Neural Computing and Applications (2022) 34:8885–8896
https://doi.org/10.1007/s00521-022-06909-y

ORIGINAL ARTICLE

Locally fitting hyperplanes to high-dimensional data

M. Hou[1] · C. Kambhampati[2]

Received: 17 April 2021 / Accepted: 4 January 2022 / Published online: 27 April 2022

Abstract

Problems such as data compression, pattern recognition and artificial intelligence often deal with a large data sample as observations of an unknown object. An effective method is proposed to fit hyperplanes to data points in each hypercubic subregion of the original data sample. Corresponding to a set of affine linear manifolds, the locally fitted hyperplanes optimally approximate the object in the sense of least squares of their perpendicular distances to the sample points. Its effectiveness and versatility are illustrated through approximation of nonlinear manifolds Möbius strip and Swiss roll, handwritten digit recognition, dimensionality reduction in a cosmological application, inter/extrapolation for a social and economic data set, and prediction of recidivism of criminal defendants. Based on two essential concepts of hyperplane fitting and spatial data segmentation, this general method for unsupervised learning is rigorously derived. The proposed method requires no assumptions on the underlying object and its data sample. Also, it has only two parameters, namely the size of segmenting hypercubes and the number of fitted hyperplanes for user to choose. These make the proposed method considerably accessible when applied to solving various problems in real applications.

1 Introduction

In many applications of scientific disciplines, it is often important to discover interrelationships among a set of variables and the patterns in them. These variables represent an unknown object, and their measurements are considered a data sample of the object. A data-driven learning mechanism is an automated process of extracting information on interrelationships of these variables from their measurements.

In an algebro-geometric framework, the data-driven unsupervised learning problem is investigated in this work. The problem is solved by fitting hyperplane patches to observed points of an object in each exclusive subset of the data sample. These exclusive subsets are obtained by subdividing the data range into hypercubic subregions of equal size and associating data points with the hypercubes. These fitted hyperplane patches intersect forming an affine linear manifold. The set of all linear manifolds form an approximation of the object. The approximation is optimal in the sense of least squares of the perpendicular distances of the data points to the fitted affine linear manifolds.

As an approximate model of the underlying unknown object, analytic expressions of the affine linear manifolds can be used for pattern recognition, dimensionality reduction, and component predictions of high-dimensional data points through inter- or extrapolation.

2 Related work

Linear approximation of an unknown nonlinear relationship is a general problem which has been investigated intensively from various perspectives. Directly used for classification in [1], one hyperplane is fitted to the data points in one class, and simultaneously furthest from the remaining data points. For dimensionality reduction, the local linear embedding [2–4] and its variation [5] firstly fit a set of weights to each subset of high-dimensional data points to reproduce one point within the subset. In a similar manner, a low-dimensional dataset is then obtained by fitting its data points to the determined weights. The philosophy of carrying certain geometric features to a low-dimensional space led to another intuitive method called Isomap [6]. It first determines a matrix containing the

✉ M. Hou
m.hou@hull.ac.uk

C. Kambhampati
C.Kambhampati@hull.ac.uk

[1] Department of Engineering, University of Hull, Hull, UK

[2] Department of Computer Science and Technology, University of Hull, Hull, UK

shortest path distance between each pair of data points in a neighbourhood of a given data point, and then, by using the classical multidimensional scaling, low-dimensional points best retaining the distances are obtained. A low-dimensional sequence produced in [2–6] reveals some intrinsic geometric properties of the unknown object, but does not indicate the extent to which it approximates the underlying object.

A largely overlooked study [7] solved the problem of fitting a linear manifold to data points in a high-dimensional space by the Eckart-Young-Mirsky theorem [8]. This also forms the basis of the method of local tangent spaces in [9] for learning nonlinear manifolds. By assuming sufficient smoothness of the underlying manifold, a tangent subspace is determined with respect to each data point and its neighbours. The subspaces associated with all data points are used to approximate the unknown nonlinear manifold. Nevertheless, the approximation uses more floating-point numbers than those needed for representing the original high-dimensional data.

Specifically addressed for the problem of linear manifold approximation of a nonlinear manifold, a sophisticated method was developed in [10] for determination of clusters of data points based on local tangent spaces. A local linear manifold called a flat was then fitted to data points in each cluster. The calculations of tangent spaces and flats were based on the Eckart–Young–Mirsky theorem. In the current study, the underlying object is not assumed to be a nonlinear manifold, and the natural hypercubic segmentation of data points is non-iterative and effective. As shown in the current work, hyperplane fitting is conceptually simple with an elementary derivation immediately leading to the linear manifold fitting [7] and the classical principal component analysis [11] through the singular value decomposition of a matrix formed by data points.

The polynomial approach [12] to the problem of subspace segmentation is a different way of grouping data points drawn from an unknown object. The subspace segmentation method is most effective when the data points of an object approximately belong to linear subspaces of an unknown number and unknown dimensions because once the segmentation is completed, each subspace can be well represented by a linear manifold. In comparison, grouping of data points in the current work is a method of spatial segmentation. When the size of the segmenting hypercubes is sufficiently small, linear subspaces can be discovered by fitting hyperplanes to data points in hypercubes.

A different and promising line of research is use of a multiple-layer neural network such as the autoencoder proposed in [13] to achieve dimensionality reduction and approximate recovery of original data. Along with determination of architecture of a neural network, initialisation and optimisation of the weights and biases associated with its neurons are normally a challenging task in applications.

The novelty of this work lies in the development of a general method for unsupervised learning of an unknown object on which no particular assumptions are made. As illustrated by several examples, the method is readily applicable to a range of non-linear approximation problems. An advantage of using least squares of perpendicular distances rather than other distance measures [14] between the hyperplanes and given points is that hyperplane fitting can be solved by the numerically reliable singular value decomposition of a matrix formed by data points in each hypercube.

3 Method

3.1 General description

Let $\mathbf{x}$ be a general point belonging to an unknown object in a subspace $\boldsymbol{X}$ of $\boldsymbol{R}^n$, the nD Euclidean real space. Given m sample points $\{\mathbf{x}_i\}$ of $\mathbf{x}$, the existence of dependencies among the components of $\mathbf{x}$ implies the existence of a kernel represented by a surjective mapping h: $\boldsymbol{X} \mapsto 0 \in \boldsymbol{R}^l$ with $0 < l < n$, and hence of an embedding specified by an injective mapping f: $\boldsymbol{Y} \subset \boldsymbol{R}^{n-l} \mapsto \boldsymbol{X}$. This kind of embedding could simply be a monomorphism with the properties such as continuity and structure-preserving, not necessarily a well-behavioured topological manifold. In many applications, maps h and f could conveniently be viewed as multivariable functions $h(\mathbf{x}) = 0$ and $\mathbf{x} = f(\mathbf{y})$ with $\mathbf{x} \in \boldsymbol{X}$ and $\mathbf{y} \in \boldsymbol{Y}$, and only $\{\mathbf{x}_i\}$ is known. The objective is to locally fit hyperplanes to $\{\mathbf{x}_i\}$ so that the unknown object is approximated by the local affine linear manifolds defined by these fitted hyperplanes. The approximation is optimal in the sense of least squares of perpendicular distances of the data points to the fitted hyperplanes.

Denote centroid $\bar{\mathbf{x}}$ of $\{\mathbf{x}_i\}$ and the associated data matrix $\bar{\mathbf{X}}$ by

$$\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i, \qquad \bar{\mathbf{X}} = [\mathbf{x}_1 - \bar{\mathbf{x}}, \quad \mathbf{x}_2 - \bar{\mathbf{x}}, \quad \ldots, \quad \mathbf{x}_m - \bar{\mathbf{x}}]', \tag{1}$$

where $'$ stands for the transpose. A hyperplane denoted by $(\mathbf{p}, d)$ is an $(n-1)$D linear subspace in $\boldsymbol{R}^n$, with unit vector $\mathbf{p}$ being its normal and scalar d its signed perpendicular distance to the origin of the reference frame. The signed perpendicular distance from point $\mathbf{x}$ to $(\mathbf{p}, d)$ is $\mathbf{x}'\mathbf{p} - d$.

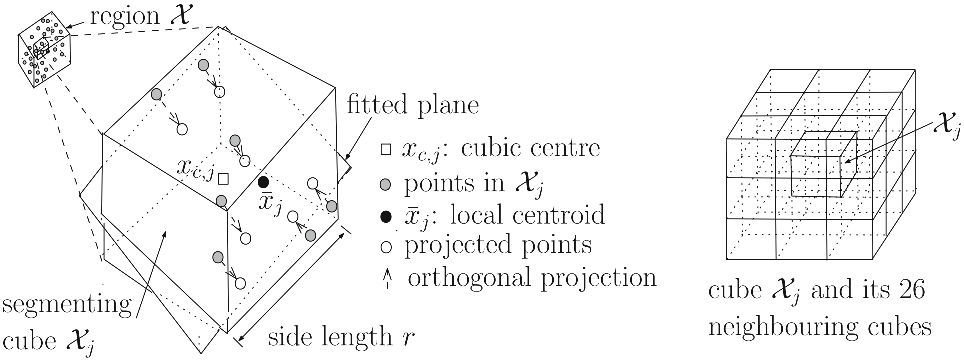

region $\mathcal{X}$
fitted plane
$x_{c,j}$
$\bar{x}_j$
□ $x_{c,j}$: cubic centre
● points in $\mathcal{X}_j$
● $\bar{x}_j$: local centroid
○ projected points
↑ orthogonal projection
segmenting cube $\mathcal{X}_j$
side length r
$\mathcal{X}_j$
cube $\mathcal{X}_j$ and its 26 neighbouring cubes

Fig. 1 An illustration of plane fitting with a cubic neighbourhood

3.2 Regional segmentation

Consider an nD hypercube bounding $\boldsymbol{X}$ and segment it with a sufficient number of equally sized and identically oriented small nD hypercubes called n-cubes. Let the length of n-cubes be $r = 2b/c$, where positive integer c is specified by the data analyst, and b is the maximum of the absolute differences between the corresponding coordinates of each point and centroid $\bar{\mathbf{x}}$.

Figure 1 illustrates a cubic neighbourhood and fitted plane in 3D space, where $\boldsymbol{X}_j$ is the subregion bounded by the jth cube. Two n-cubes are said to be neighbours to each other if they share at least one vertex. The total number of the n-cubes within the bounding nD hypercube is c^n which is prohibitively huge even for moderate integers c and n. Moreover, an n-cube has $3^n - 1$ neighbours, which is still a huge number for high-dimensional data.

To avoid searching n-cubes most of which contain no data points anyway, an effective algorithm should directly assign each point to the n-cube to which it belongs. Consider an n-cube with a vertex at $\underline{\mathbf{x}} = \bar{\mathbf{x}} - b\mathbf{e}$, where $\mathbf{e}$ is the vector of n ones. Clearly, among all c^n n-cubes and 2^n vertices of each of them, this n-cube and vertex are the furthest in the opposite directions of the reference-frame axes. They are referred to as the first n-cube and its first vertex, respectively.

Let $\mathbf{k}_c$ consist of n top-down written digits representing the base-c number of an integer counting from 0 to $c^n - 1$, and k_c be its base-10 correspondence. For example, if $n = 3$, $\mathbf{0}_2 = \mathbf{0}_3 = [0, \ 0, \ 0]'$, $\mathbf{1}_2 = \mathbf{1}_3 = [1, \ 0, \ 0]'$, $\mathbf{2}_2 = [0, \ 1, \ 0]'$, $\mathbf{2}_3 = [2, \ 0, \ 0]'$, $\mathbf{7}_2 = [1, \ 1, \ 1]'$, $\mathbf{26}_3 = [2, \ 2, \ 2]'$, and so on. The first vertex of the k_cth n-cube is at $\underline{\mathbf{x}} + r\mathbf{k}_c$. To find the n-cube containing $\mathbf{x}_i$, the rounding-down function $\lfloor x \rfloor$ of scalar x, and its extension $\lfloor \mathbf{x} \rfloor$ as the component-wise operator on vector $\mathbf{x}$ are used.

Algorithm: Point-to-hypercube assignment

```
Step 0: Calculate edge length $r$ of $n$-cubes, and the first vertex $\underline{\mathbf{x}}$ of the first
$n$-cube. In the calculations, the maximal absolute component difference $b$
is replaced by $(1+\varepsilon)b$ with a small positive number, say $\varepsilon = 10^{-10}$.
Step 1: Calculate base-$c$ number $\mathbf{k}_c = \lfloor (\mathbf{x}_i - \underline{\mathbf{x}})/r) \rfloor$, find its base-10 correspondence $k_c$, and assign point $\mathbf{x}_i$ to the $k_c$th $n$-cube.
Step 2: Repeat step 1 for all points $\{\mathbf{x}_i\}$.
Step 3: Arrange the identified non-empty $n$-cubes in the increasing order
of their $k_c$ numbers, recount them with a new integer varying from 1 to $k$,
and terminate the algorithm.
```

At step 0 of the algorithm, $(1 + \varepsilon)b$ replacing b avoids overflow of components of k_c. Without the replacement, at least one component of k_c will reach c when x_i is on the boundary of $\boldsymbol{X}$. This minor adjustment amounts to filling c^n n-cubes into an nD hypercube which is slightly bigger than that bounding$\boldsymbol{X}$.

The point-to-hypercube algorithm divides the dataset into k subsets within the identified k n-cubes. It naturally covers the case where data points scatter in disjoint subregions of $\boldsymbol{X}$. To the points within each segmenting n-cube, l hyperplane patches (hyperplanes bounded by this n-cube) can be fitted in order to obtain a local approximation of the unknown object. If subregion $\boldsymbol{X}_j$ bounding the jth n-cube contains m_j points, it is clear $m = \sum_{j=1}^{k} m_j$. The algorithm ensures that, for each $\mathbf{x}_i \in \boldsymbol{X}$, there is a unique j so that $\mathbf{x}_i \in \boldsymbol{X}_j$.

3.3 Fitting hyperplanes

For clarity of exposition, in this and next two sections, all data points are assumed to be in one n-cube, corresponding to $c = 1$ in the point-to-hypercube algorithm. The only difference between this and general case is that in the latter an n-cube normally contains a smaller number of data points.

With $1 \leq l \leq n$, fitting l hyperplanes $\{(\mathbf{p}_j, d_j)\}$ to m data points $\{\mathbf{x}_i\}$ is a constrained minimisation problem:

$$J_l = \min \sum_{j=1}^{l} \sum_{i=1}^{m} (\mathbf{x}_i' \mathbf{p}_j - d_j)^2 \qquad \text{s.t.} \quad \mathbf{p}_j' \mathbf{p}_j = 1, \quad j = 1, 2, \ldots, l. \tag{2}$$

As verified in Appendix A, (2) has the solution

$$(\mathbf{p}_j, d_j) = (\mathbf{v}_{n+1-j}, \bar{\mathbf{x}}' \mathbf{v}_{n+1-j}), \quad j = 1, \ldots, l, \tag{3}$$

where the normals are the last l vectors of $\mathbf{V}$, coming from the singular value decomposition $\bar{\mathbf{X}} = \mathbf{U}\Sigma_x \mathbf{V}'$ with $\Sigma_x =$ diag $(\sigma_1, \sigma_2, \ldots, \sigma_n)$, $\sigma_n \geq 0$, and $\sigma_i \geq \sigma_j$ for $i < j$. With the fitted l hyperplanes, the least squares criterion in (2) reaches the minimum $J_l = \sum_{j=1}^{l} \sigma_{n+1-j}^2$.

The hyperplanes in (3) are mutually perpendicular, and they all pass centroid $\bar{\mathbf{x}}$. A point simultaneously on several hyperplanes means that it is on their intersection which is also a hyperplane. Cases $l = n$ and $l = n - 1$ correspond to fitting a point and a straight line to $\{\mathbf{x}_i\}$, respectively. Figure 2a and b illustrate fitting a plane to evenly sampled 3D cylindrical surface and Möbius strip respectively, and (c) fitting (the intersection of) two mutually perpendicular planes to a randomly sampled 3D straight line with additive noise.

3.4 Affine linear manifold

As verified in Appendix B, the intersection of the hyperplanes in (3) can be described by its parametric equation or parametric inverse given, respectively, by

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{V}_{n-l}\mathbf{y}, \qquad \mathbf{y} = \mathbf{V}_{n-l}'(\mathbf{x} - \bar{\mathbf{x}}), \tag{4}$$

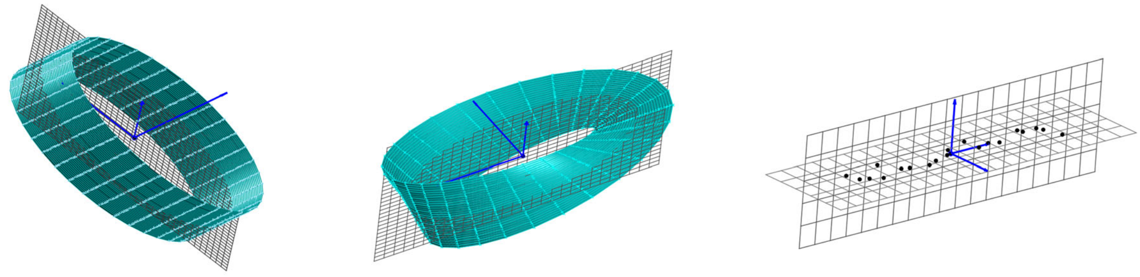
where $\mathbf{x} \in \boldsymbol{R}^n$ represents an arbitrary point on the intersection, $\mathbf{y} \in \boldsymbol{R}^{n-l}$ is its low-dimensional correspondence, and $\mathbf{V}_{n-l}$ consists of the first $n - l$ vectors of $\mathbf{V}$. The parametric equation also defines an affine linear manifold which is an $(n - l)$D subspace complementing the lD subspace spanned by the hyperplane normals. An arbitrary point belonging to the manifold is represented by $\mathbf{x}$ and $\mathbf{y}$ as its coordinates, respectively, in the global frame, and in the frame originated at $\bar{\mathbf{x}}$ with the vectors in $\mathbf{V}_{n-1}$ as its axes.

Orthogonal projections $\{\hat{\mathbf{x}}_i\}$ of $\{\mathbf{x}_i\}$ onto the manifold and low-dimensional sequence $\{\mathbf{y}_i\}$ are readily obtained from (4) as, for $i = 1, 2, \ldots, m$,

$$\hat{\mathbf{x}}_i = \bar{\mathbf{x}} + \mathbf{V}_{n-l}\mathbf{y}_i, \qquad \mathbf{y}_i = \mathbf{V}_{n-l}'(\mathbf{x}_i - \bar{\mathbf{x}}). \tag{5}$$

The difference between $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ indicates closeness of the affine linear manifold to $\mathbf{x}_i$, while $\mathbf{y}_i$ is the coordinates of $\hat{\mathbf{x}}_i$ in the $(n - l)$D subspace. The sum of the squared distances between $\{\mathbf{x}_i\}$ and $\{\hat{\mathbf{x}}_i\}$ equals the sum of the squared l smallest singular values of $\bar{\mathbf{X}}$, namely $\sum_{i=1}^{m} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 = \sum_{j=1}^{l} \sigma_{n+1-j}^2$.

Being originally addressed to fit a plane or a line to the observations of the high-dimensional vector [15], the principal component analysis is normally formulated as to determine a low-dimensional vector whose components are



(a) cylindrical surface (b) Möbius strip (c) sampled straight line

Fig. 2 Plane fitting to different objects in 3D space. In each of cases (**a**) and (**b**), a plane is fitted to 30 × 30 regularly sampled and noise-free points shown as the light strips on object's surface. In case (**c**), a line (intersection of two mutually perpendicular planes) is fitted to randomly sampled noisy points shown as black dots of a 3D straight line

linear combinations of the components of a high-dimensional vector [11], or vice versa [16], by maximising the resultant variance. The high- and low-dimensional vectors are now shown to be related to each other in (4) by the hyperplane intersection.

3.5 Determination of dimensionality

The dimensionality of an object is the number of independent components in its general point $\mathbf{x}$. Within the region of an n-cube, it can be approximated by $n - l$, where l is the largest number of hyperplanes best fitting to the data points in the n-cube. Denote the orthogonal projection of $\mathbf{x}_i$ onto (the intersection of) the l hyperplanes by $\hat{\mathbf{x}}_i^{(l)}$ obtained from (5), and recall $J_l = \sum_{i=1}^{m} \|\mathbf{x}_i - \hat{\mathbf{x}}_i^{(l)}\|^2 = \sum_{j=1}^{l} \sigma_{n+1-j}^2$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ being the singular values of data matrix $\bar{\mathbf{X}}$. It is evident that if the embedded object is of $(n-l)$ dimensions, then $\varepsilon \approx \sigma_n \approx \sigma_{n-1} \cdots \approx \sigma_{n-l+1} < < \sigma_{n-l}$ is expected, where ε is a small number whose value is dependent on the noise level of the sample points $\{\mathbf{x}_i\}$. If there is no embedding in X, then $\varepsilon < < \sigma_n$ is expected.

3.6 Efficiency of dimensionality reduction

The ideal efficiency of dimensionality reduction $(n-l)/n$ cannot be achieved. To express m nD sample points, total mn floating-point numbers are needed. The floating-point numbers used by the k groups of $(n-l)$D affine linear manifolds are the sum of $2nk$ from the centroids and n-cube centres, $(n-l)nk$ from basis vectors and $(n-l)m$ from subspace points. Hence, the reduction in usage of floating-point numbers in the approximation is given by $c_p = l/n - 2(n-l+1)k/m$ resulted from subtracting the usage from and diving the outcome by the total, and clearly $0 \leq c_p \leq 1$.

3.7 Component prediction by interor extrapolation

The information about the fitted l hyperplanes can be used to determine l components of a general nD point $\mathbf{x}$ of the underlying object as linear combinations of its remaining $n-l$ components. Let $\mathbf{P}_l$ and $\mathbf{P}_{n-l}$ be two parts of a permutation matrix $\mathbf{P}$, partitioning $\mathbf{x} - \bar{\mathbf{x}}$ into l and $n-l$ components as $\mathbf{x}_{1,l} - \bar{\mathbf{x}}_{1,l} = \mathbf{P}_l(\mathbf{x} - \bar{\mathbf{x}})$ and $\mathbf{x}_{l+1,n} - \bar{\mathbf{x}}_{l+1,n} = \mathbf{P}_{n-l}(\mathbf{x} - \bar{\mathbf{x}})$. Refer to the derivation of (4) and note, if $\mathbf{x}$ is on the intersection of the fitted hyperplanes, $\mathbf{V}_l'(\mathbf{x} - \bar{\mathbf{x}}) = 0$ with $\mathbf{V}_l$ consisting of the last l vectors of $\mathbf{V}$. Substituting $\mathbf{x} - \bar{\mathbf{x}} = \mathbf{P}_l'(\mathbf{x}_{1,l} - \bar{\mathbf{x}}_{1,l}) + \mathbf{P}_{n-l}'(\mathbf{x}_{l+1,n} - \bar{\mathbf{x}}_{l+1,n})$ into $\mathbf{V}_l'(\mathbf{x} - \bar{\mathbf{x}}) = 0$, the least squares solution of $\mathbf{x}_{1,l} - \bar{\mathbf{x}}_{1,l}$ yields

$$\mathbf{x}_{1,l} = \bar{\mathbf{x}}_{1,l} - (\mathbf{V}_l'\mathbf{P}_l')^{+}\mathbf{V}_l'\mathbf{P}_{n-l}'(\mathbf{x}_{l+1,n} - \bar{\mathbf{x}}_{l+1,n})\,, \tag{6}$$

where $(\mathbf{V}_l'\mathbf{P}_l')^{+}$ is the Moore–Penrose inverse of $\mathbf{V}_l'\mathbf{P}_l'$.

When $\mathbf{x}_{l+1,n}$ belongs to a hypercubic $(n-l)$D subregion of any non-empty segmenting n-cube, finding $\mathbf{x}_{1,l}$ is called interpolation, or extrapolation otherwise. The produced $\mathbf{x}_{1,l}$ is unreliable if the condition number of $\mathbf{P}_l\mathbf{V}_{j,l}$ is considerably greater than 1 or if $\mathbf{x}_{l+1,n}$ is faraway from any non-empty n-cube. If $\mathbf{x}_{l+1,n}$ belongs to more than one non-empty segmenting n-cube, $\mathbf{x}_{1,l}$ has multiple solutions.

In some applications, the partition of $\mathbf{x}$ is not determined a priori. In such a case, determination of $\mathbf{P}_l$ and $\mathbf{P}_{n-l}$ could be firstly based on the averaged relative errors $\{e_s\}$ defined by

$$e_s = \frac{1}{m_j}\sum_{i=1}^{m_j} \frac{|x_{si} - \hat{x}_{si}|}{|x_{si}| + |\hat{x}_{si}|}, \quad s = 1, 2, \ldots, n \tag{7}$$

where x_{si} and $\hat{x}_{si}$ are sth components of $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ in (5), respectively, and m_j is the number of data points in the jth n-cube. Let $\{e_{j_1}, e_{j_2}, \ldots, e_{j_l}\}$ be the l smallest elements in $\{e_j\}$. The rows of $\mathbf{P}_l$ are the $j_1, j_2, \ldots, j_l$ rows of the identity matrix of dimension n, while $\mathbf{P}_{n-l}$ contains the remaining $n-l$ rows.

4 Result

To show the capability of the method for linear approximation of a nonlinear object, two examples of a 2D manifold embedded in a 3D space are used. The next example is for the handwritten digit recognition using a MNIST dataset. A further example illustrates dimensionality reduction in a cosmological application. The problem of inter- and extrapolation is considered with an application in a social and economic data set. The final example deals with recidivism prediction in a judicial application.

4.1 Möbius strip and Swiss roll

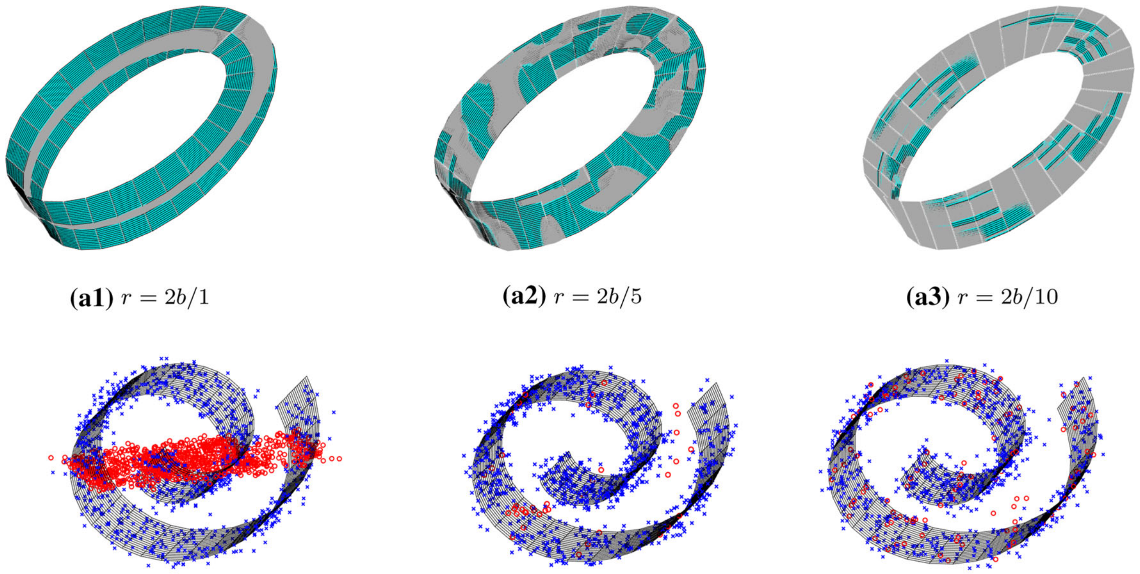
These are two simple academic examples with the parametric equation for Möbius strip as

$$p_x = (2 + v\cos u)\cos 2u, \quad p_y = (2 + v\cos u)\sin 2u, \quad p_z = v\sin u \tag{8}$$

with $u \in [0,\, \pi]$ and $v \in [-0.4,\, 0.4]$, and for Swiss roll as

$$p_x = u\cos u, \quad p_y = v \quad p_z = u\sin u \tag{9}$$

with $u \in [\pi,\, 4\pi]$ and $v \in [0,\, 20]$.



(a1) $r = 2b/1$ **(a2)** $r = 2b/5$ **(a3)** $r = 2b/10$

(a1) $r = 2b/1$ **(a2)** $r = 2b/3$ **(a3)** $r = 2b/5$

Fig. 3 Möbius strip (**a1**)–(**a3**): Approximation by fitting one hyperplane to each subset of the given points, where r is the side length of segmenting cubes and b is the maximum of the absolute differences between the corresponding coordinates of each point and the centroid of all data points. Cases (**a1**), (**a2**) and (**a3**) correspond, respectively, to 1, 29 and 86 subsets of total 30 × 30 regularly sampled and noise-free points. The original Möbius strip is in cyan with the sample points (light stripes) and the fitted planes (grey surface patches). Swiss roll (**b1**)–(**b4**): The original Swiss roll (black grids) is shown with the orthogonal projections (red circles) of the 50 × 20 irregularly sampled and noisy points (blue crosses) on to the fitted planes. The added noise is normally distributed with zero mean and standard deviation of 5% of the maximum of each component of the points. Cases (**b1**), (**b2**) and (**b3**) correspond, respectively, to 1, 27 and 109 fitted planes with the relative norm errors between the sampling points and their orthogonal projections 19%, 13% and 7%, and the reductions in usage of floating-point numbers 33%, 17% and -32%. The ratios of the averaged absolute distances between the original Swiss roll to the sampled points and to the projected points are, respectively, 16.3/19.6, 11.8/4.9 and 12.0/4.6.

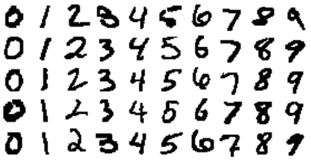


Fig. 4 Sample of five sets of handwritten digits

Table 1 Numbers of incorrect recognitions in 10 × 50 handwritten digits

Digit	0	1	2	3	4	5	6	7	8	9
Failure count	0	0	3	2	0	0	0	0	1	1

The manifold expressions (8) and (9) have been used to evaluate quality of local linear manifold approximations for the original nonlinear manifolds with clean or noisy data as shown in Fig. 3. However, in real applications such as in the following examples, the evaluation cannot be carried out because the underlying object is unknown.

4.2 Handwritten digit recognition

The MNIST dataset consists of 10,000 handwritten digits [17]. For each of numbers 0 to 9, using the first 800 handwritten digits in hyperplane fitting and the next 50 in testing, the current study has produced an error rate of 1.4% in the automated recognition. The dimension of data is 784 (28 × 28) plus 1 (index indicating 0 to 9). Figure 4 illustrates five sets of handwritten digits, while Table 1 shows numbers of incorrect recognitions in each of the ten test sets (10 × 50 handwritten digits). In this application, the segment division constant is $c = 1$ in the point-to-hypercube algorithm, and the dimensionality of the fitted linear manifolds is $n - l = 30$.

4.3 Quasar specification

As cosmological objects of great significance in study of the universe's origin [18, 19], quasars are active galactic nucleus of very high luminosity, the first of which was discovered in 1960's, and about 450,000 of them catalogued in the 14th release of the Sloan Digital Sky Survey (SDSS) in 2017. A brief introduction to the SDSS projects and broad accessibility of cosmological data was given in

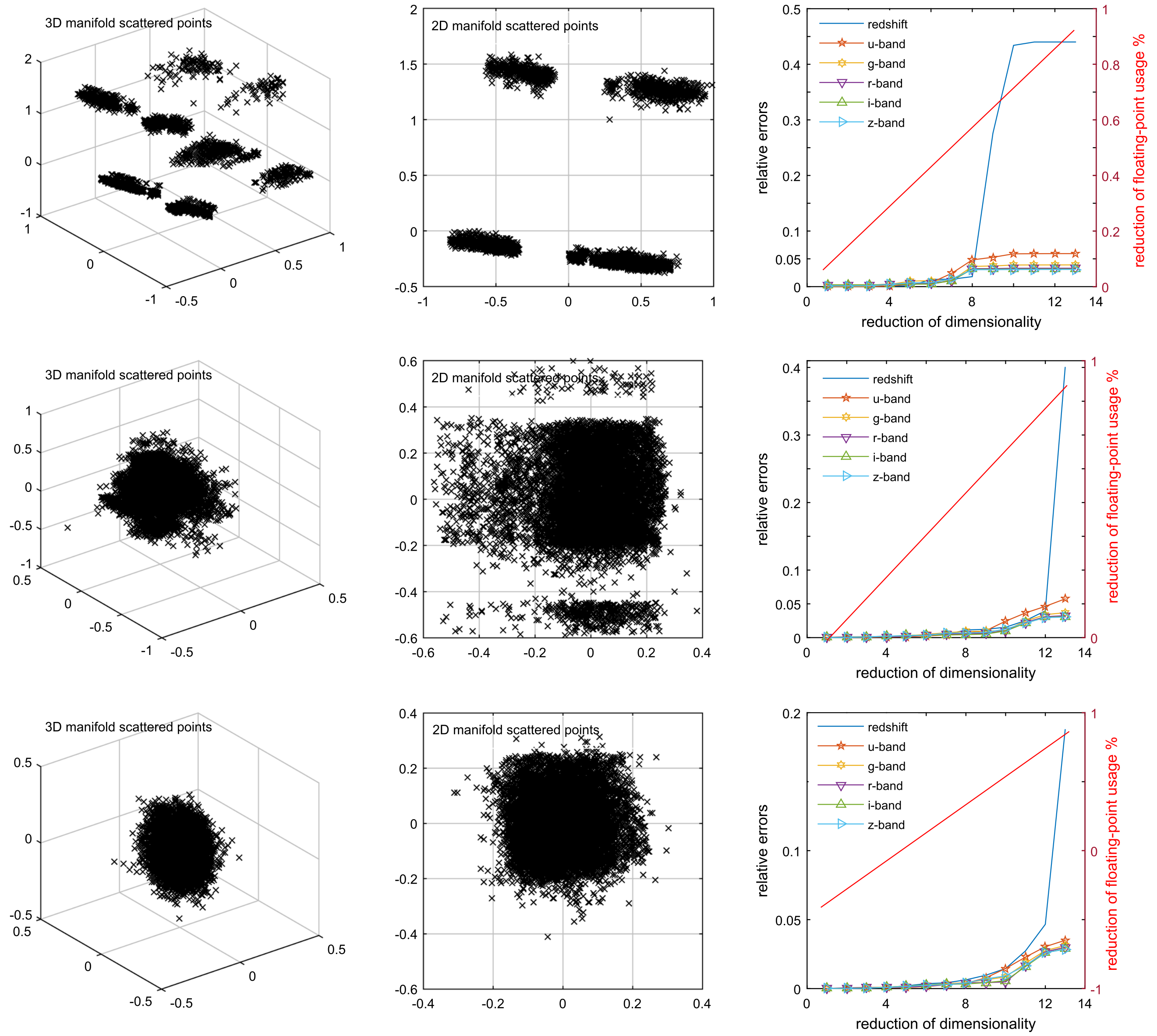


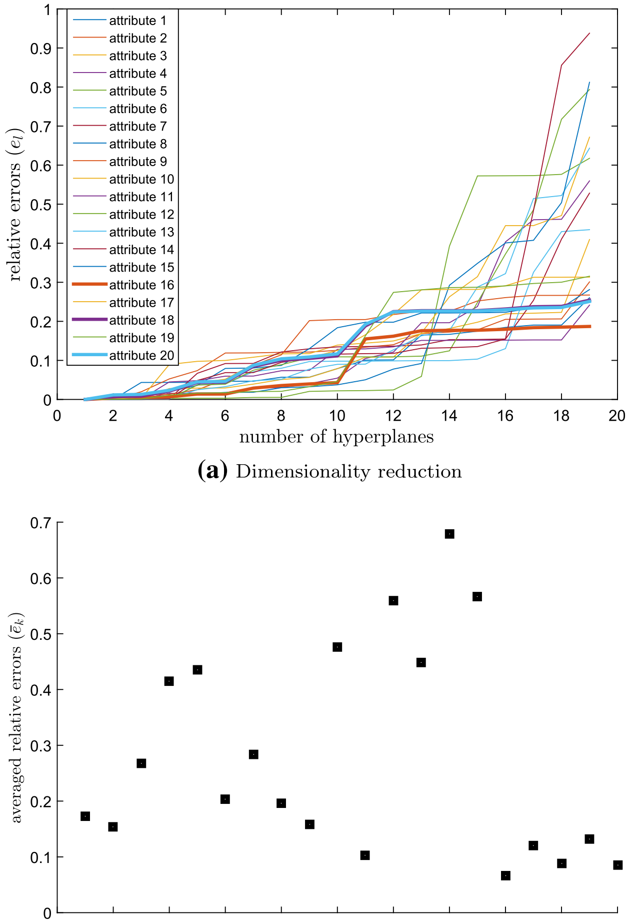
Fig. 5 Fitting of cosmological data of 10,000 sample points. The first two plots in each row show, respectively, 3D and 2D manifold scattered points with maximal magnitudes scaled to 1. A 2D plot actually shows a particular view of the corresponding 3D plot. The third, plotted against the number of the fitted hyperplanes to the points in each n-cube, shows by the left vertical axis the relative errors of the redshift (blue line) and ugriz-bands (marked lines) along with the reduction percentage of floating-point usage (red line) by the right vertical axis. The top, middle and bottom rows correspond to inclusion of the given points in 1, 29 and 169 n-cubes ($n = 14$), respectively, with cubic sizes $r = 2b$, $r = b$ and $r = 2b/3$ with b explained in Fig. 3. The relative error of any particular attribute is calculated as the norm of the difference between the original and predicted attribute vectors divided by the norm of the original attribute vector of total 10,000 components. The relative errors of the remaining eight attributes are not shown, but similar to or smaller than that of redshifts. Given the typical accuracy of the point-spread-function magnitude 0.03 (a relative error 0.15%) of the five-band (ugriz) CCD-based photometry [21] and assuming that the remaining attributes have similar accuracy, the outcomes of fitting hyperplanes suggest that these quasars are 8D objects, probably 6D or even 3D objects when lower data accuracy is assumed, embedded in the 14D space

[20]. The dataset used in this current study is the SDSS third release of 46,420 objects covering the redshift range $0.08 \leq z \leq 5.41$ [21, 22]. Each of the quasars is represented by 14 attributes: right ascension, declination, redshift, brightness in the (ugriz) five bands, radio and X-ray bands, near-infrared (JHK) three bands, and the absolute magnitude in the i band. As shown in Fig. 5, fitting local hyperplanes to the first 10,000 sample points reveals the existence of a low-dimensional object with eight or less attributes embedded in the 14-D space. This corresponds to approximate 40% reduction in the floating-point usage of the original dataset. Exploration and specification of these

low-dimensional attributes ought to be of considerable interest to cosmologists.

4.4 Component prediction applied to a social and economic dataset

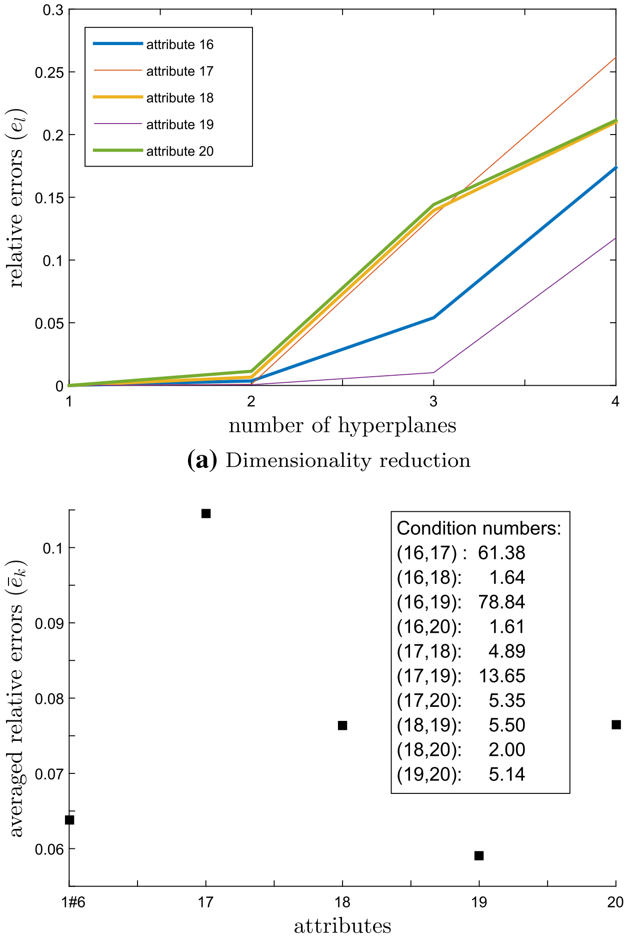
This is use of Eq. 6 to produce unknown components of a partially known data sample. Published by the Department for Communities and Local Government of UK government, the English Indices of Deprivation 2015 [23] were based on data of population, income, employment, education/skills/training, health/disability, crime, housing/services and living environment in geographical local areas in England. For each of 32,844 local authority districts, the dataset [24] records 20 primary attributes. These attributes are scores or counts in categories: 1. income; 2. employment; 3. education/skills/training; 4. health/disability; 5. crime; 6. barriers to housing/services; 7. living/environment; 8. income deprivation affecting children index; 9. income deprivation affecting older people index; 10. children/young people sub-domain; 11. adult skills sub-domain; 12. geographical barriers sub-domain; 13. wider barriers sub-domain; 14. indoors sub-domain; 15. outdoors sub-domain; 16. total population: mid 2012; 17. dependent children aged 0-15: mid-2012; 18. population aged 16-59: mid-2012; 19. older population aged 60 and over: mid 2012; 20. working age population 18-59/64. In categories 16 to 20, prisoners were excluded.



(a) Dimensionality reduction

(b) Attribute selection

Fig. 6 From the data pool, 3000 sample points were randomly selected, and the fitting/testing split was 8:2. As shown in (**a**), fitting hyperplanes to the fitting sample points indicated the existence of an 18D object embedded in the 20D space because of very small relative errors produced by fitting two hyperplanes. The curves of attributes 18 and 20 are in close proximity of each other, and due to low resolution of the plot, the former cannot been seen. By examining the averaged relative errors shown in (**b**) and the calculated condition numbers associated with several attribute pairs, pair $\{16, 20\}$ is selected. Applied to the test sample points, these two attributes are well predicted by the remaining 18 attributes



(a) Dimensionality reduction

(b) Attribute selection

Fig. 7 From the data pool, 3000 sample points were randomly selected, and the fitting/testing split was 8:2. As shown in (**a**), fitting hyperplanes to the training sample points indicated the existence of an 3D object embedded in the 5D space. By examining the averaged relative errors and the associated condition numbers shown in (**b**), attribute pair $\{16, 20\}$ is selected. Applied to the test sample points, this pair can be predicted from the remaining 3 attributes

As shown in Fig. 6, a randomly selected data sample has been used for hyperplane fitting, and attribute pair $\{16, 20\}$ has been selected for prediction with test data sample. The prediction of this pair from the remaining 18 attributes achieved the averaged relative errors at 1.12% and 1.84% with the standard deviations of 1.53 and 1.81, respectively. If other attribute pairs $\{16, 18\}$ and $\{18, 20\}$ have been selected, their predictions have similar outcomes as in the case of using pair $\{16, 20\}$. The condition numbers associated with pairs $\{16, 20\}$, $\{16, 18\}$ and $\{18, 20\}$ are, respectively, 1.61, 1.64 and 2.01.

Instead of considering all 20 attributes, Fig. 7 shows the results with consideration of only five attributes 16 to 20. The prediction of attribute pair $\{16, 20\}$ from the remaining 3 attributes achieved the averaged relative errors at 0.21% and 1.18% with the standard deviations of 0.41 and 0.90, respectively. If pair $\{16, 18\}$ is selected, the averaged relative errors for their predictions are 0.60% and 1.27% with standard deviations 0.46 and 0.98. As expected, due to large condition numbers of the coefficient matrix in Eq. 6, for pair $\{16, 17\}$, the averaged relative errors are 4.25% and 23.68% with standard deviations 1.47 and 8.01, and similarly, for $\{16, 19\}$, the outcomes are 7.72% and 33.41% with 2.67 and 10.61.

4.5 Recidivism prediction

Criminal defendant's recidivism is an important factor considered in judicial fairness and resources management. Given the ever growing number of cases, courts in the USA are increasingly using algorithmic tools to help make decisions on defendants in sentencing or bail hearings [25]. However, adding to a recent nation-wide debate on the issue of race discrimination of a prevalent algorithmic tool that has assessed more than one million defendants since 1998, efficiency of the tool is being challenged [26]. When applied to the dataset of 7214 pretrial defendants over the period of 2013-14 from Broward County, Florida [26, 27], the hyperplane fitting method has produced an overall accuracy of 68% for recidivism predictions using only two attributes (Table 2). Use of more attributes has not improved but gradually worsened the accuracy. The hyperplane fitting method does not score the attributes which are considered indiscriminately by its fitting mechanism.

5 Discussion

The general method for unsupervised machine learning is rigorously derived in this study. It shows equivalence among hyperplane fitting, affine linear manifold fitting [7] and the principal component analysis [11], all linked by the singular value decomposition of a matrix formed by the data points in a segmenting subregion.

For approximations of the Möbius strip and Swiss roll, fitting planes to regularly sampled and noise-free data points and to irregularly sampled and noise-corrupted points in each segmenting cube have been carried out, respectively. In real applications, as discussed in [28], the latter type of data are often involved. The hyperplane fitting along with spatial segmentation of the dataset behaves well in computations of the manifold approximations against changes in the number of sample points, the side length of segmenting n-cubes and the number of hyperplanes fitted to the points within each n-cube. In these

Table 2 Success rates (minimum, average, maximum %) of recidivism predictions in 10 simulations

Attributes	Overall	Recidivated	Not recidivated	l
$\{3, 6\}$	(65, 68, 71)	(62, 67, 72)	(61, 68, 73)	2
$\{3, 4, 6\}$	(65, 68, 71)	(59, 67, 72)	(61, 69, 73)	3
$\{3, 5, 6\}$	(65, 68, 71)	(58, 66, 72)	(63, 69, 73)	3
$\{3, 4, 5, 6\}$	(65, 68, 71)	(58, 66, 72)	(63, 69, 73)	4
$\{3, 4, 5, 6, 7, 8\}$	(63, 65, 68)	(60, 63, 67)	(64, 67, 70)	4
$\{1, 2, 3, 4, 5, 6\}$	(64, 66, 68)	(62, 67, 72)	(60, 65, 70)	5
$\{1, 2, 3, 4, 5, 6, 7, 8\}$	(62, 65, 69)	(60, 64, 68)	(61, 65, 70)	4

In each simulation, 1500 sample points of each of the two recidivism categories were randomly selected from 7214 defendants. The fitting/testing split was 8:2 in all simulations. Regional segmentation of fitting sample points was not needed. The number l of the fitted hyperplanes is shown in the last column of the table. In the fitting, each defendant is represented by a point with 8 (or less) attributes: race, gender, age, number of juvenile felony charges, number of juvenile misdemeanor charges, number of non-juvenile charges, current charge specification, and degree of the charge. In terms of perpendicular distances, the closeness of a test point to the two sets of the fitted hyperplanes determines its belonging to which of the two recidivism categories cases, the accuracy of the approximations is also evaluated since the nonlinear objects are exactly known. In practical applications, the accuracy can be assessed in terms of least squares values of these approximations.

For linear approximation of nonlinear objects, and dimensionality reduction of the quasar data [21, 22], the method of hyperplane fitting behaves well. A comparison of the approximations with those produced by other methods such as the local linear embedding [2–4], the tangent-based manifold approximation [10], or the autoencoder [13] would not be very indicative or objective and hence is spared. This is because the accuracy of linear approximation and efficiency of dimensionality reduction are two interrelated criteria and depend on the number of subregions and the dimensions of local linear manifolds or equivalents used in these methods. Applied to a social and economic dataset [23], the current method has predicted partial components of a high-dimensional vector, while this type of prediction in high-dimensional data application does not seem to be reported before.

For classifications, applications of the current method to the problems of hand-written digit recognition [17] and recidivism prediction [25] produced competitive results. The best reported error rate for hand-written digit recognition without data preprocessing in the literature is in the range $0.2 \sim 3.3\%$ [29, 30], where the lower end is achieved by deep learning neural networks, while the higher end by other traditional classifiers. The current method has achieved an error rate of 1.4%. Compared with hundreds or thousands of free parameters used in training neural networks, the current method has only two free parameters, namely the side length of segmenting hypercubes and number of fitted hyperplanes. For recidivism predictions, the current method has achieved the best accuracy of 68% when only two attributes for defendants were used. In comparison, untrained people and algorithmic tools achieved prediction accuracy between 65% and 67% [26]. With a slightly improved accuracy at 68%, a recent study [31] replicated the findings in [26], but also showed outperformance of computational tools over untrained peoples in different set-ups.

Fitting hyperplanes to high-dimensional data can be considered a baseline method for machine learning due to its simplicity and efficiency. There are a few obvious possibilities of improving the method developed in this work. First, the spatial segmentation of data could be combined with the tangent segmentation [10] to merge the hypercubic subregions in order to enhance efficiency of the linear approximation at a little cost of approximation accuracy. Basically, two neighbouring data subgroups are merged if their hyperplane normals are aligned to each other and the difference of their centroids is perpendicular to their normals. Second, instead of fitting one set of hyperplanes to data points in each hypercube, several hyperplane sets could be fitted on the basis of the subspace segmentation [12] of the local data points. This could enhance approximation accuracy particularly when the noise level of the data is low. Third, fitting linear manifolds can be extended to fitting nonlinear manifolds if some geometric properties of the underlying object are known a priori, for instance in geodesic applications. The extension ought to be straightforward because in calculations of the fitting, parameters of a nonlinear manifold can be formulated to appear linear in terms of sample data. Finally, using the current method, the fitted hyperplane patches are normally not connected to each other at the junctions of hypercubes, which is a drawback of a method for linear approximation of nonlinear objects. To overcome this drawback, local geometric details of the object need to be examined and incorporated in hyperplane fitting.

6 Conclusion

This paper has proposed a general approach to the problem of unsupervised machine learning, and the proposed method is rigorously derived. The method requires no assumptions on the underlying object and its data sample. Also, it has only two parameters, namely the size of segmenting hypercubes and the number of fitted hyperplanes for user to choose. These make the proposed method considerably accessible when applied to solving various problems in real applications.

A successful application of the current rigorously and transparently defined algorithm to recidivism prediction has an important implication. In line with the investigation [26], this study challenges use of any undisclosed algorithms helping make decisions in a justice system where transparency and fairness are cornerstones. This is also true about decision making processes of the public sector in a democratic society. The fundamental argument is: When a decision maker is obligated to elaborate a decision made in public services, the algorithm of a computer-aided tool used to help make such a decision must be subject to scrutiny by the scientific community and probably also the general public at large.

Based on two essential concepts of hyperplane fitting and spatial data segmentation, the developed method is conceptually simple and shows effectiveness when applied to several typical machine learning problems. It can therefore be considered a baseline method for machine learning. Several possibilities for improvements of the hyperplane method have been highlighted in the section of discussion.

Appendix A: Hyperplane fitting

Fitting hyperplane $(\mathbf{p}, d)$ to $\{\mathbf{x}_i\}$ minimises

$$J = \min \sum_{i=1}^{m} (\mathbf{x}_i'\mathbf{p} - d)^2 = \min \|\mathbf{X}\mathbf{p} - \mathbf{e}d\|^2 \quad \text{s.t.} \quad \mathbf{p}'\mathbf{p} = 1 \tag{10}$$

where $\mathbf{X} = [\mathbf{x}_1, \ \mathbf{x}_2, \ \ldots, \ \mathbf{x}_m]'$, $\mathbf{e} \in \boldsymbol{R}^m$ is the vector of ones, and $||\mathbf{a}||^2 = \mathbf{a}'\mathbf{a}$ is the squared norm of $\mathbf{a}$. With Lagrange multiplier λ, this minimisation is equivalent to minimising

$$J_\lambda = \min\left(\|\mathbf{X}\mathbf{p} - \mathbf{e}d\|^2 + \lambda(1 - \mathbf{p}'\mathbf{p}) \right). \tag{11}$$

Setting $\frac{\partial J_\lambda}{\partial d} = (\mathbf{X}\mathbf{p} - \mathbf{e}d)'\mathbf{e} = 0$ leads to $d = \bar{\mathbf{x}}'\mathbf{p}$. Substituting d into J_λ and setting $\frac{\partial J_\lambda}{\partial \mathbf{p}} = \mathbf{p}'\mathbf{Q} - \lambda\mathbf{p}' = 0$ produces $(\lambda\mathbf{I} - \mathbf{Q})\mathbf{p} = \mathbf{0}$ with $\mathbf{Q} = \bar{\mathbf{X}}'\bar{\mathbf{X}}$. This means that $\mathbf{p}$ is a unit eigenvector of $\mathbf{Q}$ associated with eigenvalue $\lambda = \mathbf{p}'\mathbf{Q}\mathbf{p}$. The singular value decomposition of $\bar{\mathbf{X}}$ in (1) implies $\mathbf{V}'\mathbf{Q}\mathbf{V} = \Sigma_x'\Sigma_x =$ diag $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$, with $\mathbf{V} = [\mathbf{v}_1, \ \mathbf{v}_2, \ \ldots, \ \mathbf{v}_n]$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}$, $\sigma_n \geq 0$, and $\sigma_i \geq \sigma_j$ for $i < j$. Hence, the n stationary points of J_λ are given by $(\mathbf{p}, d, \lambda) = (\mathbf{v}_i, \bar{\mathbf{x}}'\mathbf{v}_i, \sigma_i^2)$, and hyperplane $(\mathbf{v}_n, \bar{\mathbf{x}}'\mathbf{v}_n)$ best fitting $\{\mathbf{x}_i\}$ achieves the minimum $J = J_\lambda = \mathbf{v}_n'\mathbf{Q}\mathbf{v}_n = \sigma_n^2$.

In general, for fitting l hyperplanes $(\mathbf{p}_j, d_j)$ for $j = 1, 2, \ldots, l$ with $1 \leq l \leq n$ to $\{\mathbf{x}_i\}$, the criterion is

$$J_{\lambda,l} = \min \sum_{j=1}^{l} \left(||\mathbf{X}\mathbf{p}_j - \mathbf{e}d_j||^2 + \lambda_j(1 - \mathbf{p}_j'\mathbf{p}_j) \right). \tag{12}$$

Using similar derivations and arguments as in fitting a single hyperplane, set $\frac{\partial J_{\lambda,l}}{\partial d_j} = 0$ and then $\frac{\partial J_{\lambda,l}}{\partial \mathbf{p}_j} = 0$. $(\lambda_j\mathbf{I} - \mathbf{Q})\mathbf{p}_j = 0$ follows, which gives rise to the fitted hyperplanes in (3).

Remark 1 There is no need to make assume that the number of data points is greater than their dimension or that all data points are not coplanar. This is because the least squares fitting automatically produces hyperplanes passing all points if they are coplanar with minimisation criterion $J_{\lambda,l} = J_l = 0$.

Remark 2 If and only if $\sigma_{n-j} \neq \sigma_n$ for $j = 1, 2, \ldots, l$, the optimal hyperplanes are unique. For $i \neq j$, $\sigma_i = \sigma_j$ implies symmetric distributions of $\{\mathbf{x}_i\}$ about $\mathbf{v}_i$ and $\mathbf{v}_j$ directions and therefore non-uniqueness of optimal hyperplanes. Examples of such cases in a 3D space are evenly distributed $\{\mathbf{x}_i\}$ in a cubic, cylindrical or spherical shape. When $\sigma_i = \sigma_n$ for $1 \leq i < n$, with $\mathbf{p}_1 = \sum_{j=i}^{n} a_{1j}\mathbf{v}_j$ and arbitrary parameter set $\{a_{1i}\}$ satisfying $\sum_{j=i}^{n} a_{1j}^2 = 1$, $(\mathbf{p}_1, \bar{\mathbf{x}}'\mathbf{p}_1)$ represents a pencil of infinity of hyperplanes passing $\bar{\mathbf{x}}$, all equally best fitted to $\{\mathbf{x}_i\}$ with $J_1(\mathbf{p}_1) \equiv \sigma_n^2$. Moreover, with $\mathbf{p}_2 = \sum_{j=i}^{n} a_{2j}\mathbf{v}_j$ and anther arbitrary parameter set $\{a_{2i}\}$ satisfying $\sum_{j=i}^{n} a_{2i}^2 = 1$ and $\sum_{j=i}^{n} a_{1j}a_{2j} = 0$, two pencils of infinity of hyperplanes $(\mathbf{p}_1, \bar{\mathbf{x}}'\mathbf{p}_1)$ and $(\mathbf{p}_2, \bar{\mathbf{x}}'\mathbf{p}_2)$ ensure $J_2 \equiv 2\sigma_n^2$.

Appendix B: Affine linear manifold

Partition the right-singular vectors of $\bar{\mathbf{X}}$ into $\mathbf{V} = [\mathbf{V}_{n-l} \quad \mathbf{V}_l]$ with $\mathbf{V}_{n-l}$ and $\mathbf{V}_l$ consisting of the first $n - l$ and last l vectors, and note $\mathbf{V}_{n-l}\mathbf{V}_{n-l}' + \mathbf{V}_l\mathbf{V}_l' = \mathbf{I}$. With $\mathbf{x}$ being a general point on the intersection of the fitted l hyperplanes all passing $\bar{\mathbf{x}}$, $\mathbf{x} - \bar{\mathbf{x}}$ is perpendicular to the normals of these hyperplanes. This suggests $\mathbf{V}_l'(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{0}$ which has the general solution

$$\mathbf{x} = \bar{\mathbf{x}} + (\mathbf{I} - \mathbf{V}_l\mathbf{V}_l')\mathbf{u} = \bar{\mathbf{x}} + \mathbf{V}_{n-l}\mathbf{V}_{n-l}'\mathbf{u}, \tag{13}$$

where $\mathbf{u} \in \boldsymbol{R}^n$ is arbitrary. Arbitrariness of $\mathbf{u}$ implies arbitrariness of $\mathbf{V}_{n-l}'\mathbf{u}$ due to full row rank of $\mathbf{V}_{n-l}'$. Setting $\mathbf{y} = \mathbf{V}_{n-l}'\mathbf{u}$ leads to the parametric equation in (4). For arbitrary $\mathbf{x}$ on the intersection, the unique solution of $\mathbf{y}$ to the parametric equation is given by the parametric inverse in (4).

Setting $\mathbf{x} = \mathbf{x}_i$ in the parametric inverse in (4) produces low-dimensional point $\mathbf{y}_i$, and substituting this $\mathbf{y}_i$ back to the parametric equation in (4) yields $\hat{\mathbf{x}}_i$, all given in (5). To verify $\hat{\mathbf{x}}_i$ being the orthogonal projection of $\mathbf{x}_i$ onto the $(n - l)$D affine linear manifold, from (5) it is ready to have $\mathbf{x}_i - \hat{\mathbf{x}}_i = \mathbf{V}_l\mathbf{V}_l'(\mathbf{x}_i - \bar{\mathbf{x}})$ and hence $\mathbf{V}_{n-l}'(\mathbf{x}_i - \hat{\mathbf{x}}_i) = \mathbf{0}$. Indeed, $\sum_{i=1}^{m} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 = \sum_{i=1}^{m} ||\mathbf{V}_l\mathbf{V}_l'(\mathbf{x}_i - \bar{\mathbf{x}})||^2 =$ tr $(\bar{\mathbf{X}}\mathbf{V}_l\mathbf{V}_l'\mathbf{V}_l\mathbf{V}_l'\bar{\mathbf{X}}') =$ tr $(\mathbf{V}_l'\mathbf{Q}\mathbf{V}_l) = \sum_{j=0}^{l-1} \sigma_{n-j}^2 = J_l$, where 'tr' stands for the trace and achieves the same minimum as the original criterion for fitting l hyperplanes. This implies that simultaneously fitting several hyperplanes naturally minimises the sum of the squared perpendicular distances of their intersection to $\{\mathbf{x}_i\}$ as well.

Remark 3 Arbitrary $\bar{\mathbf{V}}_{n-l}$ satisfying $\mathbf{V}_l'\bar{\mathbf{V}}_{n-l} = \mathbf{0}$ and $\mathbf{V}_l\mathbf{V}_l' + \bar{\mathbf{V}}_{n-l}\bar{\mathbf{V}}_{n-l}' = \mathbf{I}$ is an orthogonal complement to $\mathbf{V}_l$. In general, all orthogonal complements are given by $\bar{\mathbf{V}}_{n-l} = \mathbf{V}_{n-l}\mathbf{T}$ with arbitrary orthogonal matrix $\mathbf{T}$. Use of $\mathbf{V}_{n-l}$ instead of $\bar{\mathbf{V}}_{n-l}$ in (13) makes a natural connection of the affine linear manifold to the principal component analysis.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Cevikalp H (2017) Best fitting hyperplanes for classification. IEEE Trans Pattern Anal Mach Intell 39:1076–1088
2. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
3. Saul LK, Roweis ST (2003) Think globally, fit locally: unsupervised learning of low dimensional manifolds. J Mach Learn Res 4:119–155
4. Chen J, Liu Y (2011) Locally linear embedding: a survey. Artif Intell Rev 36:29–48
5. Donoho DL, Grimes C (2002) Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. PNAS 100:5591–5596
6. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
7. Späth H (1986) Orthogonal least squares fitting with linear manifolds. Numer Math 48:441–445
8. Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. Psychometrica 1:211–218
9. Zhang Z, Zha H (2005) Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J Sci Comput 26:313–338
10. Karygianni S, Frossard P (2014) Tangent-based manifold approximation with locally linear models. Signal Process 104:232–247
11. Jolliffe IT (2004) Principal component analysis. Springer, New York
12. Vidal R, Ma Y, Sastry S (2005) Generalized principal component analysis (GPCA). IEEE Trans Pattern Anal Mach Intell 27:1945–1959
13. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507
14. Schöbel A (1999) Locating lines and hyperplanes: theory and algorithms. Kluwer Academic Publishers, New York
15. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag A 6:559–572
16. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 25:417–441
17. The dataset. http://yann.lecun.com/exdb/mnist
18. Hennawi JF, Prochaska JX, Cantalupo S, Arrigoni-Battaia1 F (2015) Quasar quartet embedded in giant nebula reveals rare massive structure in distant universe. Science 348:779–783
19. Decarli R et al (2017) Rapidly star-forming galaxies adjacent to quasars at redshifts exceeding 6. Nature 545:457–461
20. Raddick MJ, Szalay AS (2010) The universe online. Science 329:1028–1029
21. Schneider DP et al (2005) The sloan digital sky survey quasar catalog. III. Third data release. Astrophys J 130:367–380
22. The dataset. http://astrostatistics.psu.edu/datasets
23. Smith T et al (2015) The English Indices of Deprivation 2015: Research Report. Department for Communities and Local Government, UK Government
24. The dataset. https://www.cdrc.ac.uk
25. Matacic C (2018) Are algorithms good judges? Science 359:263
26. Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4, eaao5580
27. The dataset. http://www.cs.dartmouth.edu/farid/downloads/publications/scienceadvances17
28. Balasubramanian M, Schwartz EL (2002) The Isomap algorithm and topological stability. Science 295:7a
29. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient based learning applied to document recognition. Proc IEEE 86:2278–2324
30. Ciregan D, Meier U, Schmidhuber J (2012) Multi-column deep neural networks for image classification. In: IEEE Conference on Computer Vision Pattern Recognition, pp 3642–3649
31. Lin Z, Jung J, Goel S, Skeem J (2020) The limits of human predictions of recidivism. Sci Adv 7, eaaz0652