STAT 5531 Final — Fall 2010

Thanh Doan

1. (20%) Suppose that a study involves two characteristic measurements.

1. Evaluate and test the hypothesis with α = 0.05.

Comparing with critical value

We see and thus we reject at the level α = 0.05 and conclude that the population mean vector is not equal to.

1. Determine the lengths and directions for the axes of the 95% confidence ellipse.



The center is at and the half-lengths of the major/minor axes are:

The axes lie along

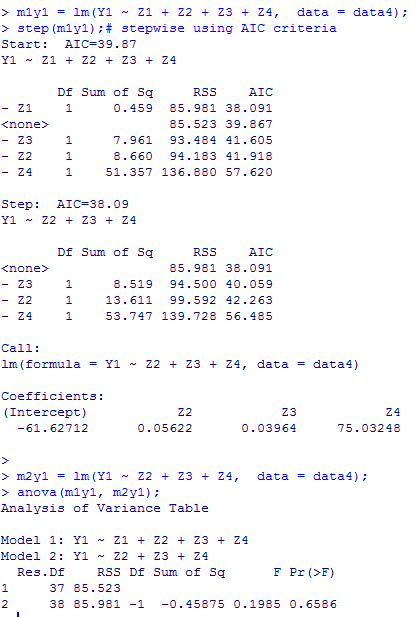
2. Suppose that a clinical study involves 3 characteristic measurements in two groups.

1. Perform the analysis of one-way MANOVA at α = 0.05.

4. (20%) In a regression study, there are three dependent variables and four independent variables.

(a) Perform a regression analysis using each of the response variables Y1, Y2 and Y3.

1. Suggest and fit appropriate linear regression models



After stepwise exploration and manual comparison.. I suggest model 2 for Y1

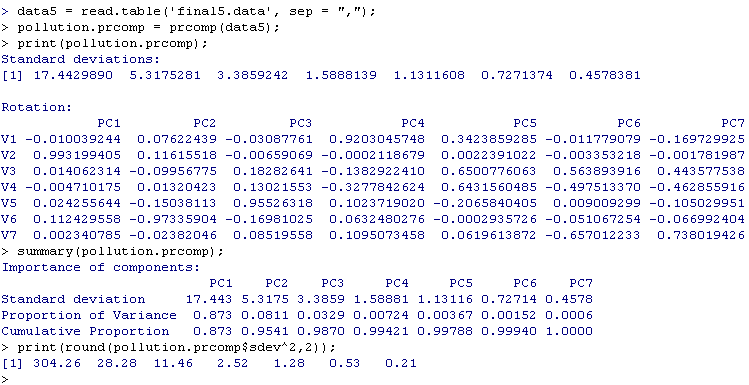
M2y1 = lm(Y1 ~ Z2 + Z3 + Z4, data = data4)

5. The data set contains measurements on seven air-pollution variables recorded

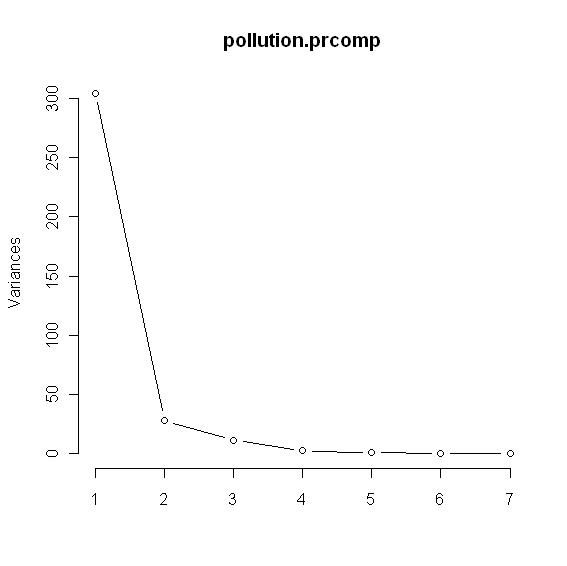
at certain time in the Houston area on different days.

(a) Principal Components Analysis from the sample covariance matrix S

i. Construct the sample principal components



ii. Determine the proportion of the total sample variance explained by the first few principal components. Interpret these components.



**Findings & Interpretation:**

* The first principal component

explains 87.3% of the total sample variance

* The first principal two components,

collectively, explain 95.41% of the total sample variance

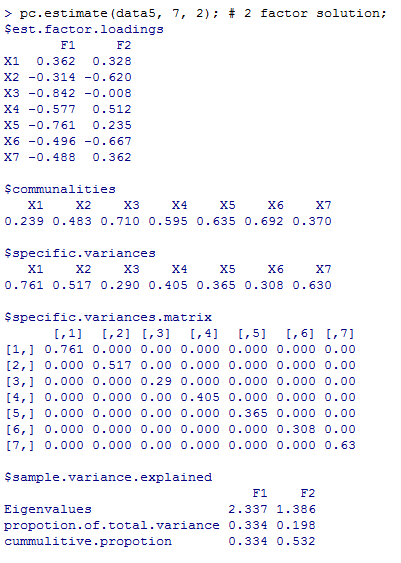
* The above findings and the *scree* plot suggest that sample variance is summarized very well by two principal components and a reduction in the data from 42 observations on 7 air-pollution variables to 42 observations on 2 principal components is reasonable.
* From the component coefficients, the first principal component appears to be essentially a weighted sum between X2 and X6. The contributed determination from other variables appears to be negligible.
* From the component coefficients, the second principal component appears to be a weighted difference between X2 and (a weighted sum X5+X6). The contributed determination from other variables appears to be negligible.

**Note:**

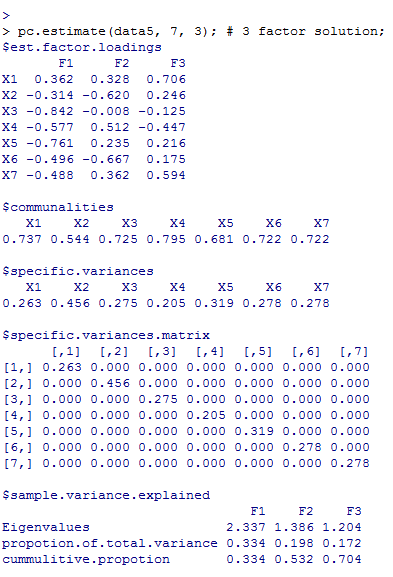
* This problem is investigated under both matlab and R computing environments to compare the result and the capability of each package. The results from both packages are identical.
* The attached matlab code contains more plots and further exploratory work than this report.

1. Factor Analysis from the sample correlation matrix R
2. Obtain the principal component solution to a factor model

**Two-factor solution**

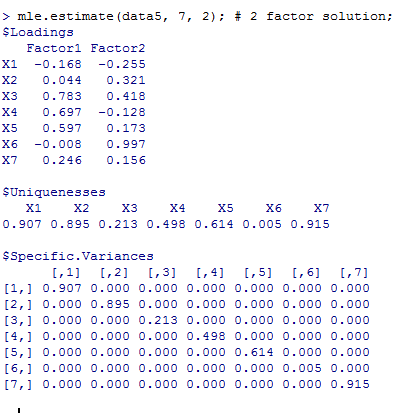
****

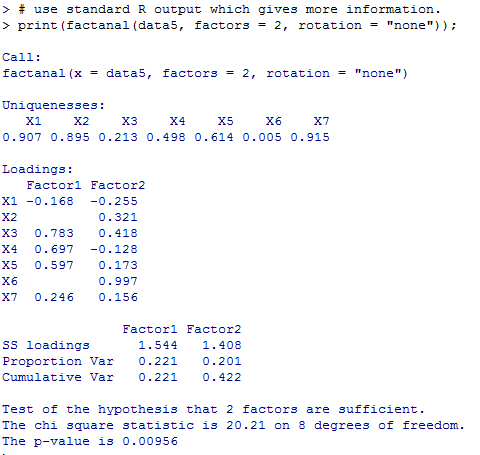
**Three-factor solution**

****

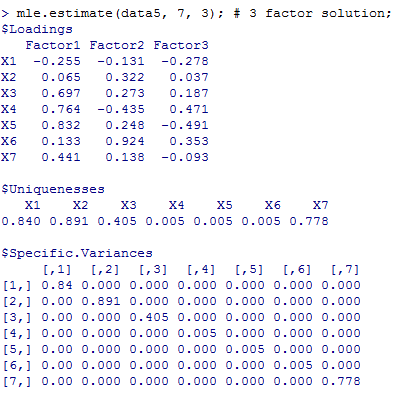
1. Find the maximum likelihood estimates of L (loadings) and (specific variances)

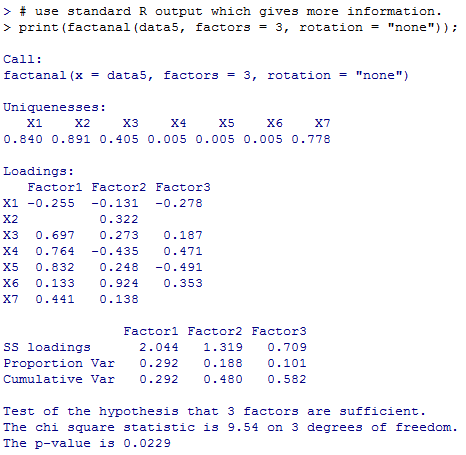
**Two-factor solution**





**Three-factor solution**





**References**

1. Johnson, A. R. and Wichern, D. W. “Applied Multivariate Statistical Analysis”
2. Li, Y. “lectures slides of Applied Multivariate Statistical Analysis”
3. Hewson, P. J. “Multivariate Statistics with R”
4. <http://www.statmethods.net/advstats/factor.html>