Chapter 3: Principal Components Analysis

- Goal of *principal components analysis* (PCA): Describe (most of) the variation in the multivariate data set $x_1, x_2, \ldots, x_q$ using a smaller (more concise) set of variables.
- PCA can be thought of as a form of *data reduction*.
- *Data Reduction*: Using a lower-dimensional summarization of the data that still contains the relevant information that is in the full data.
- The hope is to be able to summarize what makes observations similar and what makes them different using relatively few (easy to interpret?) indices.
- Also, we'd like each of these indices to say something *distinct* about how the observations vary.

Specific Ideas Behind PCA

- In PCA, we create a new set of "variables" $y_1, y_2, \ldots, y_q$, each of which is a linear combination of $x_1, x_2, \ldots, x_q$:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q \\ &\vdots \\ y_q &= a_{q1}x_1 + a_{q2}x_2 + \cdots + a_{qq}x_q \end{aligned}$$

- We further require these new "variables" (or indices) to be uncorrelated.
- This assures us that the information in y_2, say, doesn't overlap with the information in y_1.

Specific Ideas Behind PCA (continued)

- Having q of these indices doesn't give us any data reduction.
- We'd like to choose only the first m of these (where $m < q$) to focus on.
- Thus we choose the first linear combination, y_1, so that it is the linear combination of $x_1, x_2, \ldots, x_q$ that accounts for as much of the variation in the original data as possible.
- Then the second linear combination, y_2, is the linear combination that accounts for as much of the variation in the original data as possible (*provided that it is uncorrelated* with y_1 !).
- And the third linear combination, y_3, is the linear combination that accounts for as much of the variation in the original data as possible (*provided that it is uncorrelated* with y_1 and y_2), and so on.
- Eventually, it becomes a practical issue of how many of these indices (which are called *principal components*) we really need (more later on this).

Applications of PCA

- Taxologists attempt to characterize bird species on the basis of multiple measurements.
- One or two indices (based on these measurements) may be sufficient to separate the birds into useful groups.
- In economics, we may characterize states or countries based on lots of variables that are measured.
- In reality, a few well-chosen indices (based on these variables) may account for the economic differences among countries.
- In multiple regression, sometimes our explanatory (predictor) variables are highly correlated (or simply too numerous).
- It may be advantageous to consider a few *uncorrelated* indices as predictors rather than the correlated explanatory variables.

Mathematics Behind PCA

- *First step*: Choose coefficients $a_{11}, a_{12}, \ldots, a_{1q}$ such that the sample variance of $y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q$ is as large as possible.
- Of course, we could arbitrarily inflate this sample variance by making $a_{11}, a_{12}, \ldots, a_{1q}$ arbitrarily large.
- To make this maximization problem meaningful, we include the constraint that

$$\mathbf{a}_1^{'}\mathbf{a}_1 = \sum_{i=1}^{q} a_{1i}^2 = 1.$$

- *Second step*: Choose coefficients $a_{21}, a_{22}, \ldots, a_{2q}$ such that the sample variance of $y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q$ is as large as possible, subject to $\mathbf{a}_2^{'}\mathbf{a}_2 = 1$.
- But we also want y_2 to be uncorrelated with y_1, so we add the further constraint that

$$\mathbf{a}_2^{'}\mathbf{S}\mathbf{a}_1 = \sum_{j=1}^{q}\sum_{i=1}^{q} s_{ij}a_{2i}a_{1j} = 0.$$

Mathematics Behind PCA (continued)

- *j-th step*: Choose coefficients $a_{j1}, a_{j2}, \ldots, a_{jq}$ such that the sample variance of $y_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jq}x_q$ is as large as possible, subject to $\mathbf{a}_j^{'}\mathbf{a}_j = 1$ and $\mathbf{a}_j^{'}\mathbf{a}_{j'} = 0$ for all $j' < j$.
- Mathematically, we can obtain q linear combinations $y_1, y_2, \ldots, y_q$ in this way.
- This optimization problem is solved using the method of *Lagrange multipliers*.
- In practice, we allow a software package like R or SAS to find the coefficients for us.

Facts about the Principal Component Solutions

- It turns out that $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1q})'$ is the eigenvector of the sample covariance matrix $\mathbf{S}$ corresponding to its largest eigenvalue.
- Furthermore, for $j = 1, 2, \ldots, q$, $\mathbf{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jq})'$ is the eigenvector of the sample covariance matrix $\mathbf{S}$ corresponding to its j-th-largest eigenvalue.
- The variance of the i-th principal component is λ_i, where $\lambda_1, \ldots, \lambda_q$ are the eigenvalues of $\mathbf{S}$.
- The total variance of all q principal components equals the total variance of the original variables:

$$\sum_{i=1}^{q} \lambda_i = s_1^2 + \cdots + s_q^2 = trace(\mathbf{S}).$$

- Hence the first m principal components account for the following proportion of the variation in the original data:

$$\frac{\sum_{i=1}^{m} \lambda_i}{trace(\mathbf{S})}.$$

PCA based on Correlation matrix $\mathbf{R}$

- Often the q variables in the data set are very different in their scales, variability, etc.
- Basing the PCA on the covariance matrix $\mathbf{S}$ would lead to variables with large variances dominating the most important principal components.
- Also, changing the units of measurement (e.g., from ounces to pounds, or from feet to inches) would change the PCA solution.
- For this reason, it is often preferred to base the PCA solution on the eigenvectors and eigenvalues of $\mathbf{R}$ rather than $\mathbf{S}$.
- Note: This is equivalent to initially standardizing all variables and then performing the PCA based on $\mathbf{S}$.

Choosing the Number of Components

- To explain all the variation in the original data, we would need (in general) all q principal components.
- But it is practically sufficient to explain *most* of the variation in the original data.
- This can usually be done using merely the "first few" principal components.
- If we will retain the first $m < q$ components, how can we choose m?

Possible Rules for Choosing the Number of Components

1. Retain the first m components sufficient to explain a specified percentage (70%? 80%? 90%?) of the total variance of the original variables.

2. Keep only the components whose eigenvalues are at least $\sum_{i=1}^{q} \lambda_i / q$, which is the average eigenvalue and also the average sample variance of the original variables.

3. When PCA is done on the correlation matrix, this average is 1, so Kaiser (1958) suggested keeping components with eigenvalues at least 1. Jolliffe (1972) preferred using 0.7 as the threshold.

4. Cattell (1965) introduced the *scree diagram*, which plots λ_i against i, for $i = 1, \ldots, q$.

5. Look for the "elbow" in the curve and choose the corresponding number of components.

6. Jolliffe (1986) suggested a modified scree plot that plotted $\log(\lambda_i)$ against i, for $i = 1, \ldots, q$.

Principal Component Scores

- Once we obtain the coefficients (the a_{ji} values) for the principal components, we can calculate the *principal component scores* for each observation.
- To get the first PC score for observation 1, plug the observed variable values for that observation into the first PC linear combination: $y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q$.
- We can do this for each of the n observations.
- Then we can plug the observed x-values into the second component to get the second PC score for each observation, and so on.

Calculating and Plotting Principal Component Scores

- If the components are derived from $\mathbf{S}$, the first m PC scores for individual i are:

$$\begin{aligned} y_{i1} &= \mathbf{a}_1^{'}\mathbf{x}_i \\ y_{i2} &= \mathbf{a}_2^{'}\mathbf{x}_i \\ &\vdots \\ y_{im} &= \mathbf{a}_m^{'}\mathbf{x}_i \end{aligned}$$

  where $\mathbf{x}_i$ contains the observed data values for observation i.

- If the components are derived from $\mathbf{R}$, then $\mathbf{x}_i$ would contain the observed *standardized* data values for observation i.

- The PC scores can easily be scaled to have mean 0 and variance 1, if desired.

- Often we display the first two PC scores for each observation on a scatterplot.

- This allows us to visually see how the data vary according to the two most important modes of variation.

Other Issues: Correlations Between Variables and Components

- It may be of interest to examine the correlation between original variable x_i and principal component y_j, for any i, j.
- This is simply

$$\frac{a_{ji}\sqrt{\lambda_j}}{s_i}$$

- If the components are derived from $\mathbf{R}$, then all $s_i = 1$ and this correlation is

$$a_{ji}\sqrt{\lambda_j}$$

Other Issues: Rescaling and Rotating Principal Components

- The vectors of coefficients for each principal component, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q$, can be presented in rescaled form as

$$\mathbf{a}_j^* = \sqrt{\lambda_j}\mathbf{a}_j, j = 1, \ldots, q.$$

- These rescaled coefficients directly give the correlation between original variables and principal components.
- It is possible to rotate the principal components so that many of the coefficients are closer to 0 and 1.
- This can aid the interpretability of the indices, but in this case each successive component no longer represents the index with maximum variance.
- Rotation will be studied further in the next chapter on factor analysis.

Inference with Principal Components

- Our principal components solutions are based on the eigenvalues and eigenvectors of the sample covariance matrix $\mathbf{S}$ (or correlation matrix $\mathbf{R}$).
- The eigenvalues and eigenvectors of the true covariance matrix $\boldsymbol{\Sigma}$ are unknown.
- We can (for large samples) perform inference about these unknown eigenvalues and eigenvectors.
- These results assume the data vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$ are a random sample from a multivariate normal distribution.
- This assumption could be checked with plots such as the *chi-squared plot*.
- If this multivariate normal assumption is close to correct, the inferences will be approximately valid.

A Confidence Interval

- Let the variance of the i-th population principal component (which is the i-th eigenvalue of the true covariance matrix $\mathbf{\Sigma}$) be denoted by ξ_i.
- Under multivariate normality, the large-sample distribution of the variance of the i-th (sample) principal component λ_i is $N(\xi_i, 2\xi_i^2/n)$.
- Furthermore, the λ_i's are independent.
- Thus a large-sample $100(1-\alpha)$% confidence interval for ξ_i is

$$\left(\frac{\lambda_i}{1+z_{\alpha/2}\sqrt{2/n}}, \frac{\lambda_i}{1-z_{\alpha/2}\sqrt{2/n}}\right).$$

- If we want simultaneous intervals for m such eigenvalues, replace $z_{\alpha/2}$ with $z_{\alpha/2m}$ in the formula to get Bonferroni joint CIs.
- See the course web page for `R` code to perform this CI.

A Hypothesis Test for Equal Correlation Structure

- We may wish to test whether all the correlations between our q variables are the same.
- In such a case, the eigenvalues of $\boldsymbol{\Sigma}$ are all equal and the large-sample results stated previously do not hold.
- We test the null hypothesis that the true correlation matrix is equal to:

$$\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

  against the alternative that the correlations are not all the same.
- The test statistic given by Lawley (1963) involves the elements of $\mathbf{R}$; Johnson and Wichern (2002) provide details.