1 Principal Components Analysis

- The purpose of PCA is to find the new components with maximum variance.
- The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions).
- These problems do not occur in low-dimensional settings.
- The expression of the curse of dimensionality was introduced by Richard E. Bellman when considering problems in dynamic optimization in 1961. (Bellman, R. (1961), Adaptive Control Processes, Princeton, NJ: Princeton University Press.)
- When the dimensionality increases, the volume of the space increases so fast that all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient.
- This brings us to principal components analysis (PCA), which can reduce the dimensionality of a multivariate data set while accounting for as much of the original variation as possible present in the data set.
- This aim is achieved by transforming to a new set of variables, the principal components (PCs), that are linear combinations of the original variables, which are uncorrelated and are ordered so that the first few of them account for most of the variation in all the original variables.
- A set of data consisting of examination scores for several different subjects for each of a number of students. We can use the mean score for each student to consider the performance of students.
- $Var(Y_1) = \lambda_1, Var(Y_2) = \lambda_2, \ldots, Var(Y_p) = \lambda_p$.
- The total variance of PCs equals the total variance of the original variables.

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} \lambda_i = tr(\Sigma) = \sum_{i=1}^{p} \sigma_{i,i} = \sum_{i=1}^{p} Var(X_i).$$

- The jth PC accounts for a proportion P_j of the total variation of the original data, where

$$P_j = \frac{\lambda_j}{tr(\Sigma)}.$$

- The first m principal components, where $m < p$ account for a proportion $P^{(m)}$ of the total variation of the original data, where

$$P^{(m)} = \sum_{j=1}^{m} P_j = \frac{\sum_{j=1}^{m} \lambda_j}{tr(\Sigma)}.$$

- In geometrical terms, the first principal component defines the line of best fit to the p-dimensional observations in the sample.
- The first m components give the best fit in m dimensions.

1.1 Some Optimal Algebraic Properties of PCs

- The original components $X^T = (X_1, \ldots, X_p)$.
- The principal components $Y^T = (Y_1, \ldots, Y_p)$.
- $Y_j = \alpha_j^T X$.
- $Y = A^T X$, where $A = (\alpha_1, \ldots, \alpha_p)$ is a orthogonal matrix whose kth column, α_k is the eigenvector of Σ_X associated with λ_k.
- $\Sigma_X A = A\Lambda$, where Λ is the diagonal matrix whose kth diagonal element is λ_k.
- $A^T \Sigma_X A = \Lambda$ and $\Sigma_X = A\Lambda A^T$.

Property 1. *The Spectral Decomposition of Σ_X:*

$$\Sigma_X = \lambda_1 \alpha_1 \alpha_1^T + \lambda_2 \alpha_2 \alpha_2^T + \ldots + \lambda_p \alpha_p \alpha_p^T = \sum_{i=1}^{p} \lambda_i \alpha_i \alpha_i^T.$$

Proof. $\Sigma_X = A\Lambda A^T$, expanding the right-hand side matrix product. □

Property 2. *For any integer q, $1 \leq q \leq p$, consider the orthonormal linear transformation $Z = B^T X$, where Z is a $q-$element vector and $B^T \in \mathbb{R}^{q \times p}$, and let $\Sigma_Z = B^T \Sigma_X B$ be the covariance matrix for Z. Then the trace of Σ_Z is maximized by taking $B = A_q$, where A_q consists of the first q columns of A. Also, the determinant of Σ_Z is maximized by taking $B = A_q$.*

Property 3. *$Z = B^T X$, Then the trace of Σ_Z is minimized by taking $B = A_q^*$, where A_q^* consists of the last q columns of A.*

1.2 Principal Components Using a Correlation Matrix

- The derivation and properties of PCs considered in the lecture above are based on the eigenvectors and eigenvalues of the covariance matrix.
- In practice, it is more common to define PCs using the correlation matrix.
- PCA is not scale-invariant. For example, suppose the three variables in a multivariate data set are weight in pounds, height in feet, and age in years, but for some reason we would like our principal components expressed in ounces, inches, and decades. Intuitively two approaches seem feasible.
  - Multiply the variables by 16, 12, and 1/10, respectively and then carry out PCA on the covariance matrix of the three variables.

– Carry out PCA on the covariance matrix of the original variables and then multiply the elements of the relevant component by 16, 12, and 1/10.
– Unfortunately, these two procedures do not generally lead to the same result.

Suppose that we have just two variables, X_1, X_2, and that X_1 is a length variable which can equally well be measured in centimetres or millimetres. The variable X_2 is not a length measurement. The covariance matrices in the two cases are, respectively,

$$\Sigma_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}. \quad Cov(aX, bY) = abCov(X, Y).$$

- The first PC for Σ_1 is $0.707X_1 + 0.707X_2$ which has equal weight to X_1 and X_2.
- The first PC for Σ_2 is $0.998X_1 + 0.055X_2$ which is almost entirely dominated by X_1.
- So a relatively minor change in one variable has the effect of changing a PC in this way.
- The first PC accounts for 77.5 percent of the total variation for Σ_1, but 99.3 percent for Σ_2 (that is, the proportion P_1).

This example illustrates that when variables are on very different scales or have very different variances, a principal components analysis of the data should not be performed on the covariance matrix, but on the correlation matrix.

- One way is calculating PCs using the correlation matrix, not the covariance matrix.
- An equivalent way is to find the PCs of a standardized version X^* of X.
- X^* has jth element $X_j/\sqrt{\sigma_{jj}}$, where $\sigma_{jj} = Var(X_j)$.
- The covariance matrix for X^* is the correlation matrix of X.

1.3 Choosing the Number of Components

How many components are needed to provide an adequate summary of a given data set?

- Retain just enough components to explain some specified large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested, although smaller values might be appropriate as q or n, the sample size, increases.
- Exclude those PCs whose eigenvalues are less than the average, $\sum_{i=1}^{p} \frac{\lambda_i}{p}$. Since $\sum_{i=1}^{p} \lambda_i = tr(\Sigma_X)$, the average eigenvalue is also the average variance of the original variables.

1.4 Examples in Chapter 3.4 and Chapter 3.10

```
> library("tools")
> library("HSAUR2")
> library("MVA")
> demo("Ch-PCA")
```