Section 11: Factor Analysis.

Ensure you have completed all previous worksheets before advancing.

1 Model

To obtain maximum likelihood estimates of the factor loadings and the uniqueness values recall that the factor analysis model is:

\[ X - \mu = \Lambda f + \epsilon \]

It is assumed that:

\[ f \sim \text{MVN}(0, I) \]

\[ \epsilon \sim \text{MVN}(0, \Psi) \]

Here \( \Psi \) is a diagonal matrix.

When \( f \) and \( \epsilon \) are normally distributed then the observations are also normally distributed and estimates of \( \Lambda \) and \( \Psi \) can be obtained by maximizing the likelihood.

2 R Implementation

Download the cereal data set that is available in csv format from the usual location (https://www.scss.tcd.ie/Brett.Houlding/ST3011.html). This data consists of 235 observations, each of which is the perception of a customer for a certain brand of cereal. Respondents used a five point scale to indicate the extent to which the cereal brand possessed each of 25 different attributes. Load this data into R.

One objective of the survey was to try and determine which cereal brands the customer would consider purchasing as a function of the underlying characteristics of the available brands. Factor analysis can be used to reduce the dimensionality of the attribute data and uncover a smaller number of underlying (unobserved) factors which account for a large amount of the variance in the original measures.

A function in R that implements factor analysis is \texttt{factanal} (examine its help file as usual). By default \texttt{R} fits a factor analysis model using the ‘varimax’ rotation technique. However, for a first look it will probably be best to see the results without implementing a rotation. Also note that a principal components analysis suggests that a 4 or 5 component model may be sufficient for explaining variation (you could verify this yourself):

\begin{verbatim}
> f=factanal(cereal[,2:26], factors=4, rotation="none")
\end{verbatim}
Note that R standardizes the data before it fits the factor model. Hence the variance of each variable should be equal to 1 and the sum of the variances of all variables should be 25. As such, what can you infer from the result of the following command?

```r
> sum(f$uniqueness)
```

You should find that \((25 - \text{sum}(f$uniqueness))/25\) gives you the same value as the cumulative proportion of variance for the model.

**Task:** Determine the communality value for the ‘Filling’ variable:

```r
>
```

### 3 Rotations

In class we discussed how multiplying the loading matrix \(\Lambda\) by an orthogonal matrix \(G\) (so that \(GG^T = I\)) will give equivalent results. The varimax rotation that R implements by default aims to provide a rotation in which loadings are either very large or very small. This choice of rotation can then help with the interpretation of the factor results.

**Task:** Fit a factor model \(f2\) that uses the varimax rotation:

```r
>
```

You should notice that the variation accounted for by the factors has not changed. This is to be expected, as the rotation should only affect the loadings. A comparison of \(f2$loadings\) with \(f$loadings\) should also confirm that the varimax rotation has reduced the number of mid-range valued loadings.

**Exercise:** Can you give an interpretation to the unobserved latent factors resulting from the varimax rotation?

To more easily visualise the loadings on a given factor, a plot such as that produced by the following command may be useful:

```r
> plot(f2$loadings[,1],type="n",xlim=c(-5,30))
> text(c(1:25),f2$loadings[,1],names(cereal)[2:26])
```

**Exercise:** Plot the values of the other factor loadings and check if your previous intuition concerning their interpretation holds.

A pairwise plot may also ease interpretation:

```r
> plot(f2$loadings[,1],f2$loadings[,2],type="n",xlim=c(-0.5,1),ylim=c(-0.5,1))
```
4 Further analysis

Exercise: Investigate the results of using more than 4 factors? Do factors still have an intuitive interpretation in these instances?

Exercise: What effect, if any, is there in using an oblique rotation? An oblique rotation can be implement using the additional argument rotation="promax".

Sometimes the factor scores themselves will also be of interest (possibly for classification or clustering purposes). These will also be calculated when the argument scores="regression" is added to the factanal function.

Task: Find the factor scores for the cereal data when there are 4 factors and varimax rotation is applied:

> text(f2$loadings[,1],f2$loadings[,2],labels=colnames(cereal)[2:26])

Exercise: Try a factor analysis on some of the other data sets you have seen in this course. Do the factors have an intuition in these cases and do the lower dimensional factor scores aid any classification or clustering algorithm previously considered? For example, try colouring the factor scores of the olive data for four factors where the colouring is according to the region the sample was obtained.