Handout #18

Example of Principal Components: Sparrow Data

Recall the description of the "Sparrow Data" from an earlier handout: Bumpus collected data on $n = 49$ moribund female sparrows after a severe storm. Measurements were made on $p = 5$ characteristics: X_1=total length (mm); X_2=alar (wing) length (mm); X_3=length of beak and head (mm); X_4=length of humerus (mm); and X_5=length of keel of sternum (longitudinal length of breast bone, mm). Subsequently, 28 of the 49 female sparrows died. For this analysis we will ignore whether or not a sparrow died.

Principal Components Analysis of the Covariance Matrix S

Can we summarize the variance-covariance structure of this data by three or fewer dimensions (*i.e.*, principal components)?

The output of the PC analysis on the sample covariance matrix ***S*** is provided below. The eigenvalues of ***S*** in decreasing order are $\hat{\lambda}_1 = 35.33$, $\hat{\lambda}_2 = 4.62$, $\hat{\lambda}_3 = 0.63$, $\hat{\lambda}_4 = 0.31$ and $\hat{\lambda}_5 = 0.08$. The sum of the diagonal (highlighted) elements of ***S*** is 40.97 which is also (as required) the sum of the $p = 5$ eigenvalues of ***S***.

The first principal component accounts for 35.33/40.97 = 0.86, or 86% of the variation in the data. PC_1 is $\hat{Y}_1 = \hat{\underline{\mathbf{e}}}_1'\underline{\mathbf{X}} = 0.54X_1 + 0.83X_2 + 0.10X_3 + 0.07X_4 + 0.10X_5$, thus it is essentially a linear combination of X_1=total length and X_2=alar length, *i.e.*, it is approximately $\hat{Y}_1 \approx 0.54X_1 + 0.83X_2$.

The first and second principal components combined account for (35.33+4.62)/40.97 = 0.9751 or 98% of the variation in the data. PC_2 is $\hat{Y}_2 = \hat{\underline{\mathbf{e}}}_2'\underline{\mathbf{X}} = -0.83X_1 + 0.55X_2 - 0.03X_3 + 0.01X_4 - 0.10X_5$ which is approximately $\hat{Y}_2 \approx -0.83X_1 + 0.55X_2$. Thus the first two principal components summarize the 5 variables in the data by two different linear combinations of the first two variables.

Covariance Matrix

	TotalLength	AlarExtent	BeakandHead	Humerus	KeelofSternum
TotalLength	**13.353741**	13.610969	1.9220663	1.3306122	2.1922194
AlarExtent	13.610969	**25.682823**	2.7136054	2.1977041	2.6578231
BeakandHead	1.922066	2.713605	**0.6316327**	0.3422662	0.4146471
Humerus	1.330612	2.197704	0.3422662	**0.3184184**	0.3393707
KeelofSternum	2.192219	2.657823	0.4146471	0.3393707	**0.9828231**

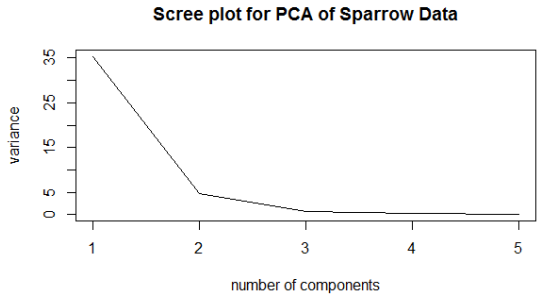
Principal Components: on Covariances

```
> PCAfitCov <- prcomp(Spar[,2:6])
> PCAfitCov
Standard deviations:
[1] 5.9435475 2.1499905 0.7943033 0.5592706 0.2784261

Rotation:
                     PC1         PC2         PC3         PC4         PC5
TotalLength   0.53650052 -0.82809990 -0.15649065 -0.04020969 -0.01765243
AlarExtent    0.82901535  0.55051223 -0.05774395 -0.06902156  0.03964203
BeakandHead   0.09649615 -0.03356237  0.23751487  0.89762653  0.35695288
Humerus       0.07435219  0.01459529  0.20324541  0.30724056 -0.92658150
KeelofSternum 0.10030441 -0.09923405  0.93512262 -0.30575979  0.11021920

> summary(PCAfitCov)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     5.9435 2.1500 0.7943 0.55927 0.27843
Proportion of Variance 0.8622 0.1128 0.0154 0.00763 0.00189
Cumulative Proportion  0.8622 0.9751 0.9905 0.99811 1.00000
```

Scree Plot

Use the scree plot to help determine the number of principal components to use to summarize a set of variables.

```
> plot(c(1,2,3,4,5),PCAfitCov$sdev^2,type="l",main="Scree plot for PCA of Sparrow
Data",xlab="number of components",ylab="variance")
```



Look for an "elbow" in the plot: two PC's are adequate to summarize most of the variation and correlation evident amongst the five variables.

Principal Components Analysis of the Correlation Matrix R

When the data under analysis exhibit large differences in sample variances across variables (the diagonal elements of ***S***), the results of a principal components analysis of ***S*** will more heavily weight the variables with the larger variances. For this reason it is often recommended that PCA be performed on standardized (centered and scaled) data. This is equivalent to performing PCA on the sample correlation matrix ***R*** (instead of ***S***) in order to remove the effects of scale differences in variables.

In this example, examination of the diagonal elements of ***S*** indicate that the largest and smallest sample variances (25.68 and 0.32) differ by a factor of 80 indicating that if we wish each variable to be considered equally regardless of the variability in its measurement, then PCA should be applied to ***R*** rather than ***S***. Can we summarize the correlation structure of this data by three or fewer directions (principal components) in 5-dimensional space?

Correlation Matrix

```
              TotalLength AlarExtent BeakandHead   Humerus KeelofSternum
TotalLength     1.0000000  0.7349642   0.6618119 0.6452841     0.6051247
AlarExtent      0.7349642  1.0000000   0.6737411 0.7685087     0.5290138
BeakandHead     0.6618119  0.6737411   1.0000000 0.7631899     0.5262701
Humerus         0.6452841  0.7685087   0.7631899 1.0000000     0.6066493
KeelofSternum   0.6051247  0.5290138   0.5262701 0.6066493     1.0000000
```

Principal Components: on Correlation

```
> PCAfitCor <- prcomp(Spar[,2:6], scale=TRUE)
> PCAfitCor
Standard deviations:
[1] 1.9015726 0.7290433 0.6216306 0.5491498 0.4056199

Rotation:            PC1         PC2        PC3         PC4         PC5
TotalLength   0.4517989 -0.05072137  0.6904702 -0.42041399  0.3739091
AlarExtent    0.4616809  0.29956355  0.3405484  0.54786307 -0.5300805
BeakandHead   0.4505416  0.32457242 -0.4544927 -0.60629605 -0.3427923
Humerus       0.4707389  0.18468403 -0.4109350  0.38827811  0.6516665
KeelofSternum 0.3976754 -0.87648935 -0.1784558  0.06887199 -0.1924341

> summary(PCAfitCor)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.9016 0.7290 0.62163 0.54915 0.40562
Proportion of Variance 0.7232 0.1063 0.07728 0.06031 0.03291
Cumulative Proportion  0.7232 0.8295 0.90678 0.96709 1.00000
```

Here the eigenvalues of $\boldsymbol{R}$ are $\hat{\lambda}_1 = 3.62$, $\hat{\lambda}_2 = 0.53$, $\hat{\lambda}_3 = 0.39$, $\hat{\lambda}_4 = 0.30$ and $\hat{\lambda}_5 = 0.16$ and the first principal component accounts for $\hat{\lambda}_1 / p = 3.62/5 = 0.72$ or 72% of the variance among the standardized variables. PC_1 is $\hat{Y}_1 = \underline{\hat{\mathbf{e}}}_1' \underline{\mathbf{z}} = 0.45z_1 + 0.46z_2 + 0.45z_3 + 0.47z_4 + 0.40z_5$, weighting each variable nearly equally. Thus the first PC may be interpreted as a measure of "overall size" of a sparrow.

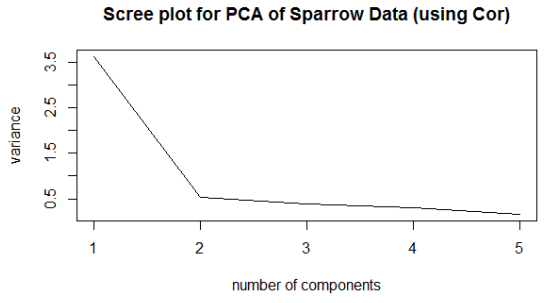
Adding in a second principle component, the two PCs together account for approximately 83% of the variance among the standardized variables. PC_2 is given by $\hat{Y}_2 = \underline{\hat{\mathbf{e}}}_2' \underline{\mathbf{z}} = -0.05z_1 + 0.30z_2 + 0.32z_3 + 0.18z_4 - 0.88z_5$. The coefficient of z_1, standardized total length, is negligible here. Roughly this component seems to be comparing (standardized) keel of sternum to a linear combination of (standardized) alar extent, beak and head, and humerus. Whether or not this has a physically meaningful interpretation could possibly be determined by an expert.

Are two components sufficient to summarize the structure of the standardized variables? Note that to achieve the same "level" of variance explained with PCA applied to $\boldsymbol{R}$ (the variance-covariance matrix of the standardized variables) as was achieved (97.5%) with two PCs for the analysis of $\boldsymbol{S}$, would require all five PCs here – no data reduction.

In addition to assessing the total proportion of variation explained by the PCs chosen, one approach to choosing the number of principal components suggested by the textbook is to examine the "scree plot" – a plot of the eigenvalues ordered from largest to smallest against their order number. Use the scree plot to choose as the number of components the point on the plot where there is an "elbow" or bend, that is where there appears a noticeable change from vertical decrease to something more horizontal – indicating that subsequent principal components are associated with small eigenvalues of roughly the same magnitude (see the discussion on page 445, just prior to Example 8.4).

Whether one is applying principal components analysis to $\boldsymbol{S}$ or $\boldsymbol{R}$, the scree plot below also seems to indicate that the elbow appears at $i = 2$, arguing for the use of just two principal components to summarize this data.

```
> plot(c(1,2,3,4,5),PCAfitCor$sdev^2,type="l",main="Scree plot for PCA of Sparrow
Data (using Cor)",xlab="number of components",ylab="variance")
```



(Courtesy of Dr. Roy St. Laurent, modified 11/07/2016)