STAT 802 Multivariate Analysis Oct. 21, 2003

Today: Comments on the Test
Ch 8 – Pritam - Principal Components
KLW comments on Principal Components

Test Comments

1. How can the limitation of a flat surface for graphics be overcome in plotting multivariate data with more than 2 variables? (Suggest two essentially different ways to do this.)

One is 1-1 plots like stars, faces, etc.
Another is coplots
Other methods limited to 3-D are
contour plots
wireframe plots
spinning 3-D scatter plots

2. A Bivariate Normal distribution has component means 0, component standard deviations 1, and correlation 0.5. Find the following statistical distances:
a) (1,1) to (0,0)
b) (1,-1) to (0,0)
c) (1,1) to (1,-1)

An easy way is to use factor of exponent of bivariate density formula p 151
squared distances are 4/3, 4, and 16/3 so distances are …
Note that part c) can be obtained from a) and b) by Pythagoras. (Why?)
Another way to do c) is to get (0,2) to (0,0). Think projections to justify.

3. Describe the geometric situation for p-variate data (p>1) when exactly one eigenvalue is 0. Refer to the eigenvectors in your description.

Hyperplane perpendicular to eigenvector corresponding to 0 eigenvalue.

4. If X_1 and X_2 are Normal, is (X_1,X_2) Bivariate Normal? Explain.

No. Here is a counter example. Let $X_1 \sim N(0,1)$ and Let $X_2 = X_1$ when $|X_1| > 1$ and $= -X_1$ otherwise. Clearly X_2 is also N(0,1) yet (X_1,X_2) is not bivariate normal.
The general idea here is that joint distributions imply marginals, but not vice-versa.

5. Refer to the boxed result on p 177 about the distribution of $n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu)$ is approximately χ^2 on p df, where p is the dimension of X. Explain why this must be so using an eigen-analysis argument. (Only words are needed here).

The distribution of the sample mean is approx N with mean mu and covariance approx

S/n. The statistic having the χ^2 distribution on p df is therefore just the (statistical distance)^2 of Xbar from mu. The projections of the vector (xbar-mu) on each eigenvector are standard normals (approx), and they are uncorrelated. So the (Euclidean distance)^2 is the sum of p squared, independent normals which must have a χ^2 distribution on p df.

6. Why does one examine a normal Q-Q plot by assessing the linearity of the plot? How do you do this assessment objectively?

If the data quantiles are linearly related to the quantiles of the standard normal, then the ecdf of the data will be a linear function of the normal quantiles. (If we had quantiles of the normal with the correct mean and SD, the line would be the 45 ° line). The objective test uses the correlation coefficient and compares with the table on p 182.

7. How are (1-α) confidence regions for p-variate normal mean related to the set of p (1-α) confidence intervals for the component means? Explain.

The component shadows of the (1-α) confidence region will include all the component means with greater than (1-α) probability, since there are points outside of the confidence region that are inside all the shadow confidence intervals, and there are no points outside the shadow confidence intervals that are in the confidence region. Thus the simultaneous CIs for the component means should be narrower than the shadow intervals, to attain the same (1-α).

8. What do you get from a multivariate regression with m dependent variables that you do not get from m multiple regressions?

You get simultaneous confidence regions for ALL the Betas. (However, we have not seen a compelling example where this would be useful!)

9. a) Use a simple hand-drawn graph to illustrate the difference between leverage and influence.
b) Explain why the diagonal elements of H (p 377) do measure leverage.
a) A typical scatter with a high leverage point on the extended regression line would be an example of low influence. A low leverage point would usually not have much influence. But a high leverage point that is far from the regression line determined by the rest of the data would be a high influence point. (easier to show in a graph).
b) Equation (7-20) on p 377 shows that residuals for each data row will be small when h_{jj} is close to 1. This means that the fit is forced through the j^{th} point, which is another way of saying that the j^{th} point has large leverage.

10. Refer to the section "Other Multivariate Test Statistics" on p 395. By analogy with multiple regression, explain why the four statistics are measures of significance of the full model compared to the reduced model. (i.e. why large values of the statistics suggest the reduced model is inadequate.)

Understand E to be the cross products matrix of residuals from the full model, and H the

extra cross products from adding the extra variables to the reduced regression model. If we use –log(Wilks) instead of Wilks lambda as defined on p 395, then all the measures will tend to be large when H is "big" compared to E. That is, large when the addition of the extra variables in the model makes a big reduction to the error matrix. (The question has a small error in the parenthetical remark – Wilks is actually small when the others are large.)

Pritam Presentation on principal components

Comments re Principal Components

1. Scaling: Use standardized variables? IE eigenanalysis of correlation matrix.

2. How many are needed to adequately describe data? Scree test.
Try it on uncorrelated data. Then on "Men" data.

3. Utility: need to interpret first few components.

4. Use of small eigenvalues to highlight features of the population studied.

5. First principal component is often similar to an average of the standardized variables.

6. Keep in mind back-pain example to avoid using simplistic criteria to throw out predictor variables in a regression.