# Correspondence Analysis for Principal Components Transformation of Multispectral and Hyperspectral Digital Images

James R. Carr and Korblaah Matanawi

## Abstract

*Correspondence analysis is introduced for principal components transformation of multispectral and hyperspectral digital images. This method relies on squared deviations between pixel values and their expected values (joint probabilities computed as the product of the sum of all pixels in one spectral band and the sum of pixel values across all bands at a given pixel position). Correspondence analysis is applied to a multispectral SPOT High Resolution Visible (HRV) image of Eleuthera, Bahamas. Correspondence analysis, principal components analysis, and factor analysis (standardized principal components) yield similar transformations. Correspondence analysis, however, compresses more image variance into fewer principal components. For the particular SPOT HRV scene chosen, correspondence analysis captures 96 percent of the original image variance in its first principal component. Used in a lossy image compression algorithm to reconstruct the original set of three SPOT HRV images, this first principal component from correspondence analysis restores spectral content better than does principal components analysis.*

## Introduction

Principal components transformation, also referred to as a Hotelling transform, of multispectral images is well known (e.g., Schowengerdt, 1997; Gonzalez and Woods, 1992; Jensen, 1996). The typical application relies on within-band variance/between-band covariance matrices, eigenvectors of which are used to rotate (transform) original image bands such that each rotated band is statistically orthogonal to (independent of) all other bands. Another application, called standardized principal components (Singh and Harrison, 1985), uses within- and between-band correlation coefficients to assemble a matrix that is eigen decomposed. Although called standardized principal components in literature related to remote sensing, this method is often known as factor analysis (see reviews in Davis (1973), Davis (1986), and Carr (1995)), referred to as such throughout this manuscript to avoid confusion with principal components analysis. Factor analysis has been used in digital image processing to minimize the numerical influence of bands having a larger variance. Realistically, principal components and factor analysis are similar because (co)variance is directly proportional to correlation coefficient.

A multivariate method of relatively recent advent, correspondence analysis (Benzecri, 1973), differs from principal components and factor analysis by using a chi-square metric for representing inter- and intra-variable relationships. For certain types of images and/or individual scenes, correspondence analysis may provide better discrimination among intercorrelated spectral bands when compared to principal components or factor analysis. Correspondence analysis may be a better method when applied to ratio data, such as pixel values, or measurements in general (Greenacre, 1984).

Correspondence analysis is more a geometrical method than a statistical one. Factor analysis and principal components analysis, for instance, are statistical methods. Both are functions of covariance among the multiple variables, where covariance is the metric for closeness (similarity). Correspondence analysis uses a different measure of closeness between variables. A chi-square metric is used to measure distance between pixel vectors of order $M$, where $M$ is the number of spectral bands. In this sense, correspondence analysis is a geometric method because its main concern is the orientation of [pixel] vectors in $M$-dimensional space, and how close these vectors are to one another. A detailed discussion of the historical and theoretical background of correspondence analysis can be found in Greenacre (1984).

A subsequent application is used to emphasize the similarities and differences among principal components analysis, factor analysis, and correspondence analysis. The primary intent of this paper is to introduce correspondence analysis for principal components transformation of multispectral and hyperspectral digital images. Depending on the image type and scene, it may yield results that are superior to those from the other two methods. An extension of correspondence analysis to digital image compression is also suggested.

## Algorithms

### Principal Components Analysis

Given $M$ intercorrelated image bands, principal components analysis proceeds by assembling a matrix $[\mathbf{S}]$, $M$ by $M$ in size and symmetrical, such that $\mathbf{S}(i,i)$ is the variance for band $i$, and $\mathbf{S}(i,j) = \mathbf{S}(j,i)$ is the covariance between bands $i$ and $j$. Variance can be computed as

$$
\begin{aligned}
\text{VAR(BV)} &= \frac{1}{N-1}\left[\sum_{i=1}^{N} \text{BV}_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N}\text{BV}_i\right)^2\right] \\
&= \frac{1}{N-1}\left[N\sum_{i=1}^{N}\text{BV}_i^2 - \left(\sum_{i=1}^{N}\text{BV}_i\right)^2\right] \quad (1)
\end{aligned}
$$

J. Carr is with the Department of Geological Sciences/172, University of Nevada, Reno, NV 89557 (carr@unr.edu).

K. Matanawi is with World Vision, Sudan Program, P.O. Box 56527, Nairobi, Kenya.

in which $N$ is the total number of pixels, BV, collectively representing a particular image band. This equation compares to that for covariance between two image bands, $BV_1$ and $BV_2$: i.e.,

$$\text{Cov}(BV_1, BV_2) = \frac{1}{N(N-1)}\left[N\sum_{i=1}^{N} BV_{1,i}\,BV_{2,i} - \sum_{i=1}^{N} BV_{1,i}\sum_{i=1}^{N} BV_{2,i}\right] \quad (2)$$

where $N$ is again the total number of pixels in each image band.

Suppose the BVs in a Landsat TM scene are to be rotated (transformed) using principal components analysis of all seven ($M$) bands. In this case, a matrix $S$ of size 7 by 7 is computed. Diagonal entries of this matrix consist of the seven variances for each band. Additionally, there are 21 off-diagonal entries, symmetrical above and below the diagonal, representing the covariances for all possible two-band combinations from the group of seven bands.

### Factor Analysis (Standardized Principal Components Analysis)
Again, assume a multispectral image consisting of $M$ spectral bands. A matrix $S$ of size $M$ by $M$ is computed such that $S(i,j) = S(j,i) = r$, the correlation coefficient between bands, $i$ and $j$. For consistency with previous equations, let the pixels in band $i$ be represented by $BV_i$, and those for band $j$ be represented by $BV_j$. Then

$$r_{ij} = \frac{\text{COV}(BV_i\,BV_j)}{(S_{BV_i})(S_{BV_j})} \quad (3)$$

In this equation, $S$ represents the standard deviation of a discrete sample. Moreover, if $i = j$, then $r = 1$; thus, all diagonal entries of $S$ are equal to 1.

Given a seven-band Landsat TM scene, $S$ is a 7 by 7 matrix. The seven diagonal entries are each equal to 1. The 21 unique off-diagonal entries, symmetrical above and below the diagonal, are equal to correlation coefficients for all possible two-band combinations (from the original group of seven bands). These correlation coefficients range numerically between $-1$ and 1, inclusive.

### Correspondence Analysis
In this development, a hypothetical notion of an image is used to demonstrate mathematical concepts. Suppose a 512 by 512, seven-band Landsat TM image is to be transformed using correspondence analysis. Each band consists of 512 by 512 pixels, or a total of 262,144 pixels. Let a matrix $Y$ be formed that is $N$ by $M$ in size; $N$, in this case, is 262,144, and $M$ is 7. Notice that $Y$ is simply the total collection of all pixels involved. Further, let the total sum of all entries in $Y$ be called $u$. The following steps are completed to yield a square, symmetrical matrix $S$ that is eigen decomposed to yield the image transformation (for more detail and computer programs, see Carr (1990; 1995; 1998)):

Step 1: normalize $Y$ by $u$: $Y' = Y/u$;
Step 2: form an $N$ by 1 vector $W$, each entry of which represents the sum of each row of $Y'$; each entry in $W$ is the total sum of [normalized] pixel values across all spectral bands considered for a given pixel position;
Step 3: form an $M$ by 1 vector $T$, each entry of which represents the sum of each column of $Y'$; each entry in $T$ is the sum of all [normalized] pixel values in a given spectral band;
Step 4: although a matrix $S$ of size $M$ by $M$ can be formed from two intermediate matrices, $S_c$ and $S_r$, such that

$$S_{c,ij} = \frac{Y_{ij}}{W_i\sqrt{T_j}}; \qquad S_{r,ij} = \frac{Y_{ij}}{\sqrt{T_j}}. \quad (4)$$

This is an inefficient approach with respect to computer memory; instead, the matrix $S$ can alternatively be formed as follows, an approach that more directly implies the chi-square analogy (Davis, 1986; Carr, 1995):

$$S_{jk} = \sum_{i=1}^{N}\left(\frac{Y_{ij} - W_iT_j}{\sqrt{W_iT_j}}\right)\left(\frac{Y_{ik} - W_iT_k}{\sqrt{W_iT_k}}\right) \quad (5)$$

For example, if $i = j = k = 1$, then

$$S_{11} = \frac{(Y_{11} - W_1T_1)^2}{W_1T_1}$$

and comparing to the equation for the chi-square statistic:

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i}$$

in which $O$ is an observed value and $E$ is the expected value, the analogy is realized if one considers that $Y$ is observed, and the product $WT$ is the expected value of $Y$. When using the second approach in correspondence analysis for computing the matrix $S$, off diagonal entries can become negative. The second approach for forming $S$ is used in this paper and in Carr (1998). As a final note, correspondence analysis always reduces the dimensional space by at least one. If $M$ image bands are considered, only $M - 1$ eigenvalues, at the most, will be significant. This has important implications for digital image compression, a concept that is explored later in this paper.

## Application
A portion (400 rows by 400 pixels per row) of a SPOT High Resolution Visible (HRV) image of Eleuthera, Bahamas (Figure 1) acquired on 21 April 1986 is chosen for an example comparison among principal components analysis, factor analysis, and correspondence analysis. The three multispectral channels of the SPOT HRV sensor are used: channel 1 (0.5 to 0.59 μm), channel 2 (0.61 to 0.68 μm), and channel 3 (0.79 to 0.89 μm). Matrices $S$, obtained from each multivariate method, are listed (Table 1), along with their associated eigenvalues and eigenvectors (Table 2).

Principal components analysis and factor analysis are identical with respect to the amount of original image variance captured (represented) in each principal component: 70 percent in the first, 28 percent in the second, and 2 percent in the third principal component. For this particular SPOT HRV scene, there seems to be no need to standardize image variance using correlation coefficients in a principal components analysis because the eigen decomposition results are identical from principal components analysis and factor analysis.

TABLE 1. MATRICES **S**, FOR PRINCIPAL COMPONENTS, FACTOR, AND CORRESPONDENCE ANALYSIS

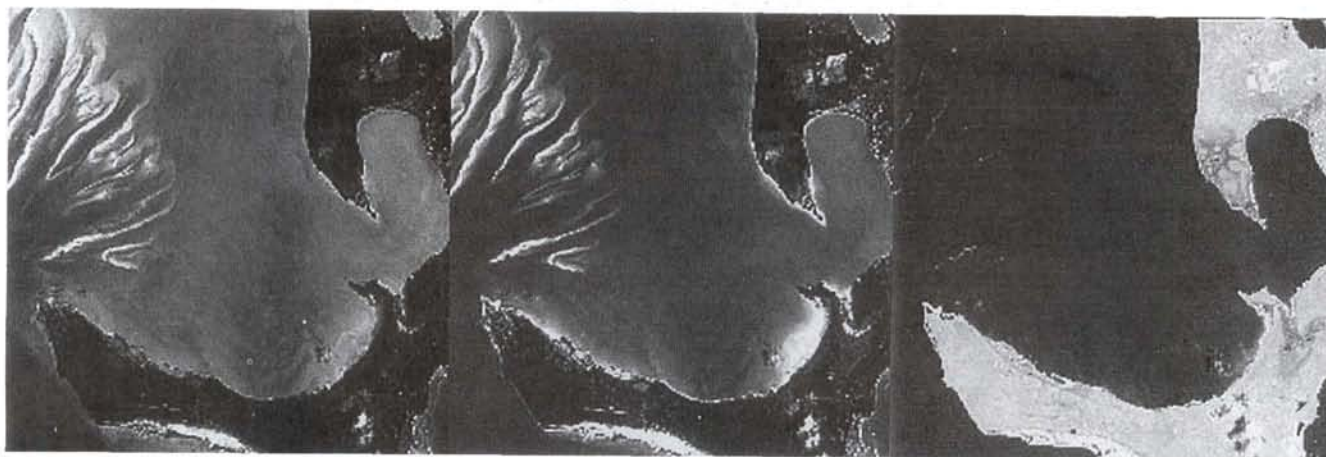| Principal Components Analysis | | |
|---|---|---|
| 2261.6 | 1546.1 | −1775.5 |
| 1546.1 | 1641.4 | −432.0 |
| −1775.5 | −432.0 | 3834.3 |
| Factor Analysis | | |
| 1.0000 | 0.8025 | −0.6029 |
| 0.8025 | 1.0000 | −0.1722 |
| −0.6029 | −0.1722 | 1.0000 |
| Correspondence Analysis | | |
| 0.0875 | 0.0370 | −0.1211 |
| 0.0370 | 0.0267 | −0.0619 |
| −0.1211 | −0.0619 | 0.1780 |

Figure 1. Mosaic of SPOT HRV spectral images of Eleuthera, Bahamas acquired on 21 April 1986. These images represent a subsampled, 400 row by 400 pixel section of the original scene. Left: Channel 1 (0.5 to 0.59 μm); center: Channel 2 (0.61 to 0.68 μm); right: Channel 3 (0.79 to 0.89 μm).

In comparison, correspondence analysis captures 96 percent of the original image variance in its first principal component. The second principal component captures most of the remaining 4 percent of the total image variance, whereas the third component represents a negligible amount of variance. The large amount of variance captured by correspondence analysis in its first principal component is used later in a discussion on digital image compression.

Eigenvectors (Table 2) from all three methods are used to transform the original suite of three bands to three mutually orthogonal (statistically independent) bands, mosaics of which are shown (Figures 2 and 3; a specific figure developed using factor analysis is not presented because its transformation is visually identical to that from principal components analysis). No scaling was used to develop the displays (Figures 2 and 3), except to transform pixel values to a range, [0, 255], to facilitate display.

Mathematically, the transformation was implemented as follows. Each of the SPOT HRV spectral band images was loaded as a single column in a matrix $O$, an $N$ by $M$ image where $N$ is the total number of pixels in each band image, and $M$ is the number of spectral bands. Eigenvectors (Table 2) from each method were loaded as columns into a matrix $E$. The transformation was then obtained as a product: $T =$

$OE$, and each column in $T$ is one of the transformed images, $N$ total pixels in size.

Principal components analysis, factor analysis, and correspondence analysis yield visually similar results for the first principal component. The second principal component image from correspondence analysis is visually similar to the third principal component image from both principal components analysis and factor analysis, probably because approximately the same amount of original image variance is being represented. The third principal component image from correspondence analysis is shown, but represents negligible variance and is not comparable to any principal component image from the other two methods.

### Extension to Digital Image Compression

Given that $OE = T$ effects the principal components transformation, then a reconstruction is afforded by reversing this process: $TE^{-1} = O$, in which $E^{-1}$ is the inverse of the eigenvector matrix. If all $M$ columns (principal component images) in $T$ are used in this process, $O$ is reconstructed exactly with no loss. In this case, $T$ and $O$ have the same information content (the same size).

A lossy reconstruction (one with error) is possible using only the most significant principal component, or compo-

TABLE 2. EIGEN DECOMPOSITION RESULTS

| | | Eigenvectors | | |
| --- | --- | --- | --- | --- |
| | | Vector 1 | Vector 2 | Vector 3 |
| Principal Components: | | −0.779 | 0.664 | 2.679 |
| | | −0.435 | 1.112 | −2.499 |
| | | 1.000 | 1.000 | 1.000 |
| | Eigenvalues: | 5463.7 | 2204.3 | 157.8 | Sum = 7825.8 |
| | Variance: | (70%) | (28%) | (2%) |
| Factor Analysis: | | −1.458 | 0.063 | 1.970 |
| | | −1.230 | 0.738 | −1.522 |
| | | 1.000 | 1.000 | 1.000 |
| | Eigenvalues: | 2.09 | 0.84 | 0.07 | Sum = 3.0 |
| | Variance: | (70%) | (28%) | (2%) |
| Correspondence Analysis: | | −0.687 | 3.948 | 0.975 |
| | | −0.342 | −5.008 | 0.968 |
| | | 1.000 | 1.000 | 1.000 |
| | Eigenvalues: | 0.28 | 0.0099 | 5.04E-15 | Sum = 0.29 |
| | Variance: | (96%) | (4%) | (0%) |
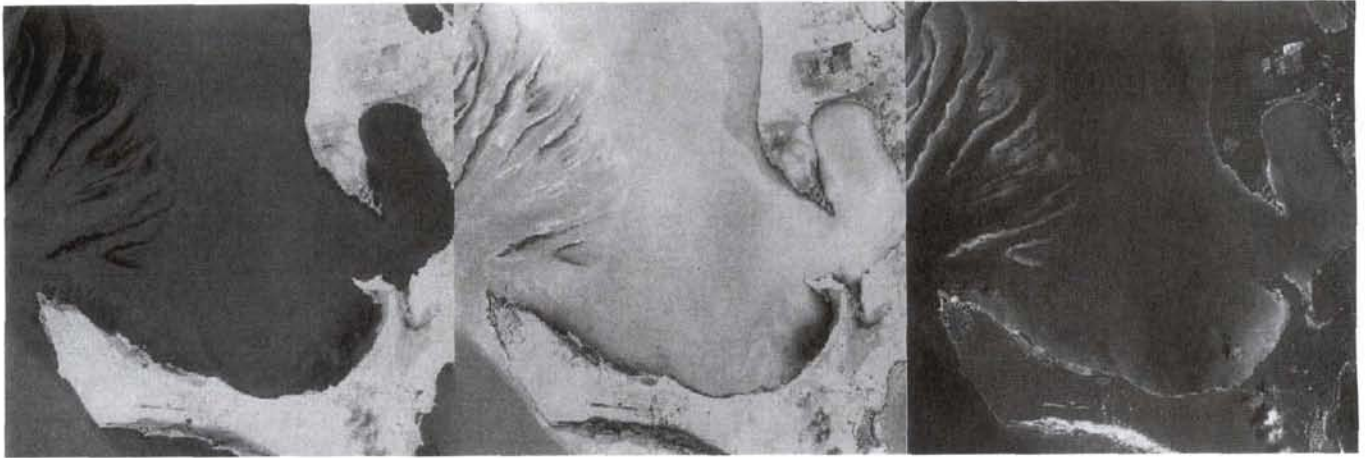
Figure 2. Mosaic of transformed images obtained using correspondence analysis. Left: first principal component image; center: second principal component image; right: third principal component image. Each image is 400 lines by 400 pixels per line.

nents, in $T$. For example, correspondence analysis of the one SPOT HRV image used in this paper resulted in the first principal component capturing 96 percent of the original data variance. An experiment is presented to reconstruct the original three images, $O$, using only the first column in $T$. In this case, the second and third columns of $T$ are discarded. Such a reconstruction is presented (Figure 4). This is an example of a *lossy compression*, because discarding the second and third columns in $T$ results in a loss of 4 percent of the original data variance. The reconstruction (Figure 4) is visually similar, but not identical to the original three spectral bands (Figure 1). For comparison, a reconstruction is attempted using only the first principal component image from principal components analysis (Figure 5). In this case, 30 percent of the original image variance is discarded, and the comparison to the original spectral images (Figure 1) is less satisfying.

In these reconstructions (Figure 4 contrasted by Figure 5), correspondence analysis restores the vegetation brightness in the infrared spectral band (SPOT HRV band 3) better than does principal components analysis when using only the first

principal component for reconstruction. In this particular case, the first principal component in correspondence analysis represents more between-band covariance than does the comparable component in principal components analysis. Whereas lossy digital image compression may not be satisfying for many applications, it is used here more to distinguish the information content in principal components developed using correspondence analysis from the information content in principal components developed using principal components analysis or factor analysis.

## Summary

Correspondence analysis is a relatively new method of multivariate analysis. Its metric is a "chi-square" deviation between a true (normalized) pixel value and its expected value. In application to a three-band, SPOT HRV image, the first two principal component images from correspondence analysis capture almost all original image variance. In contrast, all principal component images are necessary from principal components analysis or factor analysis (standardized princi-

Figure 3. Mosaic of transformed images obtained using principal components analysis. Left: first principal component image; center: second principal component image; right: third principal component image. Each image is 400 lines by 400 pixels per line.

Figure 4. Mosaic of reconstructed SPOT HRV images using the first principal component image from correspondence analysis. Channel 1 is on the left; Channel 2 is in the center; and Channel 3 is on the right. Note the restoration of vegetation brightness in Channel 3.

pal components analysis) to represent the same amount of original image variance.

Visually, correspondence analysis yields principal component images similar to those obtained from principal components analysis or factor analysis. This statement is confirmed by comparing Figures 2 and 3 again and noting that the second principal component image from correspondence analysis is almost identical to the third principal component image from principal components or factor analysis. If applied for decorrelation stretches, there is no validity to a recommendation that one of these methods be used rather than another. But, correspondence analysis captures more of the original image variance in fewer principal components, an advantage when the original image data consists of many spectral bands.

Another advantage to correspondence analysis may be for digital image compression, especially for spectral compression. Such compression is of two types: lossless (error-free reconstruction) and lossy (reconstruction with error). Principal components and factor analysis provide lossless re-construction if all principal component images are used for the reconstruction. This, however, represents no compression. Correspondence analysis will always yield some compression, even for lossless reconstruction, because it always yields $M - 1$ significant principal components (recall the negligible third principal component from correspondence analysis in the one application reviewed in this paper; Table 2). In the one example of reconstruction that is presented, correspondence analysis yielded a fair reconstruction using only its first principal component.
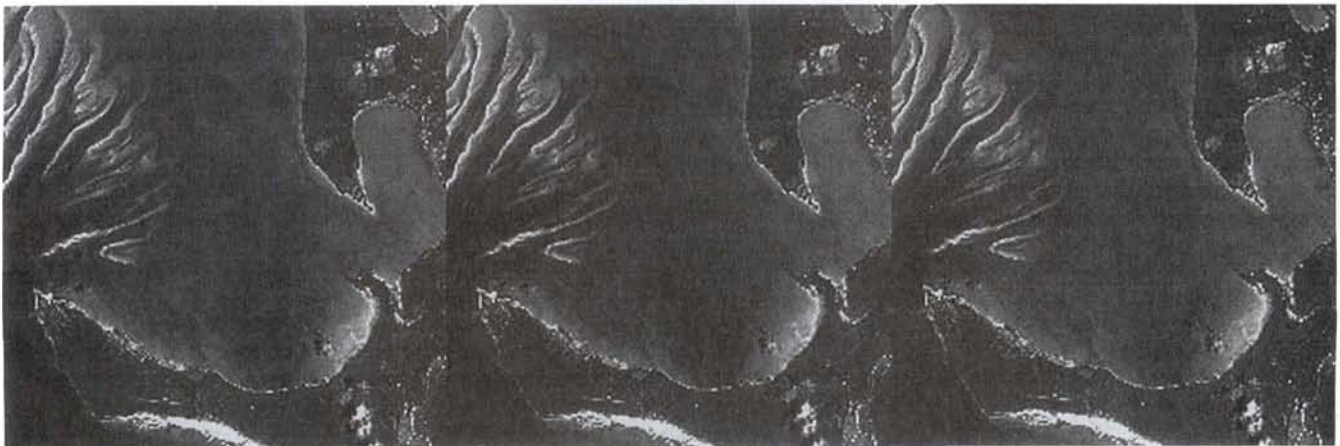
## Acknowledgments

Figure 5. Mosaic of reconstructed SPOT HRV images using the first principal component image from principal components analysis. Channels are arranged as described in the caption to Figure 4. Vegetation brightness is not reconstructed well in Channel 3.

## References

Benzecri, J.-P., 1973. *L'Analyse des Donnees. Vol. 1: La Taxinomie. Vol. 2: L'Analyse des Correspondances*, Dunod, Paris, 628 p.

Carr, J.R., 1990. CORSPOND: A Portable FORTRAN-77 Program for Correspondence Analysis, *Computers and Geosciences*, 16(3): 289–307.

———, 1995. *Numerical Analysis for the Geological Sciences*, Prentice-Hall, Englewood Cliffs, New Jersey, 592 p.

———, 1998. A Visual Basic Program for Principal Components Transformation of Digital Images, *Computers and Geosciences*, 24(3):209–218.

Davis, J.C., 1973. *Statistics and Data Analysis in Geology, First Edition*, Wiley, New York, 550 p.

———, 1986. *Statistics and Data Analysis in Geology, Second Edition*, Wiley, New York, 646 p.

Greenacre, M.J., 1984. *Theory and Applications of Correspondence Analysis*, Academic Press, London, 364 p.

Gonzalez, R.C., and R.E. Woods, 1992. *Digital Image Processing*, Addison-Wesley, Reading, Massachusetts, 716 p.

Jensen, J.R., 1996. *Introductory Digital Image Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 316 p.

Schowengerdt, R.A., 1997. *Remote Sensing: Models and Methods for Image Processing, Second Edition*, Academic Press, New York, 522 p.

Singh, A., and A. Harrison, 1985. Standardized Principal Components, *Int. J. Remote Sensing*, 6(6):885–896.