \) for \( i = 1, \ldots, k \). We are interested in the transformation of \( \mathbf{x}_1, \ldots, \mathbf{x}_{m+k} \) as \( \alpha \to 0 \).

Write \((\mathbf{X}^*)'\mathbf{X}^*\) as

\[
(\mathbf{X}^*)'\mathbf{X}^* = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} + O(\alpha) \\ \mathbf{R}_{12}' + O(\alpha) & \mathbf{R}_{22} + O(\alpha) \end{pmatrix},
\]

where each column of \( \mathbf{R}_{12} \) equals \( \mathbf{r} \) and each element of \( \mathbf{R}_{22} \) is \( r_m \). We assume \((\mathbf{X}^*)'\mathbf{X}^*\) is a positive-definite matrix for any \( \alpha > 0 \).

Let \( \langle \mathbf{C}^*, \mathbf{Q}^*, \Lambda^* \rangle \) denote the matrix triple at convergence when the algorithm for the transformation is applied to \( \mathbf{x}_1, \ldots, \mathbf{x}_{m+k} \). \( \mathbf{C}^*, \mathbf{Q}^* \) and \( \Lambda^* \) are \((m+k) \times (m+k)\) matrices that depend on \( \alpha \). Put

\[
\mathbf{C}^* = \begin{pmatrix} \mathbf{C}_{1}^{(m-1)\times(m-1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{2}^{(k+1)\times(k+1)} \end{pmatrix},
\]

and

\[
\mathbf{Q}^* = \begin{pmatrix} \mathbf{Q}_{11}^{(m-1)\times m} & \mathbf{Q}_{12}^{(m-1)\times k} \\ \mathbf{Q}_{21}^{(k+1)\times m} & \mathbf{Q}_{22}^{(k+1)\times k} \end{pmatrix},
\]

where \( \mathbf{0} \)'s denote vectors or matrices of 0s; their dimensions will only be specified if they are not clear from the context. Since, \( \mathbf{x}_{m+i} = \mathbf{x}_m + O(\alpha) \) for \( i = 1, \ldots, k \), ordering the eigenvalues from largest to smallest and we can put

\[
\Lambda^* = \begin{pmatrix} \Lambda_1^{m \times m} & \mathbf{0} \\ \mathbf{0}' & \Lambda_2^{k \times k} \end{pmatrix}.
\]
where each diagonal element \( \Lambda_2 \) is of order \( O(\alpha) \). \( \Lambda^* \) and \( C^* \) are diagonal matrices. Let \( c_0, \ldots, c_k \) denote the diagonal elements of \( C_2 \).

**Proposition 1.** Choose \( q = (q_1, \ldots, q_m)' \) so that the first row of \( Q_{21} \) is \( c_0q \). Then, as \( \alpha \to 0 \), the \((i+1)\)th row of \( Q_{21} \to c_iq' \) for \( i = 0, \ldots, k \).

**Proof.** For \( j = 1, \ldots, m \), let \( s_j \) and \( t_j \) be the \( j \)th columns of \( Q_{11} \) and \( Q_{21} \), respectively. We have that \( C^*(X^*)'X^*(s', t')' = (\lambda_j s_j', \lambda_j t_j')' \), where \( \lambda_j \) is the \( j \)th eigenvalue of \( \Lambda^* \). Hence, as \( \alpha \to 0 \),

\[
C_2 R'_{12} C_1 s_j + C_2 R'_{22} C_2 t_j \to \lambda_j t_j.
\]

Now the rows of \( R'_{12} \) are identical to each other, as are those of \( R'_{22} \). Hence, elements in the vector \( R'_{12} C_1 s_j \) are identical and so are those in \( R'_{22} C_2 t_j \). Consequently, if \( t_j = (t_{j0}, \ldots, t_{jk})' \), then \( t_{ji} t_{j0} \to c_i/c_0 \) for \( i = 0, \ldots, k \). As \( c_i/c_0 \) does not depend on which column of \( Q_{21} \) was chosen as \( t_j \), Proposition 1 follows.

Define

\[
C = \begin{pmatrix} C_1 & 0 \\ 0 & c_m \end{pmatrix}, \quad \text{and} \quad Q = \begin{pmatrix} Q_{11} \\ c_m q' \end{pmatrix},
\]

where \( c_m = (\sum_{i=0}^{k} \frac{1}{c_i})^{1/2} \). Also, let \( c = (c_0, \ldots, c_k)' \).

**Proposition 2.** As \( \alpha \to 0 \), the spectral decomposition of \( CX'XC \to Q \Lambda \alpha Q' \), where \( X = (x_1, \ldots, x_m) \).

**Proof.** Let \( \alpha \to 0 \). From Proposition 1, if \( t_j \) is the \( j \)th column of \( Q_{21} \), then \( t_j \approx q_j c \) and the \((i+1)\)th row of equation (11) gives \( c_i r C_1 s_j + c_i r c' t_j \approx c_i q_j \). Now \( c't_j \approx c_m^2 q_j \), so

\[
r C_1 s_j + r c_m^2 q_j \to \lambda_j q_j.
\]

From \( C^*(X^*)'X^*(s', t')' = (\lambda_j s_j', \lambda_j t_j')' \), we also have that \( C_1 R_{11} C_1 s_j + C_1 R_{12} C_2 t_j \to \lambda_j s_j \). As \( R_{12} C_2 t_j = r(c' t_j) \approx r(c_m^2 q_j) \),

\[
C_1 R_{11} C_1 s_j + C_1 r c_m^2 q_j \to \lambda_j t_j.
\]

From equations (13) and (14), as \( \alpha \to 0 \),

\[
\begin{pmatrix} C_1 & 0 \\ 0 & c_m \end{pmatrix} \begin{pmatrix} C_1 & 0 \\ 0 & c_m \end{pmatrix} \begin{pmatrix} r_{11} & r \\ r' & r_m \end{pmatrix} \begin{pmatrix} s_j \\ c_m q_j \end{pmatrix} \to \begin{pmatrix} \lambda_j s_j \\ \lambda_j c_m q_j \end{pmatrix}.
\]
Using equation (9), as \( \alpha \to 0 \), \( \lambda_j \) is an eigenvalue of \( CX'XC \) and \((s'_j, c_mq_j)'\) is its corresponding eigenvector. Now, \((s'_j, c_mq_j)(s'_j, c_mq_j)' = s'_js_j + c_m^2q_j^2 = s'_js_j + (q_jc')c(q_j) \approx s'_js_j + t_j't_j \). As \((s'_j, t'_j)'\) is the \( j \)th column of the orthogonal matrix \( Q' \), it follows that \((s'_j, c_mq_j)(s'_j, c_mq_j)' \approx 1 \). Hence the \( j \)th column of \( Q' \) [i.e. \((s'_j, c_mq_j)\)] is a standardized eigenvector as \( \alpha \to 0 \).

**Proposition 3.** As \( \alpha \to 0 \), diagonal(C) \( \to \) diagonal\((C^{-1}Q\Lambda_1^{1/2}Q')\).

**Proof.** As \((C^*, Q^*, \Lambda^*)\) is the matrix triple for the transformation of \( x_1, \ldots, x_{m+k} \), diagonal\((C^*C^*)\) = diagonal\((Q^*(\Lambda^*)^{1/2}(Q^*)')\). Hence as \( \alpha \to 0 \),

\[
\text{diagonal} \left( \begin{array}{cc}
C_1C_1 & 0 \\
0 & C_2C_2
\end{array} \right) \to \text{diagonal} \left( \begin{array}{cc}
Q_{11}\Lambda_1^{1/2}Q_{11}' & Q_{11}\Lambda_1^{1/2}Q_{21}' \\
Q_{21}\Lambda_1^{1/2}Q_{11}' & Q_{21}\Lambda_1^{1/2}Q_{21}'
\end{array} \right),
\]

so

\[
\text{diagonal}(C_1C_1) \to \text{diagonal}(Q_{11}\Lambda_1^{1/2}Q_{11}').
\]

Also, \( \text{trace}(C_2C_2) \to \text{trace}(Q_{21}\Lambda_1^{1/2}Q_{21}) \). Now \( \text{trace}(C_2C_2) = \sum_{i=0}^{k} c_i^2 = c_m^2 \) and \( \text{trace}(Q_{21}\Lambda_1^{1/2}Q_{21}) \approx \sum_{i=0}^{k} c_i q_i^* \Lambda_1^{1/2} q_i = (\sum_{i=0}^{k} c_i^2) q^* \Lambda_1^{1/2} q = c_m^2 q^* \Lambda_1^{1/2} q \). Thus, as \( \alpha \to 0 \),

\[
c_m^2 \to c_m^2 q^* \Lambda_1^{1/2} q.
\]

From equations (17) and (18), as \( \alpha \to 0 \),

\[
\text{diagonal} \left( \begin{array}{cc}
C_1C_1 & 0 \\
0 & c_m^2
\end{array} \right) \to \text{diagonal} \left( \begin{array}{cc}
Q_{11}\Lambda_1^{1/2}Q_{11}' & c_m q^* \Lambda_1^{1/2} q \\
c_m q^* \Lambda_1^{1/2} q & c_m q^* \Lambda_1^{1/2} q
\end{array} \right).
\]

There is a unique set of matrices that satisfy the conditions to be the matrix triple for the transformation of \( x_1, \ldots, x_m \). Propositions 2 and 3 show that \((C, Q, \Lambda_1)\) is that matrix triple as \( \alpha \to 0 \).

**Proposition 4.** Let \( \theta = q^* \Lambda_1^{1/2} q \). Then the \((m-1) \times (m-1)\) matrix \( Q_{11}(\Lambda_1^{1/2} - \Lambda_1^{1/2} q q^* \Lambda_1^{1/2} / \theta)Q_{11}' \) is of full rank.

**Proof.** Let \( \Gamma = \Lambda_1^{1/2} - \Lambda_1^{1/2} q q^* \Lambda_1^{1/2} / \theta \). As \( \Lambda_1^{1/2} \) is of rank \( m \) and \( \Lambda_1^{1/2} q q^* \Lambda_1^{1/2} \) is of rank 1, the rank of \( \Gamma \) is at least \( m - 1 \). Hence, as \( \Gamma \)
is an $m \times m$ matrix, if $g_1$ and $g_2$ are two non-zero vectors such that $g_1^t \Gamma = g_2^t \Gamma = 0$, then $g_1$ must be proportional to $g_2$. Now $q^t \Gamma = q^t A_1^{1/2} - q^t \Lambda_1^{1/2} = 0$, so for any vector $g$, if $g^t Q_{11} \Gamma = 0$, then $g^t Q_{11} = 0$ or $g^t Q_{11} = bq$ for some non-zero constant $b$. Since $(Q_{11}^t, q)^t$ is an orthogonal matrix, $(g^t, b)(Q_{11}^t, q)^t \neq 0$. Hence, if $g^t Q_{11} \Gamma = 0$, then $g^t Q_{11} = 0$. But if $g^t Q_{11} = 0$, then $(g^t, 0)(Q_{11}^t, q)^t = 0$, implying $g = 0$. Hence, if $g^t Q_{11} \Gamma = 0$, then $g = 0$, so $Q_{11} \Gamma$ is of rank $m - 1$. Since $Q_{11} \Gamma q = 0$, a similar argument shows that $Q_{11} \Gamma Q_{11}^t$ is of rank $m - 1$.

**Proposition 5.** If $u_1, \ldots, u_{m}$ and $u_1^*, \ldots, u_{m+k}^*$ denote the components resulting from the cos-square transformation when it is applied to $(x_1, \ldots, x_m)$ and $(x_1, \ldots, x_{m+k})$, respectively, then $u_i \rightarrow u_i^*$ for $i = 1, \ldots, m - 1$, as $x_{m+j} \rightarrow x_m$ for $j = 1, \ldots, k$.

**Proof.** Put $X_1 = (x_1, \ldots, x_{m-1})$, $X_2 = (x_m, \ldots, x_{m+k})$, $U_1^* = (u_1^*, \ldots, u_{m-1}^*)$, $U_2^* = (u_m^*, \ldots, u_{m+k}^*)$ and $U_1 = (u_1, \ldots, u_{m-1})$. From Theorem 1, $(U_1^*, U_2^*) = (X_1, X_2)A^*$, where $A^* = CQ^*(A^*)^{-1/2}(Q^*)^t$. Hence

$$ (X_1, X_2) = (U_1^*, U_2^*)Q^*(A^*)^{-1/2}(Q^*)^t(C^*)^{-1}. $$

Let $\alpha \rightarrow 0$. Then

$$ Q^*(A^*)^{1/2}(Q^*)^t(C^*)^{-1} \approx \begin{pmatrix} Q_{11}A_1^{1/2}Q_{11}'C_1^{-1} & Q_{11}A_1^{1/2}Q_{21}C_2^{-1} \\ Q_{21}A_1^{1/2}Q_{11}'C_1^{-1} & Q_{21}A_1^{1/2}Q_{21}C_2^{-1} \end{pmatrix}, $$

and each column of $Q_{21}C_2^{-1} \rightarrow q$. Also, each column of $X_2 \rightarrow x_m$. Hence, from (19) and (20),

$$ x_m \approx U_1^*Q_{11}A_1^{1/2}q + U_2^*Q_{21}A_1^{1/2}q $$

and

$$ X_1 \approx U_1^*Q_{11}A_1^{1/2}Q_{11}'C_1^{-1} + U_2^*Q_{21}A_1^{1/2}Q_{11}'C_1^{-1}. $$

From Proposition 1, $U_2^*Q_{21} \approx U_2^*(c_0, \ldots, c_k)^t q'$. Hence, from (21),

$$ U_2^*(c_0, \ldots, c_k)^t \approx (x_m - U_1^*Q_{11}A_1^{1/2}q)/\theta $$

where, as in Proposition 4, $\theta = q'A_1^{1/2}q$. Hence, from (22),

$$ X_1 \approx U_1^*Q_{11}A_1^{1/2}Q_{11}'C_1^{-1} + (x_m - U_1^*Q_{11}A_1^{1/2}q)q'A_1^{1/2}Q_{11}'C_1^{-1}/\theta $$

$$ = U_1^*Q_{11}(A_1^{1/2} - A_1^{1/2}qq'A_1^{1/2}/\theta)Q_{11}'C_1^{-1} + x_m q'A_1^{1/2}Q_{11}'C_1^{-1}/\theta. $$
Similar computations from the equation $(U_1, u_m) = (X_1, x_m)A$, where $A = CQ\Lambda_1^{-1/2}Q'$, give

\begin{equation}
(24) \quad x_m \approx U_1Q_{11}\Lambda_1^{1/2}q + c_mu_m\theta
\end{equation}

and

\begin{equation}
(25) \quad X_1 \approx U_1Q_{11}\Lambda_1^{1/2}Q_{11}'C^{-1} + c_mu_mq'\Lambda_1^{1/2}Q_{11}'C^{-1}.
\end{equation}

Using (24) to substitute for $c_mu_m$ in (25),

\begin{equation}
(26) \quad X_1 \approx U_1Q_{11}(\Lambda_1^{1/2} - \Lambda_1^{1/2}qq'\Lambda_1^{1/2}/\theta)Q_{11}'C^{-1} + x_mq'\Lambda_1^{1/2}Q_{11}'C^{-1}/\theta.
\end{equation}

From (23) and (26),

\begin{equation}
0 \approx (U_1^*-U_1)Q_{11}(\Lambda_1^{1/2} - \Lambda_1^{1/2}qq'\Lambda_1^{1/2}/\theta)Q_{11}'C^{-1}.
\end{equation}

From Proposition 4, $Q_{11}(\Lambda_1^{1/2} - \Lambda_1^{1/2}qq'\Lambda_1^{1/2}/\theta)Q_{11}'C^{-1}$ is of full rank. Hence $U_1^*-U_1 \approx 0$. \hfill $\square$

**Proposition 6.** As $\alpha \to 0$, $Q_{12}\Lambda_2^{-1/2}Q_{12}' \to 0$.

**Proof.** Put $Z = (x_m, \xi_1, \ldots, \xi_k)$ and

\[ H_\alpha = \begin{pmatrix} 1 & 1' \\ 0 & \alpha I_k \end{pmatrix} \]

where $I_k$ is the $k \times k$ identity matrix and $1$ is a $k \times 1$ vector of 1’s. Then $H_\alpha$ is non-singular and $X_2 = ZH_\alpha$. As $X^* = (X_1, X_2)$,

\[ [(X^*)'X^*]^{-1} = \begin{pmatrix} X_1'X_1 & X_1'ZH_\alpha \\ H_\alpha'Z'X_1 & H_\alpha'Z'ZH_\alpha \end{pmatrix}^{-1}. \]

Let $\Gamma_{11}$ denote the top-left $m \times m$ sub-matrix of $[(X^*)'X^*]^{-1}$. Then $\Gamma_{11} = (X_1'X_1 - X_1'ZH_\alpha(H_\alpha'Z'ZH_\alpha)^{-1}H_\alpha'Z'X_1)^{-1} = (X_1'X_1 - X_1'Z(Z'Z)^{-1}Z'X_1)^{-1}$. Hence $\Gamma_{11}$ does not depend on $\alpha$. Also, $[(X^*)'X^*]^{-1} = C^*Q^*\Lambda^*(Q^*)'C^*$, so $\Gamma_{11} = C_1[Q_{11}\Lambda_1^{-1}Q_{11}' + Q_{12}\Lambda_2^{-1}Q_{12}']C_1$. As $\Gamma_{11}$ does not depend on $\alpha$ and both $C_1$ and $Q_{11}\Lambda_1^{-1}Q_{11}'$ are finite, $Q_{12}\Lambda_2^{-1}Q_{12}'$ is finite for all $\alpha$. If $q_{ij}$ denotes the $(i, j)$ element of $Q^*$, the $i$th diagonal element of $Q_{12}\Lambda_2^{-1}Q_{12}'$ is

\begin{equation}
(27) \quad \sum_{j=1}^{k} q_{ij}^2/(\lambda_{m+j})
\end{equation}
for \(i = 1, \ldots, (m - 1)\). Each term in the summation in (27) is positive so each term is finite. Since \(\lambda_{m+j} \to 0\) as \(\alpha \to 0\),

\[
(28) \quad q_{i(m+j)}/\lambda_{m+j}^{1/2} \to 0
\]
as \(\alpha \to 0\) for \(i = 1, \ldots, (m - 1)\); \(j = 1, \ldots, k\). The \((h, i)\) element of \(Q_{12}A^{-1/2}_2 Q_{12}'\) is

\[
\sum_{j=1}^{k} q_{h(m+j)}/\lambda_{m+j}^{1/2} = \sum_{j=1}^{k} [(q_{h(m+j)}/\lambda_{m+j}^{1/2})(q_{i(m+j)}/\lambda_{m+j}^{1/2})]^{1/2}.
\]

Hence, from (28), \(Q_{12}A^{-1/2}_2 Q_{12}' \to 0\) as \(\alpha \to 0\).

Proposition 5 is part (i) of Theorem 3. By definition, \(c_m = (\sum_{i=0}^{k} c_i^2)^{1/2}\) so \(\sum_{i=m}^{m+k}(x_i' u_i)^2 \to (x_m' u_m)^2\). Consequently, \(\sum_{i=m}^{m+k}(x_i' u_i)^2 \to \sum_{i=1}^{m}(x_i' u_i)^2\), giving (ii). As \(A = CQ\Lambda_1^{-1/2}Q'\), it follows from equation (13) that \(A_{11} = C_1 Q_{11} A_1^{-1/2} Q_{11}'\). Also, \(A^* = C^* Q^*(A^*)^{-1/2}(Q^*)'\) so \(A^*_{11} = C_1 (Q_{11} A_1^{-1/2} Q_{11}' + Q_{12} A_2^{-1/2} Q_{12}')\). Thus, from Proposition 6, \(A^*_{11} \to A_{11}\) as \(\alpha \to 0\). This gives part (iii), completing the proof of the theorem.

REFERENCES


