CHAPTER 6 High-dimensional Data

In data mining typically the data is very high dimensional, as the number of attributes can easily be in the hundreds or thousands. Understanding the nature of high-dimensional space, or *hyperspace*, is very important, especially because hyperspace does not behave like the more familiar geometry in two or three dimensions.

6.1 HIGH-DIMENSIONAL OBJECTS

Consider the $n \times d$ data matrix

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

where each point $\mathbf{x}_i \in \mathbb{R}^d$ and each attribute $X_j \in \mathbb{R}^n$.

Hypercube

Let the minimum and maximum values for each attribute X_j be given as

$$\min(X_j) = \min_i \{x_{ij}\} \qquad \max(X_j) = \max_i \{x_{ij}\}$$

The data hyperspace can be considered as a d-dimensional *hyper-rectangle*, defined as

$$\begin{aligned} R_d &= \prod_{j=1}^{d} \Big[\min(X_j), \max(X_j)\Big] \\ &= \Big\{\mathbf{x} = (x_1, x_2, \ldots, x_d)^T \;\Big|\; x_j \in [\min(X_j), \max(X_j)], \text{for } j = 1, \ldots, d\Big\} \end{aligned}$$

Assume the data is centered to have mean $\boldsymbol{\mu} = \mathbf{0}$. Let m denote the largest absolute value in $\mathbf{D}$, given as

$$m = \max_{j=1}^{d} \max_{i=1}^{n} \Big\{ |x_{ij}| \Big\}$$

The data hyperspace can be represented as a *hypercube*, centered at $\mathbf{0}$, with all sides of length $l = 2m$, given as

$$H_d(l) = \Big\{ \mathbf{x} = (x_1, x_2, \ldots, x_d)^T \;\big|\; \forall i,\ x_i \in [-l/2, l/2] \Big\}$$

The hypercube in one dimension, $H_1(l)$, represents an interval, which in two dimensions, $H_2(l)$, represents a square, and which in three dimensions, $H_3(l)$, represents a cube, and so on. The *unit hypercube* has all sides of length $l = 1$, and is denoted as $H_d(1)$.

Hypersphere

Assume that the data has been centered, so that $\boldsymbol{\mu} = \mathbf{0}$. Let r denote the largest magnitude among all points:

$$r = \max_{i} \Big\{ \|\mathbf{x}_i\| \Big\}$$

The data hyperspace can also be represented as a d-dimensional *hyperball* centered at $\mathbf{0}$ with radius r, defined as

$$B_d(r) = \Big\{ \mathbf{x} \;|\; \|\mathbf{x}\| \leq r \Big\}$$

$$\text{or } B_d(r) = \Big\{ \mathbf{x} = (x_1, x_2, \ldots, x_d) \;\Big|\; \sum_{j=1}^{d} x_j^2 \leq r^2 \Big\}$$

The surface of the hyperball is called a *hypersphere*, and it consists of all the points exactly at distance r from the center of the hyperball, defined as

$$S_d(r) = \Big\{ \mathbf{x} \;|\; \|\mathbf{x}\| = r \Big\}$$

$$\text{or } S_d(r) = \Big\{ \mathbf{x} = (x_1, x_2, \ldots, x_d) \;\Big|\; \sum_{j=1}^{d} (x_j)^2 = r^2 \Big\}$$

Because the hyperball consists of all the surface and interior points, it is also called a *closed hypersphere*.

Example 6.1. Consider the 2-dimensional, centered, Iris dataset, plotted in Figure 6.1. The largest absolute value along any dimension is $m = 2.06$, and the point with the largest magnitude is $(2.06, 0.75)$, with $r = 2.19$. In two dimensions, the hypercube representing the data space is a square with sides of length $l = 2m = 4.12$. The hypersphere marking the extent of the space is a circle (shown dashed) with radius $r = 2.19$.

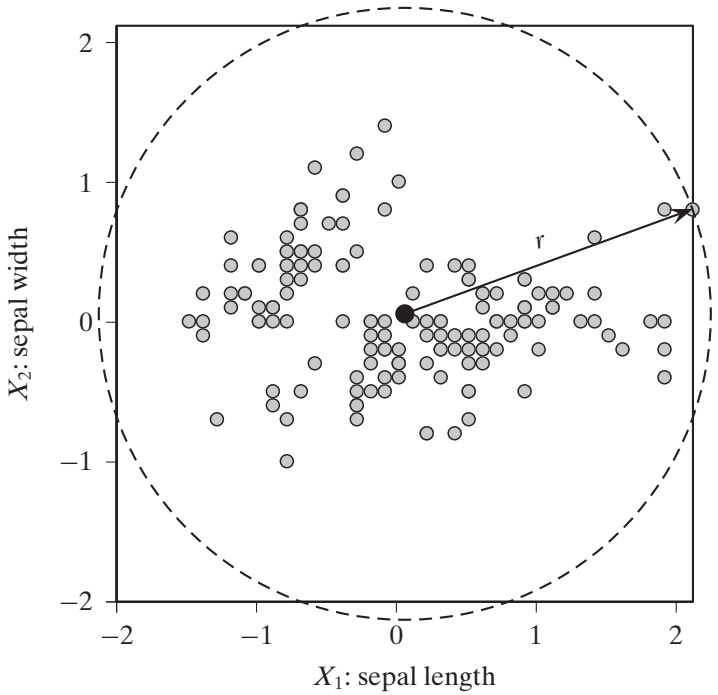

Figure 6.1. Iris data hyperspace: hypercube (solid; with $l = 4.12$) and hypersphere (dashed; with $r = 2.19$).

6.2 HIGH-DIMENSIONAL VOLUMES

Hypercube

The volume of a hypercube with edge length l is given as

$$\text{vol}(H_d(l)) = l^d$$

Hypersphere

The volume of a hyperball and its corresponding hypersphere is identical because the volume measures the total content of the object, including all internal space. Consider the well known equations for the volume of a hypersphere in lower dimensions

$$\text{vol}(S_1(r)) = 2r \tag{6.1}$$

$$\text{vol}(S_2(r)) = \pi r^2 \tag{6.2}$$

$$\text{vol}(S_3(r)) = \frac{4}{3}\pi r^3 \tag{6.3}$$

As per the derivation in Appendix 6.7, the general equation for the volume of a d-dimensional hypersphere is given as

$$\text{vol}(S_d(r)) = K_d r^d = \left(\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}\right) r^d \tag{6.4}$$

where

$$K_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} \tag{6.5}$$

is a scalar that depends on the dimensionality d, and Γ is the gamma function [Eq. (3.17)], defined as (for $\alpha > 0$)

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \tag{6.6}$$

By direct integration of Eq. (6.6), we have

$$\Gamma(1) = 1 \qquad \text{and} \qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \tag{6.7}$$

The gamma function also has the following property for any $\alpha > 1$:

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \tag{6.8}$$

For any integer $n \geq 1$, we immediately have

$$\Gamma(n) = (n-1)! \tag{6.9}$$

Turning our attention back to Eq. (6.4), when d is even, then $\frac{d}{2}+1$ is an integer, and by Eq. (6.9) we have

$$\Gamma\left(\frac{d}{2}+1\right) = \left(\frac{d}{2}\right)!$$

and when d is odd, then by Eqs. (6.8) and (6.7), we have

$$\Gamma\left(\frac{d}{2}+1\right) = \left(\frac{d}{2}\right)\left(\frac{d-2}{2}\right)\left(\frac{d-4}{2}\right)\cdots\left(\frac{d-(d-1)}{2}\right)\Gamma\left(\frac{1}{2}\right) = \left(\frac{d!!}{2^{(d+1)/2}}\right)\sqrt{\pi}$$

where $d!!$ denotes the double factorial (or multifactorial), given as

$$d!! = \begin{cases} 1 & \text{if } d = 0 \text{ or } d = 1 \\ d \cdot (d-2)!! & \text{if } d \geq 2 \end{cases}$$

Putting it all together we have

$$\Gamma\left(\frac{d}{2}+1\right) = \begin{cases} \left(\frac{d}{2}\right)! & \text{if } d \text{ is even} \\ \sqrt{\pi}\left(\frac{d!!}{2^{(d+1)/2}}\right) & \text{if } d \text{ is odd} \end{cases} \tag{6.10}$$

Plugging in values of $\Gamma(d/2+1)$ in Eq. (6.4) gives us the equations for the volume of the hypersphere in different dimensions.

Example 6.2. By Eq. (6.10), we have for $d = 1$, $d = 2$ and $d = 3$:

$$\Gamma(1/2+1) = \frac{1}{2}\sqrt{\pi}$$
$$\Gamma(2/2+1) = 1! = 1$$
$$\Gamma(3/2+1) = \frac{3}{4}\sqrt{\pi}$$

Thus, we can verify that the volume of a hypersphere in one, two, and three dimensions is given as

$$\text{vol}(S_1(r)) = \frac{\sqrt{\pi}}{\frac{1}{2}\sqrt{\pi}} r = 2r$$
$$\text{vol}(S_2(r)) = \frac{\pi}{1} r^2 = \pi r^2$$
$$\text{vol}(S_3(r)) = \frac{\pi^{3/2}}{\frac{3}{4}\sqrt{\pi}} r^3 = \frac{4}{3}\pi r^3$$

which match the expressions in Eqs. (6.1), (6.2), and (6.3), respectively.

Surface Area The *surface area* of the hypersphere can be obtained by differentiating its volume with respect to r, given as

$$\text{area}(S_d(r)) = \frac{d}{dr}\text{vol}(S_d(r)) = \left(\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}\right) d r^{d-1} = \left(\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}\right) r^{d-1}$$

We can quickly verify that for two dimensions the surface area of a circle is given as $2\pi r$, and for three dimensions the surface area of sphere is given as $4\pi r^2$.

Asymptotic Volume An interesting observation about the hypersphere volume is that as dimensionality increases, the volume first increases up to a point, and then starts to decrease, and ultimately vanishes. In particular, for the unit hypersphere with $r = 1$,

$$\lim_{d\to\infty} \text{vol}(S_d(1)) = \lim_{d\to\infty} \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \to 0$$

Example 6.3. Figure 6.2 plots the volume of the unit hypersphere in Eq. (6.4) with increasing dimensionality. We see that initially the volume increases, and achieves the highest volume for $d = 5$ with $\text{vol}(S_5(1)) = 5.263$. Thereafter, the volume drops rapidly and essentially becomes zero by $d = 30$.

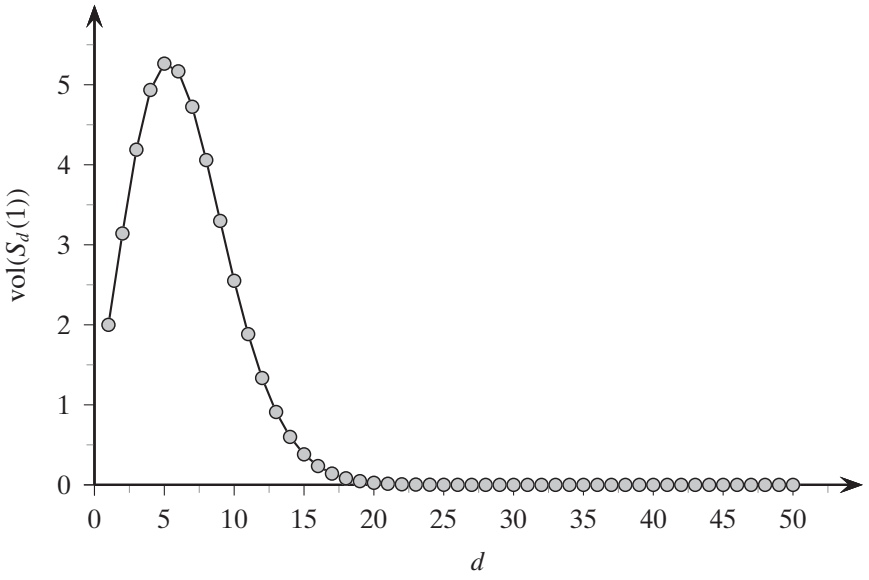

Figure 6.2. Volume of a unit hypersphere.

6.3 HYPERSPHERE INSCRIBED WITHIN HYPERCUBE

We next look at the space enclosed within the largest hypersphere that can be accommodated within a hypercube (which represents the dataspace). Consider a hypersphere of radius r inscribed in a hypercube with sides of length $2r$. When we take the ratio of the volume of the hypersphere of radius r to the hypercube with side length $l = 2r$, we observe the following trends.

In two dimensions, we have

$$\frac{\text{vol}(S_2(r))}{\text{vol}(H_2(2r))} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} = 78.5\%$$

Thus, an inscribed circle occupies $\frac{\pi}{4}$ of the volume of its enclosing square, as illustrated in Figure 6.3a.

In three dimensions, the ratio is given as

$$\frac{\text{vol}(S_3(r))}{\text{vol}(H_3(2r))} = \frac{\frac{4}{3}\pi r^3}{8r^3} = \frac{\pi}{6} = 52.4\%$$

An inscribed sphere takes up only $\frac{\pi}{6}$ of the volume of its enclosing cube, as shown in Figure 6.3b, which is quite a sharp decrease over the 2-dimensional case.

For the general case, as the dimensionality d increases asymptotically, we get

$$\lim_{d\to\infty} \frac{\text{vol}(S_d(r))}{\text{vol}(H_d(2r))} = \lim_{d\to\infty} \frac{\pi^{d/2}}{2^d \Gamma(\frac{d}{2}+1)} \to 0$$

This means that as the dimensionality increases, most of the volume of the hypercube is in the "corners," whereas the center is essentially empty. The mental picture that

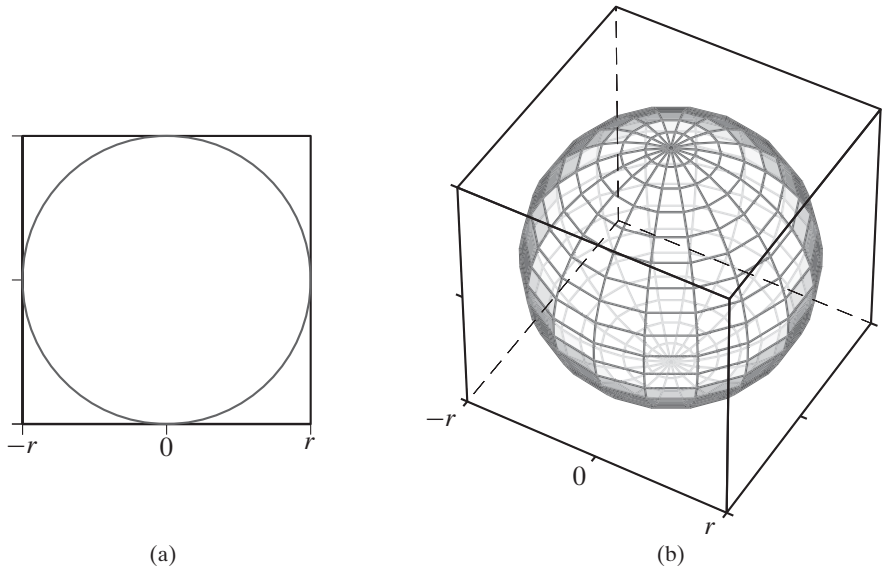

Figure 6.3. Hypersphere inscribed inside a hypercube: in (a) two and (b) three dimensions.

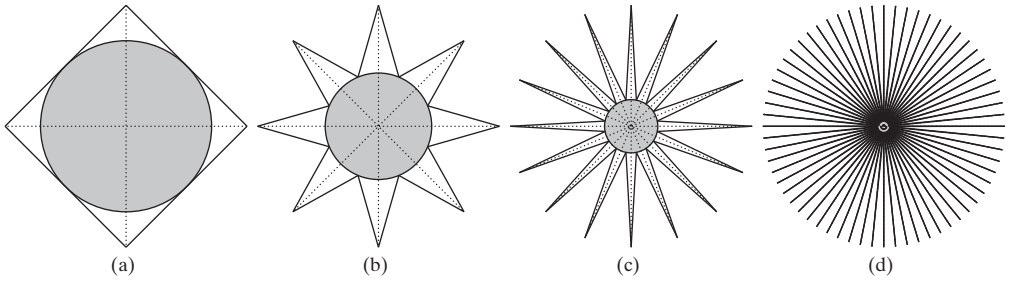

Figure 6.4. Conceptual view of high-dimensional space: (a) two, (b) three, (c) four, and (d) higher dimensions. In d dimensions there are 2^d "corners" and 2^{d-1} diagonals. The radius of the inscribed circle accurately reflects the difference between the volume of the hypercube and the inscribed hypersphere in d dimensions.

emerges is that high-dimensional space looks like a rolled-up porcupine, as illustrated in Figure 6.4.

6.4 VOLUME OF THIN HYPERSPHERE SHELL

Let us now consider the volume of a thin hypersphere shell of width ϵ bounded by an outer hypersphere of radius r, and an inner hypersphere of radius $r - \epsilon$. The volume of the thin shell is given as the difference between the volumes of the two bounding hyperspheres, as illustrated in Figure 6.5.

Let $S_d(r, \epsilon)$ denote the thin hypershell of width ϵ. Its volume is given as

$$\text{vol}(S_d(r, \epsilon)) = \text{vol}(S_d(r)) - \text{vol}(S_d(r - \epsilon)) = K_d r^d - K_d (r - \epsilon)^d.$$

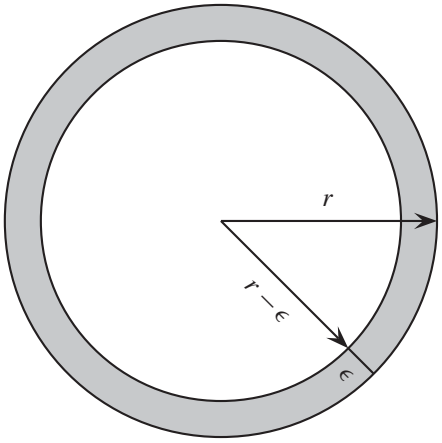

Figure 6.5. Volume of a thin shell (for $\epsilon > 0$).

Let us consider the ratio of the volume of the thin shell to the volume of the outer sphere:

$$\frac{\text{vol}(S_d(r,\epsilon))}{\text{vol}(S_d(r))} = \frac{K_d r^d - K_d (r-\epsilon)^d}{K_d r^d} = 1 - \left(1 - \frac{\epsilon}{r}\right)^d$$

Example 6.4. For example, for a circle in two dimensions, with $r = 1$ and $\epsilon = 0.01$ the volume of the thin shell is $1 - (0.99)^2 = 0.0199 \simeq 2\%$. As expected, in two-dimensions, the thin shell encloses only a small fraction of the volume of the original hypersphere. For three dimensions this fraction becomes $1 - (0.99)^3 = 0.0297 \simeq 3\%$, which is still a relatively small fraction.

Asymptotic Volume

As d increases, in the limit we obtain

$$\lim_{d\to\infty} \frac{\text{vol}(S_d(r,\epsilon))}{\text{vol}(S_d(r))} = \lim_{d\to\infty} 1 - \left(1 - \frac{\epsilon}{r}\right)^d \to 1$$

That is, almost all of the volume of the hypersphere is contained in the thin shell as $d \to \infty$. This means that in high-dimensional spaces, unlike in lower dimensions, most of the volume is concentrated around the surface (within ϵ) of the hypersphere, and the center is essentially void. In other words, if the data is distributed uniformly in the d-dimensional space, then all of the points essentially lie on the boundary of the space (which is a $d-1$ dimensional object). Combined with the fact that most of the hypercube volume is in the corners, we can observe that in high dimensions, data tends to get scattered on the boundary and corners of the space.

6.5 DIAGONALS IN HYPERSPACE

Another counterintuitive behavior of high-dimensional spaces deals with the diagonals. Let us assume that we have a d-dimensional hypercube, with origin $\mathbf{0}_d = (0_1, 0_2, \ldots, 0_d)$, and bounded in each dimension in the range $[-1, 1]$. Then each "corner" of the hyperspace is a d-dimensional vector of the form $(\pm 1_1, \pm 1_2, \ldots, \pm 1_d)^T$. Let $\mathbf{e}_i = (0_1, \ldots, 1_i, \ldots, 0_d)^T$ denote the d-dimensional canonical unit vector in dimension i, and let $\mathbf{1}$ denote the d-dimensional diagonal vector $(1_1, 1_2, \ldots, 1_d)^T$.

Consider the angle θ_d between the diagonal vector $\mathbf{1}$ and the first axis $\mathbf{e}_1$, in d dimensions:

$$\cos\theta_d = \frac{\mathbf{e}_1^T\mathbf{1}}{\|\mathbf{e}_1\|\,\|\mathbf{1}\|} = \frac{\mathbf{e}_1^T\mathbf{1}}{\sqrt{\mathbf{e}_1^T\mathbf{e}_1}\sqrt{\mathbf{1}^T\mathbf{1}}} = \frac{1}{\sqrt{1}\sqrt{d}} = \frac{1}{\sqrt{d}}$$

Example 6.5. Figure 6.6 illustrates the angle between the diagonal vector $\mathbf{1}$ and $\mathbf{e}_1$, for $d = 2$ and $d = 3$. In two dimensions, we have $\cos\theta_2 = \frac{1}{\sqrt{2}}$ whereas in three dimensions, we have $\cos\theta_3 = \frac{1}{\sqrt{3}}$.

Asymptotic Angle

As d increases, the angle between the d-dimensional diagonal vector $\mathbf{1}$ and the first axis vector $\mathbf{e}_1$ is given as

$$\lim_{d\to\infty} \cos\theta_d = \lim_{d\to\infty} \frac{1}{\sqrt{d}} \to 0$$

which implies that

$$\lim_{d\to\infty} \theta_d \to \frac{\pi}{2} = 90^\circ$$

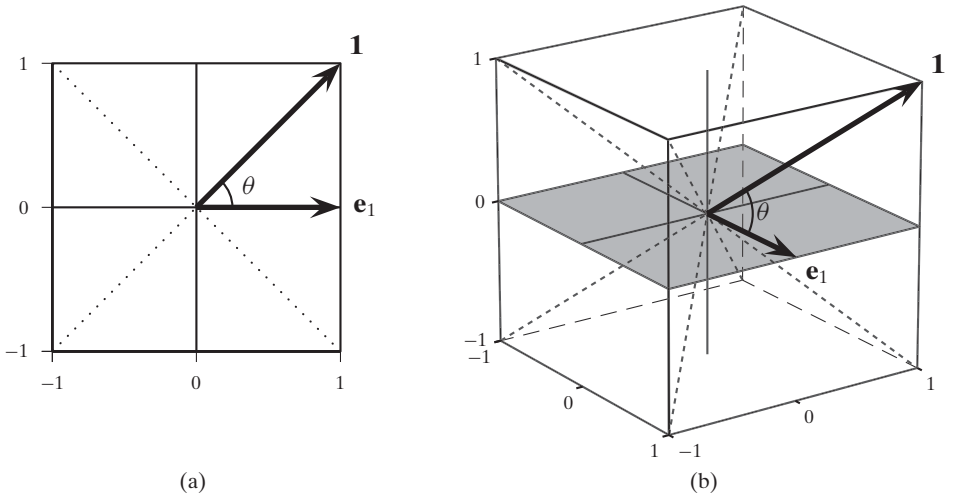

Figure 6.6. Angle between diagonal vector $\mathbf{1}$ and $\mathbf{e}_1$: in (a) two and (b) three dimensions.

This analysis holds for the angle between the diagonal vector $\mathbf{1}_d$ and any of the d principal axis vectors $\mathbf{e}_i$ (i.e., for all $i \in [1, d]$). In fact, the same result holds for any diagonal vector and any principal axis vector (in both directions). This implies that in high dimensions all of the diagonal vectors are perpendicular (or orthogonal) to all the coordinates axes! Because there are 2^d corners in a d-dimensional hyperspace, there are 2^d diagonal vectors from the origin to each of the corners. Because the diagonal vectors in opposite directions define a new axis, we obtain 2^{d-1} new axes, each of which is essentially orthogonal to all of the d principal coordinate axes! Thus, in effect, high-dimensional space has an exponential number of orthogonal "axes." A consequence of this strange property of high-dimensional space is that if there is a point or a group of points, say a cluster of interest, near a diagonal, these points will get projected into the origin and will not be visible in lower dimensional projections.

6.6 DENSITY OF THE MULTIVARIATE NORMAL

Let us consider how, for the standard multivariate normal distribution, the density of points around the mean changes in d dimensions. In particular, consider the probability of a point being within a fraction $\alpha > 0$, of the peak density at the mean.

For a multivariate normal distribution [Eq. (2.33)], with $\boldsymbol{\mu} = \mathbf{0}_d$ (the d-dimensional zero vector), and $\boldsymbol{\Sigma} = \mathbf{I}_d$ (the $d \times d$ identity matrix), we have

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^d} \exp\left\{-\frac{\mathbf{x}^T\mathbf{x}}{2}\right\} \tag{6.11}$$

At the mean $\mu = \mathbf{0}_d$, the peak density is $f(\mathbf{0}_d) = \frac{1}{(\sqrt{2\pi})^d}$. Thus, the set of points $\mathbf{x}$ with density at least α fraction of the density at the mean, with $0 < \alpha < 1$, is given as

$$\frac{f(\mathbf{x})}{f(\mathbf{0})} \geq \alpha$$

which implies that

$$\exp\left\{-\frac{\mathbf{x}^T\mathbf{x}}{2}\right\} \geq \alpha$$

$$\text{or } \mathbf{x}^T\mathbf{x} \leq -2\ln(\alpha)$$

$$\text{and thus } \sum_{i=1}^{d}(x_i)^2 \leq -2\ln(\alpha) \tag{6.12}$$

It is known that if the random variables X_1, X_2, …, X_k are independent and identically distributed, and if each variable has a standard normal distribution, then their squared sum $X^2 + X_2^2 + \cdots + X_k^2$ follows a χ^2 distribution with k degrees of freedom, denoted as χ_k^2. Because the projection of the standard multivariate normal onto any attribute X_j is a standard univariate normal, we conclude that $\mathbf{x}^T\mathbf{x} = \sum_{i=1}^{d}(x_i)^2$ has a χ^2 distribution with d degrees of freedom. The probability that a point $\mathbf{x}$ is within α times the density at the mean can be computed from the χ_d^2 density function using Eq. (6.12),