THE UNIVERSITY OF MINNESOTA

Statistics 5401 November 14, 2005

Assignment Sheet No. 7

Reading

Week of November 14 - 18; J&W, Sec 9.3-9.6, 10.1-10.2
Week of November 21 - 23; J&W, Sec. 10.3-10.5,
Week of November 28 - December 2 J&W, Chapter 11
Week of December 5 - 9 J&W, Sec. 12.1-12.3
Week of December 12 - 14 J&W, remainder of Chapter 12

Final Examination: A **take home** final examination will be handed out in the next to final class, Monday, December 12. It will be due by 12 noon on Tuesday December 20.

Written Assignment (due in class Friday November 25)

1. Here are the first few lines of matrix `T09_12` in file `JW5Data.txt` containing data from Table 9.12 on page 539-540 of Johnson and Wichern.

```
T09_12    50    7 format
) Data from Table 9.12 p. 539 in
) Applied Multivariate Statistical Analysis, 5th Edition
) by Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002
) These data were edited from file T9-12.DAT on disk from book
) Salespeople data
) Col. 1: x1 = Index of sales growth
) Col. 2: x2 = Index of sales profitability
) Col. 3: x3 = Index of new-account sales
) Col. 4: x4 = Score on creativity test
) Col. 5: x5 = Score on mechanical reasoning test
) Col. 6: x6 = Score on abstract reasoning test
) Col. 7: x7 = Score on mathematics test
(3f6.1,4f3.0)
  93.0  96.0  97.8  9 12  9 20
  88.8  91.8  96.8  7 10 10 15
  95.0 100.3  99.0  8 12  9 26
 101.3 103.8 106.8 13 14 12 29
 102.0 107.8 103.0 10 15 12 32
  95.8  97.5  99.3 10 14 11 21
  95.5  99.5  99.0  9 12  9 25
  . . . . . . . . . . . . . . .
```

Assume an orthogonal factor model for the standardized variables $X_i/\sqrt{\sigma_{ii}}$. This implies you should factor analysis starting with the correlation matrix.

(a) Obtain unrotated factor analysis estimates for both 1 factor and 2 factor models using the *principal component* method, the *iterated principal factor* method, and the *maximum likelihood* method. Use `facanal()` for all but the principal component method. Your results should include the estimated factor loadings, uniquenesses, and communalities. Note that the factor extraction macros in file `mulvar.mac` save the most recent results in variables `CRITERION` (the quantity being minimized, computed from eigenvalues), `PSI` (uniquenesses) and `LOADING` (p by m matrix of unrotated loadings).

(b) Compare the 1 and 2 factor maximum likelihood solutions. Which choice of m do you prefer?

(c) For both m = 1 and m = 2, test the null hypothesis that $\mathbf{\Sigma}$ is of factor analytic form. Do the results modify your opinion about the choice of m?

(d) Based the m = 2 maximum likelihood solution, estimate the actual factor scores for each of the 50 salespersons. Make a scatter plot of factor 2 scores against factor 1 scores. Note you will need to standardize the X_i's or unstandardize your estimated factor structure to compute estimated factor scores..

(e) Use MacAnova function `rotation()` to do a varimax (`rotation(loadings, method:"varimax",kaiser:T`) and quartimax (`rotation(loadings, method: "quartimax",kaiser:T`) rotations of the maximum likelihood estimated loading matrix in the m = 2 case. Or redo the MLE estimation using keyword phrase `rotation:"quartimax"` as an argument to `facanal()`. Make a scatter plots of the columns of the original loading matrix and the rotated loading matrix. Comment on the differences.

2. Here are the first few lines of data set `pollendata.txt` available from
`http://www.stat.umn.edu/~kb/classes/5401/datafiles.html/#Data7`

```
pollen         481     5 LABELS NOTES
) This is part of a synthetic data set generated by David Coleman at
) RCA Laboratories in Princeton, N.J. that purpots to represent
) measurements made on pollen grains.  The first three variables are
) the lengths of geometric features observed in sampled pollen grains
) - in the x, y, and z dimensions: a "ridge" along x, a "nub" in the y
) direction, and a "crack" along the z dimension.  The fourth variable
) is pollen grain weight, and the fifth is density.
) Retrieved from http://lib.stat.cmu.edu/datasets/pollen.data
) Col. 1: Ridge (x dimension)
) Col. 2: Nub (y dimension)
) Col. 3: Crack (z dimension)
) Col. 4: Weight
) Col. 5: Density
)"%lf %lf %lf %lf %lf"
 -2.3482   3.6314   5.0289  10.8721  -1.3852
 -1.1520   1.4805   3.2375  -0.5939   2.1235
 -2.5245  -6.8633  -2.8037   8.4631  -3.4126
  5.7523  -6.5091  -5.1510   4.3480 -10.3261
 . . . . . . . . . . . . . . . . . . . . . .
```

(a) Perform a maximum likelihood factor analysis with m = 1 and m = 2 factors, using varimax rotation. Start with the correlation matrix.

(b) Describe in words the pattern of loadings found in the m = 2 analysis . Use phrases like "Weight loads strongly on both F1 and F2" or "Weight loads strongly on F1 but not on F2".

(c) Use a formal test to test the null hypotheses H_0: The correlation matrix has factor analytic structure with 1 factor and H_0: The correlation matrix has factor analytic structure with 2 factors.

MacAnova notes

Use `facanal()` for any estimation other than principal components. This can sometimes be a little tricky although the updated version should work better than the old. You may have to "tweak" things a little bit to get it to converge to something sensible. If all the uniqueness are essentially 0, it may have converged to something not sensible.

Things to tweak:
The convergence criterion. The default settings are quite stringent making convergence hard to achieve, especially when near a Heywood case. Iteration stops the first time one of the following three things happens:

1. The change in `psihat` (actually, `log(psihat)`) is small enough
2. The change in the objective function (quantity being minimized) is small enough
3. The gradient (vector of derivatives of the objective function with respect to the $\log(\psi_i)$) is small enough.

You can specify what "small enough" means using keyword phrase `crit:vector(nsig1,nsig2,delta)`, where each number defines "small enough" for one of the three convergence criteria. A negative value specifies the criterion is ignored.

`nsig1` should be a positive integer specifying the number of decimals in any estimated log uniquenesses that should be accurate. Thus when `nsig1` = 5, iteration stops when all the changes in log uniquenesses are less than .00001 = 10^{-5}.

`nsig2` should be a positive integer specifying the number of decimals in the minimized objective function. Thus when `nsig2` = 5, iteration stops when the change in the objective function is less than .00001 = 10^{-5}.

`delta` should be a small positive quantity δ. Iteration stops when $||\gamma|| < \delta$ where γ is the gradient vector.

The default is `crit:vector(5,8,-1)`, that is log uniquenesses should be accurate to 5 decimals, the minimized objective function should be accurate to about 8 decimals accuracy and the gradient is ignored.

If you have difficulty with convergence, you may do better with `crit:vector(5,5,-1)`.

The minimizer used. `facanal()` uses one of three minimizing methods. You use keyword phrase `minimizer:minmeth`, where `minmeth` is one of `"dfp"` (Davidon-Fletcher-Powell), `"dfp"`), `"bfs"` (Broyden-Fletcher-Shanno), or `"broyden"` (Broyden), to choose the minimizer. The default is `minimizer:"bfs"`. If that doesn't work, try `minimizer:"broyden"` or `minimizer:"dfp"`.