must be square if it is to have an **inverse matrix**. Thus, for an m by m matrix $\mathbf{A}$, we are asking if we can find a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_m$. The answer is that some square matrices have inverses and some do not.

More abstractly, an m by m matrix has an inverse only if both of its null spaces contain only the 0 vector, or if, equivalently, the row and column spaces are both of dimension m. (This is equivalent to the rank of the matrix being m.) Conceptually, an m by m matrix has an inverse if and only if it uniquely maps every non-zero m-dimensional row (column) vector onto a unique, non-zero m-dimensional row (column) vector.

The existence of an inverse matrix is important when solving various matrix equations.

A.2.5 Eigenvalue and Singular Value Decomposition

We now discuss a very important area of linear algebra: eigenvalues and eigenvectors. Eigenvalues and eigenvectors, along with the related concept of singular values and singular vectors, capture the structure of matrices by allowing us to factor or decompose matrices and express them in a standard format. For that reason, these concepts are useful in the solution of mathematical equations and for dimensionality and noise reduction. We begin with the definition of eigenvalues and eigenvectors.

Definition A.8 (Eigenvectors and Eigenvalues). The eigenvalues and eigenvectors of an m by n matrix $\mathbf{A}$ are, respectively, the scalar values λ and the vectors $\mathbf{u}$ that are solutions to the following equation.

$$A\mathbf{u} = \lambda\mathbf{u} \tag{A.13}$$

In other words, **eigenvectors** are the vectors that are unchanged, except for magnitude, when multiplied by $\mathbf{A}$. The **eigenvalues** are the scaling factors. This equation can also be written as $(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0}$.

For square matrices, it is possible to decompose the matrix using eigenvalues and eigenvectors.

Theorem A.1. *Assume that* $\mathbf{A}$ *is an* n *by* n *matrix with* n *independent (orthogonal) eigenvectors,* $u_1, \ldots, u_n$ *and* n *corresponding eigenvalues,* $\lambda_1, \ldots, \lambda_n$. *Let* $\mathbf{U}$ *be the matrix whose columns are these eigenvectors, i.e.,* $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ *and let* $\mathbf{\Lambda}$ *be a diagonal matrix, whose diagonal entries are the* λ_i, $1 \leq i \leq n$. *Then* $\mathbf{A}$ *can be expressed as*

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}. \tag{A.14}$$

Thus, $\mathbf{A}$ can be decomposed into a product of three matrices. $\mathbf{u}$ is known as the **eigenvector matrix** and $\mathbf{\Lambda}$ as the **eigenvalue matrix.**

More generally, an arbitrary matrix can be decomposed in a similar way. Specifically, any m by n matrix $\mathbf{A}$ can be factored into the product of three matrices as described by the following theorem.

Theorem A.2. *Assume that* $\mathbf{A}$ *is an* m *by* n *matrix. Then* $\mathbf{A}$ *can be expressed as follows*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{A.15}$$

Where $\mathbf{U}$ *is* m *by* m, Σ *is* m *by* n, *and* $\mathbf{V}$ *is* n *by* n. $\mathbf{U}$ *and* $\mathbf{V}$ *are orthonormal matrices, i.e., their columns are of unit length and are mutually orthogonal. Thus,* $\mathbf{U}\mathbf{U}^T = \mathbf{I}_m$ *and* $\mathbf{V}\mathbf{V}^T = \mathbf{I}_n$. $\mathbf{\Sigma}$ *is a diagonal matrix whose diagonal entries are non-negative and are sorted so that the larger entries appear first, i.e.,* $\sigma_{i,i} \geq \sigma_{i+1,i+1}$

The column vectors of $\mathbf{V}$, $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the **right singular vectors**, while the columns of $\mathbf{U}$ are the **left singular vectors**. The diagonal elements of $\mathbf{\Sigma}$, the **singular value matrix**, are typically written as $\sigma_1, \ldots, \sigma_n$ and are called the **singular values** of $\mathbf{A}$. (This use of σ should not be confused with the use of σ to represent the standard deviation of a variable.) There are at most $rank(A) \leq \min(m, n)$ non-zero singular values.

It can be shown that the eigenvectors of $\mathbf{A}^T\mathbf{A}$ are the right singular vectors (i.e., the columns of $\mathbf{V}$), while the eigenvectors of $\mathbf{A}\mathbf{A}^T$ are the left singular vectors (i.e., the columns of $\mathbf{U}$). The non-zero eigenvalues of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are the σ_i^2, i.e., the squares of the singular values. Indeed, the eigenvalue decomposition of a square matrix can be regarded as a special case of singular value decomposition.

The singular value decomposition (SVD) of a matrix can also be expressed with the following equation. Note that while $\mathbf{u}_i\mathbf{v}_i^T$ might look like a dot product, it is not, and the result is a rank 1 m by n matrix.

$$\mathbf{A} = \sum_{i=1}^{rank(\mathbf{A})} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{A.16}$$

The importance of the above representation is that every matrix can be expressed as a sum of rank 1 matrices that are weighted by singular values. Since singular values, which are sorted in non-increasing order, often decline rapidly in magnitude, it is possible to obtain a good approximation of a matrix by using only a few singular values and singular vectors. This is useful for dimensionality reduction and will be discussed further in Appendix B.

B

Dimensionality Reduction

This appendix considers various techniques for dimensionality reduction. The goal is to expose the reader to the issues involved and to describe some of the more common approaches. We begin with a discussion of Principal Components Analysis (PCA) and Singular Value Decomposition (SVD). These methods are described in some detail since they are among the most commonly used approaches and we can build on the discussion of linear algebra in Appendix A. However, there are many other approaches that are also employed for dimensionality reduction, and thus, we provide a quick overview of several other techniques. We conclude with a short review of important issues.

B.1 PCA and SVD

PCA and SVD are two closely related techniques. For PCA, the mean of the data is removed, while for SVD, it is not. These techniques have been widely used for decades in a number of fields. In the following discussion, we will assume that the reader is familiar with linear algebra at the level presented in Appendix A.

B.1.1 Principal Components Analysis (PCA)

The goal of PCA is to find a new set of dimensions (attributes) that better captures the variability of the data. More specifically, the first dimension is chosen to capture as much of the variability as possible. The second dimension is orthogonal to the first, and, subject to that constraint, captures as much of the remaining variability as possible, and so on.

PCA has several appealing characteristics. First, it tends to identify the strongest patterns in the data. Hence, PCA can be used as a pattern-finding technique. Second, often most of the variability of the data can be captured by a small fraction of the total set of dimensions. As a result, dimensionality reduction using PCA can result in relatively low-dimensional data and it may be possible to apply techniques that don't work well with high-dimensional data. Third, since the noise in the data is (hopefully) weaker than the patterns, dimensionality reduction can eliminate much of the noise. This is beneficial both for data mining and other data analysis algorithms.

We briefly describe the mathematical basis of PCA and then present an example.

Mathematical Details

Statisticians summarize the variability of a collection of multivariate data; i.e., data that has multiple continuous attributes, by computing the covariance matrix **S** of the data.

Definition B.1. Given an m by n data matrix $\mathbf{D}$, whose m rows are data objects and whose n columns are attributes, the covariance matrix of $\mathbf{D}$ is the matrix $\mathbf{S}$, which has entries s_{ij} defined as

$$s_{ij} = covariance(\mathbf{d}_{*i}, \mathbf{d}_{*j}). \tag{B.1}$$

In words, s_{ij} is the covariance of the i^{th} and j^{th} attributes (columns) of the data.

The covariance of two attributes is defined in Appendix C, and is a measure of how strongly the attributes vary together. If $i = j$, i.e., the attributes are the same, then the covariance is the variance of the attribute. If the data matrix $\mathbf{D}$ is preprocessed so that the mean of each attribute is 0, then $\mathbf{S} = \mathbf{D}^T\mathbf{D}$.

A goal of PCA is to find a transformation of the data that satisfies the following properties:

1. Each pair of new attributes has 0 covariance (for distinct attributes).
2. The attributes are ordered with respect to how much of the variance of the data each attribute captures.
3. The first attribute captures as much of the variance of the data as possible.

4. Subject to the orthogonality requirement, each successive attribute captures as much of the remaining variance as possible.

A transformation of the data that has these properties can be obtained by using eigenvalue analysis of the covariance matrix. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\mathbf{S}$. The eigenvalues are all non-negative and can be ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_{m-1} \geq \lambda_m$. (Covariance matrices are examples of what are called **positive semidefinite matrices**, which, among other properties, have non-negative eigenvalues.) Let $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ be the matrix of eigenvectors of $\mathbf{S}$. These eigenvectors are ordered so that the i^{th} eigenvector corresponds to the i^{th} largest eigenvalue. Finally, assume that data matrix $\mathbf{D}$ has been preprocessed so that the mean of each attribute (column) is 0. We can make the following statements.

- The data matrix $\mathbf{D}' = \mathbf{DU}$ is the set of transformed data that satisfies the conditions posed above.

- Each new attribute is a linear combination of the original attributes. Specifically, the weights of the linear combination for the i^{th} attribute are the components of the i^{th} eigenvector. This follows from the fact that the j^{th} column of $\mathbf{D}'$ is given by $\mathbf{D}\mathbf{u}_j$ and the definition of matrix-vector multiplication given in Equation A.12.

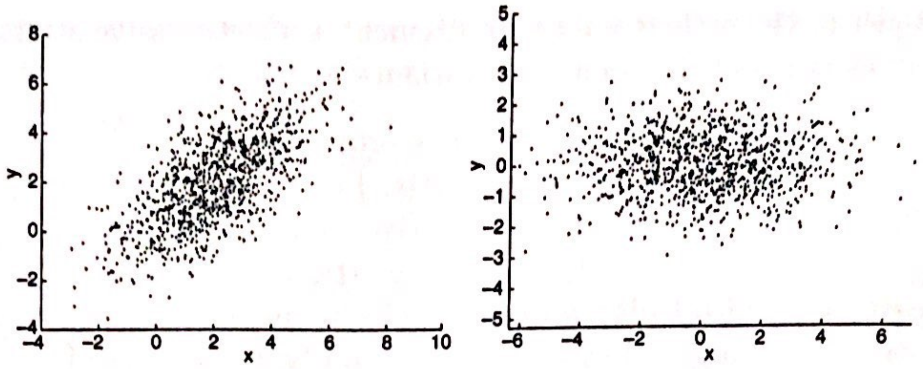
- The variance of the i^{th} new attribute is λ_i.

- The sum of the variance of the original attributes is equal to the sum of the variance of the new attributes.

- The new attributes are called **principal components**; i.e., the first new attribute is the first principal component, the second new attribute is the second principal component, and so on.

The eigenvector associated with the largest eigenvalue indicates the direction in which the data has the most variance. In other words, if all of the data vectors are projected onto the line defined by this vector, the resulting values would have the maximum variance with respect to all possible directions. The eigenvector associated with the second largest eigenvalue is the direction (orthogonal to that of the first eigenvector) in which the data has the largest remaining variance.

The eigenvectors of $\mathbf{S}$ define a new set of axes. Indeed, PCA can be viewed as a rotation of the original coordinate axes to a new set of axes that are aligned with the variability in the data. The total variability of the data is preserved, but the new attributes are now uncorrelated.



(a) Original points.

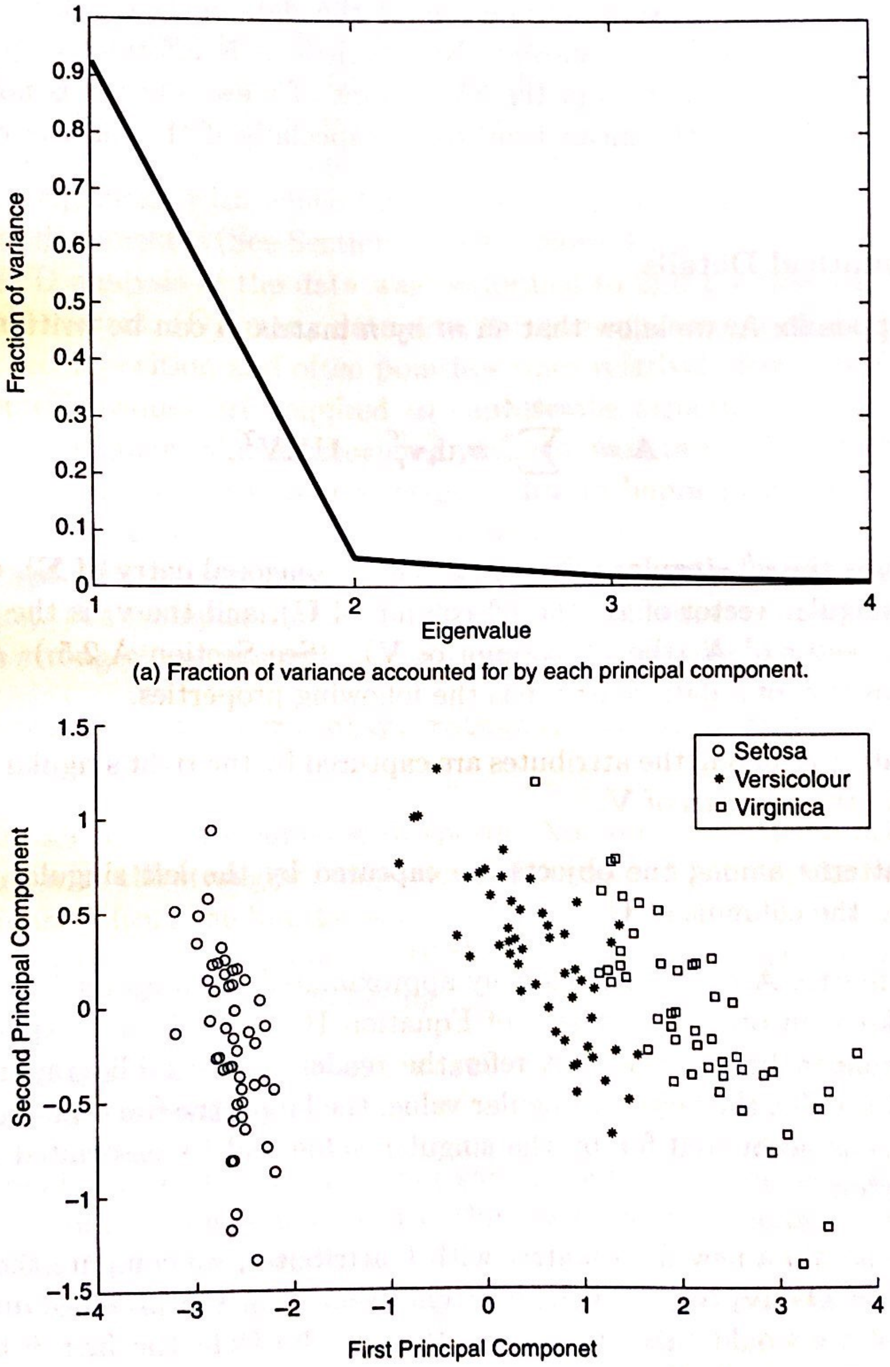
(b) Points after transformation.

Figure B.1. Using PCA to transform the data.

Example B.1 (Two-Dimensional Data). We illustrate the use of PCA for aligning the axes in the directions of the maximum variability of the data. Figure B.1 shows a set of 1000 two-dimensional data points, before and after a PCA transformation. The total variance for the original set of points is the sum of the variance of the x and y attributes, which is equal to $2.84 + 2.95 = 5.79$. After transformation, the variance is $4.81 + 0.98 = 5.79$. ■

Example B.2 (Iris Data). This example uses the Iris data set to demonstrate the use of PCA for dimensionality reduction. This data set contains 150 data objects (flowers); there are 50 flowers from each of three different Iris species: Setosa, Versicolour, and Virginica. Each flower is described by four attributes: sepal length, sepal width, petal length, and petal width. See Chapter 3 for more details.

Figure B.2(a) shows a plot of the fraction of the overall variance accounted for by each eigenvalue (principal component) of the covariance matrix. This type of plot is known as a **scree plot** and is useful for determining how many principal components need to be kept to capture most of the variability of the data. For the Iris data, the first principal component accounts for most of the variation (92.5%), the second for only 5.3%, and the last two components for just 2.2%. Thus, keeping only the first two principal components preserves most of the variability in the data set. Figure B.2(b) shows a scatter plot of the Iris data based on the first two principal components. Note that the Setosa flowers are well separated from the Versicolour and Virginica flowers. The latter two sets of flowers, while much closer to each other, are still relatively well separated.



(a) Fraction of variance accounted for by each principal component.

(b) Plot of first two principal components of Iris data.

Figure B.2. PCA applied to the Iris data set.

B.1.2 SVD

PCA is equivalent to an SVD analysis of the data matrix, once the mean of each variable has been removed. Nonetheless, it is informative to look at dimensionality reduction from the SVD point of view, since it is not always desirable to remove the mean from data, especially if the data is relatively sparse.

Mathematical Details

From Appendix A, we know that an m by n matrix A can be written as

$$\mathbf{A} = \sum_{i=1}^{rank(A)} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{B.2}$$

where σ_i is the i^{th} singular value of $\mathbf{A}$ (the i^{th} diagonal entry of $\mathbf{\Sigma}$), $\mathbf{u}_i$ is the i^{th} left singular vector of $\mathbf{A}$ (the i^{th} column of $\mathbf{U}$), and the $\mathbf{v}_i$ is the i^{th} right singular vector of $\mathbf{A}$ (the i^{th} column of $\mathbf{V}$). (See Section A.2.5.) An SVD decomposition of a data matrix has the following properties.

- Patterns among the attributes are captured by the right singular vectors, i.e., the columns of $\mathbf{V}$.

- Patterns among the objects are captured by the left singular vectors, i.e., the columns of $\mathbf{U}$.

- A matrix $\mathbf{A}$ can be successively approximated in an optimal manner by taking, in order, the terms of Equation B.2. We do not explain what we mean by optimal, but refer the reader to the bibliographic notes. Informally, the larger a singular value, the larger the fraction of a matrix that is accounted for by the singular value and its associated singular vectors.

- To obtain a new data matrix with k attributes, we compute the matrix $\mathbf{D}' = \mathbf{D} * [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$. It might seem from the previous discussion that we would take the matrix that results from the first k terms of Equation A.12. However, while the resulting matrix is of rank k, it still has n columns (attributes).

Example B.3 (Document Data). SVD decomposition can be used to analyze document data. The data for this example consists of 3204 newspaper articles from the *Los Angeles Times*. These articles come from 6 different sections: Entertainment, Financial, Foreign, Metro, National, and Sports. The data matrix is a document-term matrix, where each row represents a document and each column is a term (word). The value of the ij^{th} entry is the number of times the j^{th} term occurs in the i^{th} document. The data was processed using standard techniques to remove common words, to adjust for the different frequencies with which terms appear, and to adjust for the different lengths of documents. (See Section 2.3.7 for more details.)

An SVD analysis of the data was performed to find the first 100 singular values and vectors. (For many data sets, it is too expensive to find a full SVD or PCA decomposition and often pointless since relatively few of the singular values or eigenvalues are required to capture the structure of the matrix.) The largest singular value is associated with common terms that are frequent, but not eliminated by the preprocessing. (It can happen that the strongest patterns represent noise or uninteresting patterns.)

However, the patterns associated with other singular values were more interesting. For example, the following are the top 10 terms (words) associated with the strongest components in the second right singular vector:

```
game, score, lead, team, play, rebound, season, coach, league,
goal
```

These are all terms associated with sports. Not surprisingly, the documents associated with the strongest components of the second left singular vector are predominantly from the Sports section.

The top 10 terms associated with the strongest components in the third right singular vector are the following:

```
earn, million, quarter, bank, rose, billion, stock, company,
corporation, revenue
```

These are all financial terms, and, not surprisingly, the documents associated with the strongest components in the third left singular vector are predominantly from the Financial section.

We reduced the dimensionality of the data using the second and third singular vectors, i.e., $\mathbf{D}' = \mathbf{D} * [\mathbf{v}_2, \mathbf{v}_3]$. In other words, all documents were expressed in terms of two attributes, one relating to Sports and one relating to Finance. A scatter plot of documents is given by Figure B.3. For clarity, non-Sports, non-Financial documents have been eliminated. The Sports documents are shown in a lighter shade of gray, while the Financial documents are a darker gray. The two different categories of documents are well separated for

C

Probability and Statistics

This appendix presents some of the basic concepts in probability and statistics used throughout this book.

C.1 Probability

A **random experiment** is the act of measuring a process whose outcome is uncertain. Examples include rolling a die, drawing from a deck of cards, and monitoring the types of traffic across a network router. The set of all possible outcomes of a random experiment is known as the **sample space**, Ω. For example, $\Omega = \{1, 2, 3, 4, 5, 6\}$ is the sample space for rolling a die. An **event** E corresponds to a subset of these outcomes, i.e., $E \subseteq \Omega$. For example $E = \{2, 4, 6\}$ is the event of observing an even number when rolling a die.

A probability P is a real-valued function defined on the sample space Ω that satisfies the following properties:

1. For any event $E \subseteq \Omega$, $0 \leq P(E) \leq 1$.
2. $P(\Omega) = 1$.
3. For any set of disjoint events, $E_1, E_2, \ldots, E_k \in \Omega$,

$$P(\bigcup_{i=1}^{k} E_i) = \sum_{i=1}^{k} P(E_i).$$

The probability of an event E, which is written as $P(E)$, is the fraction of times event E is observed in a potentially unlimited number of experiments.

In a random experiment, there is often a quantity of interest we want to measure; e.g., counting the number of times a tail turns up when tossing a coin fifty times or measuring the height of a person taking a roller coaster ride at a theme park. Since the value of the quantity depends on the outcome of a random experiment, the quantity of interest is known as a **random variable.** The value of a random variable can be discrete or continuous. A Bernoulli random variable, for example, is a discrete random variable whose only possible values are 0 and 1.

For a discrete random variable X, the probability X takes on a particular value ν is given by the total probability of all outcomes e in which $X(e) = \nu$:

$$P(X = \nu) = P(E = \{e | e \in \Omega, X(e) = \nu\}). \tag{C.1}$$

The probability distribution of a discrete random variable X is also known as its **probability mass function.**

Example C.1. Consider a random experiment where a fair coin is tossed four times. There are 16 possible outcomes of this experiment: HHHH, HHHT, HHTH, HTHH, THHH, HHTT, HTHT, THHT, HTTH, THTH, TTHH, HTTT, THTT, TTHT, TTTH, and TTTT, where H (T) indicates that a head (tail) is observed. Let X be a random variable that measures the number of times a tail is observed in the experiment. The five possible values for X are 0, 1, 2, 3, and 4. The probability mass function for X is given by the following table:

X	0	1	2	3	4
P(X)	1/16	4/16	6/16	4/16	1/16

For example, $P(X = 2) = 6/16$ because there are six outcomes in which the tail is observed twice during the four tosses. ■

On the other hand, if X is a continuous random variable, then the probability that X has a value between a and b is

$$P(a < x < b) = \int_a^b f(x)dx \tag{C.2}$$

The function $f(x)$ is known as the **probability density function** (pdf). Because f is a continuous distribution, the probability that X takes a particular value x is always zero.

Table C.1. Examples of probability functions. ($\Gamma(n+1) = n\Gamma(n)$ and $\Gamma(1) = 1$)

	Probability Function	Parameters
Gaussian	$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	μ, σ
Binomial	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$	n, p
Poisson	$p(x) = \frac{1}{x!}\theta^x \exp^{-\theta}$	θ
Exponential	$p(x) = \theta \exp^{-\theta x}$	θ
Gamma	$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp^{-\lambda x}$	λ, α
Chi-square	$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp^{-x/2}$	k

Table C.1 shows some of the well-known discrete and continuous probability functions. The notion of a probability (mass or density) function can be extended to more than one random variable. For example, if X and Y are random variables, then $p(X, Y)$ denotes their **joint** probability function. The random variables are **independent** of each other if $P(X, Y) = P(X) \times P(Y)$. If two random variables are independent, it means that the value for one variable has no impact on the value for the other.

Conditional probability is another useful concept for understanding the dependencies among random variables. The conditional probability for variable Y given X, denoted as $P(Y|X)$, is defined as

$$P(Y|X) = \frac{P(X,Y)}{P(X)}. \tag{C.3}$$

If X and Y are independent, then $P(Y|X) = P(Y)$. The conditional probabilities $P(Y|X)$ and $P(X|Y)$ can be expressed in terms of one another using a formula known as the **Bayes theorem:**

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}. \tag{C.4}$$

If $\{X_1, X_2, \ldots, X_k\}$ is the set of mutually exclusive and exhaustive outcomes of a random variable X, then the denominator of the above equation can be expressed as follows:

$$P(X) = \sum_{i=1}^{k} P(X, Y_i) = \sum_{i=1}^{k} P(X|Y_i)P(Y_i). \tag{C.5}$$

Equation C.5 is called the **law of total probability**.

C.1.1 Expected Values

The **expected value** of a function g of a random variable X, denoted as $E[g(X)]$, is the weighted-average value of $g(X)$, where the weights are given by the probability function for X. If X is a discrete random variable, then the expected value can be computed as follows:

$$E[g(X] = \sum_{i} g(x_i)P(X = x_i). \tag{C.6}$$

On the other hand, if X is a continuous random variable,

$$E[g(X)] = \int_{-\infty}^{\infty} g(X)f(X)dX, \tag{C.7}$$

where $f(X)$ is the probability density function for X. The remainder of this section considers only the expected values for discrete random variables. The corresponding expected values for continuous random variables are obtained by replacing the summation with an integral.

There are several particularly useful expected values in probability theory. First, if $g(X) = X$, then

$$\mu_X = E[X] = \sum_{i} x_i \; P(X = x_i). \tag{C.8}$$

This expected value corresponds to the **mean** value of the random variable X. Another useful expected value is when $g(X) = (X - \mu_X)$. The expected value of this function is

$$\sigma_X^2 = E[(X - \mu_X)^2] = \sum_{i} (x_i - \mu_X)^2 \; P(X = x_i). \tag{C.9}$$

This expected value corresponds to the **variance** of the random variable X. The square root of the variance corresponds to the **standard deviation** of the random variable X.

Example C.2. Consider the random experiment described in Example C.1. The average number of tails expected to show up when a fair coin is tossed four times is

$$\mu_X = 0 \times 1/16 + 1 \times 4/16 + 2 \times 6/16 + 3 \times 4/16 + 4 \times 1/16 = 2. \tag{C.10}$$

The variance for the number of tails expected to show up is

$$\begin{aligned} \sigma_X^2 &= (0-2)^2 \times 1/16 + (1-2)^2 \times 4/16 + (2-2)^2 \times 6/16 \\ &\quad + (3-2)^2 \times 4/16 + (4-2)^2 \times 1/16 = 1. \end{aligned}$$

■

For pairs of random variables, a useful expected value to compute is the **covariance** function, *Cov*, which is defined as follows:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{C.11}$$

Note that the variance of a random variable X is equivalent $Cov(X, X)$. The expected value of a function also has the following properties:

1. $E[a] = a$, if a is a constant.

2. $E[aX] = aE[X]$.

3. $E[aX + bY] = aE[X] + bE[Y]$.

Based on these properties, Equations C.9 and C.11 can be rewritten as follows:

$$\sigma_X^2 = E[(X - \mu_X)^2] = E[X^2] - E[X]^2 \tag{C.12}$$

$$Cov(X, Y) = E[XY] - E[X]E[Y] \tag{C.13}$$

C.2 Statistics

To draw conclusions about a population, it is generally not feasible to gather data from the entire population. Instead, we must make reasonable conclusions about the population based on evidence gathered from sampled data. The process of drawing reliable conclusions about the population based on sampled data is known as **statistical inference**.

D

Regression

Regression is a predictive modeling technique where the target variable to be estimated is continuous. Examples of applications of regression include predicting a stock market index using other economic indicators, forecasting the amount of precipitation in a region based on characteristics of the jet stream, projecting the total sales of a company based on the amount spent for advertising, and estimating the age of a fossil according to the amount of carbon-14 left in the organic material.

D.1 Preliminaries

Let D denote a data set that contains N observations,

$$D = \{(\mathbf{x}_i, y_i) |\ i = 1, 2, \ldots, N\}.$$

Each $\mathbf{x}_i$ corresponds to the set of attributes of the ith observation (also known as the **explanatory variables**) and y_i corresponds to the **target** (or response) **variable.** The explanatory attributes of a regression task can be either discrete or continuous.

Definition D.1 (Regression). Regression is the task of learning a **target function** f that maps each attribute set $\mathbf{x}$ into a continuous-valued output y.

The goal of regression is to find a target function that can fit the input data with minimum error. The **error function** for a regression task can be expressed in terms of the sum of absolute or squared error:

$$\text{Absolute Error} = \sum_i |y_i - f(\mathbf{x}_i)| \tag{D.1}$$

$$\text{Squared Error} = \sum_i (y_i - f(\mathbf{x}_i))^2 \tag{D.2}$$

D.2 Simple Linear Regression

Consider the physiological data shown in Figure D.1. The data corresponds to measurements of heat flux and skin temperature of a person during sleep. Suppose we are interested in predicting the skin temperature of a person based on the heat flux measurements generated by a heat sensor. The two-dimensional scatter plot shows that there is a strong linear relationship between the two variables.

Heat Flux	Skin Temperature
10.858	31.002
10.617	31.021
10.183	31.058
9.7003	31.095
9.652	31.133
10.086	31.188
9.459	31.226
8.3972	31.263
7.6251	31.319
7.1907	31.356
7.046	31.412
6.9494	31.468
6.7081	31.524

Heat Flux	Skin Temperature
6.3221	31.581
6.0325	31.618
5.7429	31.674
5.5016	31.712
5.2603	31.768
5.1638	31.825
5.0673	31.862
4.9708	31.919
4.8743	31.975
4.7777	32.013
4.7295	32.07
4.633	32.126
4.4882	32.164

Heat Flux	Skin Temperature
4.3917	32.221
4.2951	32.259
4.2469	32.296
4.0056	32.334
3.716	32.391
3.523	32.448
3.4265	32.505
3.3782	32.543
3.4265	32.6
3.3782	32.657
3.3299	32.696
3.3299	32.753
3.4265	32.791

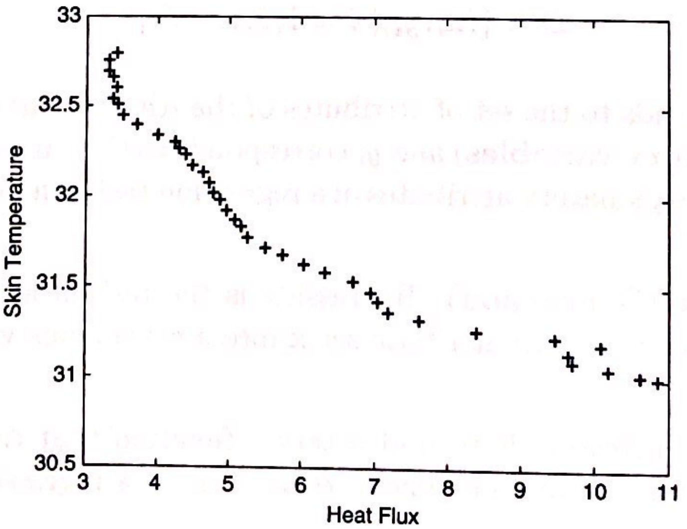


Figure D.1. Measurements of heat flux and skin temperature of a person.

D.2.1 Least Square Method

Suppose we wish to fit the following linear model to the observed data:

$$f(x) = \omega_1 x + \omega_0, \tag{D.3}$$

where ω_0 and ω_1 are parameters of the model and are called the **regression coefficients**. A standard approach for doing this is to apply the **method of least squares**, which attempts to find the parameters (ω_0, ω_1) that minimize the sum of the squared error

$$SSE = \sum_{i=1}^{N}[y_i - f(x_i)]^2 = \sum_{i=1}^{N}[y_i - \omega_1 x - \omega_0]^2, \tag{D.4}$$

which is also known as the **residual sum of squares.**

This optimization problem can be solved by taking the partial derivative of E with respect to ω_0 and ω_1, setting them to zero, and solving the corresponding system of linear equations.

$$\begin{aligned} \frac{\partial E}{\partial \omega_0} &= -2\sum_{i=1}^{N}[y_i - \omega_1 x_i - \omega_0] = 0 \\ \frac{\partial E}{\partial \omega_1} &= -2\sum_{i=1}^{N}[y_i - \omega_1 x_i - \omega_0]x_i = 0 \end{aligned} \tag{D.5}$$

These equations can be summarized by the following matrix equation, which is also known as the **normal equation**:

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}. \tag{D.6}$$

Since $\sum_i x_i = 229.9$, $\sum_i x_i^2 = 1569.2$, $\sum_i y_i = 1242.9$, and $\sum_i x_i y_i = 7279.7$, the normal equations can be solved to obtain the following estimates for the parameters.

$$\begin{aligned} \begin{pmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{pmatrix} &= \begin{pmatrix} 39 & 229.9 \\ 229.9 & 1569.2 \end{pmatrix}^{-1} \begin{pmatrix} 1242.9 \\ 7279.7 \end{pmatrix} \\ &= \begin{pmatrix} 0.1881 & -0.0276 \\ -0.0276 & 0.0047 \end{pmatrix} \begin{pmatrix} 1242.9 \\ 7279.7 \end{pmatrix} \\ &= \begin{pmatrix} 33.1699 \\ -0.2208 \end{pmatrix} \end{aligned}$$

Thus, the linear model that best fits the data in terms of minimizing the SSE is

$$f(x) = 33.17 - 0.22x.$$

Figure D.2 shows the line corresponding to this model.

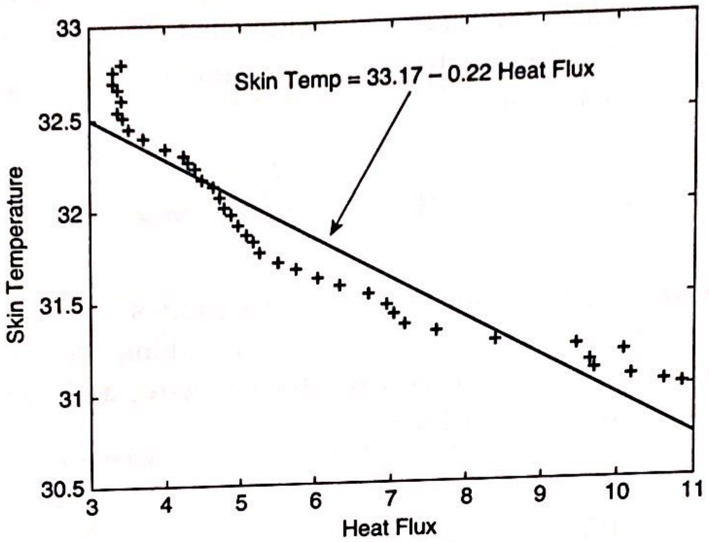


Figure D.2. A linear model that fits the data given in Figure D.1.

We can show that the general solution to the normal equations given in D.6 can be expressed as follow:

$$\begin{aligned} \hat{\omega}_0 &= \overline{y} - \hat{\omega}_1 \overline{x} \\ \hat{\omega}_1 &= \frac{\sigma_{xy}}{\sigma_{xx}} \end{aligned} \tag{D.7}$$

where $\overline{x} = \sum_i x_i/N$, $\overline{y} = \sum_i y_i/N$, and

$$\sigma_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) \tag{D.8}$$

$$\sigma_{xx} = \sum_i (x_i - \overline{x})^2 \tag{D.9}$$

$$\sigma_{yy} = \sum_i (y_i - \overline{y})^2 \tag{D.10}$$

Thus, linear model that results in the minimum squared error is given by

$$f(x) = \overline{y} + \frac{\sigma_{xy}}{\sigma_{xx}}[x - \overline{x}]. \tag{D.11}$$

In summary, the least squares method is a systematic approach to fit a linear model to the response variable y by minimizing the squared error between the true and estimated value of y. Although the model is relatively simple, it seems to provide a reasonably accurate approximation because a linear model is the first-order Taylor series approximation for any function with continuous derivatives.

D.2.2 Analyzing Regression Errors

Some data sets may contain errors in their measurements of $\mathbf{x}$ and y. In addition, there may exist confounding factors that affect the response variable y, but are not included in the model specification. Because of this, the response variable y in regression tasks can be non-deterministic, i.e., it may produce a different value even though the same attribute set $\mathbf{x}$ is provided.

We can model this type of situation using a probabilistic approach, where y is treated as a random variable:

$$\begin{aligned} y &= f(\mathbf{x}) + [y - f(\mathbf{x})] \\ &= f(\mathbf{x}) + \epsilon. \end{aligned} \tag{D.12}$$

Both measurement errors and errors in model specification have been absorbed into a random noise term, ϵ. The random noise present in data is typically assumed to be independent and follow a certain probability distribution.

For example, if the random noise comes from a normal distribution with zero mean and variance σ^2, then

$$P(\epsilon|\mathbf{x}, \Omega) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{[y - f(\mathbf{x},\Omega)]^2}{2\sigma^2}} \tag{D.13}$$

$$\log[P(\epsilon|\mathbf{x}, \Omega)] = -\frac{1}{2}(y - f(\mathbf{x}, \Omega))^2 + \text{constant} \tag{D.14}$$

This analysis shows that minimizing the SSE, $[y - f(\mathbf{x}, \Omega]^2$, implicitly assumes that the random noise follows a normal distribution. Furthermore, it can be shown that the constant model, $f(\mathbf{x}, \Omega) = c$, that best minimizes this type of error is the mean, i.e., $c = \overline{y}$.

Another typical probability model for noise uses the Laplacian distribution:

$$\begin{aligned} P(\epsilon|\mathbf{x},\Omega) &= c\exp^{-c|y-f(\mathbf{x},\Omega)|} \quad &(D.15)\\ \log[P(\epsilon|\mathbf{x},\Omega)] &= -c|y-f(\mathbf{x},\Omega)| + \text{constant} \quad &(D.16)\end{aligned}$$

This suggests that minimizing the absolute error $|y - f(\mathbf{x},\Omega)|$ implicitly assumes that the random noise follows a Laplacian distribution. The best constant model for this case corresponds to $f(\mathbf{x},\Omega) = \tilde{y}$, the median value of y.

Besides the SSE given in Equation D.4, we can also define two other types of errors:

$$\begin{aligned} SST &= \sum_i (y_i - \overline{y})^2 \quad &(D.17)\\ SSM &= \sum_i (f(x_i) - \overline{y})^2 \quad &(D.18)\end{aligned}$$

where SST is known as the total sum of squares and SSM is known as the regression sum of squares. SST represents the prediction error when the average value $\overline{y}$ is used as an estimate for the response variable. SSM, on the other hand, represents the amount of error in the regression model. The relationship among SST, SSE, and SSM is derived as follows:

$$\begin{aligned} SSE &= \sum_i [y_i - \overline{y} + \overline{y} - f(x_i)]^2\\ &= \sum_i [y_i - \overline{y}]^2 + \sum_i [f(x_i) - \overline{y}]^2 + 2\sum_i (y_i - \overline{y})(\overline{y} - f(x_i))\\ &= \sum_i [y_i - \overline{y}]^2 + \sum_i [f(x_i) - \overline{y}]^2 - 2\sum_i (y_i - \overline{y})\omega_1(x_i - \overline{x})\\ &= \sum_i [y_i - \overline{y}]^2 + \sum_i [f(x_i) - \overline{y}]^2 - 2\sum_i \omega_1^2(x_i - \overline{x})^2\\ &= \sum_i [y_i - \overline{y}]^2 - \sum_i [f(x_i) - \overline{y}]^2\\ &= SST - SSM \qquad (D.19)\end{aligned}$$

where we have applied the following relationships:

$$\begin{aligned} \overline{y} - f(x_i) &= -\omega_1(x_i - \overline{x})\\ \sum_i [y_i - \overline{y}][x_i - \overline{x}] &= \sigma_{xy} = \omega_1\sigma_{xx} = \omega_1 \sum_i [x_i - \overline{x}]^2.\end{aligned}$$

Thus, we can write $SST = SSE + SSM$.

D.2.3 Analyzing Goodness of Fit

One way to measure the goodness of the fit is by computing the following measure:

$$R^2 = \frac{SSM}{SST} = \frac{\sum_i [f(x_i) - \overline{y}]^2}{\sum_i [y_i - \overline{y}]^2} \tag{D.20}$$

The R^2 (or *coefficient of determination*) for a regression model may range between 0 and 1. Its value is close to 1 if most of the variability observed in the response variable can be explained by the regression model.

R^2 is also related to the correlation coefficient, r, which measures the strength of the linear relationship between the explanatory and response variables

$$r = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{xy}}}. \tag{D.21}$$

From Equations D.9, D.10, and D.11, we can write

$$\begin{aligned} R^2 &= \frac{\sum_i [f(x_i) - \overline{y}]^2}{\sum_i [y_i - \overline{y}]^2} \\ &= \frac{\sum_i [\frac{\sigma_{xy}}{\sigma_{xx}}(x_i - \overline{x})]^2}{\sigma_{yy}} \\ &= \frac{\sigma_{xy}^2}{\sigma_{xx}^2 \sigma_{yy}} \sum_i (x_i - \overline{x})^2 \\ &= \frac{\sigma_{xy}^2}{\sigma_{xx}^2 \sigma_{yy}} \sigma_{xx} \\ &= \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}}. \end{aligned} \tag{D.22}$$

The above analysis shows that the correlation coefficient is equivalent to the square root of the coefficient of determination (except for its sign, which depends on the direction of the relationship, whether positive or negative).

It is worth noting that R^2 increases as we add more explanatory variables into the model. One way to correct for the number of explanatory variables added to the model is by using the following adjusted R^2 measure:

$$\text{Adjusted } R^2 = 1 - \left(\frac{N-1}{N-d}\right)(1 - R^2), \tag{D.23}$$

where N is the number of data points and $d+1$ is the number of parameters of the regression model.

D.3 Multivariate Linear Regression

The normal equations can be written in a more compact form using the following matrix notation. Let $\mathbf{X} = (\mathbf{1}\ \mathbf{x})$, where $\mathbf{1} = (1,1,1,\ldots)^T$ and $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$. Then, we can show that

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{x} \\ \mathbf{x}^T\mathbf{1} & \mathbf{x}^T\mathbf{x} \end{pmatrix} = \begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}, \qquad \text{(D.24)}$$

which is equivalent to the left-hand side matrix of the normal equation. Similarly, if $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$, we can show that

$$\begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix}^T \mathbf{y} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{x}^T\mathbf{y} \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}, \qquad \text{(D.25)}$$

which is equivalent to the right-hand side matrix of the normal equation. Substituting Equations D.24 and D.25 into Equation D.6 we obtain the following equation:

$$\mathbf{X}^T\mathbf{X}\Omega = \mathbf{X}^T\mathbf{y}, \qquad \text{(D.26)}$$

where $\Omega = (\omega_0, \omega_1)^T$. We can solve for the parameters in Ω can as follows:

$$\Omega = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \qquad \text{(D.27)}$$

The above notation is useful because it allows us to extend the linear regression method to the multivariate case. More specifically, if the attribute set consists of d explanatory attributes $(x_1, x_2, \ldots, x_d)$, $\mathbf{X}$ becomes an $N \times d$ **design matrix**:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1d} \\ 1 & x_{21} & x_{22} & \ldots & x_{2d} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{Nd} \end{pmatrix}, \qquad \text{(D.28)}$$

while $\Omega = (\omega_0, \omega_1, \ldots, \omega_{d-1})^T$ is a d-dimensional vector. The parameters can be computed by solving the matrix equation given in Equation D.26.

D.4 Alternative Least-Square Regression Methods

The least squares method can also be used to find other types of regression models that minimize the SSE. More specifically, if the regression model is

$$\begin{aligned} y &= f(\mathbf{x}, \Omega) + \epsilon \qquad &\text{(D.29)}\\ &= \omega_0 + \sum_i \omega_i g_i(\mathbf{x}) + \epsilon, \qquad &\text{(D.30)} \end{aligned}$$

and the random noise is normally distributed, then we can apply the same methodology as before to determine the parameter vector Ω. The g_i's can be any type of basis functions, including polynomial, kernel, and other nonlinear functions.

For example, suppose $\mathbf{x}$ is a two-dimensional feature vector and the regression model is a polynomial function of degree 2

$$f(x_1, x_2, \Omega) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1 x_2 + \omega_4 x_1^2 + \omega_5 x_2^2. \tag{D.31}$$

If we create the following design matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} & x_{11}^2 & x_{22}^2 \\ 1 & x_{21} & x_{22} & x_{21}x_{22} & x_{21}^2 & x_{22}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{N1} & x_{N2} & x_{N1}x_{N2} & x_{N1}^2 & x_{N2}^2 \end{pmatrix}, \tag{D.32}$$

where x_{ij} is the jth attribute of the ith observation, then the regression problem becomes equivalent to solving Equation D.26. The least-square solution to the parameter vector Ω is given by Equation D.27. By choosing the appropriate design matrix, we can extend this method to any type of basis functions.