BSCB 694: Theory of Multivariate Statistics

# Homework 2

**Due: Tuesday, April 8**

1. Simultaneous Confidence Intervals

   (a) Prediction intervals.

      i. Consider making a prediction for a new observation in a regression problem. That is

      $$\hat{\mathbf{y}}_{new} = \mathbf{x}_{new}^T \hat{B}$$

      Show that this prediction is unbiased, and derive an expression for $\text{cov}(\mathbf{y}_{new} - \hat{\mathbf{y}}_{new})$.

      ii. Derive a test procedure that will allow you to produce in prediction intervals for $\mathbf{a}^T\mathbf{y}_{new}$ for all $\mathbf{a}$.

   (b) Extended Scheffé confidence regions.

   Union-Intersection tests for MANOVA can be confidence regions can be derived as tests of $H_0 B\mathbf{m} = 0$ that hold over all vectors $\mathbf{m}$. We extend this $H_0 : \mathbf{c}^T B\mathbf{m} = 0$ for all vectors $\mathbf{c}$ as well.

      i. Show that for all $\|\mathbf{c}\| = 1$,

      $$|\mathbf{c}^T(\hat{B} - B)\mathbf{m}| \leq \mathbf{m}^T(\hat{B} - B)^T(\hat{B} - B)\mathbf{m}$$

      ii. Show that simultaneous confidence intervals for $H_0 : \mathbf{c}^T B\mathbf{m} = 0$ may be derived based on the critical values of Roy's greatest root test for $B = 0$.

2. Limits of the General Linear Hypothesis

   Demonstrate that the general linear hypothesis does not include tests for all combinations of the elements of $B$. Provide an example of a contrast that is not included in the general linear hypothesis that may represent a question of scientific interest.

3. Extended Linear Hypothesis

   Mudholkar, Davidson and Subaiah (1974) proposed the *Extended Linear Hypothesis* for more general questions as being

   $$H_0 : \text{Tr}(G^T B) = 0$$

   Here we follow their reasoning. You may assume the identity that for matrices $A$ and $B$,

   $$\sup_{A \neq 0} \frac{AB^T}{\text{Tr}(AA^T)} = \text{Tr}(BB^T)$$

(a) Give an example of a contrast that is expressable as an extended linear hypothesis, but not as a general linear hypothesis.

(b) For a general linear hypothesis $C_1BM_1$, show that there are matrices $F$ and $K$ such that

$$t(G) = \frac{\mathrm{Tr}(G^T B)}{\mathrm{Tr}(K^T G^T F^T F G K)}$$

is bounded above by the Hotelling-Lawley Trace statistic for the test $H_0 : C_1BM_1 = 0$.

(c) An *intermediate* Extended Linear Hypothesis is defined for a collection $G_1, \ldots, G_k$ by

$$H_0 : \mathrm{Tr}\big((\sum \gamma_i G_i)^T B\big) = 0, \ \forall \gamma_1, \ldots, \gamma_k$$

Derive a union intersection test statistic for this hypothesis using the test above. Observe that this is also bounded by a Hotelling-Lawley Trace statistic.

4. Timm Exercizes 5.3 1 and 2. Example files for the text are linked on the class website.

5. Netflix Data and MDS: Data from the Netflix competition are available on the class website. These data provide the ratings that 10000 users gave to 100 movies. A list of movie titles is also available on the class website. No all movies were rated by all users.

   In the R statistical language, non-metric MDS can be obtained through the `isoMDS` function.

   (a) Suggest an appropriate distance (or similarity) metric between movies based on their ratings. Perform classical multidimensional scaling on the distance that you create. How many dimensions appear to be sufficient? Do the canonical co-ordinates appear to be interpretable.

   (b) Repeat the exercise above with users.

   (c) Explore non-metric scaling for distances between movies. Does this yield substantially different results?

6. (MKB 14.2.3) Let $D$ be an $(n \times n)$ Euclidean distance matrix with $X$ the classical solution to the MDS problem in $p$-dimensional space. Suppose we wish to add an additional point with distances $d_{r,n+1}$, known to be Euclidean.

   We consider the projection into $(p+1)$-dimensional space with the rows of the new representation given by $\mathbf{x}_r^* = (\mathbf{x}_r, 0)$. Let the representation of the $(n+1)st$ point be given by $\mathbf{x}_{n+1}^* = (\mathbf{x}^*, y)$. Show that $\mathbf{x}^*$ is determined uniquely (and give an expression for it) but that $y$ is only determined up to its sign.