Factor Analysis

Rebecca C. Steorts, Duke University

STA 325, Chapter 10 ISL

Agenda

Add later

From PCA to Factor Analysis

Our data in practice won't be altogether accurate — it will have noise in it which may be due to measurement error.

What is measurement error?

Measurement error results from the fact that the person who created the data set didn't create them perfectly and hence they have intristic noise in them.

Examples: transcription errors, collection errors, etc.

PCA versus Factor Analysis

Let's start with PCA.

- PCA doesn't care about measurement error.
- PCA will try to reproduce true-value-plus-noise from a small number of components.
- Can we do something like PCA, where we reduce a large number of features to additive combinations of a smaller number of variables, but which allows for noise (measurement error)?

Starting simple

The simplest model starting with PCA would be the following:

- Suppose each object or data point has p features.
- That is X_{ij} is the value of feature j for object i.
- Let's assume all observations are centered.

Factor analysis

- Let's assume there are q factor variables and each observation is a linear combination of the factor scores F_{ir} plus some noise:

$$X_{ij} = \sum_{r=1}^{k} F_{ir} w_{rj} + \epsilon_{ij} \tag{1}$$

- w_{rj} are factor loadings of the observed features. They say how much feature j changes (on average) in response to a one-unit change in factor score r.
- ϵ_{ij} is the noise for feature j on object i. We assume $\epsilon_{ij} \sim (0, \phi_j)$.
- We further assume that $E[\epsilon_{ij}\epsilon_{\ell m}] = 0$ unless $i = \ell, j = m$ (each object and feature has uncorrelated noise).

Factor analysis

We can rewrite the model as

$$X_i = \epsilon_i + F_i w_{q \times p} \tag{2}$$

We can stack the vector into a matrix and get

$$X = \epsilon + F w_{q \times p} \tag{3}$$

Our task is to estimate the factor loadings w, factor scores \$, and the variances ϕ_j.

General question: where did this model come from? We made hypothesized it and we check if it works. We do this in practice with all models.

Preserving correlations

- PCA aims to preserve variance, or (what comes to the same thing) minimize mean-squared residuals (reconstruction error).
- But it doesn't preserve correlations.
- That is, the correlations of the features of the image vectors are not the same as the correlations among the features of the original vectors (unless $q = p$, and we're not really doing any data reduction).
- We might ask for a set of vectors whose image in the feature space will have the same correlation matrix as the original vectors, or as close to the same correlation matrix as possible while still reducing the number of dimensions.
- This also leads to the factor analysis model, but we need to take a somewhat circuitous root to get there.

Factor Estimation

- Assume all the factor scores are uncorrelated with each other and have variance 1.
- also that they are uncorrelated with the noise terms.
- We'll solve the estimation problem for factor analysis by reducing it to an eigenvalue problem again.
- Since, again this requires linear algebrea and is beyond the scope of ISLR and the pre-reqs, we omit the details.

Application to states data

We will consider first an application of PCA and compare this to Factor Analysis (FA) on state data in R.

The data contains information on all 50 US states from 1977.

Application to states data

```
data(state)
summary(state.x77)
```

```
##    Population        Income       Illiteracy       Life
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.
##  Median : 2838   Median :4519   Median :0.950   Median
##  Mean   : 4246   Mean   :4436   Mean   :1.170   Mean
##  3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.
##  Max.   :21198   Max.   :6315   Max.   :2.800   Max.
##      Murder          HS Grad          Frost
##  Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.
##  1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Q
##  Median : 6.850   Median :53.25   Median :114.50   Media
##  Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean
##  3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Q
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.
```

Interpretation of the first two PCs

- Says all the variables except population, illiteracy, and murder have a negative project on the first PC.
- Area hardly projects onto the first PC at all.
- What about the second PC?

Let's look at a biplot

```
biplot(state.pca)
```