T1

Monday, August 2, 2021 9:15 PM

Applied

1. Complete Tasks 6 and 7 in lecture note Week 1. That is, perform the PCA analysis on the 'eurojob.txt' data set.

Example 5.

In this example we are going to look at a new data set on employment in 26 European countries but you are going to be doing all the work!

The data gives for each of 26 European countries the percentage of the total workforce employed in nine different industries in 1979 (Hand et al, 1994).

Variable name	Description
Agriculture	% employed in agriculture
Mining	% employed in mining
Manufacture	% employed in manufacturing
Power	% employed in power supply industries
Construction	% employed in construction
Service	% employed in service industries
Finance	% employed in finance
Social	% employed in social and personal services
Transport	% employed in transport & communications

```
employ<-read.table("eurojob.txt",header=T,row.names=1);head(employ,4)

        AGRIC MINING MANU POWER CONSTR SERVICE FINANCE SOCIAL TRANS
Belgium   3.3    0.9 27.6   0.9    8.2    19.1     6.2   26.6   7.2
Denmark   9.2    0.1 21.8   0.6    8.3    14.6     6.5   32.2   7.1
France   10.8    0.8 27.5   0.9    8.9    16.8     6.0   22.6   5.7
WGerm     6.7    1.3 35.8   0.9    7.3    14.4     5.0   22.3   6.1
```

Task 6.

- Produce numerical summaries and comment on them.
- Produce a sensible plot or plots for these data and comment on them.
- Produce two important numerical summaries for deciding on how to run PCA and to tell how successful it is likely to be. Comment on these.
- Run PCA on the appropriate matrix and look at the output.
- Assuming we are most concerned with preserving information, how many coefficients should we retain if we want to have 90% of the original variability kept?
- Assuming we want to use Cattell's method, how many components would we retain?
- Assuming we want to use Kaiser's method, how many components would we retain?
- Assuming we have decided to retain 2 components, is there any useful interpretation to be had for these?

Task 7.

Say we have the following entries for our observations

```
obs1<-c(5.1,0.5,32.3,0.8,8.1,16.7,4.3,21.2,6.3)
obs2<-c(4.2,0.7,25.4,0.7,9.3,15.0,5.8,31.0,6.9)
```

Calculate the scores for the two new observations and produce a scatterplot of the data's scores for the first 2 PCs and comment.

Conceptual

1. Complete Task 5 in lecture note Week 1. That is, calculate the first component score for the new observation by hand.

Task 5.

Try to calculate the first component score for the new observation $(12, 4, 3, 25, 100, 2, 1, 0.4, 2, 4, 1, 2, 600)$ first by hand (using R as a calculator) and then using the `predict` command. You will need the centring vector, `wine.pca$center`, and the scaling vector, `wine.pca$scale` as well as the first component loadings, `wine.pca$loadings[,1]`.

You will have to centre the new observation by taking away the centre vector. Then, because we used the correlation matrix and so we were working with standardised data, you have to scale the resulting centred vector by dividing by the scale vector. Finally you should take the inner product of the resulting standardised version of the new observation with the vector of first principal component loadings, resulting in the score.

Answer to Task 5. To calculate the first component score

```
#By hand using R as a calculator
new.x<-matrix(c(12,4,3,25,100,2,1,0.4,2,4,1,2,600),nrow=1)
colnames(new.x)<-colnames(wine.new)
centre.wine<-wine.pca$center
scale.wine<-wine.pca$scale
first.load<-wine.pca$loadings[,1]

new.x.cent<-new.x-centre.wine
new.x.stand<-new.x.cent/scale.wine
new.x.stand%*%first.load

          [,1]
[1,] -1.852166

#Using the predict command
predict(wine.pca,as.data.frame(new.x))

        Comp.1   Comp.2    Comp.3    Comp.4     Comp.5    Comp.6    Comp.7
[1,] -1.852166 0.196511 -2.909757 -0.290249 -0.5356328 0.1049906 -0.488592
        Comp.8    Comp.9    Comp.10   Comp.11   Comp.12   Comp.13
[1,] 0.1793053 -1.296705 -0.7004643 0.4829155 0.1600023 0.7121765
```

We can see the answer by hand is the same as the first element of the `predict` result. If we were using PCA on the covariance matrix, we would skip the scaling stage (only centering before taking the inner product).

2. Calculate the correlation between the k^{th} original variable X_k and the i^{th} principal component Y_i.

$e_i = (e_{1i}, \cdots e_{pi})^T$ is the ith eigenvector, Σ is cov matrix w/ diag elemnt of $(\sigma_{11}, \cdots \sigma_{pp})$

let $v_k = (0, \cdots 0, 1, 0 \cdots 0)$ be the selector of the k^{th} elemnt from a $p \times 1$ vector

$$\begin{aligned} Cov(Y_i, X_k) &= Cov(e_i^T X, v_k X) \\ &= e_i^T \Sigma v_k = v_k^T \Sigma e_i \\ &= v_k^T \lambda_i e_i = v_k^T e_i \lambda_i \\ &= e_{ki} \lambda_i \end{aligned}$$

$$\rho_{Y_i, X_k} = \frac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}} = \frac{e_{ki}\lambda_i}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

3. (optional) Show that principal components for standardised variables can be obtained from the eigenvectors of the correlation matrix $\boldsymbol{R}$ of the random vector $\boldsymbol{X}$.

let standardised variable $X = (X_1, \cdots, X_p)^T$ to $Z = (Z_1 \cdots Z_p)^T$ where $Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$

using a diagonal matrix $V^{\frac{1}{2}}$ for standard deviation $\sqrt{\sigma_{ii}}$, we can write the standardisation in matrix notation

$$Z = (V^{\frac{1}{2}})^{-1}(X - \mu)$$

$$E(Z) = 0$$

$$\Rightarrow Cov(Z) = (V^{\frac{1}{2}})^{-1} Cov(X) (V^{\frac{1}{2}})^{-1} = (V^{\frac{1}{2}})^{-1} \Sigma (V^{\frac{1}{2}})^{-1} = R$$