STAT 575 – Assignment 3 – due date is on course outline

For each of questions 2,4,6 – please e-mail me your R programs. The three files should be named `yourname_Qx.R`*, where* $x \in \{2, 4, 6\}$*. I should be able to run them exactly as I receive them, and see the output nicely displayed (see the help files for the* `cat` *function). Round the numbers to 2 or 3 decimal places, as appropriate. For these three questions your written assignment then need contain only the discussions of your solutions.*

1. Here you are asked to derive an alternate interpretation of the sample principal components – that the k linear combinations of the columns of the data matrix $\mathbf{X}_{n\times p}$, which yield minimum mean squared error (mse) linear predictors of $\mathbf{X}$, are the first k sample pc's. To formalize and prove this statement, assume that the data matrix has been standardized, so that $\mathbf{X}^T\mathbf{X} = (n-1)\,\mathbf{R}$, where $\mathbf{R}$ is the sample correlation matrix. Let

$$\mathbf{R} = \mathbf{\Gamma\Lambda\Gamma}^T$$

be the spectral decomposition (with the eigenvalues in decreasing order of magnitude). Now consider the problem of finding k linear combinations $\mathbf{XA}_{p\times k}$ of the columns of $\mathbf{X}$, where $\mathbf{A}$ is chosen to minimize the mse when $\mathbf{X}$ is predicted by linear combinations of the form $\hat{\mathbf{X}} = \mathbf{XA}\hat{\mathbf{B}}_{k\times p}$.

   (a) With this mse defined as $tr\left(\mathbf{X}-\hat{\mathbf{X}}\right)^T\left(\mathbf{X}-\hat{\mathbf{X}}\right)$, verify that

$$\hat{\mathbf{B}} = \left(\mathbf{A}^T\mathbf{RA}\right)^{-1}\mathbf{A}^T\mathbf{R},$$

   and that the resulting mse is

$$mse = (n-1)\left\{p - tr\mathbf{RA}\left(\mathbf{A}^T\mathbf{RA}\right)^{-1}\mathbf{A}^T\mathbf{R}\right\}.$$

   [Hint: think about regressing $\mathbf{X}$ on $\mathbf{XA}$.]

   (b) Show that this mse is minimized by choosing $\mathbf{A}$ in such a way that the k columns of $\mathbf{XA}$ are the first k sample principal components. What then is the optimal $\hat{\mathbf{B}}$? [There is no doubt a number of ways to do this; my method made crucial use of the various properties, from Lec. 11, of idempotent matrices.]

2. In the radiotherapy data at T1-7.DAT the $n = 98$ observations on $p = 6$ variables represent patients' reactions to radiotherapy.

   (a) Obtain the covariance and correlation matrices $\mathbf{S}$ and $\mathbf{R}$ for these data. Explain why it probably makes more sense to carry out a pc analysis using $\mathbf{R}$ rather than $\mathbf{S}$. (These data are discussed more fully in exercise 1.15 in the text.)

... over

(b) Determine the eigenvalues and eigenvectors. Prepare a table showing, in decreasing order, the percent that each eigenvalue contributes to the total sample variance.

(c) Given the results in (b), decide (with the aid of an appropriate plot) on the number of important sample principal components. Is it possible to summarize the radiotherapy data with a single reaction index component? Explain.

(d) Prepare a table of the correlation coefficients between each principal component you decide to retain and the original variables. If possible, interpret the components.

3. (a) Establish the inequality (16.2) in the class notes – that $\left\|\mathbf{S} - \left(\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}}\right)\right\|^2 \leq \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2$.

(b) Establish the identity, for matrices $\mathbf{A}_{p\times q}$ and $\mathbf{B}_{q\times p}$ and assuming the existence of the inverse:

$$\mathbf{B}\left(\mathbf{I}_p + \mathbf{AB}\right)^{-1} = \left(\mathbf{I}_q + \mathbf{BA}\right)^{-1}\mathbf{B},$$

and use this to establish (17.1) in the class notes.

(c) Exhibit and prove a similar identity allowing one to evaluate $(\mathbf{I}_p + \mathbf{AB})^{-1}$ from $(\mathbf{I}_q + \mathbf{BA})^{-1}$; use it to invert $\mathbf{I}_n - \mathbf{ab}^T$, where $\mathbf{a}$ and $\mathbf{b}$ are $n \times 1$ vectors. What is the condition required for the existence of this last inverse?

4. A firm is attempting to evaluate the quality of its sales staff and is trying to find an examination or series of tests that may reveal the potential for good performance in sales. The firm has selected a random sample of 50 sales people and has evaluated each on 3 measures of performance: growth of sales, profitability of sales, and new-account sales. These measures have been converted to a scale, on which 100 indicates 'average' performance. Each of the 50 individuals took each of 4 tests, which purported to measure creativity, mechanical reasoning, abstract reasoning, and mathematical ability, respectively. The $n = 50$ observations on $p = 7$ variables are listed in T9-12.DAT.

(a) Assume an orthogonal factor model for the <u>standardized</u> variables. Obtain the (unrotated) principal component solution and the (unrotated) maximum likelihood solution for $m = 2$ and $m = 3$ common factors. Given these four solutions, obtain the rotated solutions. For all eight of these fits, list (i) the proportions of variance accounted for by each of the m factors, (ii) estimated communalities, (iii) specific variances, and (iv) the sum of squares of the residuals for the $m = 2$ and $m = 3$ solutions. When I run your R program I should see eight pieces of output displayed, each similar to the following.

... over

```
rotated pc solution, m = 2:
prop.  variance = 0.45 0.41
communalities are 0.96 0.89 0.89 0.85 0.69 0.81 0.87
specific variances are 0.04 0.11 0.11 0.15 0.31 0.19 0.13
ss.resids = 0.24
```

(b) Choose what you feel is the 'best' of the eight fits (explain why), and interpret the factor solutions in this model. (As is almost always the case, there is no one 'right' answer here.)

(c) Suppose a new salesperson, selected at random, obtains the (unstandardized) test scores $\mathbf{x}^T = (110, 98, 105, 15, 18, 12, 35)$. Calculate the salesperson's factor scores using the rotated ml model with $m = 2$, together with each of the weighted least squares and regression methods. (As a check, you might let $\mathbf{x}$ be the first row of the original data matrix, and check that in this case your answer agrees with R's 'scores' output.)

5. (a) Establish the inequality (18.4) in the class notes – that $\sum_{i=1}^{p} \mathrm{VAR}[X_i] \geq \sum_{i=1}^{p} \mathrm{VAR}[U_i]$.

   (b) Establish equality (19.1) in the class notes – that $\log\left(\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|}\right) = -\log \prod_{i=1}^{p} \left(1 - \hat{\rho}_i^{*2}\right)$.

6. For the data set in Question 4, investigate the canonical correlations between the three measures of performance and the four tests.

   (a) Obtain the canonical correlations.

   (b) Carry out large sample tests to determine the number of significant canonical correlations.

   (c) Exhibit all significant canonical pairs $(U_i, V_i) = \left(\mathbf{a}_i^T \mathbf{x}, \mathbf{b}_i^T \mathbf{y}\right)$ with $\|\mathbf{a}_i\| = \|\mathbf{b}_i\| = 1$.