Exam 2: Is This Test Really Necessary?

36-402, Advanced Data Analysis

Due at 10:30 am on Tuesday, 17 April 2012

You can use your notes, the textbooks, and indeed anything you find in the library or online, if that is properly acknowledged. However, **all your work must be your own**. You cannot work with classmates, friends, a tutor, or anyone else. If you are unclear about what is allowed and what is not, please check the university policy on cheating and plagiarism (`http://www.cmu.edu/policies/documents/Cheating.html`), or ask the professor.

Please include the following text in your write-up:

> I, YOUR NAME, have completed this examination honestly, without giving prohibited assistance to anyone, or receiving it from anyone.

If, for reasons of conscience, you are unable to make such an affirmation, let the professor know at once, to arrange for an oral mid-term.

The dataframe `mathmarks` in the library `SMPracticals`[1] contains scores for 88 university students in five mathematical subjects: vectors, algebra, analysis, statistics, and mechanics. All subjects were scored out of 100 points. We will use multivariate methods to explore how this data, and the relationships between the variables.

When asked to report numbers, give only two significant figures (not decimal places!) unless told otherwise.

1. (5) What is the sample correlation matrix among the grades?

2. (5) Find the least-squares coefficients for predicting the grade in statistics as a linear combination of the other four grades. What is the (in-sample) root-mean-squared error of this regression?

3. (10) Fit a factor model with one common factor, using `factanal`. Report the factor loadings and the uniquenesses, and explain what the numbers mean.

4. (10) Calculate the correlation matrix among the grades implied by the factor model. Why is it not the same as the sample correlation matrix? How does it differ?

5. (5) Calculate the covariance matrix implied by the factor model. How does it compare to the sample covariance matrix?

6. (10) Calculate the coefficients for predicting the grade in statistics as a linear combination of the other four grades, under the factor model. What is the (in-sample) root-mean-squared error? *Hint:* §14.2.2.

7. (5) The `factanal` function reports a test which compares the factor model to an unrestricted multivariate Gaussian. What is the p-value of this test, and what does it tell you about how well the factor model fits?

8. (10) Under the factor model, each variable should have a Gaussian distribution. Check this for each of the five grades, using `ddst.norm.test`, and describe your results.

9. (10) Using `mvnormalmixEM` from the `mixtools` package, fit a multivariate Gaussian mixture model with two components. Report the mean vector of each Gaussian, its covariance matrix, and the two mixing weights.

10. *Mixture of Gaussians vs. factor model* (30 total)

    (a) (10) Write a function to draw n vectors from a mixture of two multivariate Gaussians. *Hint:* select a random Gaussian (how?) and then call `rmvnorm`.

[1] You can also get the data file from `http://statwww.epfl.ch/davison/SM/`.

(b) (5) Write a function which takes a data frame, fits a one-factor model to it, and returns the p-value. Check that it works by making sure it gives the same answer as what you got in Problem 7

(c) (10) Repeatedly simulate data sets of size 88 from the two-Gaussian mixture you fit in problem 9. What fraction of them fit a one-factor model at least as well as the data?

(d) (5) What can you conclude about whether to prefer a one-factor or two-mixture model for this data?