STAT 802 Applied Multivariate Analysis Oct. 3, 2005

Today:

Review of Assignments
Comment on Ch 6 and 7 (Return Later)
Intro to Ch 8: Principle Components

Assignment about data Cleaning (Section 4.7)
dotplots
use of chi-sq distribution to assess squared distances.
(star plots)

Assignment about outliers and Ex 4.21 and 4.26

Does a univariate outlier have to be a multivariate outlier?

Define "outlier" to be in the .0013 tail of the distribution. Assume normality.

Suppose $x_1=3$, $x_2=0$ and $\mu_1=\mu_2=0$, $\sigma_1=\sigma_2=1$ and $\rho=0$.
(assume large enough sample so ests accurate)

x_1 is a univariate outlier.

But multivariate statistical distance^2 is 9, and $P(d^2 > 9) > .0013$
(Because the value of $\chi^2(.9987)$ on 2 df is 13.29.)

So x_1 is a univariate outlier but (x_1, x_2) is not a bivariate outlier.

The general idea is that, if the other components of a multivariate observation are not extreme, the sum of squared component distances can be a moderate value compared with the distribution of this sum.

4.21 and 4.26 were well done – no further comment ...

Ch 8: Principal Components

Data: n observations on p variables

Construction of a few (<p) indices that

1. contains most of the information (variability) in the p-variables
2. is easier to describe (and communicate) than the original p variables

Main results 8.1 and 8.2 p 428.

8.1 principal components are lengths of projections of data on (ordered) eigenvectors (projection? Review p 55 in Ch 2.)
variance of principal components is given by eigenvalues

8.2 sum of variances of variables = sum of eigenvalues =trace(Σ=covariance matrix)

Note: Ordering by size of eigenvalue, which is the same as ordering by successive variance maximization. Matrix result proven in section 2.7 of text, and specialized to data matrix in Ch 8.

Since eigenvectors form an orthonormal basis, principal components are just coordinates in rotated basis.

Typically we choose the top 1, 2 or 3 principal components to represent the data. Several questions arise:

0. Correlation matrix or Covariance matrix (Standardized Data?)
1. How do you choose number of components? (Look at sequence of eigenvalues)
2. How do you describe the components? (Look at correlations with original variables, Result 8.3 p 429)
3. Sampling variation (p 438). Just use estimates of variances.