1

High-Dimensional Data

Overview. This chapter introduces the difficulties raised by the analysis of high-dimensional data and motivates the use of appropriate methods. Both practical and theoretical motivations are given. The former ones mainly translate the need to solve real-life problems, which naturally involve high-dimensional feature vectors. Image processing is a typical example. On the other hand, theoretical motivations relate to the study of high-dimensional spaces and distributions. Their properties prove unexpected and completely differ from what is usually observed in low-dimensional spaces. The empty space phenomenon and other strange behaviors are typical examples of the so-called curse of dimensionality. Similarily, the problem of data visualization is shortly dealt with. Regarding dimensionality reduction, the chapter gives two directions to explore: the relevance of variables and the dependencies that bind them. This chapter also introduces the theoretical concepts and definitions (topology, manifolds, etc.) that are typically used in the field of nonlinear dimensionality reduction. Next, a brief section presents two simple manifolds that will be used to illustrate how the different methods work. Finally, the chapter ends with an overview of the following chapters.

1.1 Practical motivations

By essence, the world is multidimensional. To persuade yourself, just look at human beings, bees, ants, neurons, or, in the field of technology, computer networks, sensor arrays, etc. In most cases, combining a large number of simple and existing units allows us to perform a great variety of complex tasks. This solution is cheaper than creating or designing a specific device and is also more robust: the loss or malfunction of a few units does not impair the whole system. This nice property can be explained by the fact that units are often partially redundant. Units that come to failure can be replaced with others that achieve the same or a similar task.

Redundancy means that parameters or features that could characterize the set of various units are not independent from each other. Consequently, the efficient management or understanding of all units requires taking the redundancy into account. The large set of parameters or features must be summarized into a smaller set, with no or less redundancy. This is the goal of *dimensionality reduction* (DR), which is one of the key tools for analyzing high-dimensional data.

1.1.1 Fields of application

The following paragraphs present some fields of technology or science where high-dimensional data are typically encountered.

Processing of sensor arrays

These terms encompass all applications using a set of several identical sensors. Arrays of antennas (e.g., in radiotelescopes) are the best example. But to this class also belong numerous biomedical applications, such as electrocardiagram or electroencephalograph acquisition, where several electrodes record time signals at different places on the chest or the scalp. The same configuration is found again in seismography and weather forecasting, for which several stations or satellites deliver data. The problem of geographic positioning using satellites (as in the GPS or Galileo system) may be cast within the same framework too.

Image processing

Let's consider a picture as the output of a digital camera; then its processing reduces to the processing of a sensor array, like the well-known photosensitive CCD or CMOS captors used in digital photography. However, image processing is often seen as a standalone domain, mainly because vision is a very specific task that holds a priviliged place in information science.

Multivariate data analysis

In contrast with sensor arrays or pixel arrays, multivariate data analysis rather focuses on the analysis of measures that are related to each other but come from different types of sensors. An obvious example is a car, wherein the gearbox connecting the engine to the wheels has to take into account information from rotation sensors (wheels and engine shaft), force sensors (brake and gas pedals), position sensors (gearbox stick, steering wheel), temperature sensors (to prevent engine overheating or to detect glaze), and so forth. Such a situation can also occur in psychosociology: a poll often gathers questions for which the answers are from different types (true/false, percentage, weight, age, etc.).

Data mining

At first sight, data mining seems to be very close to multivariate data analysis. However, the former has a broader scope of applications than the latter, which is a classical subdomain of statistics. Data mining can deal with more exotic data structures than arrays of numbers. For example, data mining encompasses text mining. The analysis of large sets of text documents aims, for instance, at detecting similarities between texts, like common vocabulary, same topic, etc. If these texts are Internet pages, hyperlinks can be encoded in graph structures and analyzed using tools like graph embedding. Cross references in databases can be analyzed in the same way.

1.1.2 The goals to be reached

Understanding large amounts of multidimensional data requires extracting information out of them. Otherwise, data are useless. For example, in electroencephalography, neurologists are interested in finding among numerous electrodes the signals coming from well-specified regions of the brain. When automatically processing images, computers should be able to detect and estimate the movement of objects in the scene. In a car with an automatic gearbox, the on-board computer must be able to select the most appropriate gear ratio according to data from the car sensors.

In all these examples, computers have to help the user to discover and extract information that lies hidden in the huge quantity of data. Information discovery amounts to detecting which variables are relevant and how variables interact with each other. Information extraction then consists of reformulating data, using less variables. Doing so may considerably simplify any further processing of data, whether it is manual, visual, or even automated. In other words, information discovery and extraction help to

- Understand and classify the existing data (by using a “data set” or “learning set”), i.e., assign a class, a color, a rank, or a number to each data sample.
- Infer and generalize to new data (by using a “test set” or “validation set”), i.e., get a continuous representation of the data, so that the unknown class, colour, rank, or number of new data items can be determined, too.

1.2 Theoretical motivations

From a theoretical point of view, all difficulties that occur when dealing with high-dimensional data are often referred to as the “curse of dimensionality”. When the data dimensionality grows, the good and well-known properties of the usual 2D or 3D Euclidean spaces make way for strange and annoying phenomena. The following two subsections highlight two of these phenomena.

1.2.1 How can we visualize high-dimensional spaces?

Visualization is a task that regards mainly two classes of data: *spatial* and *temporal*. In the latter case, the analysis may resort to the additional information given by the location in time.

Spatial data

Quite obviously, a high dimensionality makes the visualization of objects rather uneasy. Drawing one- or two-dimensional objects on a sheet of paper seems very straightforward, even for children. Things becomes harder when three-dimensional objects have to represented. The knowledge of perspective, and its correct mastering, are still recent discoveries (paintings before the Renaissance are not very different from Egyptian papyri!). Even with today's technology, a smooth, dynamic, and realistic representation of our three-dimensional world on a computer screen requires highly specialized chips. On the other hand, three-dimensional objects can also be sculptured or carved. To replace the chisel and the hammer, computer representations of 3D objects can be materialized in a polymer bath: on the surface a laser beam is solidifying the object, layer per layer.

But what happens when more than three dimensions must be taken into account? In this case, the computer screen and the sheet of paper, with only two dimensions, become very limited. Nevertheless, several techniques exist: they use colors or multiple linear projections. Unfortunately, all these techniques are not very intuitive and are often suited only for 4D objects. As an example, Fig. 1.1 shows the projection of a 4D cube that has been projected on a plane in a linear way; the color indicates the depth. Regardless of the projection method is, it is important to remark that the human eye attempts to understand high-dimensional objects in the same way as 3D objects: it seeks distances from one point to another, tries to distinguish what is far and what is close, and follows discontinuities like edges, corners, and so on. Obviously, objects are understood by identifying the relationships between their constituting parts.

Temporal data

When it is known that data are observed in the course of time, an additional piece of information is available. As a consequence, the above-mentioned geometrical representation is no longer unique. Instead of visualizing all dimensions simultaneously in the same coordinate system, one can draw the evolution of each variable as a function of time. For example, in Fig. 1.2, the same data set is displayed "spatially" in the first plot, and "temporally" in the second one: the time structure of data is revealed by the temporal representation only. In constrast with the spatial representation, the temporal representation easily generalizes to more than three dimensions. Nevertheless,

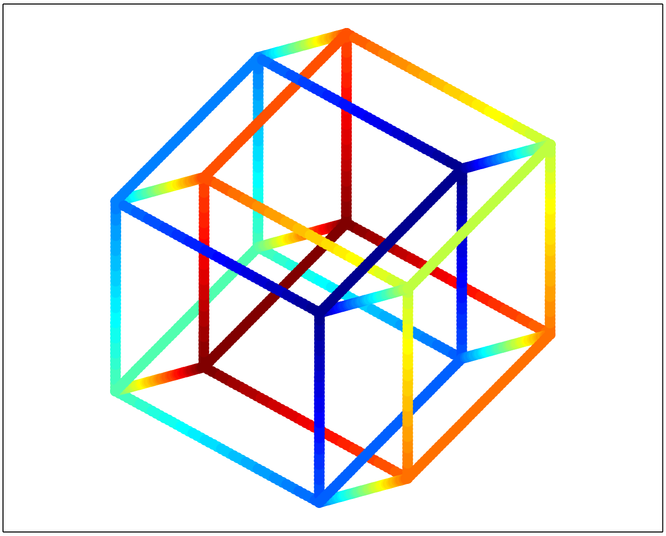

Fig. 1.1. Two-dimensional representation of a four-dimensional cube. In addition to perspective, the color indicates the depth in the fourth dimension.

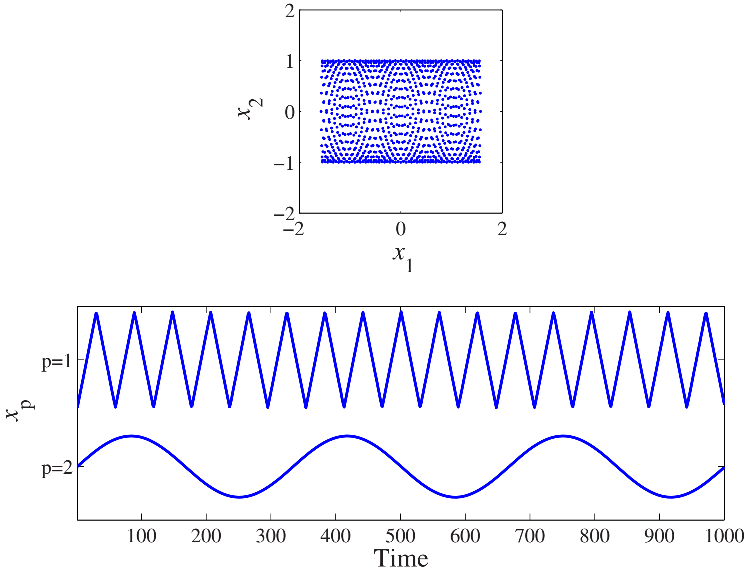

Fig. 1.2. Two plots of the same temporal data. In the first representation, data are displayed in a single coordinate system (spatial representation). In the second representation, each variable is plotted in its own coordinate system, with time as the abscissa (time representation).

when dimensionality increases, it becomes harder and harder to perceive the similarities and dissimilarities between the different variables: the eye is continually jumping from one variable to another, and finally gets lost! In such a case, a representation of data using a smaller set of variables is welcome, as for spatial data. This makes the user's perception easier, especially if each of these variables concentrates on a particular aspect of data. A compact representation that avoids redundancy while remaining trustworthy proves to be the most appealing.

1.2.2 Curse of dimensionality and empty space phenomenon

The colorful term "curse of dimensionality" was apparently first coined by Bellman [14] in connection with the difficulty of optimization by exhaustive enumeration on product spaces. Bellman underlines the fact that considering a Cartesian grid of spacing 1/10 on the unit cube in 10 dimensions, the number of points equals 10^{10}; for a 20-dimensional cube, the number of points further increases to 10^{20}. Accordingly, Bellman's interpretation is the following: if the goal consists of optimizing a function over a continuous domain of a few dozen variables by exhaustively searching a discrete search space defined by a crude discretization, one could easily be faced with the problem of making tens of trillions of evaluations of the function. In other words, the curse of dimensionality also refers to the fact that in the absence of simplifying assumptions, the number of data samples required to estimate a function of several variables to a given accuracy (i.e., to get a reasonably low-variance estimate) on a given domain grows exponentially with the number of dimensions. This fact, responsible for the curse of dimensionality, is often called the "empty space phenomenon" [170]. Because the amount of available data is generally restricted to a few observations, high-dimensional spaces are inherently sparse. More concretely, the curse of dimensionality and the empty space phenomenon give unexpected properties to high-dimensional spaces, as illustrated by the following subsections, which are largely inspired by Chapter 1 of [169].

Hypervolume of cubes and spheres

In a D-dimensional space, a sphere and the corresponding circumscripted cube (all edges equal the sphere diameter) lead to the following volume formulas:

$$V_{\text{sphere}}(r) = \frac{\pi^{D/2} r^D}{\Gamma(1+D/2)} \ , \tag{1.1}$$

$$V_{\text{cube}}(r) = (2r)^D \ , \tag{1.2}$$

where r is the radius of the sphere. Surprisingly, the ratio $V_{\text{sphere}}/V_{\text{cube}}$ tends to zero when D increases:

$$\lim_{D\to\infty} \frac{V_{\text{sphere}}(r)}{V_{\text{cube}}(r)} = 0 \ . \tag{1.3}$$

Intuitively, this means that as dimensionality increases, a cube becomes more and more spiky, like a sea urchin: the spherical body gets smaller and smaller while the number of spikes increases, the latter occupying almost all the available volume. Now, assigning the value 1/2 to r, V_{cube} equals 1, leading to

$$\lim_{D\to\infty} V_{\text{sphere}}(r) = 0 \ . \tag{1.4}$$

This indicates that the volume of a sphere vanishes when dimensionality increases!

Hypervolume of a thin spherical shell

By virtue of Eq. (1.1), the relative hypervolume of a thin spherical shell is

$$\frac{V_{\text{sphere}}(r) - V_{\text{sphere}}(r(1-\epsilon))}{V_{\text{sphere}}(r)} = \frac{1^D - (1-\epsilon)^D}{1^D} \ , \tag{1.5}$$

where ϵ is the thickness of the shell ($\epsilon \ll 1$). When D increases, the ratio tends to 1, meaning that the shell contains almost all the volume [194].

Tail probability of isotropic Gaussian distributions

For any dimension D, the probability density function (pdf) of an isotropic Gaussian distribution (see Appendix B) is written as

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi\sigma^2)^D}} \exp(-\frac{1}{2}\frac{\|\mathbf{y}-\mu_{\mathbf{y}}\|^2}{\sigma^2}) \ , \tag{1.6}$$

where $\mathbf{y}$ is a D-dimensional vector, $\mu_{\mathbf{y}}$ its D-dimensional mean, and σ^2 the isotropic (scalar) variance. Assuming the random vector $\mathbf{y}$ has zero mean and unit variance, the formula simplifies into

$$f_{\mathbf{y}}(\mathbf{y}) = K(r) = \frac{1}{\sqrt{(2\pi)^D}} \exp(-\frac{r^2}{2}) \ , \tag{1.7}$$

where $r = \|\mathbf{y}\|$ can be interpreted as a radius. Indeed, because the distribution is isotropic, the equiprobable contours are spherical. With the previous examples in mind, it can thus be expected that the distribution behaves strangely in high dimensions.

This is confirmed by computing $r_{0.95}$ defined as the radius of a hypersphere that contains 95% of the distribution [45]. The value of $r_{0.95}$ is such that

$$\frac{\int_0^{r_{0.95}} S_{\text{sphere}}(r)K(r)dr}{\int_0^{\infty} S_{\text{sphere}}(r)K(r)dr} = 0.95 \ , \tag{1.8}$$

where $S_{\text{sphere}}(r)$ is the surface of a D-dimensional hypersphere of radius r:

$$S_{\text{sphere}}(r) = \frac{2\pi^{D/2} r^{D-1}}{\Gamma(D/2)} \quad . \tag{1.9}$$

The radius $r_{0.95}$ grows as the dimensionality D increases, as illustrated in the following table:

D	1	2	3	4	5	6
$r_{0.95}$	1.96σ	2.45σ	2.80σ	3.08σ	3.33σ	3.54σ

This shows the weird behavior of a Gaussian distribution in high-dimensional spaces.

Concentration of norms and distances

Another problem encountered in high-dimensional spaces regards the weak discrimination power of a metric. As dimensionality grows, the contrast provided by usual metrics decreases, i.e., the distribution of norms in a given distribution of points tends to concentrate. This is known as the *concentration phenomenon* [20, 64].

For example, the Eulidean norm of vectors consisting of several variables that are i.i.d. (independent and identically distributed) behaves in a totally unexpected way. The explanation can be found in the following theorem (taken from [45], where the demonstration can be found as well):

Theorem 1.1. *Let* $\mathbf{y}$ *be a* D*-dimensional vector* $[y_1, \ldots, y_d, \ldots, y_D]^T$*; all components* y_d *of the vector are independent and identically distributed, with a finite eighth order moment. Then the mean* $\mu_{\|\mathbf{y}\|}$ *and the variance* $\sigma^2_{\|\mathbf{y}\|}$ *of the Euclidean norm (see Subsection 4.2.1) are*

$$\mu_{\|\mathbf{y}\|} = \sqrt{aD - b} + \mathcal{O}(D^{-1}) \tag{1.10}$$

$$\sigma^2_{\|\mathbf{y}\|} = b + \mathcal{O}(D^{-1/2}) \quad , \tag{1.11}$$

where a *and* b *are parameters depending only on the central moments of order 1, 2, 3, and 4 of the* x_i*:*

$$a = \mu^2 + \mu_2 \tag{1.12}$$

$$b = \frac{4\mu^2\mu_2 - \mu_2^2 + 4\mu\mu_3 + \mu_4}{4(\mu^2 + \mu_2)} \quad , \tag{1.13}$$

where $\mu = E\{x_d\}$ *is the common mean of all components* x_d *and* μ^k *their common central* k*-th order moment (*$\mu_k = E\{(x_d - \mu)^k\}$*).*

In other words, the norm of random vectors grows proportionally to $\sqrt{D}$, as naturally expected, but the variance remains more or less constant for a sufficiently large D. This also means that the vector $\mathbf{y}$ seems to be normalized in high dimensions. More precisely, thanks to Chebychev's inequality, one has

$$P\left(\left|\|\mathbf{y}\| - \mu_{\|\mathbf{y}\|}\right| \geq \varepsilon\right) \leq \frac{\sigma^2_{\|\mathbf{y}\|}}{\varepsilon^2} \ , \tag{1.14}$$

i.e., the probability that the norm of $\mathbf{y}$ falls outside an interval of fixed width centered on $\mu_{\|\mathbf{y}\|}$ becomes approximately constant when D grows. As $\mu_{\|\mathbf{y}\|}$ also grows, the relative error made by taking $\mu_{\|\mathbf{y}\|}$ instead of $\|\mathbf{y}\|$ becomes negligible. Therefore, high-dimensional random i.i.d. vectors seem to be distributed close to the surface of a hypersphere of radius $\mu_{\|\mathbf{y}\|}$. This means not only that successive drawings of such random vectors yield almost the same norm, but also that the Euclidean distance between any two vectors is approximately constant. The Euclidean distance is indeed the Euclidean norm of the difference of two random vectors (see Subsection 4.2.1), and this difference is also a random vector.

In practice, the concentration phenomenon makes the nearest-neighbor search problem difficult to solve in high-dimensional spaces [20, 26]. Other results about the surprising behavior of norms and distances measured in high-dimensional spaces are given, for instance, in [1, 64] and references therein.

Diagonal of a hypercube

Considering the hypercube $[-1,+1]^D$, any segment from its center to one of its 2^D corners, i.e., a half-diagonal, can be written as $\mathbf{v} = [\pm 1, \ldots, \pm 1]^T$. The angle between a half-diagonal $\mathbf{v}$ and the dth coordinate axis

$$\mathbf{e}_d = [0, \ldots, 0, 1, 0, \ldots, 0]^T$$

is computed as

$$\cos\theta_D = \frac{\mathbf{v}^T \mathbf{e}_d}{\|\mathbf{v}\| \, \|\mathbf{e}_d\|} = \frac{\pm 1}{\sqrt{D}} \ . \tag{1.15}$$

When the dimensionality D grows, the cosine tends to zero, meaning that half-diagonals are nearly orthogonal to all coordinates axes [169]. Hence, the visualization of high-dimensional data by plotting a subset of two coordinates on a plane can be misleading. Indeed, a cluster of points lying near a diagonal line of the space will be surprisingly plotted near the origin, whereas a cluster lying near a coordinate axis is plotted as intuitively expected.

1.3 Some directions to be explored

In the presence of high-dimensional data, two possibilities exist to avoid or at least attenuate the effects of the above-mentioned phenomena. The first one focuses on the separation between relevant and irrelevant variables. The second one concentrates on the dependencies between the (relevant) variables.

1.3.1 Relevance of the variables

When analyzing multivariate data, not necessarily all variables are related to the underlying information the user wishes to catch. Irrelevant variables may be eliminated from the data set.

Most often, techniques to distinguish relevant variables from irrelevant ones are supervised: the "interest" of a variable is given by an "oracle" or "professor". For example, in a system with many inputs and outputs, the relevance of an input can be measured by computing the correlations between known pairs of input/output. Input variables that are not correlated with the outputs may then be eliminated.

Techniques to determine whether variables are (ir)relevant are not further studied in this book, which focuses mainly on non-supervised methods. For the interested reader, some introductory references include [2, 96, 139].

1.3.2 Dependencies between the variables

Even when assuming that all variables are relevant, the dimensionality of the observed data may still be larger than necessary. For example, two variables may be highly correlated: knowing one of them brings information about the other. In that case, instead of arbitrarily removing one variable in the pair, another way to reduce the number of variables would be to find a new set of *transformed* variables. This is motivated by the facts that dependencies between variables may be very complex and that keeping one of them might not suffice to catch all the information content they both convey.

The new set should obviously contain a smaller number of variables but should also preserve the interesting characteristics of the initial set. In other words, one seeks a transformation of the variables with some well-defined properties. These properties must ensure that the transformation does not alter the information content conveyed by the initial data set, but only represents it in a different form. In the remainder of this book, linear as well as nonlinear transformations of observed variables will often be called *projections*, mainly because many transformations are designed for the preservation of characteristics that are geometrical or interpreted as such.

The type of projection must be chosen according to the model that underlies the data set. For example, if the given variables are assumed to be mixtures of a few unobserved ones, then a projection that inverts the mixing process is very useful. In other words, this projection tracks and eliminates dependencies between the observed variables. These dependencies often result from a lack of knowledge or other imperfections in the observation process: the interesting variables are not directly accessible and are thus measured in several different but largely redundant ways. The determination of a projection may also follow two different goals.

The first and simplest one aims to just detect and eliminate the dependencies. For this purpose, the projection is determined in order to reduce the number of variables. This task is traditionally known as *dimensionality reduction* and attempts to eliminate any redundancy in the initial variables. Principal component analysis (PCA) is most probably the best-known technique for dimensionality reduction.

The second and more complex goal of a projection is not only to reduce the dimensionality, but also to retrieve the so-called *latent variables*, i.e., those that are at the origin of the observed ones but cannot be measured directly. This task, in its most generic acceptation, is often called *latent variable separation.* Blind source separation (BSS), in signal processing, or Independent component analysis (ICA), in multivariate data analysis, are particular cases of latent variable separation.

As can be deduced, dimensionality reduction only focuses on the *number* of latent variables and attempts to give a low-dimensional representation of data according to this number. For this reason, dimensionality reduction does not care for the latent variables themselves: any *equivalent* representation will do. By comparison, latent variable separation is more difficult since it aims, beyond dimensionality reduction, at recovering the unknown latent variables as well as possible.

1.4 About topology, spaces, and manifolds

From a geometrical point of view, when two or more variables depend on each other, their joint distribution — or, more accurately, the *support* of their joint distribution — does not span the whole space. Actually, the dependence induces some structure in the distribution, in the form of a geometrical locus that can be seen as a kind of object in the space. The hypercube illustrated in Fig. 1.1 is an example of such a structure or object. And as mentioned above, dimensionality reduction aims at giving a new representation of these objects while preserving their structure.

In mathematics, topology studies the properties of objects that are preserved through deformations, twistings, and stretchings. Tearing is the only prohibited operation, thereby guaranteeing that the intrinsic "structure" or connectivity of objects is not altered. For example, a circle is topologically equivalent to an ellipse, and a sphere is equivalent to an ellipsoid.[1] However, subsequent chapters of this book will show that tearing still remains a very interesting operation when used carefully.

One of the central ideas of topology is that spatial objects like circles and spheres can be treated as objects in their own right: the knowledge of objects does not depend on how they are represented, or *embedded*, in space. For example, the statement, "If you remove a point from a circle, you get a (curved) line segment" holds just as well for a circle as for an ellipse, and even for

[1] Of course, this does *not* mean that soccer is equivalent to rugby!

tangled or knotted circles. In other words, topology is used to abstract the intrinsic connectivity of objects while ignoring their detailed form. If two objects have the same topological properties, they are said to be *homeomorphic*.

The "objects" of topology are formally defined as topological spaces. A *topological space* is a set for which a *topology* is specified [140]. For a set $\mathcal{Y}$, a topology T is defined as a collection of subsets of $\mathcal{Y}$ that obey the following properties:

- Trivially, $\emptyset \in T$ and $\mathcal{Y} \in T$.
- Whenever two sets are in T, then so is their intersection.
- Whenever two or more sets are in T, then so is their union.

This definition of a topology holds as well for a Cartesian space ($\mathbb{R}^D$) as for graphs. For example, the natural topology associated with $\mathbb{R}$, the set of real numbers, is the union of all open intervals.

From a more geometrical point of view, a topological space can also be defined using neighborhoods and Haussdorf's axioms. The neighborhood of a point $\mathbf{y} \in \mathbb{R}^D$, also called a ϵ-neighborhood or infinitesimal open set, is often defined as the open ϵ-ball $B_\epsilon(\mathbf{y})$, i.e. the set of points inside a D-dimensional hollow sphere of radius $\epsilon > 0$ and centered on $\mathbf{y}$. A set containing an open neighborhood is also called a neighborhood. Then, a topological space is such that

- To each point $\mathbf{y}$ there corresponds at least one neighborhood $\mathcal{U}(\mathbf{y})$, and $\mathcal{U}(\mathbf{y})$ contains $\mathbf{y}$.
- If $\mathcal{U}(\mathbf{y})$ and $\mathcal{V}(\mathbf{y})$ are neighborhoods of the same point $\mathbf{y}$, then a neighborhood $\mathcal{W}(\mathbf{y})$ exists such that $\mathcal{W}(\mathbf{y}) \subset \mathcal{U}(\mathbf{y}) \cup \mathcal{V}(\mathbf{y})$.
- If $\mathbf{z} \in \mathcal{U}(\mathbf{y})$, then a neighborhood $\mathcal{V}(\mathbf{z})$ of $\mathbf{z}$ exists such that $\mathcal{V}(\mathbf{z}) \subset \mathcal{U}(\mathbf{y})$.
- For two distinct points, two disjoint neighborhoods of these points exist.

Within this framework, a (topological) *manifold* $\mathcal{M}$ is a topological space that is locally Euclidean, meaning that around every point of $\mathcal{M}$ is a neighborhood that is topologically the same as the open unit ball in $\mathbb{R}^D$. In general, any object that is nearly "flat" on small scales is a manifold. For example, the Earth is spherical but looks flat on the human scale.

As a topological space, a manifold can be compact or noncompact, connected or disconnected. Commonly, the unqualified term "manifold" means "manifold without boundary". Open manifolds are noncompact manifolds without boundary, whereas closed manifolds are compact manifolds without boundary. If a manifold contains its own boundary, it is called, not surprisingly, a "manifold with boundary". The closed unit ball $\bar{B}_1(\mathbf{0})$ in $\mathbb{R}^D$ is a manifold with boundary, and its boundary is the unit hollow sphere. By definition, every point on a manifold has a neighborhood together with a homeomorphism of that neighborhood with an open ball in $\mathbb{R}^D$.

An *embedding* is a representation of a topological object (a manifold, a graph, etc.) in a certain space, usually $\mathbb{R}^D$ for some D, in such a way that its topological properties are preserved. For example, the embedding of a manifold preserves open sets. More generally, a space $\mathcal{X}$ is embedded in another space $\mathcal{Y}$ when the properties of $\mathcal{Y}$ restricted to $\mathcal{X}$ are the same as the properties of $\mathcal{X}$.

A smooth manifold, also called an (infinitely) differentiable manifold, is a manifold together with its "functional structure" (e.g., parametric equations). Hence, a smooth manifold differs from a simple topological manifold, as defined above, because the notion of differentiability exists on it. Every smooth manifold is a topological manifold, but the reverse statement is not always true. Moreover, the availability of parametric equations allows us to relate the manifold to its latent variables, namely its parameters or degrees of freedom.

A smooth manifold $\mathcal{M}$ without boundary is said to be a *submanifold* of another smooth manifold $\mathcal{N}$ if $\mathcal{M} \subset \mathcal{N}$ and the identity map of $\mathcal{M}$ into $\mathcal{N}$ is an embedding. However, it is noteworthy that, while a submanifold $\mathcal{M}$ is just a subset of another manifold $\mathcal{N}$, $\mathcal{M}$ can have a dimension from a geometrical point of view, and the dimension of $\mathcal{M}$ may be lower than the dimension of $\mathcal{N}$. With this idea in mind, and according to [175], a *P-manifold* or *P-dimensional manifold* $\mathcal{M}$ is defined as a submanifold of $\mathcal{N} \subset \mathbb{R}^D$ if the following condition holds for all points $\mathbf{y} \in \mathcal{M}$: there exist two open sets $\mathcal{U}, \mathcal{V} \subset \mathcal{M}$, with $\mathbf{y} \in \mathcal{U}$, and a diffeomorphism $\mathbf{h} : \mathcal{U} \rightarrow \mathcal{V}, \mathbf{y} \mapsto \mathbf{x} = \mathbf{h}(\mathbf{y})$ such that

$$\mathbf{h}(\mathcal{U} \cap \mathcal{M}) = \mathcal{V} \cap (\mathbb{R}^P \times \{\mathbf{0}\}) = \{\mathbf{x} \in \mathcal{V} : x_{P+1} = \cdots = x_D = 0\} \ .$$

As can be seen, $\mathbf{x}$ can trivially be reduced to P-dimensional coordinates. If $\mathcal{N} = \mathbb{R}^D$ in the previous definition, then

- A point $\mathbf{y} \in \mathbb{R}^D$ is a manifold.
- A P-dimensional vector subspace (a P-dimensional hyperplane) is a P-manifold.
- The hollow D-dimensional hypersphere is a $(D-1)$-manifold.
- Any open subset is a D-manifold.

Whitney [202] showed in the 1930s that any P-manifold can be embedded in $\mathbb{R}^{2P+1}$, meaning that $2P+1$ dimensions *at most* are *necessary* to embed a P-manifold. For example, an open line segment is an (open) 1-manifold that can already be embedded in $\mathbb{R}^1$. On the other hand, a circle is a (compact) 1-manifold that can be embedded in $\mathbb{R}^2$ but not in $\mathbb{R}^1$. And a knotted circle, like a trefoil knot, reaches the bound of Whitney's theorem: it can be embedded only in $\mathbb{R}^D$, with $D \geq 2P+1 = 3$.

In the remainder of this book, the word *manifold* used alone typically designates a P-manifold embedded in $\mathbb{R}^D$. In the light of topology, dimensionality reduction amounts to re-embedding a manifold from a high-dimensional space to a lower-dimensional one. In practice, however, a manifold is nothing more than the underlying support of a data distribution, which is known only through a finite sample. This raises two problems. First, dimensionality reduction techniques must work with partial and limited data. Second, assuming the existence of an underlying manifold allows us to take into account the support of the data distribution but not its other properties, such as its density. This may be problematic for latent variable separation, for which a model of the data density is of prime importance.

Finally, the manifold model does not account for the noise that may corrupt data. In that case, data points no longer lie on the manifold: instead fly nearby. Hence, regarding terminology, it is correct to write that dimensionality reduction re-embeds a manifold, but, on the other hand, it can also be said that noisy data points are (nonlinearly) *projected* on the re-embedded manifold.

1.5 Two benchmark manifolds

In order to illustrate the advantages and drawbacks of the various methods of dimensionality reduction to be studied in Chapters 4 and 5, the manifolds shown in Fig. 1.3 will be used repeatedly as running examples. The first manifold, on the left in Fig. 1.3, is called the Swiss roll, according to the name of a Swiss-made cake: it is composed of a layer of airy pastry, which is spread with jam and then rolled up. The manifold shown in the figure represents the thin layer of jam in a slice of Swiss roll. The challenge of the Swiss roll consists of finding a two-dimensional embedding that “unrolls” it, in order to avoid superpositions of the successive turns of the spiral and to obtain a bijective mapping between the initial and final embeddings of the manifold. The Swiss roll is a noncompact, smooth, and connected manifold.

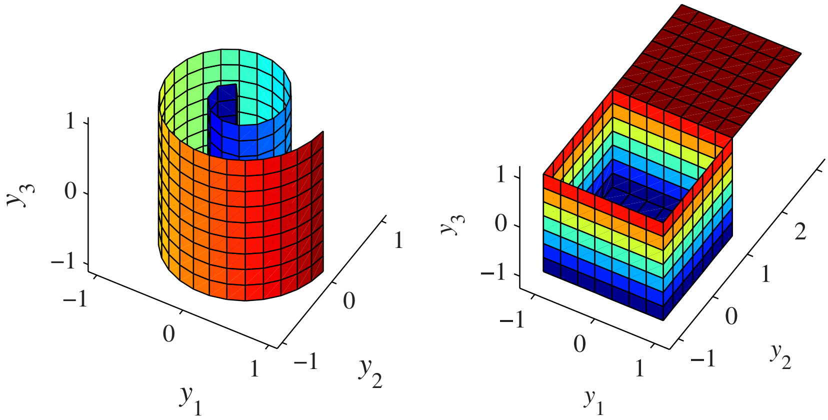

Fig. 1.3. Two benchmark manifold: the ‘Swiss roll’ and the ‘open box’.

The second two-manifold of Fig. 1.3 is naturally called the “open box”. As for the Swiss roll, the goal is to reduce the embedding dimensionality from three to two. As can be seen, the open box is connected but neither compact (in contrast with a cube or closed box) nor smooth (there are sharp edges and corners). Intuitively, it is not so obvious to guess what an embedding of the open box should look like. Would the lateral faces be stretched? Or torn? Or would the bottom face be shrunk? Actually, the open box helps to show the way each particular method behaves.

In practice, all DR methods work with a discrete representation of the manifold to be embedded. In other words, the methods are fed with a finite subset of points drawn from the manifold. In the case of the Swiss roll and open box manifolds, 350 and 316 points are selected, respectively, as shown in Fig. 1.4. The 350 and 316 available points are regularly spaced, in order to be as representative of the manifold as possible. Moreover, points are connected and displayed with different colors (indicating the height in the box or the radius in the Swiss roll). In the case of the box, points also have different shapes (small circles inside the faces, larger squares on the edges). All these features are intended to improve the readability once the manifold is mapped onto a plane, although the three-dimensional representation of Fig. 1.4 looks a bit overloaded.

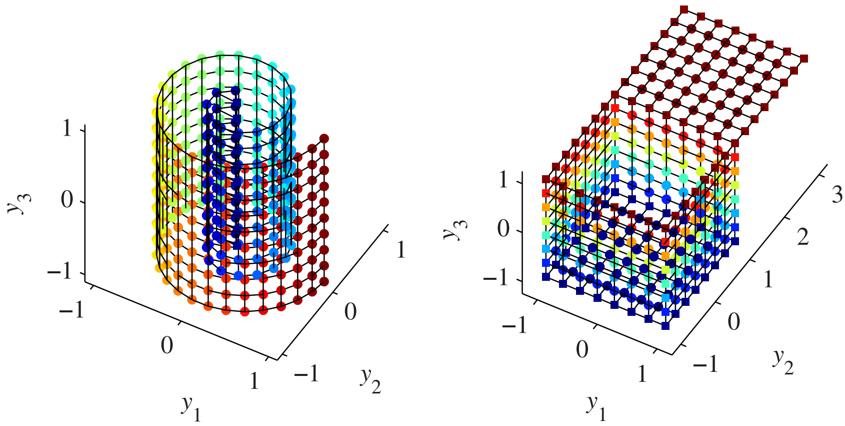

Fig. 1.4. A subset of points drawn from the “Swiss roll” and “open box” manifolds displayed in Fig. 1.3. These points are used as data sets for DR methods in order to assess their particular behavior. Corners and points on the edges of the box are shown with squares, whereas points inside the faces are shown as smaller circles. The color indicates the height of the points in the box or the radius in the Swiss roll. A lattice connects the points in order to highlight their neighborhood relationships.

1.6 Overview of the next chapters

This chapter has quickly reviewed some of the practical and theoretical reasons that raise interest toward methods of analyzing high-dimensional data. Next, Chapter 2 details the most common characteristics of such a method:

- Which functionalities are expected by the user?
- How is the underlying data model defined?
- Which criterion is to be optimized?

In order to illustrate the answers to these questions, Chapter 2 contains a description of principal component analysis (PCA), which is probably the most-known and used method of analyzing high-dimensional data. The chapter ends by listing several properties that allows us to categorize methods of nonlinear dimensionality reduction.

Because numerous DR methods do not integrate an estimator of the intrinsic dimensionality of the data, Chapter 3 describes some usual estimators of the intrinsic dimensionality. A good estimation of the intrinsic dimensionality spares a lot of time when the method takes it as an external hyperparameter. This chapter is necessary for completeness, but the reader familiar with the subject may easily skip it.

The next two chapters are dedicated to the study of two main families of DR techniques. Those techniques can be viewed as replacements, evolutions, or specializations of PCA. On one side, Chapter 4 details methods based on distance preservation. On the other side, Chapter 5 concentrates on the more elegant but more difficult principle of topology preservation. Each of these browses a wide range of classical and more recent methods, and describes them extensively. Next, Chapter 6 gives some examples and compares the results of the various methods.

Finally, Chapter 7 draws the conclusions. It summarizes the main points of the book and outlines a unifying view of the data flow for a typical method of analyzing high-dimensional data. Chapter 7 is followed by several appendices that deal with mathematical or technical details.