Math 189: Homework 7

Factor Analysis of USDA Women's Health Survey

In this assignment you will study the USDA Women's Health data set. In 1985, the USDA commissioned a study of women's nutrition. Nutrient intake was measured for a random sample of 737 women aged 25-50 years. The variables may together represent facets of *health*. We seek a factor model, where the latent factors will explain the major nutritional features.

The task is to develop a factor model to better explain the main drivers of nutrition, and to see whether dimension reduction is possible. This data can be found on GitHub.

Metadata for Nutrient Dataset

The following variables were measured:

1. Calcium (mg)
2. Iron (mg)
3. Protein (g)
4. Vitamin A (μg)
5. Vitamin C (mg)

Tasks

Analyze the dataset according to the following steps:

1. Explore the data graphically in order to investigate the correlations between variables. Make the case for correlation to a non-technical audience by using a level plot.
2. Fit the factor model using both PCA and MLE, and compare the parameter estimates. Discuss the underlying assumptions for each method. Which results do you prefer, and why?
3. Use a scree plot to decide on a dimension reduction, and justify your choice.
4. Examine the factor loadings, and discuss in your report which variables have high or low loadings. Can you associate an interpretation to your factors?
5. Examine the factor scores by scatter plots or pairwise scatter plots. Is there a story to tell from these results?
6. Summarize your findings and try to tell a nice story with this data analysis.

Remarks

Your R Markdown Notebook report should have a introduction, body, conclusion (and optional appendix). Importantly, your code should run!