STAT 802 Multivariate Analysis September 9, 2003

Today: Examples of Multivariable Strategies
Review of Statistical Distance Idea
Some details of Eigen-analysis (pp 60 ff)

Last week I asked you to consider:

Multivariate Strategies

See pp 3-4

Data Reduction or Simplification
Sorting and Grouping
Investigation of the dependence among variables
Prediction
Hypothesis Testing

Exercise for Sept 9 class: In a context of interest to you, suggest a hypothetical application of each of these strategies.

Statistical Distance

Heuristic Explanation – See Fig 1.20 p 31 and Fig 1.24 p 36.

Some details (Ch 2)

Vector Representation and Arithmetic.

L_x = Length of a vector x = $(x'x)^{1/2}$ (Note x'x is a scalar).
Angle between a vector x = (x_1,x_2) and the x_1 axis is $\cos^{-1}(x_1/L_x)$ (By definition of cos)
Inner Product of vectors x and y is denoted x'y and is scalar.
If □ is the angle between vectors x and y, then cos(□) = $x'y/(L_xL_y)$
Generalizes to k dimensions. (not k variables!)

See example p 54

Recall definition of Linear Dependence
Vectors $x_1, x_2, \ldots, x_p$ are linearly dependent if you can express one as a linear function of the others. (See (2.7) p 54 bottom)

Projection of vector x on y – must be a multiple of the vector y – but what multiple?
Ans: x'y/y'y (see equation 2-8) so the projection is (x'y/y'y)y

If we want to write the projection of x on y as a multiple of a unit vector, we re-scale y to be unit by y/L_y and multiply it by $|x'y|/L_y$
Can infer that length of this projection is $|x'y|/y'y = L_x \cos(\square)$

Matrices (p 55 ff)
Arithmetic: Sums, Products, Inverses and the Identity Matrix
Symmetric Matrices are Important for Statistics (Why?)
Orthogonal (Square) Matrices: $A' = A^{-1}$ so $A'A = I$
(Note: Orthonormal might be better, with Orthogonal meaning A'A=Diagonal – but we will stick to the terminology of the book).

Eigenvalues and Eigenvectors

(Synonyms Latent and Characteristic have other meanings as well – confusing)

Square Matrix A – x is Eigen vector of A if $Ax=\square x$ where □ is a scalar. □ is called an eigenvalue of A corresponding to the eigenvector x. There are usually many such x and □, but never more than the rank of A (≤ the dimension of A).

Now concentrate on (square) symmetric matrices (like covariance matrices) – see theorem p 61. Eigenvectors are mutually perpendicular unless some eigenvalues are multiple.
By convention, usually specify eigenvalues to have unit length (wlog see definition).

Positive Definite Matrices

A symmetric matrix A is Positive Definite if $x'Ax > 0$ for all $x \neq 0$
Can show this is equivalent to all eigenvalues >0 (p 64)

Statistical Distance of vector x from vector 0 (origin) has the form x'Ax where A is pos definite.
From x to □ is $(x-\square)'A(x-\square)$. But we still have to make A specific to call it statistical distance.

To find appropriate A: See p 66 Fig 2.6 If the ellipse shown were a contour of the bivariate normal density, then we would find e_1, e_2, $\square_1$ and $\square_2$ from the eigenanalysis of the $\square^{\square 1}$ where □ is the covariance matrix. So the distance of points from the origin would be $x'\square^{\square 1} x$ (In the diagram, the variables are centered so the mean is the origin). This generalizes to p dimensions (See Theorem bottom p 81) . It implies that if you keep extracting maximal variance projections, each one perpendicular to the previous one, then the eigenvalues will be the maximal variances at each stage, and the eigenvectors will be the directions that the projections must be taken.

The interpretation depends on the multivariate normality, but the procedure does not. The means and covariances of the data will produce statistical distances from the centroid for every "case".

 can be expressed as E[(x-)’(x-)] and is estimated by the equivalent sample formula (p 125). Because is positive definite usually, there is a spectral representation based on the eigenanalysis (see p 67 , (2-21)). This represeantation shows that the statistical distance is indeed found by an orthogonal rotation of the axes followed by the Euclidean distance in the new coordinates. The A in this formula can be replaced by ^{1} or its sample estimate.

Here is a MINITAB macro for computing mean and covariance and correlation matrices: You need to specify k3 (variables) and k4 (cases) . I used this on exercise 3.10 p 147.

```
Gmacro
Meancov.mac
#Data matrix is k4 rows of k3 variables
let k3=3
let k4=5
copy c1-ck3 m1     #  is X which is k4 x k3
trans m1 m2              #  is X' which is k3 x k4
set c10              # 1 which is k4 x 1
k4(1)
end
copy c10 m3        #  just the matrix version of c10 k4 x 1
tran m3 m4         # 1' is 1 x k4
let k1=1/k4        # this is 1/n
mult k1 m2 m5     # starting to compute xbar
mult m5 m3 m16          #  Xbar from (3-24)
let k2=1/(k4-1)         #  1/(n-1)
mult m3 m4 m6     #  1 1', starting to compute S from (3-27)
mult k1 m6 m7
diag c10 m8          # the identity of dim n=k4
subt m7 m8 m9       # inside brackets of (3-27)
mult m2 m9 m10
mult m10 m1 m11
mult k2 m11 m12    # this is S
diag m12 c11           # pick off variances from diagonal
let c12=1/(c11**.5)     # get SDs from variances and invert
diag c12 m13           # form D^(-.5) matrix.
mult m13 m12 m14
mult m14 m13 m15   # this is R
name m16 'MEANS', m12 'COVAR', m15 'CORR'
print m16, m12, m15
```

endmacro

Exercise: Use the data of Ex 3.10 to compute the distances of the data rows to the centroid, in two ways:

1. Using $\sqrt{(x - \bar{x})' \hat{\Sigma}^{-1} (x - \bar{x})}$
2. Using eigenanalysis – find eigenvectors, project a data row on the eigenvectors, and the mean vector, find the Euclidean distance from the projected row to the projected mean vector, and check that you get the same answer as in 1. (just do it for one row).

Data of Ex 3.10

Row	X1	X2	X3
1	3	1	0
2	6	4	6
3	4	2	2
4	7	0	3
5	5	3	4