1

Pattern Analysis as Data Reduction

Patterns may be found everywhere in nature where there is not total disorder. The essence of what we refer to as a pattern is indeed a reflection of coherence, or organization of information. The tendency for physical systems to *self-organize* provides an opportunity to examine such a process through the resulting patterns. For example, the apparent order observed in financial markets and weather systems provide ample evidence that our ability to understand, manipulate, predict and control patterns is extremely important and potentially rewarding.

1.1 DATA ACQUISITION AND APPLICATIONS

In this section we discuss various problems which naturally lead to the investigation of high-dimensional data sets, or models. In each instance the investigator is confronted with a process which is difficult to understand given the data, or information, associated with the phenomenon is to massive to be readily digested.

1.1.1 Digital Imaging Systems

A digital photograph or a sequence of digital images may provide detailed information about many varied phenomena. A *framegrabber* produces an array of pixel values which correspond to the light intensity reflected off the image. A typical output of such a device is a matrix of integer valued gray levels. One may take any such matrix and concatenate the rows (or columns) to make a vector. Specifically, consider an $M \times N$ array of pixels and let each pixel have an integer value $s_{ij} \in \{0, 1, 2, \ldots, 254, 255]$ as shown in Figure 1.1. It is interesting to make a simple count of the number of possible images which might be generated by such an array. Every pixel has 256 possible values so we have

$$\text{Total number of configurations} = 256^{M \times N}.$$

Naturally this huge number of configurations is capable of depicting a great variety of different images. All human faces, all the trees and all the clouds may be represented (as two-dimensional projections of course) by these pixel matrices.

We may view the values of the grey levels in an image vector as the coefficients of a finite orthogonal expansion with respect to the standard basis. This description is extremely general, as well it should be to represent both trees and faces. Yet, we may be presented with a face recognition task and not be interested in the description of trees at all. In this case the number of possible vectors (faces) is greatly reduced. In fact the total world population, say 10^{10}, is much less than the number of configurations stated above, even for modest values of M and N. Thus we are led to the conclusion that the

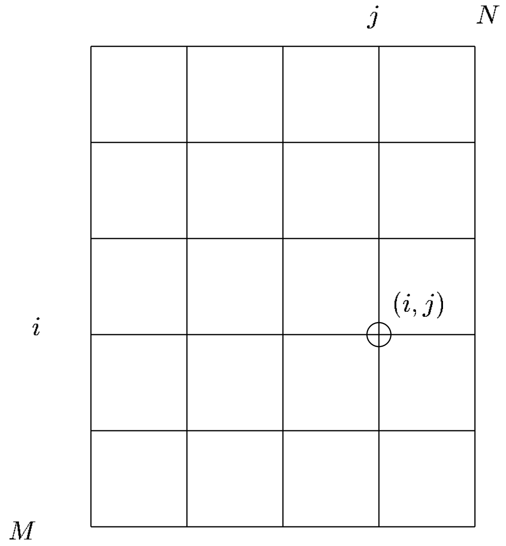

Fig. 1.1 The data array.

standard coordinate system is much too general for a specific type of pattern. See Section 4.1 for a more detailed treatment of this application.

Digital images may represent a family of different patterns or a time-evolving sequence of highly correlated patterns. In the latter case the quantities of data to be managed are extreme.

1.1.2 Experimental Apparatus

Laboratory experiments of physical phenomena may generate massive data sets from which the investigator seeks to develop general principles or theorys. The analysis of large data sets is a time-honored means of applying the scientific method. Kepler developed his laws of planetary motion from examining data associated with orbits.

Wind tunnels are extensively used for simulating air speeds up to Mach 20 and are used for the development of aeropropulsion technology including the space shuttle and the national aerospace plane. Quantities such as air lift, drag and temperature are measured via high-volume disk recording systems. It is not unusual to record 1000 time-dependent variables to monitor the behavior of the object in the tunnel.

The visualization of temporally evolving fluid flows has reached the point where highly resolved 3-dimensional velocity and temperature fields may be captured using particle image velocimetry (PIV).

Fig. 1.2 Snapshot of an experimental data set.

Funtional Magnetic Resonance Imaging (fMRI) is a new technology capable of creating 3-dimensional movies of the mind and is capable of identifying regions of the brain that are involved in performing specific tasks. In a typical fMRI experiment 256^2 images are captured at the rate of 50 images per second.

Financial and economic data sets are routinely exceeding terabyte sizes for ten year periods. It is an enormous challenge for government to mine such data for relevant facts to help construct policy.

Method of Delays Often in experiments it is not possible to collect spatially distributed data sets. Instead, a single probe for instance, may be all that is available for measuring temperature or velocity. The method of delays was proposed by [28] as a technique to reconstruct the attractor of a dynamical system when only limited measurements are available. The generality of this procedure is rather surprising. It forms the basis of the *lagged* vector approach for modeling multi-variable systems where only one observable is available.

For instance, logistic map $x_{n+1} = rx_n(1 - x_n)$ is a deterministic rule for producing a sequence of numbers which appear random, or *chaotic*, as shown in Figure 1.3, However, if we plot the $n + 1$'th iteration x_{n+1} as a function of the previous value x_n, we see the nice parabolic structure revealed. Given a set of apparently random numbers this plotting device, or visualization of the lagged vector (x_n, x_{n+1}) clearly indicates that the dynamics are the result of a quadratic nonlinearity. Note that the mapping from x_n to x_{n+1} is readily approximated numerically while the mapping n to x_{n+1} is far more difficult, if not impossible.

Given a single time series of discrete observations $y(n\delta t)$, $t = n\delta t$ the k-dimensional delay vector is given by

$$\mathbf{y}_n = (y((n - k + 1)\delta t), \ldots, y((n - 1)\delta t), y(n\delta t))$$

Taken's theorem established that the delay (or lag) vector might have to be as large as $k = 2m + 1$ to accurately reconstruct the data on the attractor,

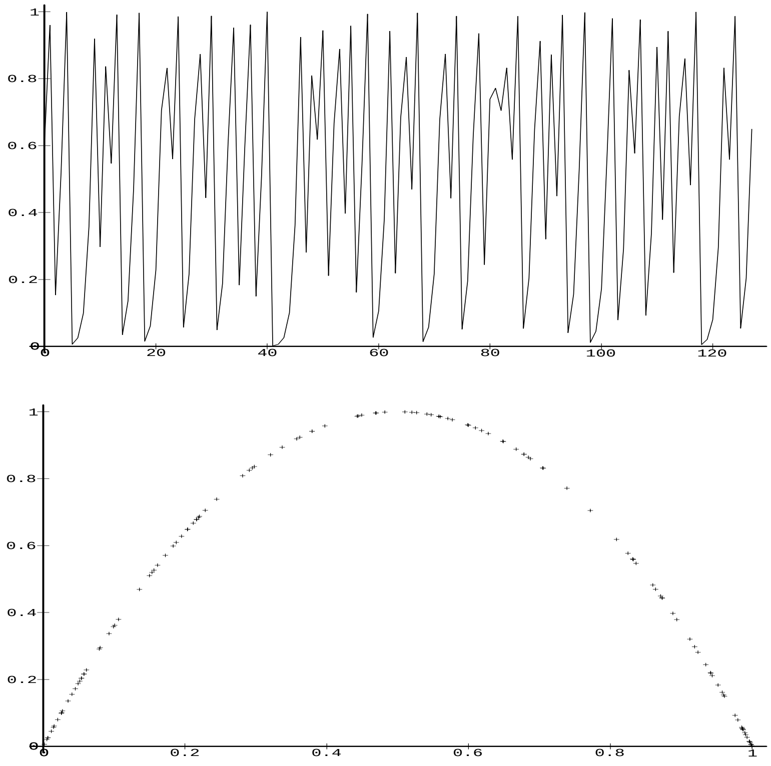

Fig. 1.3 The top graph displays $(t, x(t))$ while the bottom shows the 2-dimensional lag vector $(x(t-1), x(t))$.

where m is the topological dimension of the attractor [35]. It is interesting to view this as a data augmentation process given that the one observed variable is being mapped to many. Certainly this idea fits in naturally to the data reduction and reconstruction procedures to be considered here.

1.1.3 Numerical Simulations

In science and engineering researchers are often confronted with models of physical systems which cannot be solved exactly. Often these models are in the form of systems of nonlinear partial differential equations. Symbolically we may write

$$\frac{\partial u}{\partial t} = D(u)$$

Fig. 1.4 Numerical simulation of the Navier-Stokes equation.

where $D(u)$ is some nonlinear partial differential operator. While these equations may be based on coordinate free conservation laws, determining numerical approximations of their solutions involves the introduction of a coordinate system. A classical example of a nonlinear system of partial differential equations (PDEs) is given by the Navier-Stokes (N-S) equation governing the motion of an incompressible ($\text{div}\mathbf{u} = 0$) fluid

$$\rho(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u}) = -\nabla p + \nu \nabla^2 \mathbf{u}$$

where $\mathbf{u}(\mathbf{x}, t) = (u(\mathbf{x}, t), v(\mathbf{x}, t), w(\mathbf{x}, t))$, i.e., 3 spatio-temporally dependent velocities, $\mathbf{x} = (x, y, z)$, $p(\mathbf{x}, t)$ is the pressure and ν is the dynamic viscosity **??**. When the equation for the internal energy $\theta(\mathbf{x}, t)$ is included the motion of a fluid is governed by 5 scalar variables. These equations have no known closed form solution except for very special subcases.

A standard approach for solving such an equation numerically is to exchange the PDE with a finite system of ODEs by projecting the PDE onto a (truncated set) of complete eigenfunctions [6]. Specifically, $u(x, t)$ is approximated by the truncated expansion

$$u(x, t) \approx u_N(x, t) = \sum_{k=1}^{N} a_k(t)\phi_k(x).$$

Fig. 1.5 The pattern processing paradigm.

Then, by applying the Galerkin procedure, i.e., by requiring

$$(\phi_k, \frac{\partial u_N}{\partial t} - D(u_N)) = 0$$

for $k = 1, \ldots, N$ we produce a finite set of ODEs

$$\frac{da_k}{dt} - F_k(a_1, \ldots, a_N) = 0$$

where again $k = 1, \ldots, N$.

We observe that the introduction of a basis $\{\phi_k\}$ introduces a coordinatization of the coordinate free PDE. The central issue now becomes how do we choose a coordinate system such that the number of equations N needed to reproduce the dynamics of the original PDE is as small as possible? In general it is not uncommon that a numerical simulation of the Navier-Stokes equation based on a Fourier-Galerkin spectral method may require a very large number of terms, e.g., $N = O(10^6)$, for each variable. This may be true even if the dynamics are simple, i.e., periodic. The problem is that the numerical method must compute the full geometry of phase-space every iteration, rather than simply stepping along a closed curve. This is a reflection of a non-optimal coordinate system. This fact suggests that the dimensionality reduction procedures may provide a means to discover the appropriate low-dimensional parametrization of a system of differential equations.

1.2 DIMENSIONALITY REDUCTION

The common feature of the problems in the preceding section is the fact that the data is collected in a high-dimensional space and yet represents patterns which actually may be mapped to a lower dimensional space without loss. We will refer to this initial space as the *ambient space* and the artificially high-dimension as the *ambient dimension.*

Our goal is to find a new representation for the data which reflects its actual, or intrinsic dimension. This new representation is obtained via a *dimensionality reducing transformation* which produces, in effect, a reduced space. Hence, we view, in general terms at least, the analysis of patterns as consisting of the application of a reduction mapping to data for the purposes of extracting salient information, or features, not readily available by a direct approach.

Let's summarize the general features of the approach for pattern analysis that we will pursue in the sequel. Our patterns may be regarded as members of a set $U \subset \mathbb{R}^n$, where the ambient dimension n is large enough that this initial space is too cumbersome, or opaque, for a meaningful direct study of the patterns. Thus we seek a dimensionality reducing mapping $\mathbf{G}$ which takes U to its image, i.e., a set $V \subset \mathbb{R}^m$ of lower dimension m which retains the *essential information* about the patterns. Again, what information is essential may be highly problem dependent.

Symbolically we have

$$U \overset{\mathbf{G}}{\overset{\frown}{\rightarrow}} V \overset{\mathbf{G}^{-1}}{\overset{\frown}{\rightarrow}} U.$$

Let $\mathbf{u} \in U, \mathbf{v} \in V$. Then by the above we mean $\mathbf{G}$ is a mapping from the high-dimensional space U, i.e.,

$$\mathbf{G} : U \to V$$

$$\mathbf{u} \rightsquigarrow \mathbf{v} = \mathbf{G}(\mathbf{u}).$$

The need to be able to reconstruct a pattern to accurately resemble the same or modified pattern in the ambient coordinate system requires that we construct a mapping $\mathbf{G}^{-1}$, i.e.,

$$\mathbf{G}^{-1} : V \to U$$

$$\mathbf{v} \rightsquigarrow \mathbf{u} = \mathbf{G}^{-1}(\mathbf{v}).$$

It is implicit in this discussion that the mapping $\mathbf{G}$ is injective. Thus, the composition of mappings produces the identity as

$$\mathbf{u} = (\mathbf{G}^{-1} \circ \mathbf{G})(\mathbf{u})$$

for all $\mathbf{u} \in U$. Similarly,

$$\mathbf{v} = (\mathbf{G} \circ \mathbf{G}^{-1})(\mathbf{v})$$

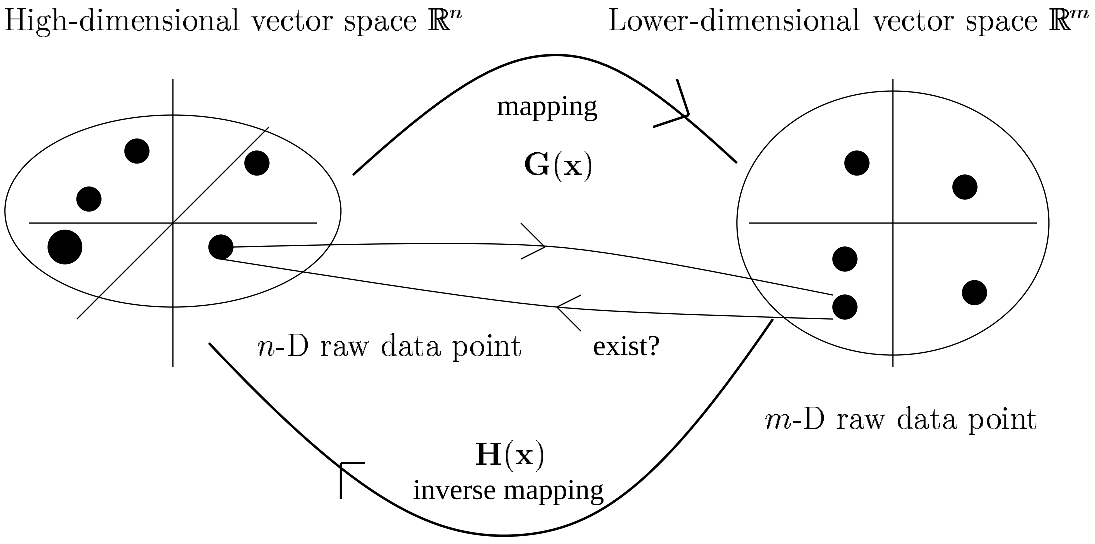

Fig. 1.6 The dimensionality reduction mapping permits the analysis of data in a potentially simpler setting.

This approach for dimensionality reduction is summarized in Figure 1.6.

We interpret that the information has been retained if there is a reconstruction mapping $\mathbf{G}^{-1}$ which behaves as the inverse of $\mathbf{G}$ on V. This interpretation can be made more rigorous as follows [10]:

Theorem 1.1. *Let $U \subset \mathbb{R}^n$. If $\mathbf{G} : U \to \mathbb{R}^m$ is a bi-Lipschitz transformation, i.e.,*

$$c_1\|\mathbf{x}-\mathbf{y}\| \leq \|\mathbf{G}(\mathbf{x})-\mathbf{G}(\mathbf{y})\| \leq c_2\|\mathbf{x}-\mathbf{y}\|$$

where $0 < c_1 \leq c_2 < \infty$, then $\dim U = \dim \mathbf{G}(U)$.

If U is smooth m-dimensional submanifold of $\mathbb{R}^n$ then dim means topological dimension. In general, dim is take to be the Hausdorff dimension, see [10] for details.

Note that this restrictive invertibility condition may not be necessary, as for instance in many classification problems. However, we take the perspective that we seek a new coordinate system to represent the data without loss in some optimal sense. This step may be viewed as a preprocessing step for other pattern analytic applications such as classification.

1.2.1 Intrinsic Dimensionality

To this point we have only alluded to the notion of dimension and have not provided any precise definition. Characterizing the dimensionality of a data set, except at the most basic level, is surprisingly involved and several definitions and measures are available, including basis dimension, topological dimension, fractal dimension, information dimension, and degrees of freedom [10]. There is in fact a whole branch of mathematics dedicated to dimension theory, see [20].

We propose to use the term *intrinsic* dimension to mean the fewest number of parameters requires to model the data without loss. For instance, if

$$\mathbf{u} = \mathbf{H}(v_1, \ldots, v_m)$$

for all $\mathbf{u} \in U \subset \mathbb{R}^n$ we say that the intrinsic dimension of the data is m. This number m may change depending on whether the mapping $\mathbf{H}$ is local or global. In such cases we refer to the global intrinsic dimension or the local intrinsic dimension.

Basis Dimension. This is the standard defintion in elementary linear algebra and indicates the number of vectors required to have a basis for the data in the vector space. Hence, a general digital image with $N \times M$ pixels has dimension $N \times M$ since this is the number of vectors required to form a basis for the space. It might be postulated that a collection of images of faces does not require the generality of this standard basis and we may suppose that a smaller basis exists to represent the data. Therefore we say that the intrinsic dimensionality, or the minimum number degrees of freedom required to characterize the data set, is smaller than the original ambient or measurement dimensionality.

Topological Dimension. This is the basis dimension of the local linear approximation of the hypersurface on which the data resides, i.e., the tangent space. For example, if the data set resides on an m-dimensional submanifold it has an m-dimensional tangent space at every point in the set. For instance, a sphere has a 2-dimensional tangent space at every point and may be viewed as a 2-dimensional manifold.

1.2.2 Empirical Mappings

Given the nature of the data, the size of the ambient space, and the way the data resides in the ambient space are all highly problem dependent it is appropriate to examine *empirical* reduction mappings, i.e., mappings which are data dependent. For example, the Karhunen-Loève expansion, radial basis functions, sigmoidal neural networks and clustering schemes are all data driven. We distinguish these empirical mappings from their *analytical* counterparts such as the Fourier transform, the wavelet transform and many other well-known transformations. Semi-analytical mappings, i.e., mappings which combine analytical and empirical portions are also an emerging area of interest.

The empirical nonlinear functions $\mathbf{G}$ and $\mathbf{G}^{-1}$, may be found by solving the *interpolation problem* [5]. In general, this involves constructing an approximation function $\tilde{\mathbf{f}}(\mathbf{x})$ capable of mapping a collection of input vectors to a set of associated output vectors.

Given an ensemble of P input vectors $\{\mathbf{x}^{(\mu)}\}_{\mu=1}^{P}$, with each $\mathbf{x}^{(\mu)} \in \mathbb{R}^n$, and an associated ensemble of P output vectors, $\{\mathbf{y}^{(\mu)}\}_{\mu=1}^{P}$, with each $\mathbf{y}^{(\mu)} \in \mathbb{R}^m$, find a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ such that the interpolation condition

$$\mathbf{f}(\mathbf{x}^{(\mu)}) = \mathbf{y}^{(\mu)} \tag{1.1}$$

is satisfied for all $\mu = 1 \ldots P$.

In practice we seek an approximation $\tilde{\mathbf{f}}$ to $\mathbf{f}$ which minimizes the interpolating error.

For us it is of central interest to construct approximations $\tilde{\mathbf{G}}, \tilde{\mathbf{G}^{-1}}$ to the desired the reduction and reconstruction mappings $\mathbf{G}$ and $\mathbf{G}^{-1}$ with minimum error and such that the dimension of V is a mimimum. In particular, the reconstruction error

$$\text{error} = \|\mathbf{u} - (\tilde{\mathbf{G}^{-1}} \circ \tilde{\mathbf{G}})(\mathbf{u})\|^2$$

be as small as possible. Typically the problem is further formulated so that an average error is minimized over many pattern realizations, which we write

$$\text{error} = \langle \|\mathbf{u} - (\tilde{\mathbf{G}^{-1}} \circ \tilde{\mathbf{G}})(\mathbf{u})\|^2 \rangle.$$

How we measure the error $\|\cdot\|$ is problem dependent. A related point concerns the required smoothness of the mappings. We will examine instances where the mappings are continuous, Lipschitz and continuously differentiable.

1.2.3 On the Nature of Reduction Mappings

We propose to consider the issue of developing transformations in as general a context as possible. As such, it is useful to consider the totality of possible types of transformation.

- empirical or analytical
- linear or nonlinear
- global or local
- well-conditioned
- piecewise continuous, continuous, Lipschitz, differentiable...

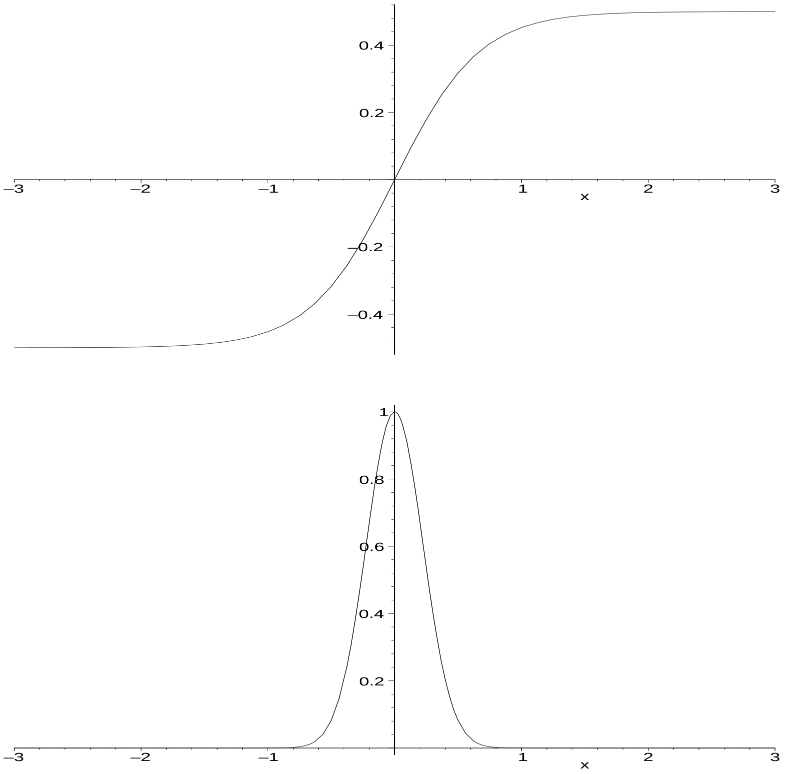

Fig. 1.7 The top graph displays the global sigmoidal function $\sigma(x) = (1 + exp(-3x))^{-1} - 0.5$ while the bottom graph shows the local Gaussian $\phi(x) = \exp(-10x^2)$.

We will be primarily interested in lossless rather than lossy transformations and deterministic rather than stochastic. In addition, we may consider hybrid combinations of transformations from any of the types described below.

Given to sets U and V, a mapping f_α is said to be a *local* mapping if it is defined to act only over subsets domain, i.e, $f_\alpha : S_\alpha \subset U \rightarrow T_\alpha \subset V$. A mapping defined to act over the entire domain, i.e., $f : U \rightarrow V$ is said to be *global* mapping. Global maps include the discrete Fourier transform, the KL transform and the multilayer perceptron. A mapping is said to be *local* if only a subset of the domain contibutes to the image of the map. Gaussian radial basis functions, bump functions and wavelets are examples local maps.

The nature of the reduction mapping appropriate for a given data set is a function of the manner in which the data resides in the ambient space. If the data is best viewed as a subset of a *linear subspace* of smaller dimension than the ambient dimension then one may seek an appropriate change of basis which reveals this structure. On the other hand, if the data is more effectively modeled as a sampling of an m-dimensional sub-manifold $\mathcal{M}$, then it is appropriate to represent the data as a graph from a suitable linear subspace to its orthogonal compliment. The former procedure may be viewed as data *encapsulation* while the latter procedure may be referred to as data *parametrization*.

One can interpret the encapsulation of the data set by a basis as a linear parameterization $\mathbf{u} = \mathbf{H}(\mathbf{v})$, i.e., it is implicit that in this case $\mathbf{H}$ is restricted to be a linear mapping; in practice this means $\mathbf{H}$ is a matrix. As a result, the number of parameters required to encapsulate, or span, a data set is typically significantly larger than the number of dimensions required to nonlinearly parameterize a data set. Therefore, except in special cases, even an optimal linear parameterization (i.e., data encapsulation) will not produce a representation of the data which reflects the intrinsic dimensionality of the data which resides on a submaifold $\mathcal{M}$.

1.2.3.1 $\mathbf{G}$ *global linear,* $\mathbf{G}^{-1}$ *global linear.* Data reduction via the encapsulation procedure seeks to prescribe a new basis which can be used to map, typically via a global linear transformation, e.g. orthogonal or unitary, the data to a subspace of reduced dimension. Under the assumption that the reduced data may be perfectly reconstructed via an inverse linear transformation the new coordinate system may be viewed as encapsulating, or spanning, the entirety of the data in the new basis. In this setting, every data point in lies in the span of the new basis; if a component of a newly observed data point does not lie in this span, it appears as the residual, or error, when the data is reconstructed.

Given a data set which fills a subspace of the ambient space how is a good, or even *optimal*, linear transformation to be found? The Karhunen-Loève transform will provide a prototype procedure for empirically constructing an optimal orthogonal transformation $\mathbf{G}$ and associated inverse $\mathbf{G}^T$. For further discussion of this approach see Chapter 3.

1.2.3.2 $\mathbf{G}$ *global linear,* $\mathbf{G}^{-1}$ *global nonlinear.* The parameterization of a data set may be achieved by determining a new coordinate system which is the result of a dimensionality reducing mapping $\mathbf{v} = \mathbf{G}(\mathbf{u}) \in V \subset \mathbb{R}^d$. The associated reconstruction mapping $\mathbf{v} = \mathbf{G}^{-1}(\mathbf{v}) \in U \subset \mathbb{R}^n$ takes the data and maps it back to the original ambient coordinates. Thus, given $\mathbf{v}$ is a d-tuple, $\mathbf{G}^{-1}$ provides a global d-dimensional parameterization of the data.

The first *architecture* we consider for parametrization based reduction is the case of a global, linear reduction mapping inverted via a global, nonlinear mapping. This particular case is appealing given Whitney's (easy) embedding theorem from differential topology [16] (see also [15]). Paraphrasing, this theorem says that every m dimensional manifold may be diffeomorphically mapped to the Euclidean space of dimension $\mathbb{R}^{2m+1}$. Hence the Euclidean space $\mathbb{R}^{2m+1}$ is large enough to contain a diffeomorphic copy of every m-dimensional manifold [15]. (Whitney's theorem has also been extended to fractal sets [32] removing the requirement that the data set lie on a manifold.)

Whitney's theorem further specifies that the reduced coordinate system may be obtained via a projection $\mathbf{G}$, i.e., a special global linear mapping, Naturally the question arises concerning how to obtain *good* projections in some sense. In practice, we shall approximate the nonlinear inverse $\mathbf{G}^{-1}$ empirically by fitting data, using, for example, radial basis functions or multilayer perceptrons.

We will return to the basic questions concerning this particular reduction approach in Chapter 9 after developing the necessary groundwork in Chapter 7. The methods of Chapter 8 will provide further tools for this problem.

1.2.3.3 $\mathbf{G}$ *global nonlinear,* $\mathbf{G}^{-1}$ *global nonlinear.* When both $\mathbf{G}$ and $\mathbf{G}^{-1}$ are fully nonlinear it may be possible to improve upon the Whitney limit as discussed above where the reduction mapping is linear. For example, consider the special (but important) case of data residing on a closed curve in a high-dimensional space. This curve may always be mapped to a circle in the plane (the sets are *homeomorphic*) so $d = 2$ while the object itself is a one-dimensional manifold. Whitney's theorem suggests that there are some circles which may not be projected to the plane without some loss, i.e., $d = 3$ dimensions may be required. For further discussion of this approach see Section 9.2 in Chapter 9.

Note that the improvement in reduced dimension obtained by using the nonlinear reduction over the linear reduction (with nonlinear reconstruction) will never be greater than a factor of two.

1.2.3.4 $\mathbf{G}$ *local linear,* $\mathbf{G}^{-1}$ *local nonlinear.* For *locally* defined mappings $\mathbf{G}$ and $\mathbf{G}^{-1}$ it is possible to obtain a reduction of the data to the topological dimension m [15]. The nonlinearity of $\mathbf{G}^{-1}$ reduces the number of local regions required to reconstruct $\mathcal{M}$ to any desired accuracy.

Specifically, given that the data lies on a submanifold $\mathcal{M}$ of $\mathbb{R}^n$, we know that it can be represented locally everywhere as the graph of a smooth function. In this situtation there is a locally invertible, linear projection of $\mathcal{M}$ into $\mathbb{R}^m$. In our context, this reduces the data to its intrinsic dimension. Associated with this local linear reduction, there is an inverse mapping, which is generally *nonlinear* [15]. Thus, a local partitioning of the data may be obtained to generate mappings which locally reproduce the identity, while achieving a reduction to the topological dimension m. Hence, in theory, the local approach described above characterizes the data optimally well. For further discussion of this approach see 9.4.

1.2.3.5 $\mathbf{G}$ *local linear,* $\mathbf{G}^{-1}$ *local linear.* This case permits a fast reduction scheme which requires a dense partitioning of the ambient space to approach the quality of reduction possible when a nonlinear inverse is used.

1.2.3.6 Summary. In view of the above remarks we see that the nature of the mappings $\mathbf{G}, \mathbf{G}^{-1}$ plays a central role in the reduction procedure. If the data set resides on an m-dimensional manifold in an ambient space of n-dimensions then the reduction capacity of the various methods can be compared with Whitney's limit $d = 2m + 1$ and the optimal reduction to the m dimensional manifold which is attainable locally. The spectrum of reduction dimensions is then

$$n > n' \geq 2m + 1 \geq m' \geq m$$

where n' is the reduction of a global linear transformation and m' is the reduction dimension of a global nonlinear transformation, in general. More will be made of these comparisons in Chapter 9. In an absolute sense, we view the reduction mapping as optimal if the representation involves only a number of parameters equivalent to the intrinsic dimensionality of the data. Typically this optimal reduction is only possible for local mappings. Of course, given well-defined criteria, we will refer to mappings as optimal over a given class, e.g., orthogonal transformations.

1.3 ON THE NATURE OF PATTERNS IN DATA

In this section we briefly outline the properties of the various types of pattern which are ubiquitous in nature. We also consider the nature of the data which is a measure of the pattern.

1.3.1 Pattern Types

It will be convenient to distinguish between three broad classes of patterns, namely *temporal, spatial* and *spatio-temporal* patterns.

- On a worldwide scale, the temperature over a 24 hour period at Denver International Airport is a temporal pattern $T(t)$.
- The temperature distribution over the entire surface of the earth at any given instant is a spatial pattern $T(\mathbf{x})$.
- The evolution of the earth's surface temperature is a spatio-temporal pattern $T(\mathbf{x}, t)$.

We consider all data which is generated by an equation, or set of equations, to be *deterministic.* In the case of modeling, we assume that there exist explicit mathematical equations governing the system's evolution. We now present basic examples of deterministic data.

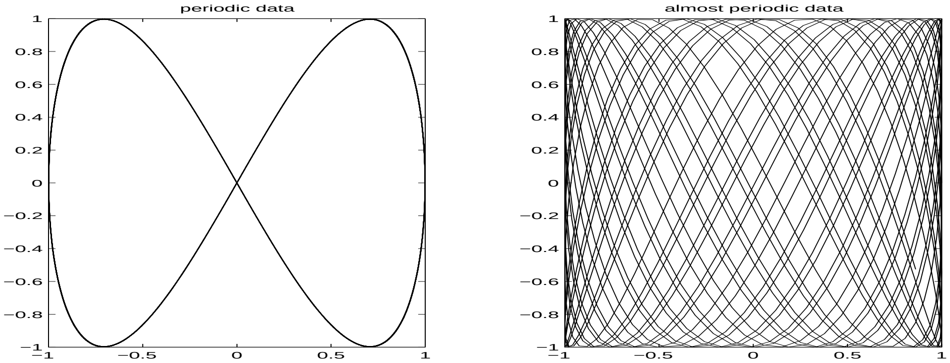

Fig. 1.8 Left: The peridoic curve $(\cos t, \sin 2t)$. Right: the aperiodic curve $(\cos t, \sin\sqrt{5}t)$.

Example 1.1. *Periodic Data.* Data is said to be periodic with period T if the relation

$$x(t) = x(t + nT)$$

holds for $n = 0, \pm 1, \pm 2, \ldots$ and all $t \in \mathbb{R}$. If the independent variable is time then the function is said to be temporally periodic and if the independent variable is space then the function is said to be spatially periodic. A simple example of a periodic function is given by $x(t) = \cos(t) + \sin(2t)$. If the data is time-discrete then the definition of periodicity becomes $x_m = x_{m+nT}$ for all $m, n \in \mathbb{Z}$.

Example 1.2. *Almost Periodic Data.* A function is almost periodic if it is made up of the superposition of 2 or more periodic functions whose frequencies are almost commensurate. If the ratio of the frequencies is rational, then the function falls into the category of periodic function as above. In other words, for an almost periodic function the relationship $x(t) = x(t+nT)$ is not satisfied for any finite value of T, i.e., the period of $x(t)$ is infinitely long. An example of an almost periodic function is provided by

$$x(t) = \cos(2t) + \sin(\sqrt{5}t).$$

See Figure 1.8 to compare this with a periodic function.

Often it may not be an easy question to determine whether a function is periodic or almost periodic. We will return to these issues in our study of Fourier analysis.

Example 1.3. *Non-periodic data.* If a function $x(t)$ is neither periodic nor almost periodic we refer to is as non-periodic. An example of a non-periodic functional relation is given by

$$x(t) = \exp|t|.$$

Example 1.4. *Chaotic data.* The distinction between deterministic and stochastic data becomes somewhat blurred when data produced by an explicit set of mathematical equations *appears* to be random. This behavior of nonlinear systems reflects the mixing properties of deterministic models. Specifically, particles which are initially close together in phase space diverge at an exponentially fast rate.

Example 1.5. *Stochastic data.* Data is said to be stochastic if no mathematical equations exist to model the data production. It is useful to distinguish between what are considered to be truly random data, e.g., the interval between emissions from a radioactive substance, and data which are only modeled as random because an attempt to model them explicitly proves too difficult.

The main dichotomy between deterministic and stochastic data should not be viewed as rigid. Sometimes it is convenient when one is confronted with the study of extremely complex spatio-temporal behavior to model the data as stochastic while fully believing that the underlying processes are deterministic. While strictly speaking chaos is deterministic, it serves to bridge that gap between random and stochastic models. Distinguishing between chaotic and stochastic phenomena is still an active area of research.

1.3.2 Continuous and Discrete Data

Consider the function being measured $v(t)$ to be defined on the domain $t \in T \subset \mathbb{R}$ taking values in $v(t) \in V\mathbb{R}$. We distinguish between the possibilities that the values assumed in T or V may be discrete, i.e. come from a finite set, or continuous, i.e, are drawn from a continuum.

- both t and v are continuous, e.g., the sound of an orchestra.
- t is discrete and v is continuous, e.g., currency exchange rates.
- both t and v are discrete, e.g., digital images.

Data Sampling We note that it is often the case that data to be analyzed is actually discrete-time and/or -space. In some cases, e.g., for computer imaging, that the data is *digital*, i.e., both the dependent and independent variables are discrete.

Discrete sampling also permits the passage from continuous to discrete data sets. For example, if $x(\xi, t)$ is a continuous spatio-temporal pattern then we obtain the discrete-time/space variable

$$x_j^{(k)} = x(j\Delta\xi, k\Delta t)$$

If we further assume that the spatial domain has finite extent, i.e., $j = 1 \ldots N$, then we may view the discretized data as an ensemble of pattern vectors $\{\mathbf{x}^{(k)}\}_{k=1}^{P}$ each lying in a vector space $\mathbb{R}^N$ where $\dim \mathbf{x}^{(k)} = N$ may be large. Often we will compile this data into an $N \times P$ *data matrix* $X = [\mathbf{x}^{(1)} | \cdots | \mathbf{x}^{(k)}]$.

The starting point of the analysis is now a collection of vectors lying in a vector space, possibly of high dimension. An ensemble of spatial patterns may be viewed geometrically as clouds or clusters of points with no temporal line connecting or ordering them. A dynamically evolving pattern on the other hand has a unique time line, or trajectory, connecting the points in space.

Approaches for the analysis of patterns can be divided into two groups, *probabilistic* and *deterministic*. We will be concerned almost exclusively with techniques which fall into the latter category. Our emphasis will be centered on methodologies which all serve as approaches for data reduction and/or function approximation. For statistical approaches to pattern analysis see, e.g., [8, 11, 9].

Problems

1.1 Human beings are amazing information processors. We continuously process phenomonal quantites of data using many naturally developed or physiolgically inherent pattern compression schemes.

(a) List five ways the human brain processes sensory inputs in a fashion which might be interpreted as dimensionality reduction.

(b) Estimate the amount of information (in bytes) a digital television with resolution 640 × 480 produces in one minute assuming a refresh rate of 60 images/second. For a color image assume each pixel is encoded by 3 bytes. This simple calculation will give you an immediate idea of what is meant by massive quanitities of data.

1.2 Numerically integrate the Lorenz equations

$$\dot{x}_1 = a(-x_1 + x_2) \tag{1.2}$$

$$\dot{x}_2 = rx_1 - x_2 - x_1x_3 \tag{1.3}$$

$$\dot{x}_3 = -bx_3 + x_1x_2 \tag{1.4}$$

for $a = 10, r = 28, b = 8/3$. Using the resulting data prepare the following plots:

(a) $x(t), y(t), z(t)$ all as a function of t.

(b) $(x(t), y(t), z(t))$ as a graph in $\mathbb{R}^3$.

(c) $(x(t), x(t - \Delta t))$ as a graph in $\mathbb{R}^2$.

(d) $(x(t), x(t - \Delta t), x(t - 2\Delta t))$ as a graph in $\mathbb{R}^3$.

(e) $(y(t), y(t - \Delta t), y(t - 2\Delta t))$ as a graph in $\mathbb{R}^3$.

(f) $(z(t), z(t - \Delta t), z(t - 2\Delta t))$ as a graph in $\mathbb{R}^3$.

Experiment with various values of Δt. Comment on these graphs in view of the remarks in Section 1.1.2.

1.3 Consider the $m \times n$ real matrix A. By considering the action of A on an element x of $\mathbb{R}^n$, argue whether it is an example of a global or local mappping.

1.4 Describe how local mappings may be used to construct a global mapping.

1.5 Consider the action of an $m \times n$ matrix A.

1.5.1. Under what conditions is this mapping injective, surjective and bijective? (Specify the domain and range of A.)

1.5.2. Provide an example of a specific $m \neq n$ matrix A and subspace U such that A is bijective. (Hint: think geometrically).

1.6 Consider a torus residing in a 3-dimensional ambient space. What is the

(a) basis dimension

(b) local dimension

(c) intrinsic dimension

1.7 Let $U \subset \mathbb{R}^n$ and $\mathbf{G} : U \to \mathbb{R}^m$. Given only

$$c_1 \|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{y})\|$$

or

$$\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{y})\| \leq c_2 \|\mathbf{x} - \mathbf{y}\|$$

where $0 < c_1 \leq c_2 < \infty$, i.e., the function *is not* bi-Lipschitz, show examples of mappings which do not preserve the dimension of U.

2

Mathematical Preliminaries