Chapter 8

Descriptive statistics

In this chapter we describe several techniques for visualizing data, as well as for computing quantities that summarize it effectively. Such quantities are known as descriptive statistics. As we will see in the following chapters, these statistics can often be interpreted within a probabilistic framework, but they are also useful when probabilistic assumptions are not warranted. Because of this, we present them from a **deterministic** point of view.

8.1 Histogram

We begin by considering data sets containing one-dimensional data. One of the most natural ways of visualizing 1D data is to plot their histogram. The histogram is obtained by binning the range of the data and counting the number of instances that fall within each bin. The width of the bins is a parameter that can be adjusted to yield higher or lower resolution. If we interpret the data as corresponding to samples from a random variable, then the histogram would be a piecewise constant approximation to their pmf or pdf.

Figure 8.1 shows two histograms computed from temperature data gathered at a weather station in Oxford over 150 years.[1] Each data point represents the maximum temperature recorded in January or August of a particular year. Figure 8.2 shows a histogram of the GDP per capita of all countries in the world in 2014 according to the United Nations.[2]

8.2 Sample mean and variance

Averaging the elements in a one-dimensional data set provides a one-number summary of the data, which is a deterministic counterpart to the mean of a random variable (recall that we are making no probabilistic assumptions in this chapter). This can be extended to multi-dimensional data by averaging over each dimension separately.

Definition 8.2.1 (Sample mean)**.** *Let* $\{x_1, x_2, \ldots, x_n\}$ *be a set of real-valued data. The sample*

[1] The data is available at http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt.

[2] The data is available at http://unstats.un.org/unsd/snaama/selbasicFast.asp.

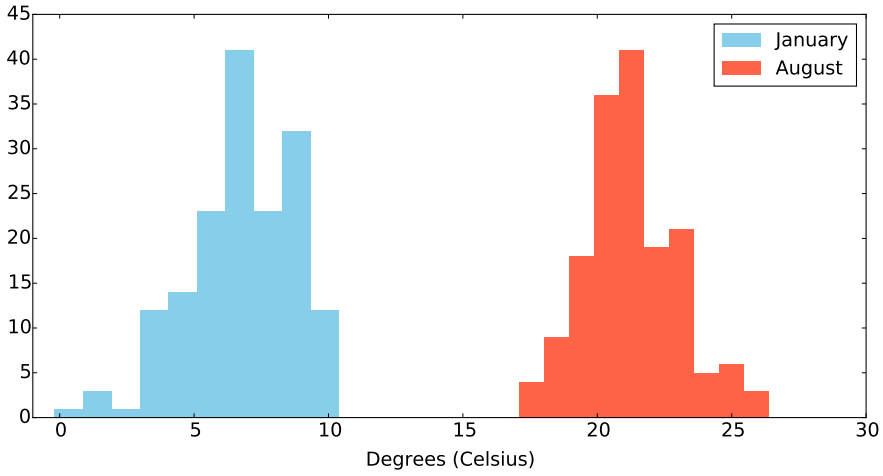

Figure 8.1: Histograms of temperature data taken in a weather station in Oxford over 150 years. Each data point equals the maximum temperature recorded in a certain month in a particular year.

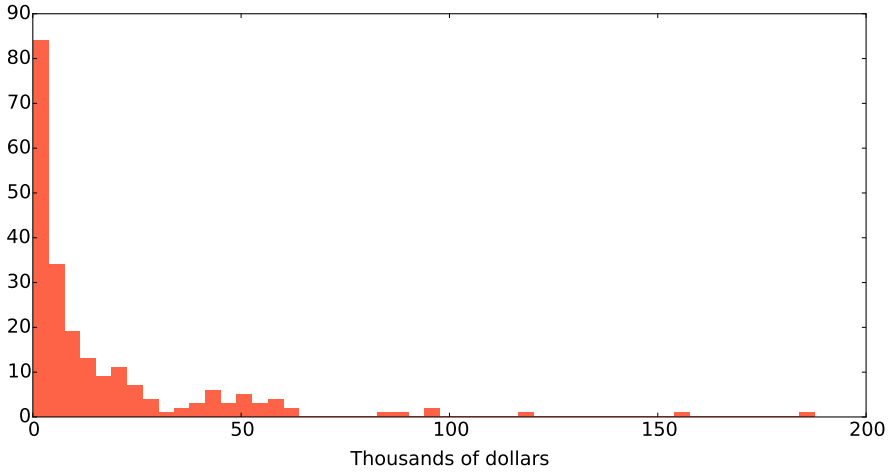

Figure 8.2: Histogram of the GDP per capita of all countries in the world in 2014.

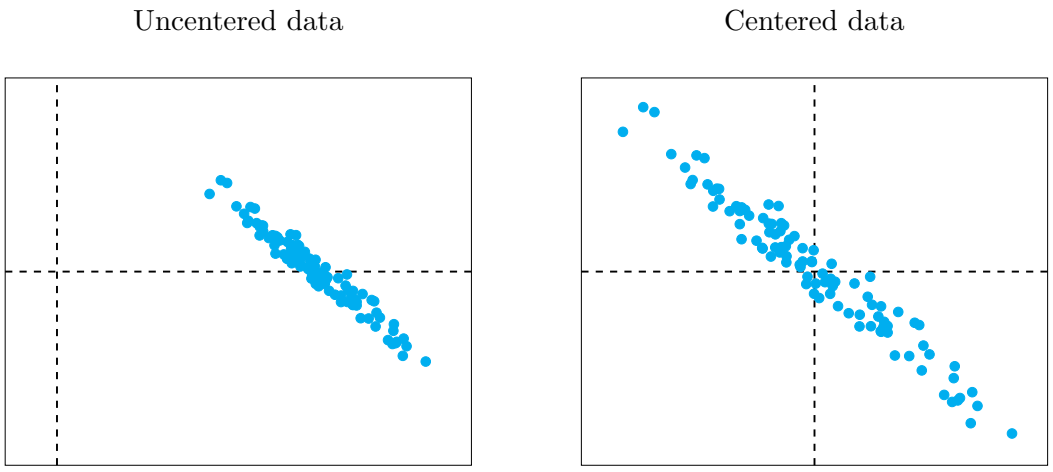

Figure 8.3: Effect of centering a two-dimensional data set. The axes are depicted using dashed lines.

mean of the data is defined as

$$\text{av}\left(x_1, x_2, \ldots, x_n\right) := \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{8.1}$$

Let $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ *be a set of d-dimensional real-valued data vectors. The sample mean is*

$$\text{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right) := \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i. \tag{8.2}$$

The sample mean of the data in Figure 8.1 is 6.73 °C in January and 21.3 °C in August. The sample mean of the GDPs per capita in Figure 8.2 is $16,500.

Geometrically, the average, also known as the sample mean, is the center of mass of the data. A common preprocessing step in data analysis is to **center** a set of data by subtracting its sample mean. Figure 8.3 shows an example.

Algorithm 8.2.2 (Centering)**.** *Let* $\vec{x}_1, \ldots, \vec{x}_n$ *be a set of d-dimensional data. To center the data set we:*

1. *Compute the sample mean following Definition 8.2.1.*
2. *Subtract the sample mean from each vector of data. For* $1 \leq i \leq n$

$$\vec{y}_i := \vec{x}_i - \text{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right). \tag{8.3}$$

The resulting data set $\vec{y}_1, \ldots, \vec{y}_n$ *has sample mean equal to zero; it is centered at the origin.*

The sample variance is the average of the squared deviations from the sample mean. Geometrically, it quantifies the average variation of the data set around its center. It is a deterministic counterpart to the variance of a random variable.

Definition 8.2.3 (Sample variance and standard deviation)**.** *Let* $\{x_1, x_2, \ldots, x_n\}$ *be a set of real-valued data. The sample variance is defined as*

$$\text{var}\left(x_1, x_2, \ldots, x_n\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \text{av}\left(x_1, x_2, \ldots, x_n\right)\right)^2 \tag{8.4}$$

The sample standard deviation is the square root of the sample variance

$$\text{std}\left(x_1, x_2, \ldots, x_n\right) := \sqrt{\text{var}\left(x_1, x_2, \ldots, x_n\right)}. \tag{8.5}$$

You might be wondering why the normalizing constant is $1/\left(n-1\right)$ instead of $1/n$. The reason is that this ensures that the expectation of the sample variance equals the true variance when the data are iid (see Lemma 9.2.5). In practice there is not much difference between the two normalizations.

The sample standard deviation of the temperature data in Figure 8.1 is 1.99 °C in January and 1.73 °C in August. The sample standard deviation of the GDP data in Figure 8.2 is $25,300.

8.3 Order statistics

In some cases, a data set is well described by its mean and standard deviation.

In January the temperature in Oxford is around 6.73 °C give or take 2 °C.

This a pretty accurate account of the temperature data from the previous section. However, imagine that someone describes the GDP data set in Figure 8.2 as:

Countries typically have a GDP per capita of about $16 500 give or take $25 300.

This description is pretty terrible. The problem is that most countries have very small GDPs per capita, whereas a few have really large ones and the sample mean and standard deviation don't really convey this information. Order statistics provide an alternative description, which is usually more informative when there are extreme values in the data.

Definition 8.3.1 (Quantiles and percentiles)**.** *Let* $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ *denote the ordered elements of a set of data* $\{x_1, x_2, \ldots, x_n\}$*. The* q *quantile of the data for* $0 < q < 1$ *is* $x_{([q(n+1)])}$*, where* $[q\left(n+1\right)]$ *is the result of rounding* $q\left(n+1\right)$ *to the closest integer. The* $100\,p$ *quantile is known as the* p *percentile.*

The 0.25 *and* 0.75 *quantiles are known as the first and third* ***quartiles****, whereas the* 0.5 *quantile is known as the* ***sample median****. A quarter of the data are smaller than the* 0.25 *quantile, half are smaller (or larger) than the median and three quarters are smaller than the* 0.75 *quartile. If* n *is even, the sample median is usually set to*

$$\frac{x_{(n/2)} + x_{(n/2+1)}}{2}. \tag{8.6}$$

The difference between the third and the first quartile is known as the ***interquartile range*** *(IQR).*

It turns out that for the temperature data set in Figure 8.1 the sample median is 6.80 °C in January and 21.2 °C in August, which is essentially the same as the sample mean. The IQR is

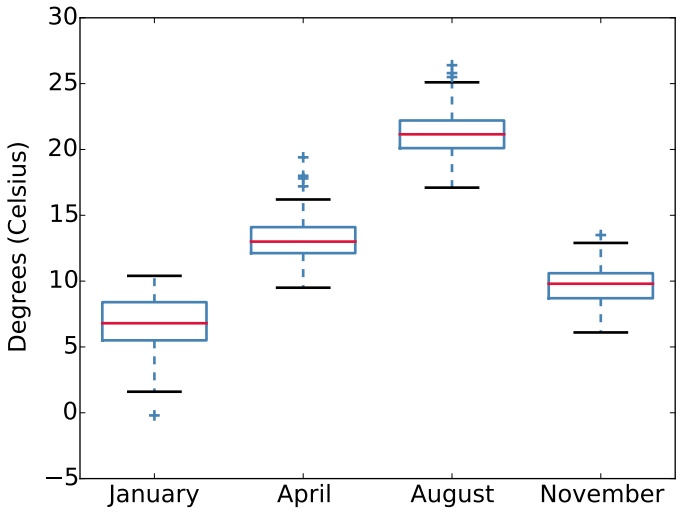

Figure 8.4: Box plots of the Oxford temperature data set used in Figure 8.1. Each box plot corresponds to the maximum temperature in a particular month (January, April, August and November) over the last 150 years.

2.9 °C in January and 2.1 °C in August. This gives a very similar spread around the median, as the sample mean. In this particular example, there does not seem to be an advantage in using order statistics.

For the GDP data set, the median is $6,350. This means that half of the countries have a GDP of less than $6,350. In contrast, 71% of the countries have a GDP per capita lower than the sample mean! The IQR of these data is $18,200. To provide a more complete description of the data set, we can list a **five-number summary** of order statistics: the minimum $x_{(1)}$, the first quartile, the sample median, the third quartile and the maximum $x_{(n)}$. For the GDP data set these are $130, $1,960, $6,350, $20,100, and $188,000 respectively.

We can visualize the main order statistics of a data set by using a **box plot**, which shows the median value of the data enclosed in a box. The bottom and top of the box are the first and third quartiles. This way of visualizing a data set was proposed by the mathematician John Tukey. Tukey's box plot also includes *whiskers*. The lower whisker is a line extending from the bottom of the box to the smallest value within 1.5 IQR of the first quartile. The higher whisker extends from the top of the box to the highest value within 1.5 IQR of the third quartile. Values beyond the whiskers are considered **outliers** and are plotted separately.

Figure 8.4 applies box plots to visualize the temperature data set used in Figure 8.1. Each box plot corresponds to the maximum temperature in a particular month (January, April, August and November) over the last 150 years. The box plots allow us to quickly compare the spread of temperatures in the different months. Figure 8.5 shows a box plot of the GDP data from Figure 8.2. From the box plot it is immediately apparent that most countries have very small GDPs per capita, that the spread between countries increases for larger GDPs per capita and that a small number of countries have very large GDPs per capita.

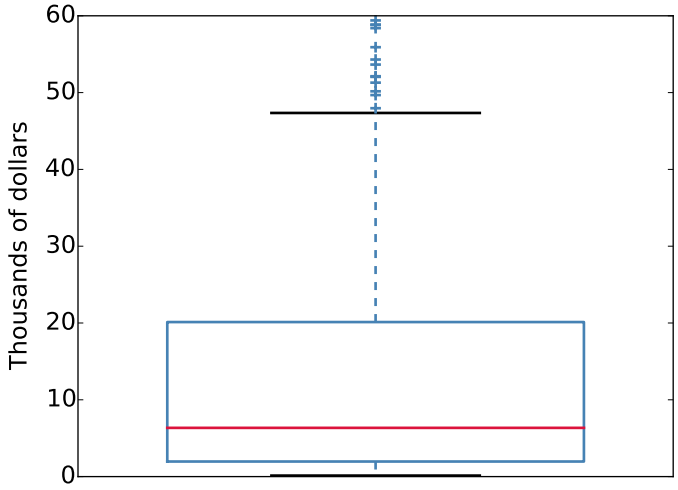

Figure 8.5: Box plot of the GDP per capita of all countries in the world in 2014. Not all of the outliers are shown.

8.4 Sample covariance

In the previous sections we mostly considered data sets consisting of one-dimensional data (except when we discussed the sample mean of a multidimensional data set). In machine-learning lingo, there was only one feature per data point. We now study a multidimensional scenario, where there are several features associated to each data point.

If the dimension of the data set equals to two (i.e. there are two features per data point), we can visualize the data using a **scatter plot**, where each axis represents one of the features. Figure 8.6 shows several scatter plots of temperature data. These data are the same as in Figure 8.1, but we have now arranged them to form two-dimensional data sets. In the plot on the left, one dimension corresponds to the temperature in January and the other dimension to the temperature in August (there is one data point per year). In the plot on the right, one dimension represents the minimum temperature in a particular month and the other dimension represents the maximum temperature in the same month (there is one data point per month). The sample covariance quantifies whether the two features of a two-dimensional data set tend to vary in a similar way on average, just as the covariance quantifies the expected joint variation of two random variables.

Definition 8.4.1 (Sample covariance)**.** *Let* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ *be a data set where each example consists of a measurement of two different features. The sample covariance is defined as*

$$\text{cov}\left((x_1, y_1), \ldots, (x_n, y_n)\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \text{av}\left(x_1, \ldots, x_n\right)\right) \left(y_i - \text{av}\left(y_1, \ldots, y_n\right)\right). \tag{8.7}$$

In order to take into account that each individual feature may vary on a different scale, a common preprocessing step is to *normalize* each feature, dividing it by its sample standard deviation.

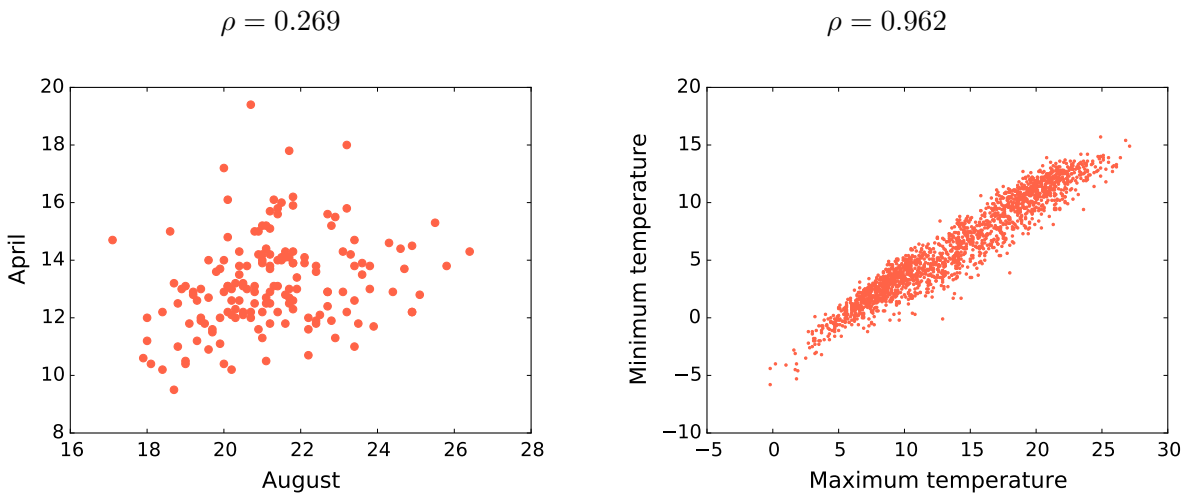

Figure 8.6: Scatterplot of the temperature in January and in August (left) and of the maximum and minimum monthly temperature (right) in Oxford over the last 150 years.

If we normalize before computing the covariance, we obtain the sample correlation coefficient of the two features. One of the advantages of the correlation coefficient is that we don't need to worry about the units in which the features are measured. In contrast, measuring a feature representing distance in inches or miles can severely distort the covariance, if we don't scale the other feature accordingly.

Definition 8.4.2 (Sample correlation coefficient)**.** *Let* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ *be a data set where each example consists of two features. The sample correlation coefficient is defined as*

$$\rho\left((x_1, y_1), \ldots, (x_n, y_n)\right) := \frac{\operatorname{cov}\left((x_1, y_1), \ldots, (x_n, y_n)\right)}{\operatorname{std}\left(x_1, \ldots, x_n\right) \operatorname{std}\left(y_1, \ldots, y_n\right)}. \tag{8.8}$$

By the Cauchy-Schwarz inequality (Theorem B.2.4), which states that for any vectors $\vec{a}$ and $\vec{b}$

$$-1 \leq \frac{\vec{a}^T \vec{b}}{||a||_2 \, ||b||_2} \leq 1, \tag{8.9}$$

the magnitude of the sample correlation coefficient is bounded by one. If it is equal to 1 or -1, then the two centered data sets are collinear. The Cauchy-Schwarz inequality is related to the Cauchy-Schwarz inequality for random variables (Theorem 4.3.7), but here it applies to deterministic vectors.

Figure 8.6 is annotated with the sample correlation coefficients corresponding to the two plots. Maximum and minimum temperatures within the same month are highly correlated, whereas the maximum temperature in January and August within the same year are only somewhat correlated.

8.5 Sample covariance matrix

8.5.1 Definition

We now consider sets of multidimensional data. In particular, we are interested in analyzing the variation in the data. The sample covariance matrix of a data set contains the pairwise sample covariance between every pair of features.

Definition 8.5.1 (Sample covariance matrix)**.** *Let $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ be a set of d-dimensional real-valued data vectors. The sample covariance matrix of these data is the $d \times d$ matrix*

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right) \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T. \tag{8.10}$$

The (i, j) entry of the covariance matrix, where $1 \leq i, j \leq d$, is given by

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)_{ij} = \begin{cases} \text{var}\left(\left(\vec{x}_1\right)_i, \ldots, \left(\vec{x}_n\right)_i\right) & \text{if } i = j, \\ \text{cov}\left(\left(\left(\vec{x}_1\right)_i, \left(\vec{x}_1\right)_j\right), \ldots, \left(\left(\vec{x}_n\right)_i, \left(\vec{x}_n\right)_j\right)\right) & \text{if } i \neq j. \end{cases} \tag{8.11}$$

In order to characterize the variation of a multidimensional data set around its center, we consider its variation in different directions. The average variation of the data in a certain direction is quantified by the sample variance of the projections of the data onto that direction. Let $\vec{v}$ be a unit-norm vector aligned with a direction of interest, the sample variance of the data set in the direction of $\vec{v}$ is given by

$$\text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{v}^T\vec{x}_i - \text{av}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right)\right)^2 \tag{8.12}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{v}^T\left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\right)^2 \tag{8.13}$$

$$= \vec{v}^T \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right) \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T\right) \vec{v}$$

$$= \vec{v}^T \Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) \vec{v}. \tag{8.14}$$

Using the sample covariance matrix we can express the variation in every direction! This is a deterministic analog of the fact that the covariance matrix of a random vector encodes its variance in every direction.

8.5.2 Principal component analysis

Consider the eigendecomposition of the covariance matrix

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}^T. \tag{8.15}$$

By definition, $\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)$ is symmetric, so its eigenvectors u_1, u_2, ..., u_n are orthogonal. By equation (8.14) and Theorem B.7.2, the eigenvectors and eigenvalues completely characterize the variation of the data in every direction.

$\lambda_1/n = 0.497$	$\lambda_1/n = 0.967$	$\lambda_1/n = 1.820$
$\lambda_2/n = 0.476$	$\lambda_2/n = 0.127$	$\lambda_2/n = 0.021$

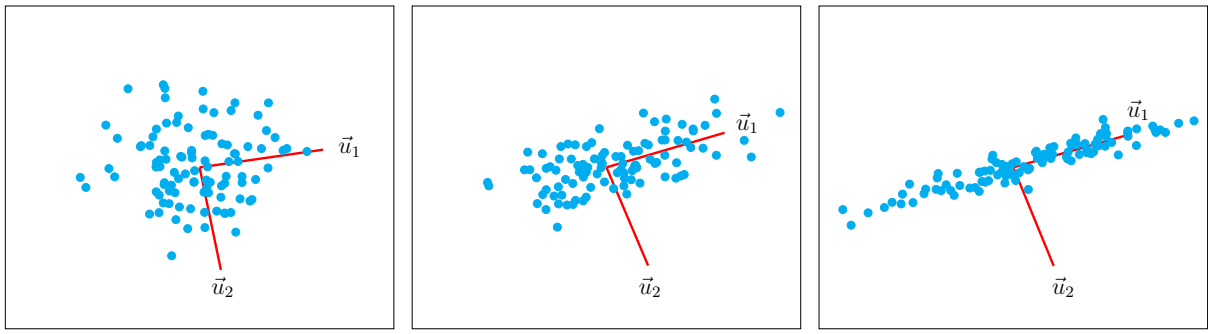

Figure 8.7: PCA of a set consisting of $n = 100$ two-dimensional data points with different configurations.

Theorem 8.5.2. *Let the sample covariance of a set of vectors $\Sigma(\vec{x}_1, \ldots, \vec{x}_n)$ have an eigendecomposition given by (8.15) where the eigenvalues are ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then,*

$$\lambda_1 = \max_{||\vec{v}||_2=1} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.16}$$

$$\vec{u}_1 = \arg\max_{||\vec{v}||_2=1} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.17}$$

$$\lambda_k = \max_{||\vec{v}||_2=1, \vec{u}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.18}$$

$$\vec{u}_k = \arg\max_{||\vec{v}||_2=1, \vec{u}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right). \tag{8.19}$$

This means that $\vec{u}_1$ is the direction of maximum variation. The eigenvector $\vec{u}_2$ corresponding to the second largest eigenvalue λ_2 is the direction of maximum variation that is orthogonal to $\vec{u}_1$. In general, the eigenvector $\vec{u}_k$ corresponding to the kth largest eigenvalue λ_k reveals the direction of maximum variation that is orthogonal to $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_{k-1}$. Finally, $\vec{u}_n$ is the direction of minimum variation.

In data analysis, the eigenvectors of the sample covariance matrix are usually called principal directions. Computing these eigenvectors to quantify the variation of a data set in different directions is called **principal component analysis** (PCA). Figure 8.7 shows the principal directions for several 2D examples.

Figure 8.8 illustrates the importance of centering before applying PCA. Theorem 8.5.2 still holds if the data are not centered. However, the norm of the projection onto a certain direction no longer reflects the variation of the data. In fact, if the data are concentrated around a point that is far from the origin, the first principal direction tends be aligned with that point. This makes sense as projecting onto that direction captures more energy. As a result, the principal directions do not reflect the directions of maximum variation *within* the cloud of data. Centering the data set before applying PCA solves the issue.

The following example explains how to apply principal component analysis to dimensionality reduction. The motivation is that in many cases directions of higher variation are more informative about the structure of the data set.

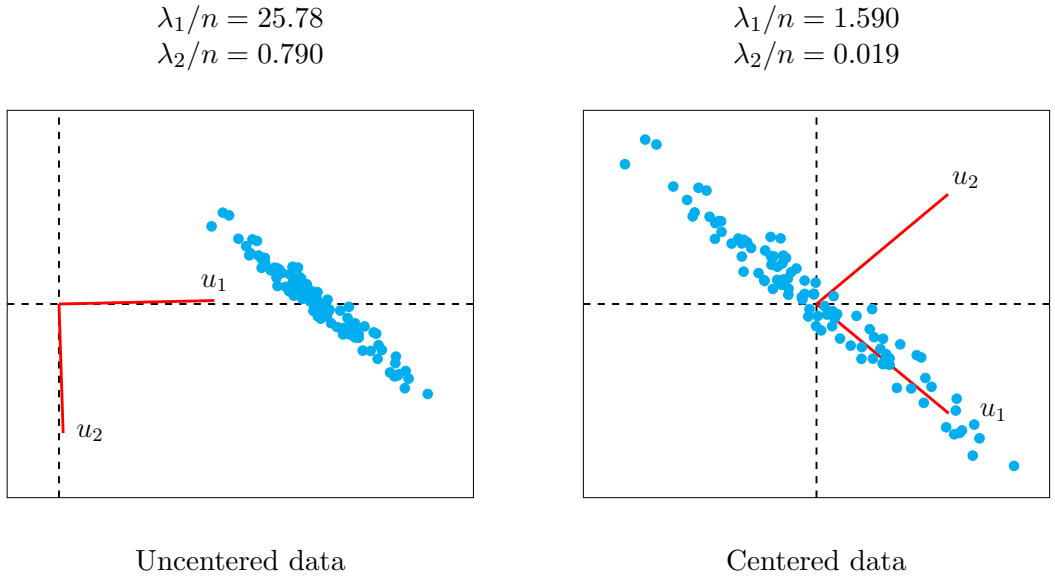

Figure 8.8: PCA applied to $n = 100$ 2D data points. On the left the data are not centered. As a result the dominant principal direction u_1 lies in the direction of the mean of the data and PCA does not reflect the actual structure. Once we center, u_1 becomes aligned with the direction of maximal variation.

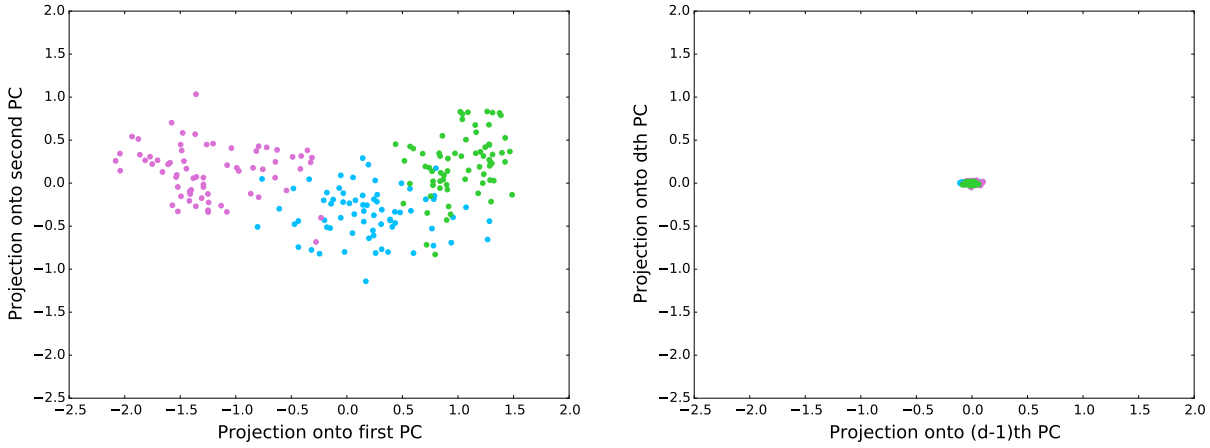

Figure 8.9: Projection of 7-dimensional vectors describing different wheat seeds onto the first two (left) and the last two (right) principal directions of the data set. Each color represents a variety of wheat.

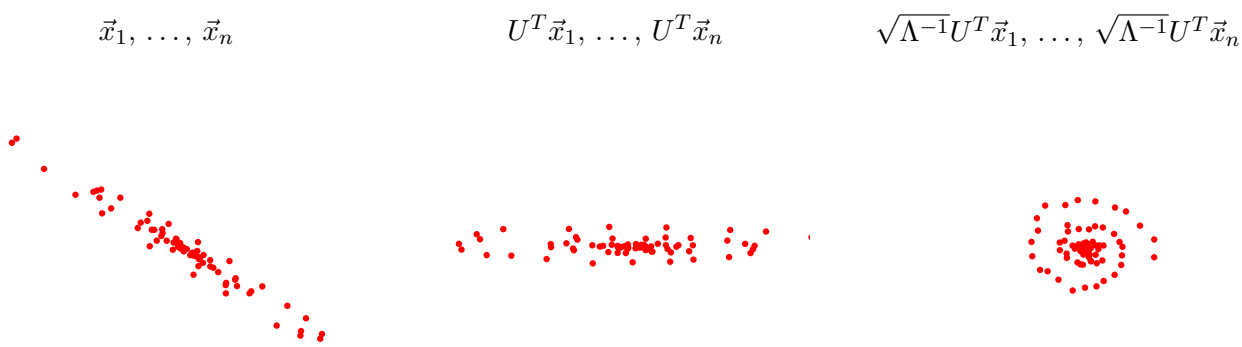

Figure 8.10: Effect of whitening a set of data. The original data are dominated by a linear skew (left). Applying U^T aligns the axes with the eigenvectors of the sample covariance matrix (center). Finally, $\sqrt{\Lambda^{-1}}$ reweights the data along those axes so that they have the same average variation, revealing the nonlinear structure that was obscured by the linear skew (right).

Example 8.5.3 (Dimensionality reduction via PCA)**.** We consider a data set where each data point corresponds to a seed which has seven features: area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. The seeds belong to three different varieties of wheat: Kama, Rosa and Canadian.[3] Our aim is to visualize the data by projecting the data down to two dimensions in a way that preserves as much variation as possible. This can be achieved by projecting each point onto the two first principal dimensions of the data set.

Figure 8.9 shows the projection of the data onto the first two and the last two principal directions. In the latter case, there is almost no discernible variation. The structure of the data is much better conserved by the two first directions, which allow to clearly visualize the difference between the three types of seeds. Note however that projection onto the first principal directions only ensures that we preserve as much variation as possible, but it does not necessarily preserve useful features for tasks such as classification. △

8.5.3 Whitening

Whitening is a useful procedure for preprocessing data that contains nonlinear patterns. The goal is to eliminate the linear skew in the data by rotating and contracting the data along different directions in order to reveal its underlying nonlinear structure. This can be achieved by applying a linear transformation that essentially inverts the sample covariance matrix, so that the result is uncorrelated. The process is known as **whitening**, because random vectors with uncorrelated entries are often referred to as white noise. It is closely related to Algorithm 8.5.4 for coloring random vectors.

Algorithm 8.5.4 (Whitening)**.** *Let $\vec{x}_1, \ldots, \vec{x}_n$ be a set of d-dimensional data, which we assume to be centered and to have a full-rank covariance matrix. To whiten the data set we:*

1. *Compute the eigendecomposition of the sample covariance matrix $\Sigma(\vec{x}_1, \ldots, \vec{x}_n) = U\Lambda U^T$.*

[3]The data can be found at https://archive.ics.uci.edu/ml/datasets/seeds.

2. *Set* $\vec{y}_i := \sqrt{\Lambda}^{-1} U^T \vec{x}_i$, *for* $i = 1, \ldots, n$, *where*

$$\sqrt{\Lambda} := \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}, \tag{8.20}$$

so that $\Sigma(\vec{x}_1, \ldots, \vec{x}_n) = U\sqrt{\Lambda}\sqrt{\Lambda}U^T$.

The whitened data set $\vec{y}_1, \ldots, \vec{y}_n$ has a sample covariance matrix equal to the identity,

$$\Sigma(\vec{y}_1, \ldots, \vec{y}_n) := \frac{1}{n-1} \sum_{i=1}^{n} \vec{y}_i \vec{y}_i^T \tag{8.21}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \sqrt{\Lambda}^{-1} U^T \vec{x}_i \left(\sqrt{\Lambda}^{-1} U^T \vec{x}_i \right)^T \tag{8.22}$$

$$= \sqrt{\Lambda}^{-1} U^T \left(\frac{1}{n-1} \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \right) U \sqrt{\Lambda}^{-1} \tag{8.23}$$

$$= \sqrt{\Lambda}^{-1} U^T \Sigma(\vec{x}_1, \ldots, \vec{x}_n) U \sqrt{\Lambda}^{-1} \tag{8.24}$$

$$= \sqrt{\Lambda}^{-1} U^T U \sqrt{\Lambda} \sqrt{\Lambda} U^T U \sqrt{\Lambda}^{-1} \tag{8.25}$$

$$= I. \tag{8.26}$$

Intuitively, whitening first rotates the data and then shrinks or expands it so that the average variation is the same in every direction. As a result, nonlinear patterns become more apparent, as illustrated by Figure 8.10.