Homework 3

testdata.txt concerns the results of 88 students on a five part exam:
1) Closed Book on Mechanics
2) Closed Book on Vectors;
3) Open Book on Algebra;
4) Open Book on Analysis
5) Open Book in Statistics.

Homework 3

a) Choose to either use the correlation matrix or the covariance matrix for your principal components analysis and justify your choice.

b) Describe how much information would be lost if the data was summarized using only 1, 2, 3, or 4 components.
Homework 3

c) Find an interpretation for what each of the first four components measure.

d) Based on your answers to B and C, what information in particular seems to be lost by using only the first four components.

Homework 3

e) Verify your answer in d (or maybe refute and revise it) by using multiple regression to predict each of the five original scores from only the first four principal components.

Factor Analysis

Assuming the X have been standardized:

\[
X_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1k}F_k + u_1
\]

\[
X_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2k}F_k + u_2
\]

\[
X_q = \lambda_{q1}F_1 + \lambda_{q2}F_2 + \cdots + \lambda_{qk}F_k + u_q
\]
Assumptions

- The $u_i$ are independent of each other and of the $F_i$
- The $F_i$ are independent of each other
- Usually set the $F_i$ to have mean 0 and variance 1

PC-FA on the Testdata

In PC we have that

$$Y_{(nxq)} = X_{(nxq)}A_{(qxq)}$$

We could just try reversing it, and use the $Y$'s (the Principal Components) as our factors:

$$X_{(nxq)} = Y_{(nxq)}A^{-1}_{(qxq)}$$

So $A^{-1}_{(qxq)}$ is one idea for $\Gamma_{(qxq)}$

A Fact that Makes it Easy

The matrix $A$ for principal components is an orthogonal matrix, so that

$$A^{-1}_{(qxq)} = A^T_{(qxq)}$$
And so….

The only problem is that the Y have variances equal to the eigenvalues and we want the F to have variance 1.

So before taking the transpose, multiply each column of A by the standard deviation of that component (the square root of the eigen value).

How Well Does it Work?

• What are the communalities?  
  (Sum of the $\lambda_i^2$)
• What are the specificities?  
  (1-Sum of the $\lambda_i^2$)
• Does the model give the right correlation?  
  ($\Lambda_{qk}^T\Lambda_{kq}^T+\text{diag}(\text{specificities})$)

Can We Do Better

Principal Factor Factor Analysis says to estimate the specificity and subtract that from the correlation matrix first.

Then do principal components on that “reduced correlation matrix”.

For example, could use the largest correlation a variable has with any other.