Machine Learning

Module 6

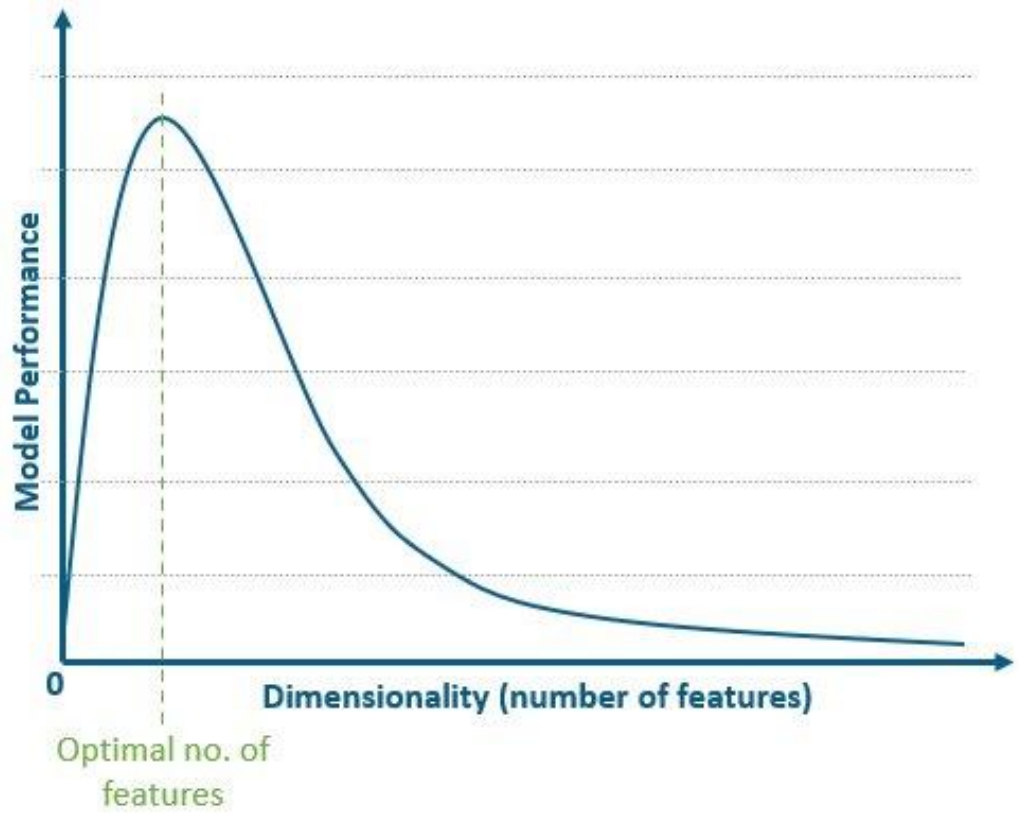
Dimensionality reduction

Curse of Dimensionality

- In many learning problems, the datasets have large number of variables.
- Sometimes, the number of variables is more than the number of observations.
- For example, in image processing, mass spectrometry, time series analysis, internet search engines, and automatic text analysis among others.
- Statistical and machine learning methods have some difficulty when dealing with such high-dimensional data.
- Normally the number of input variables is reduced before the machine learning algorithms can be successfully applied.

Introduction

- In statistical and machine learning, dimensionality reduction or dimension reduction is the process of reducing the number of variables under consideration by obtaining a smaller set of principal variables.



Introduction

Dimensionality reduction may be implemented in two ways.

- Feature selection
  - In feature selection, we are interested in finding k of the total of n features that give us the most information and we discard the other (n−k) dimensions.
  - We are going to discuss subset selection as a feature selection method.
- Feature extraction
  - In feature extraction, we are interested in finding a new set of k features that are the combination of the original n features.
  - The best known and most widely used feature extraction methods are Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA).

Measures of error

- In both methods we require a measure of the error in the model.
- **In regression problems**, we may use the Mean Squared Error (MSE) or the Root Mean Squared Error (RMSE) as the measure of error.
- MSE is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of
- If y1; .....; yn are the o $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ nd yi ^; ..... ; yn ^ are the predicted values, tnen

- **In classification problems**, we may use the misclassification rate as a measure of the error.
- This is defined as follows:

Misclassification rate = (no. of misclassified examples)/total no. of examples

Why dimensionality reduction is useful?

- In most learning algorithms, the complexity depends on the number of input dimensions, d, as well as on the size of the data sample, N, and for reduced memory and computation, we are interested in reducing the dimensionality of the problem.
- Decreasing d also decreases the complexity of the inference algorithm during testing.
- When an input is decided to be unnecessary, we save the cost of extracting it. Simpler models are more robust on small datasets. Simpler models have less variance, that is, they vary less depending on the particulars of a sample, including noise, outliers, and so forth.

Why dimensionality reduction is useful?

- When data can be explained with fewer features, we get a better idea about the process that underlies the data, which allows knowledge extraction.
- When data can be represented in a few dimensions without loss of information, it can be plotted and analyzed visually for structure and outliers.

Subset selection

- In machine learning subset selection, sometimes also called feature selection, or variable selection, or attribute selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.
- Feature selection techniques are used for four reasons:
  - simplification of models to make them easier to interpret by researchers/users
  - shorter training times,
  - to avoid the curse of dimensionality
  - enhanced generalization by reducing overfitting
- The central premise when using a feature selection technique is that the data contains many features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information.

Subset selection

- **Forward selection**
- In forward selection, we start with no variables and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not decrease the error (or decreases it only sightly).

Subset selection

- Procedure:

We use the following notations:

n	:	number of input variables
$x_1, \ldots, x_n$	:	input variables
F_i	:	a subset of the set of input variables
$E(F_i)$	:	error incurred on the validation sample when only the inputs in F_i are used

1. Set $F_0 = \varnothing$ and $E(F_0) = \infty$.
2. For $i = 0, 1, \ldots$, repeat the following until $E(F_{i+1}) \geq E(F_i)$:
   (a) For all possible input variables x_j, train the model with the input variables $F_i \cup \{x_j\}$ and calculate $E(F_i \cup \{x_j\})$ on the validation set.
   (b) Choose that input variable x_m that causes the least error $E(F_i \cup \{x_j\})$:

$$m = \arg\min_j E(F_i \cup \{x_j\})$$

   (c) Set $F_{i+1} = F_i \cup \{x_m\}$.
3. The set F_i is outputted as the best subset.

Subset selection

- **Backward selection**

  In sequential backward selection, we start with the set containing all features and at each step remove the one feature that causes the least error.

Procedure: We use the following notations:

n	:	number of input variables
$x_1, \ldots, x_n$	:	input variables
F_i	:	a subset of the set of input variables
$E(F_i)$	:	error incurred on the validation sample when only the inputs in F_i are used

1. Set $F_0 = \{x_1, \ldots, x_n\}$ and $E(F_0) = \infty$.
2. For $i = 0, 1, \ldots$, repeat the following until $E(F_{i+1}) \geq E(F_i)$:
   (a) For all possible input variables x_j, train the model with the input variables $F_i - \{x_j\}$) and calculate $E(F_i - \{x_j\})$ on the validation set.
   (b) Choose that input variable x_m that causes the least error $E(F_i - \{x_j\})$:
   (c) Set $F_{i+1} = F_i - \{x_m\}$.
3. The set F_i is outputted as the best subset.

Principal Component Analysis

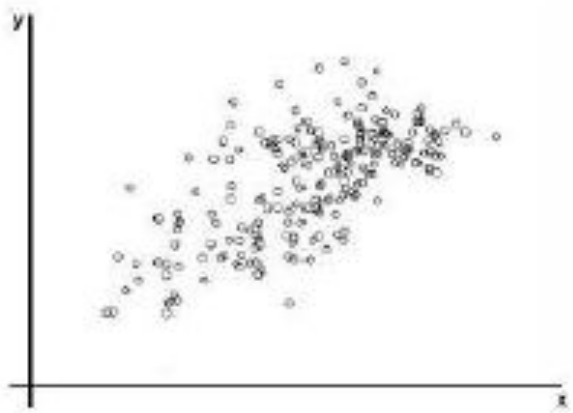
- Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations.
- This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

Principal Component Analysis

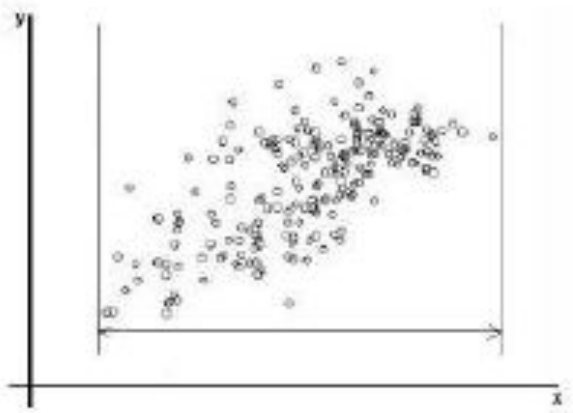
- Consider a two-dimensional data, that is, a dataset consisting of examples having two features.
- Let each of the features be numeric data. So, each example can be plotted on a coordinate plane (x-coordinate indicating the first feature and y-coordinate indicating the second feature).
- Plotting the example, we get a scatter diagram of the data.
- Now let us examine some typical scatter diagram and make some observations regarding the directions in which the points in the scatter diagram are spread out.

- Figure a shows a scatter diagram of a two-dimensional data.
- Figure b shows spread of the data in the x direction and Figure c shows the spread of the data in the y-direction. We note that the spread in the x-direction is more than the spread in the y direction.
- Figures d and e, we note that the maximum spread occurs in the direction shown in Figure e. Figure e also shows the point whose coordinates are the mean values of the two features in the dataset. This direction is called the direction of the first principal component of the given dataset.
- The direction which is perpendicular (orthogonal) to the direction of the first principal component is called the direction of the second principal component of the dataset. This direction is shown in Figure f. (This is only with reference to a two-dimensional dataset.)
- The unit vectors along the directions of principal components are called the principal component vectors, or simply, principal components. These are shown in Figure g.

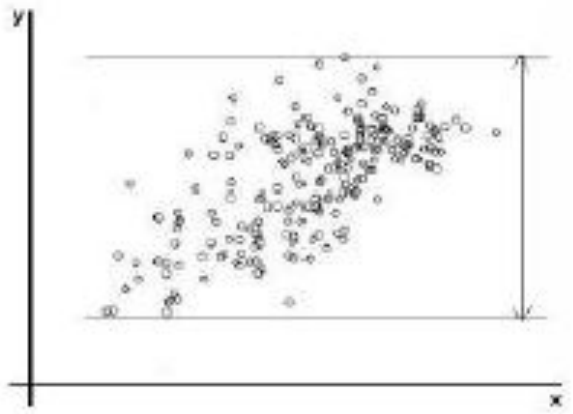
Principal Component Analysis



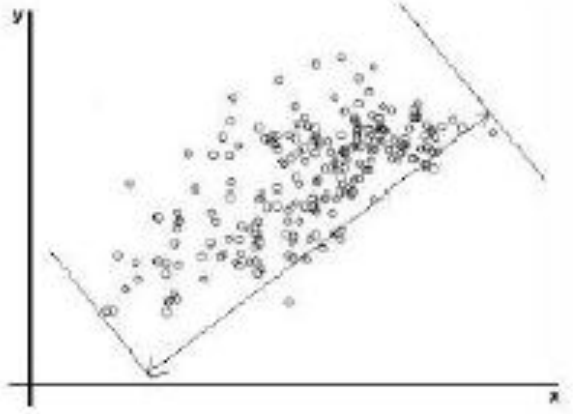
(a) Scatter diagram



(b) Spread along x-direction

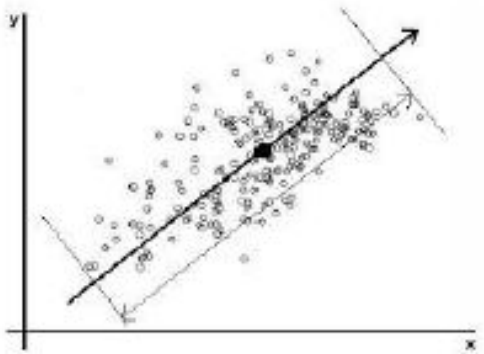


(c) Spread along y-direction

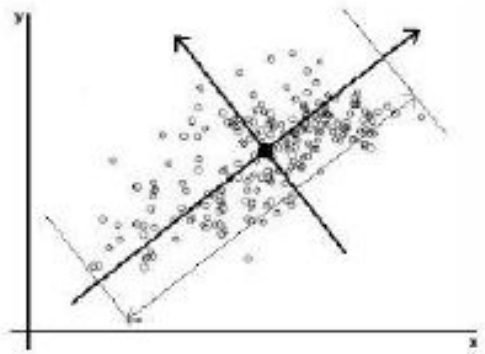


(d) Largest spread

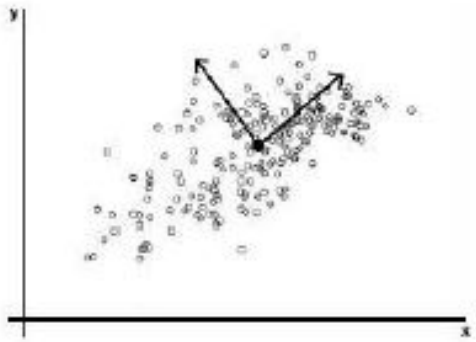
Principal Component Analysis



(e) Direction of largest spread : Direction of the first principal component (solid dot is the point whose coordinates are the means of x and y)



(f) Directions of principal components



(g) Principal component vectors (unit vectors in the directions of principal components)

Computation of the principal component vectors (PCA algorithm)

Step 1. Data

We consider a dataset having n features or variables denoted by $X_1, X_2, \ldots, X_n$. Let there be N examples. Let the values of the i-th feature X_i be $X_{i1}, X_{i2}, \ldots, X_{iN}$ (see Table 4.1).

Features	Example 1	Example 2	⋯	Example N
X_1	X_{11}	X_{12}	⋯	X_{1N}
X_2	X_{21}	X_{22}	⋯	X_{2N}
⋮				
X_i	X_{i1}	X_{i2}	⋯	X_{iN}
⋮				
X_n	X_{n1}	X_{n2}	⋯	X_{nN}

Table 4.1: Data for PCA algorithm

Computation of the principal component vectors (PCA algorithm)

Step 2. Compute the means of the variables

We compute the mean $\bar{X}_i$ of the variable X_i:

$$\bar{X}_i = \frac{1}{N}(X_{i1} + X_{i2} + \cdots + X_{iN}).$$

Step 3. Calculate the covariance matrix

Consider the variables X_i and X_j (i and j need not be different). The covariance of the ordered pair (X_i, X_j) is defined as[1]

$$\text{Cov}(X_i, X_j) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j). \tag{4.1}$$

We calculate the following $n \times n$ matrix S called the covariance matrix of the data. The element in the i-th row j-th column is the covariance $\text{Cov}(X_i, X_j)$:

$$S = \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & & & \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Cov}(X_n, X_n) \end{bmatrix}$$

Computation of the principal component vectors (PCA algorithm)

Step 4. Calculate the eigenvalues and eigenvectors of the covariance matrix

Let S be the covariance matrix and let I be the identity matrix having the same dimension as the dimension of S.

i) Set up the equation:

$$\det(S - \lambda I) = 0. \tag{4.2}$$

This is a polynomial equation of degree n in λ. It has n real roots (some of the roots may be repeated) and these roots are the eigenvalues of S. We find the n roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of Eq. (4.2).

ii) If $\lambda = \lambda'$ is an eigenvalue, then the corresponding eigenvector is a vector

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

such that

$$(S - \lambda' I)U = 0.$$

(This is a system of n homogeneous linear equations in $u_1, u_2, \ldots, u_n$ and it always has a nontrivial solution.) We next find a set of n orthogonal eigenvectors $U_1, U_2, \ldots, U_n$ such that U_i is an eigenvector corresponding to λ_i.[2]

Computation of the principal component vectors (PCA algorithm)

iii) We now normalise the eigenvectors. Given any vector X we normalise it by dividing X by its length. The length (or, the norm) of the vector

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

is defined as

$$\|X\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

Given any eigenvector U, the corresponding normalised eigenvector is computed as

$$\frac{1}{\|U\|} U.$$

We compute the n normalised eigenvectors $e_1, e_2, \ldots, e_n$ by

$$e_i = \frac{1}{\|U_i\|} U_i, \quad i = 1, 2, \ldots, n.$$

Computation of the principal component vectors (PCA algorithm)

Step 5. Derive new data set

Order the eigenvalues from highest to lowest. The unit eigenvector corresponding to the largest eigenvalue is the first principal component. The unit eigenvector corresponding to the next highest eigenvalue is the second principal component, and so on.

i) Let the eigenvalues in descending order be $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and let the corresponding unit eigenvectors be $e_1, e_2, \ldots, e_n$.

ii) Choose a positive integer p such that $1 \leq p \leq n$.

iii) Choose the eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and form the following $p \times n$ matrix (we write the eigenvectors as row vectors):

$$F = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_p^T \end{bmatrix},$$

where T in the superscript denotes the transpose.

Computation of the principal component vectors (PCA algorithm)

iv) We form the following $n \times N$ matrix:

$$X = \begin{bmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_1 & \cdots & X_{1N} - \bar{X}_1 \\ X_{21} - \bar{X}_2 & X_{22} - \bar{X}_2 & \cdots & X_{2N} - \bar{X}_2 \\ \vdots & & & \\ X_{n1} - \bar{X}_n & X_{n2} - \bar{X}_n & \cdots & X_{nN} - \bar{X}_n \end{bmatrix}$$

v) Next compute the matrix:

$$X_{\text{new}} = FX.$$

Note that this is a $p \times N$ matrix. This gives us a dataset of N samples having p features.

Computation of the principal component vectors (PCA algorithm)

Step 6. New dataset

The matrix X_{new} is the new dataset. Each row of this matrix represents the values of a feature. Since there are only p rows, the new dataset has only features.

Step 7. Conclusion

This is how the principal component analysis helps us in dimensional reduction of the dataset. Note that it is not possible to get back the original n-dimensional dataset from the new dataset.

Example:

Given the data in Table 4.2, use PCA to reduce the dimension from

Feature	Example 1	Example 2	Example 3	Example 4
X_1	4	8	13	7
X_2	11	4	5	14

Table 4.2: Data for illustrating PCA

Solution

1. Scatter plot of data

We have

$$\bar{X}_1 = \frac{1}{4}(4+8+13+7) = 8,$$
$$\bar{X}_2 = \frac{1}{4}(11+4+5+14) = 8.5.$$

Figure 4.2 shows the scatter plot of the data together with the point $(\bar{X}_1, \bar{X}_2)$.

Example:

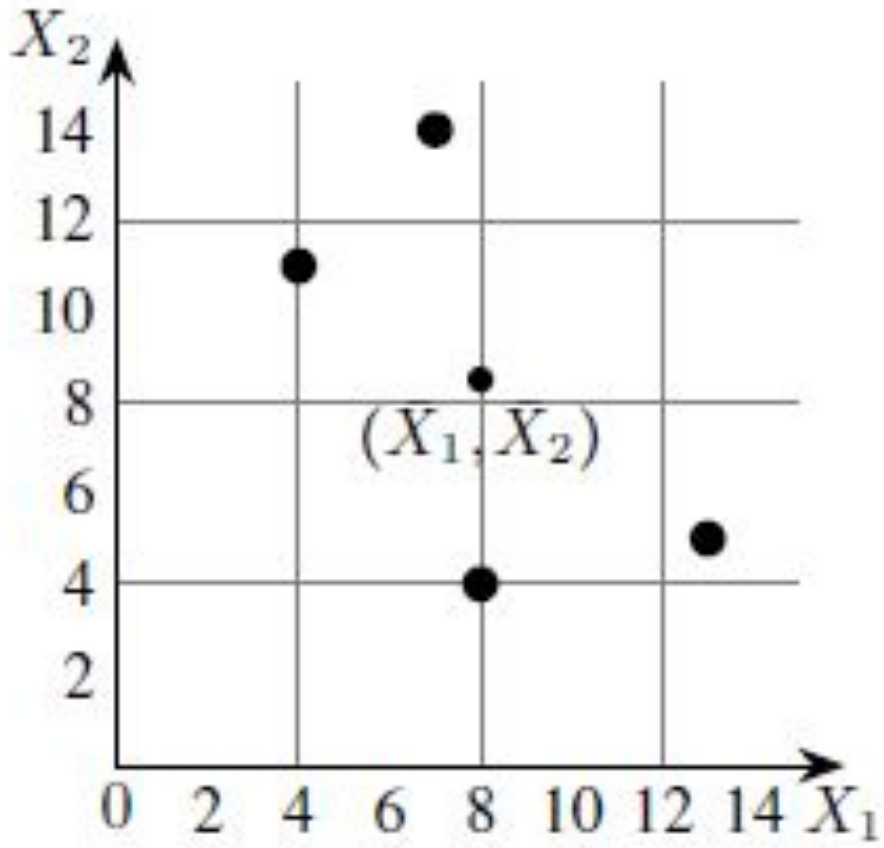


Figure 4.2: Scatter plot of data in Table 4.2

2. Calculation of the covariance matrix:

$$\text{Cov}(X_1, X_2) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{1k} - \bar{X}_1)^2$$
$$= \tfrac{1}{3}\left((4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2\right)$$
$$= 14$$

$$\text{Cov}(X_1, X_2) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{1k} - \bar{X}_1)(X_{2k} - \bar{X}_2)$$
$$= \tfrac{1}{3}((4-8)(11-8.5) + (8-8)(4-8.5)$$
$$+ (13-8)(5-8.5) + (7-8)(14-8.5)$$
$$= -11$$

$$\text{Cov}(X_2, X_1) = \text{Cov}(X_1, X_2)$$
$$= -11$$

$$\text{Cov}(X_2, X_2) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{2k} - \bar{X}_2)^2$$
$$= \tfrac{1}{3}\left((11-8.5)^2 + (4-8.5)^2 + (5-8.5)^2 + (14-8.5)^2\right)$$
$$= 23$$

The covariance matrix is

$$S = \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) \end{bmatrix}$$
$$= \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

3. Eigenvalues of the covariance matrix

The characteristic equation of the covariance matrix is

$$
\begin{aligned}
0 &= \det(S - \lambda I) \\
&= \begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix} \\
&= (14-\lambda)(23-\lambda) - (-11) \times (-11) \\
&= \lambda^2 - 37\lambda + 201
\end{aligned}
$$

Solving the characteristic equation we get

$$
\begin{aligned}
\lambda &= \tfrac{1}{2}(37 \pm \sqrt{565}) \\
&= 30.3849,\ 6.6151 \\
&= \lambda_1,\ \lambda_2 \quad \text{(say)}
\end{aligned}
$$

4. Computation of the eigenvectors

To find the first principal components, we need only compute the eigenvector corresponding to the largest eigenvalue. In the present example, the largest eigenvalue is λ_1 and so we compute the eigenvector corresponding to λ_1.

The eigenvector corresponding to $\lambda = \lambda_1$ is a vector $U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ satisfying the following equation:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = (S - \lambda_1 I)X$$

$$= \begin{bmatrix} 14 - \lambda_1 & -11 \\ -11 & 23 - \lambda_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$= \begin{bmatrix} (14 - \lambda_1)u_1 - 11u_2 \\ -11u_1 + (23 - \lambda_1)u_2 \end{bmatrix}$$

This is equivalent to the following two equations:

$$(14 - \lambda_1)u_1 - 11u_2 = 0$$
$$-11u_1 + (23 - \lambda_1)u_2 = 0$$

Using the theory of systems of linear equations, we note that these equations are not independent and solutions are given by

$$\frac{u_1}{11} = \frac{u_2}{14 - \lambda_1} = t,$$

that is

$$u_1 = 11t, \quad u_2 = (14 - \lambda_1)t,$$

where t is any real number. Taking $t = 1$, we get an eigenvector corresponding to λ_1 as

$$U_1 = \begin{bmatrix} 11 \\ 14 - \lambda_1 \end{bmatrix}.$$

To find a unit eigenvector, we compute the length of X_1 which is given by

$$\begin{aligned} \|U_1\| &= \sqrt{11^2 + (14 - \lambda_1)^2} \\ &= \sqrt{11^2 + (14 - 30.3849)^2} \\ &= 19.7348 \end{aligned}$$

Therefore, a unit eigenvector corresponding to $lambda_1$ is

$$e_1 = \begin{bmatrix} 11/\|U_1\| \\ (14 - \lambda_1)/\|U_1\| \end{bmatrix}$$
$$= \begin{bmatrix} 11/19.7348 \\ (14 - 30.3849)/19.7348 \end{bmatrix}$$
$$= \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

By carrying out similar computations, the unit eigenvector e_2 corresponding to the eigenvalue $\lambda = \lambda_2$ can be shown to be

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}.$$

5. Computation of first principal components

Let $\begin{bmatrix} X_{1k} \\ X_{2k} \end{bmatrix}$ be the k-th sample in Table 4.2. The first principal component of this example is given by (here "T" denotes the transpose of the matrix)

$$e_1^T \begin{bmatrix} X_{1k} - \bar{X}_1 \\ X_{2k} - \bar{X}_2 \end{bmatrix} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} X_{1k} - \bar{X}_1 \\ X_{2k} - \bar{X}_2 \end{bmatrix}$$
$$= 0.5574(X_{1k} - \bar{X}_1) - 0.8303(X_{2k} - \bar{X}_2).$$

For example, the first principal component corresponding to the first example $\begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix} = \begin{bmatrix} 4 \\ 11 \end{bmatrix}$ is calculated as follows:

$$\begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} X_{11} - \bar{X}_1 \\ X_{21} - \bar{X}_2 \end{bmatrix} = 0.5574(X_{11} - \bar{X}_1) - 0.8303(X_{21} - \bar{X}_2)$$
$$= 0.5574(4 - 8) - 0.8303(11 - 8,5)$$
$$= -4.30535$$

The results of calculations are summarised in Table 4.3.

X_1	4	8	13	7
X_2	11	4	5	14
First principal components	-4.3052	3.7361	5.6928	-5.1238

Table 4.3: First principal components for data in Table 4.2

6. Geometrical meaning of first principal components

As we have seen in Figure 4.1, we introduce new coordinate axes. First we shift the origin to the "center" $(\bar{X}_1, \bar{X}_2)$ and then change the directions of coordinate axes to the directions of the eigenvectors e_1 and e_2 (see Figure 4.3).

Next, we drop perpendiculars from the given data points to the e_1-axis (see Figure 4.4). The first principal components are the e_1-coordinates of the feet of perpendiculars, that is, the projections on the e_1-axis. The projections of the data points on e_1-axis may be taken as approximations of the given data points hence we may replace the given data set with these points. Now, each of these approximations can be unambiguously specified by a single number, namely, the e_1-coordinate of approximation. Thus the two-dimensional data set given in Table 4.2 can be represented approximately by the following one-dimensional data set (see Figure 4.5):

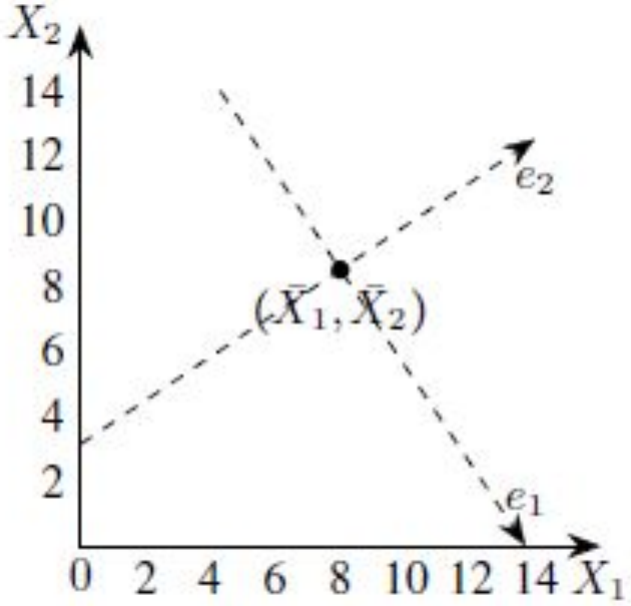


Figure 4.3: Coordinate system for principal components

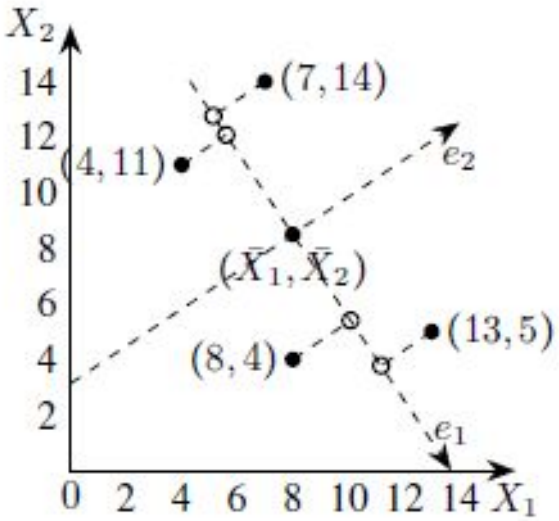


Figure 4.4: Projections of data points on the axis of the first principal component

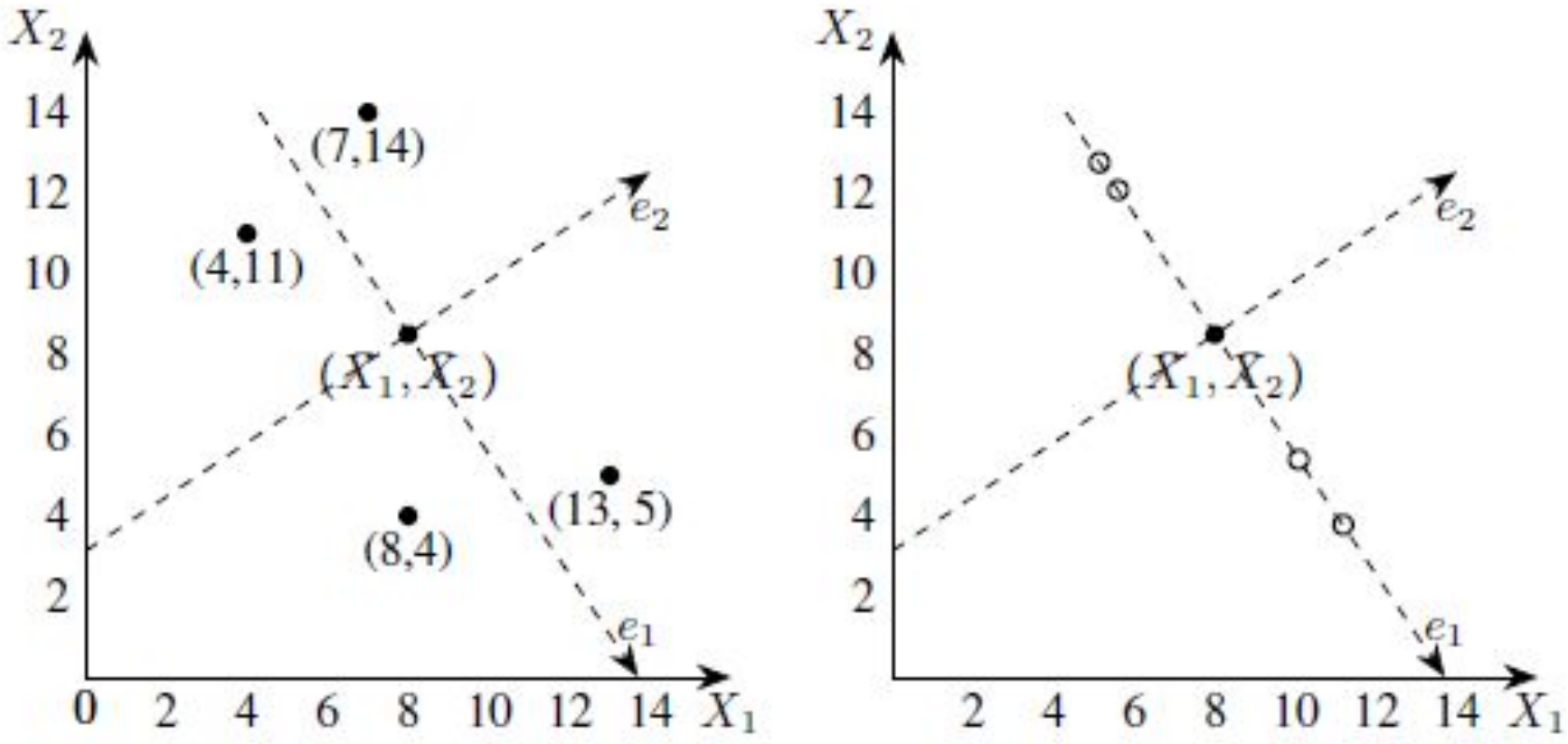


Figure 4.5: Geometrical representation of one-dimensional approximation to the data in Table 4.2