STAT 802 Applied Multivariate Analysis Sept. 21, 2005

Today:

A few items from Ch 3: (pp 120-148)
1. Sample Covariance Matrix (p 124)
2. Generalized Sample Variance (p 124 for S, p 136 for R)
3. Total Sample Variance (p 138)
4. Mean and Covariance of Linear Combinations (p 142)

Ch 4 Multivariate Normal: (pp 149-167)
1. Details for Bivariate Normal
2. p-Normal, contours, distances, properties.

Ch 3:

1. Sample Covariance Matrix

Def of S and S_n p 124. Issue of unbiasedness (not as important as MSE).

Note n denominator has lower mse for var and for sd.

2. Generalized Sample Variance

The GSV is just the determinant of S = |S|. For standardized variables =|R|.
There is a sense in which this is proportional to the volume spanned by the p variables – This depends on the "p-vectors in n-dimensions" described at the beginning of Ch 3. Positively correlated variables are a smaller angle from each other and the volume enclosed is less. In a sense, two correlated variables do not have as much "variance" as two uncorrelated variables. GSV measures this for any p-variables.
In practice, not much used….

3. Total Sample Variance.

We sum the p-variances to get the TSV (p 138). This is the trace of the covariance matrix. It just measures a sort of average variation per variable, and does not discount variables that are positively correlated (for example).
Also, not too useful in practice. But it is good to know that such a thing has been thought of and deleted from further consideration!

4. Mean and Covariance of Linear Combinations (Result 3.5 p 142)

The main thing is sample covariance of b'X is b'Sb.

Ch 4: Multivariate Normal Distribution

General formula for density of p-variate Normal: Eqn 4-4 p 150.

Note power coefficient is $-\delta^2/2$, where δ is statistical distance of x from the mean μ. The connection of statistical distance and normal density contours is detailed in para 1 of p 153. Para 2 describes the role of the eigen vectors in describing these density contours. See also the result (4-7) on p 153, and the Remark pp 164-5 re statistical distance.

Details of Bivariate Normal (p=2) See Example 4.1 p 151

Useful to review regression prediction (predict Y given X) for bivariate normal.

Dist of Y|X is $N(\mu=\mu_{Y|X}, \sigma^2_{Y|X})$ where $\mu_{Y|X}=\mu_Y + \rho\,(X-\mu_X)/\sigma_X$ and $\sigma^2_{Y|X}=\sigma_Y^2(1-\rho^2)$ where ρ is the correlation between X and Y.

Compare this with Result 4.6 on p 160 for the p-variate Normal.

Note that the prediction variance does not depend on X in our bivariate case, and that the covariance of the conditional distribution does not depend on the predictor variables in general. This is a nice property of the p-variate Normal.

The distribution of the statistical distance is chi-square on p df (if mean and variance are known). See box p 155. Estimating the mean and variance reduces the df but the statistical distance still has an approximate chi-square distribution. See box p 177.

Properties of p-variate Normal see p 156

X p-Normal. What is dist of a'X?

We know mean and covariance from Ch 3. See consequence Result 4.2 p 156 and Result 4.3 p 157.

Next day – Ch 4 pp 168-end of Ch 4 and some exercises. Reminder: Ex assigned Sept 19 for hand in Monday Sept 26. (This is a small exercise.)

Exercise for Wed, Sept 28, to hand in.

1. Using examples show that
i) a univariate outlier is not necessarily a multivariate outlier.
ii) a multivariate outlier is not necessarily a univariate outlier.

2. Ex 4.21

3. Ex 4.26