Statistics 135

Chapter 8

Principal Components

`Chris Drake`

Department of Statistics

University of California, Davis

Population Principal Components

One aspect of data, covered in chapters 5 and 6, is the mean structure. Another aspect is the variance-covariance structure. Principal components is concerned with explaining this structure through a few linear combinations of the variables.

1 $\mathbf{X}' = (X_1, ..., X_p)$ has covariance matrix $\mathbf{\Sigma}$

2 $\lambda_1 > \lambda_2 > ... > \lambda_p$ are the eigenvalues of $\mathbf{\Sigma}$

3 Consider linear combinations

$$Y_i = \mathbf{a}_i'\mathbf{X} = \sum_{j=1}^{p} a_{ij} X_j$$

$$Y_k = \mathbf{a}_k'\mathbf{X} = \sum_{j=1}^{p} a_{kj} X_j$$

$$Var(Y_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i \qquad Cov(Y_i, Y_k) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_k$$

for $i, k = 1, .., p$.

1 Def: the first principal component is defined as the linear combination $Y_1 = \mathbf{a}_1'\mathbf{X}$ such that the variance of Y_1 is maximized subject to $\| \mathbf{a}_1 \| = 1$.

2 Def: the second principal component is defined as the linear combination $Y_2 = \mathbf{a}_2'\mathbf{X}$ such that the variance of Y_2 is maximized subject to $\| \mathbf{a}_2 \| = 1$ and $Cov(\mathbf{a}_1\mathbf{X}, \mathbf{a}_2\mathbf{X}) = 0$.

3 Def: the k^{th} principal component is defined as the linear combination $Y_k = \mathbf{a}_k'\mathbf{X}$ such that the variance of Y_k is maximized, $Cov(\mathbf{a}_j\mathbf{X}, \mathbf{a}_k\mathbf{X}) = 0$ for $j = 1, .., k-1$.

4 If $\mathbf{\Sigma}$ is $Cov(\mathbf{X})$ and $\lambda_1 > \lambda_2 > ... > \lambda_p$ with associated eigenvectors $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_p$, then

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p$$

and

$$\begin{aligned} Var(Y_i) &= Var(\mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i) = \lambda_i \qquad i = 1, ..., p \\ Cov(\mathbf{e}_i'X_i, \mathbf{e}_k'X_k) &= \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_k = 0 \qquad i \neq k \end{aligned}$$

5 If $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$ then

$$tr(\mathbf{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Y_i)$$

since

$$tr(\mathbf{\Sigma}) = tr(\mathbf{P\Lambda P'} = tr(\mathbf{\Lambda P'P}) = tr(\mathbf{\Lambda}) = \lambda_1 + ... + \lambda_p$$

6 Total population variance is $\sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i$ and

$$p_k = \frac{\lambda_i}{\lambda_1 + ... + \lambda_p} \qquad k = 1, .., p$$

is the proportion of the total variance explained by the k^{th} principal component. The magnitude of e_{ki} the i^{th} component of the k^{th} eigenvector $\mathbf{e}_k$ measures the importance of the i^{th} variable X_i to the k^{th} principal component Y_k.

7 The correlation between Y_k and X_i is given by

$$\rho_{Y_k,X_i} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{ii}}} \qquad i,k = 1,..,p$$

8 The principal components lie in the directions of the axes of a constat density ellipsoid when $\mu = 0$

$$(\mathbf{x} - \mu)'(\mathbf{\Sigma})^{-1}(\mathbf{x} - \mu) = c^2$$

When $\mu \neq 0$ the centered principal components $Y_i = \mathbf{e}_i'(\mathbf{x} - \mu)$

9 The standardized variables $Z_i = (X_i - \mu_i)/\sqrt{\sigma_{ii}}$ have covariance matrix ρ, the correlation matrix of the original variables. The principal components can also be obtained for the correlation matrix. They are normally not the same as the principal components derived from the covariance matrix.

Summarizing Sample Variation

Typically, $\mathbf{\Sigma}$ and μ are unknown and the population principal components cannot be obtained. The sample equivalent of the covariance matrix is $\mathbf{S}$.

1 The i^{th} sample principal component is given by

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x} = \hat{e}_{i1}x_1 + .. + \hat{e}_{ip}x_p$$

and has variance $\hat{\lambda}_i$ where $\hat{\lambda}_i$ is the i^{th} eigenvalue of $\mathbf{S}$ and $\hat{\mathbf{e}}_\mathbf{i}$ is the associated eigenvector.

2 The sample covariance $Cov(\hat{y}_i, \hat{y}_k) = 0$.

3 The total sample variance is $\sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + .. + \hat{\lambda}_p$

4 The correlation between x_i and $\hat{y}_k$ is given by

$$r_{\hat{y}_k, x_i} = \frac{\hat{e}_{ki}\sqrt{\hat{\lambda}_k}}{\sqrt{s_{ii}}}$$

5 Principal components can also be obtained form the sample correlation matrix. The principal components obtained from the sample correlation are not the same s those obtained from the covariance matrix.

6 If the data are normally distributed, the sample principal components can also be obtained from the maximum likelihood estimate of the sample covariance matrix. In this case, the sample principal components are the maximum likelihood estimates of the population principal components.

7 The principal components can also be obtained from the centered data. It gives the i^{th} component $\hat{y}_i = \mathbf{\hat{e}_i'}(\mathbf{x} - \mathbf{\bar{x}})$. Note, centering does not affect the sample covariance matrix. $\hat{y}_{ji} = \mathbf{\hat{e}_i'}(\mathbf{x_j} - \mathbf{\bar{x}})$ is the i^{th} principal component for the j^{th} sample vector.

8 The sample mean of the centered sample principal components is zero.

$$\bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \mathbf{\hat{e}_i'}(\mathbf{x_j} - \mathbf{\bar{x}}) = \frac{1}{n}\ \mathbf{\hat{e}_i'}\sum_{j=1}^{n}(\mathbf{x_j} - \mathbf{\bar{x}}) = 0$$

The number of principal components to select

1 There is NO definitive answer to the questions of how many principal components to select. There is NO statistical test!

2 Things to consider are
   a the relative sizes of the eigenvalues
   b the interpretation of the components
   c the total amount of variation explained

3 A visual aid to select the number of principal components is a scree plot.
   a Order the eigenvalues form largest to smallest $\hat{\lambda}_1 > \hat{\lambda}_2 > ... > \hat{\lambda}_p$
   b Plot the eigenvalues against the index ($\hat{\lambda}_1$ vs 1 etc).
   c Connect the points and look for a bend in the curve at which the remaining eigenvalues are relatively small.

Graphing Principal Components

1 Note, that each observation can be expressed a linear combination of the principle components and eigenvectors:

$$\mathbf{x}_j = \hat{y}_{j1}\hat{\mathbf{e}}_1 + \hat{y}_{j2}\hat{\mathbf{e}}_2 + ... + \hat{y}_{jp}\hat{\mathbf{e}}_p$$

The magnitude of the last principal components determine how well the first few approximate (fit) the data. Recall: one goal of principal components is data reduction; the expression of a p-variate vector with a q-variate vector where $q < p$

2 Q-Q plots of the principal components for the data; remember, we have a value for the principal component for each data point; if our data our normally distributed, the principal components should also be normally distributed; they are linear combinations of the data.

3 Scatterplots of $\hat{y}_i$ vs $\hat{y}_j$ to identify potential outliers

Large Sample Inference

The large sample properties of the sample principal components

1 Assumption:

$$\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n \sim N_p(\mu, \mathbf{\Sigma})$$

2 $\hat{\lambda}$ is the vector of eigenvalues of $\mathbf{S}$; the eigenvectors are given by $\hat{\mathbf{e}}_i$, for $i = 1, .., n$

3 $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{\Sigma}$, then

$$\sqrt{n}(\hat{\lambda} - \lambda) \approx N_p(0, 2\mathbf{\Lambda}^2)$$

4 If

$$\mathbf{E}_i = \lambda_i \sum_{k=1, k \neq i}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{e}_k \mathbf{e}_k' \qquad then \qquad \sqrt{n}(\hat{\mathbf{e}}_i - \mathbf{e}_i) \approx N_p(0, \mathbf{E}_i)$$

5 Each $\hat{\lambda}_i$ is independent of $\hat{\mathbf{e}}_i$, ie of the components of the associated sample eigenvector.