STAT 802 Multivariate Analysis September 11, 2003

Today: More details about exercise
Ch 3: Sample Means and Covariances
Some Ch 3 exercises

[Recall

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).]

Please hand in Tuesday, Sept 16

Some calculation details:

$S^{-1} = \hat{\Sigma}^{-1}$ can be computed from (3-11) on p 124. Or, you can use my MINITAB program. To run the program in MINITAB, just issue the MTB command %meancov. Of course the file has to be in the macros folder of MINITAB. (It can be elsewhere if give the path to where the file is, as in %path/meancov, but the default is the MINITAB macros folder). The program gives the mean too but in this case that can be written down by a quick look at the data.

Note how a mean is computed in matrix algebra:

Form a column vector of n 1's (if n is the sample size= no. of rows in data)
Transpose the data matrix (if X is n by p, X' is p by n, where p is the number of variables)
Then multiply $X'1_n$, where "1_n" represents the column vector of n 1's.
Mult by (1/n).
The result is a row vector of means (dimension 1 by p), m' say.

For the covariance, (look at the formula p 124), the first thing we need to do is realize that the vectors X_j in that formula are column vectors, but they are of dimension p. The text uses X as an n by p data matrix, as usual, but chooses X_j to represent the row of the data matrix as a column vector! Of course, X_j' is a row vector representing the jth row of our data matrix - See (3-8) p 120. So when the formula for S is given on p 124, the $(X_j - \bar{X})$ there is a column vector (p by 1) which represents the mean-adjusted vector of the p measurements on row j. Thus $(X_j - \bar{X})(X_j - \bar{X})'$ is a p by p matrix containing the cross-products and squares of the measurements for the jth case. There are n such matrices so multiplying by 1/n or 1/(n-1) does make sense to get an average cross-product. And remember that the covariance of x and y is $E(x-\mu_x)(y-\mu_y)$, so we really do want average cross-products from our sample to estimate the population covariances. (For more on this see pp139-140)

To find the coordinates of a projection of x on y … We did this on 030909.
The result is (x’y/y’y)y
If y is a unit vector (eg – an eigenvector) then the length of the projection is just x’y

Does it help to consider the geometry of a data matrix by considering p vectors in n-space? J&W says it does. We will re-visit if we find we need it!

Section 3.4 Generalized variance. (don’t spend too much time on this)

Single scalar value representing variation in all the variables of a multivariate data set. Defined as the determinant of the sample covariance matrix = |S|. See p 126 for interpretation as proportional to the square of the volume spanned by the p deviation vectors.

When would this be useful? Coherent variables, comparison of two groups or times?

Section 3.5 – S related to R (sample correlation matrix):

$$S=D^{1/2}RD^{1/2}$$

$D^{1/2}$ is the diagonal matrix of sample standard deviations.

Section 3.6

Sample Means and Covariance of a Linear Combo of Variables

Consider $Y= c'X = c_1X_1 + c_2X_2 +\ldots+ c_pX_p$ as on p 141.

Observe $y_j = c'x_j = c_1x_{j1} + c_2x_{j2} +\ldots+ c_px_{jp}$ for $j = 1,\ldots,n$

What are mean and variance estimates for Y?

Mean = $c'\,\bar{x}$
Variance = $c'Sc$

(cf p 77 for population parameters)

Suppose $Z = b'X = b_1X_1 + b_2X_2 +\ldots+ b_pX_p$

What is Cov (Y,Z)?

$= b'Sc = c'Sb$

See summary p 142 bottom