HUDM5124: Introduction to Multidimensional Scaling, Clustering, and Related Methods

Session 2: Brief review of matrix algebra
& Principal Components Analysis (PCA)

Matrix Algebra: Notation

We denote scalar quantities by a letter (often lowercase, italic); vectors by an underlined letter; matrices by a capital letter

Examples:

Y = $n \times m$ matrix of multivariate data

(e.g. scores of n subjects on m subtests)

R = $m \times m$ matrix of observed correlations among the m variables

$\underline{y}$ = $n \times 1$ column vector

Operations on matrices and vectors

- The transpose of a matrix
- Matrix addition & subtraction
- Matrix multiplication

Rank of a matrix

- The rank of a matrix (r) can be thought of as the dimensionality of the data space spanned by its columns (rows)
- An m x m matrix may be of full rank (i.e., $r = m$), or it has rank $r < m$
- For an n x m (rectangular) matrix, the rank cannot exceed MIN(n,m)
- A square symmetric m x m matrix can be factored (more on this later). If it has $r = m$ positive roots, it is said to be *positive definite*. If it has $r<m$ positive roots, and the remaining roots are 0 (that is, all roots are nonnegative), it is said to be *positive semidefinite.*

The trace of a matrix

- The trace of a symmetric matrix A is simply the sum of its diagonal elements
- The trace of a symmetric matrix A is also equal to the sum of its eigenvalues

Review: Principal Components Analysis (PCA)

- PCA is a data reduction technique that can summarize and express relationships among variables in multivariate data (or a correlation/covariance matrix)
- Basic model: R = PP', where R is an observed correlation matrix, and P is an *m* (variables) by *m* (components) matrix of "component loadings". Each column of P defines a principal component.

  Sometimes used for data reduction: $R_c = P_c P_c'$, where R_c is a "reproduced" correlation matrix, and P_c is an *m* (variables) by *k* ($k<m$) matrix of "important" components. So we are *modeling* R with R_c.

PCA and eigendecomposition

A principal components factorization can be defined in terms of the basic eigenvalue-eigenvector decomposition of a matrix

${}_mR_m = {}_mP_mP'_m$ This is the basic PCA factorization of R. P is the "component loadings" matrix, which we interpret.

$P = E\Lambda^{½}$, where Λ = an *m* x *m* <u>diagonal</u> matrix of *eigenvalues* and E = an *m* x *m* matrix where each column is an *eigenvector* of R

Therefore,

$R = PP' = E\Lambda E'$ i.e. the PCA can be expressed in terms of an eigenvalue-eigenvector decomposition of R

NOTE: R above may be thought of as the observed correlation matrix. If we use all *m* components, we should be able to <u>perfectly</u> reproduce this matrix R.

If we use fewer than *m* components (say, $k < m$) , we can only <u>approximately</u> reproduce R: $= R_c = {}_mP_k P'_m$

PCA (cont.)

- The TRACE of a square symmetric matrix is defined as the sum of its diagonal values (which is equivalent to the sum of its roots)
- Thus, for an m x m correlation matrix R, Trace(R) = m (the size of the matrix)
- This means that the sum of the eigenvalues must equal m, and the average-sized eigenvalue is = 1

  (this is the origin of the eigenvalue>1 criterion for how many roots/components to extract, rotate & interpret)

Summary: goals of PCA

- To understand the structure in a correlation or covariance matrix among n variables or entities
- This is accomplished by explaining the pattern of correlations in terms of n underlying components (“factors”) and the weights or “loadings” of the n variables or entities on these components
- We can obtain a more parsimonious “explanation” by approximating the correlation with a reduced number of components
- <Example>