

Chapter 3

High Dimensional Data Analysis

Last Modified: 2008-09-12 16:58

In data mining typically the data is very high dimensional, since 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 since hyperspace does not behave like the more familiar geometry in two or three dimensions.

3.1 High-Dimensional Objects

Let the data consists of d numeric dimensions or attributes $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^d$. Let $dom(\mathbf{x}^i)$ denote the domain of attribute \mathbf{x}^i , and let $l^i = \min\{dom(\mathbf{x}^i)\}$ and $u^i = \max\{dom(\mathbf{x}^i)\}$ denote the lower and upper bounds for attribute \mathbf{x}^i . The overall hyper-space is defined by the cross-product:

$$\prod_{i=1}^d [l^i, u^i] \subseteq \prod_{i=1}^d dom(\mathbf{x}^i) = \mathbb{R}^d \quad (3.1)$$

Hyper-rectangle and Hypercube In one dimension (1D) the space corresponds to a line segment. In two dimensions (2D) the space corresponds to a rectangle. In three dimensions (3D) the space is a cuboid. In higher dimensions the space corresponds to an *orthotope*, or a *hyper-rectangle*. A special case of the hyper-rectangle is a *hypercube*, which has all sides of length l , defined as

$$H_d(l) = \left\{ \mathbf{x} = (x^1, x^2, \dots, x^d) \mid \forall i, x^i \in [-l/2, l/2] \right\} \quad (3.2)$$

Thus $H_2(l)$ represents a square in 2D space, whereas $H_3(l)$ represents a cube in 3D space. The unit hypercube has all sides of length $l = 1$, denoted as $H_d(1)$.

Hypersphere A hypersphere is a generalization of the 3D sphere to higher dimensions. A sphere is given by the equation $x^2 + y^2 + z^2 = r^2$ (assuming it is centered at the origin). Note that in 2D the analog is a circle with the equation $x^2 + y^2 = r^2$.

In general, a hypersphere in d dimensions, with radius r , is defined as:

$$S_d(r) = \left\{ \mathbf{x} = (x^1, x^2, \dots, x^d) \mid \sum_{i=1}^d (x^i)^2 = r^2 \right\} \quad (3.3)$$

This equation describes the boundary or surface of the hypersphere, and as such it is a $d - 1$ dimensional surface, though it is a 3D object, since it can only be embedded only in 3D space. The solid hypersphere, also called a **closed hyperball**, includes all interior points as well, and is given as:

$$B_d(r) = \left\{ \mathbf{x} = (x^1, x^2, \dots, x^d) \mid \sum_{i=1}^d (x^i)^2 \leq r^2 \right\} \quad (3.4)$$

3.2 High Dimensional Volumes

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

$$\boxed{\mathbb{V}(H_d(l)) = l^d} \quad (3.5)$$

The volume of the hyperball and hypersphere are identical, since the volume measures the total “content” of the object, including all internal space.

To determine the general equation for the volume of a hypersphere in d dimensions, let us first look at lower dimensions first.

$$\mathbb{V}(S_1(r)) = 2r \quad (3.6)$$

$$\mathbb{V}(S_2(r)) = \pi r^2 \quad (3.7)$$

$$\mathbb{V}(S_3(r)) = \frac{4}{3}\pi r^3 \quad (3.8)$$

The general form for a hypersphere in d dimensions is given as

$$\mathbb{V}(S_d(r)) = K_d r^d \quad (3.9)$$

where K_d is a constant that depends on the dimensionality d . The equation for the volume of the hypersphere with radius r is then given as

$$\boxed{\mathbb{V}(S_d(r)) = \left(\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} \right) r^d} \quad (3.10)$$

where the Gamma function Γ was first encountered in (2.9):

$$\Gamma(\alpha > 0) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad (3.11)$$

By direct integration of (3.11), we have

$$\Gamma(1) = 1 \quad \text{and} \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \quad (3.12)$$

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

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \quad (3.13)$$

Combined with (3.12), for any integer $n \geq 1$, we have

$$\Gamma(n) = (n - 1)! \quad (3.14)$$

Turning our attention back to (3.10), when d is even, then $\frac{d}{2} + 1$ is an integer, and by (3.14) we have

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

and when d is odd, then by (3.13) and (3.12), 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:

$$d!! = \begin{cases} 1 & \text{if } d = 0 \text{ or } d = 1 \\ d \cdot (d-2)!! & \text{if } d \geq 2 \end{cases} \quad (3.15)$$

Putting it all together we have

$$\Gamma\left(\frac{d}{2} + 1\right) = \begin{cases} \left(\frac{d}{2}\right)! & \text{if } d \text{ is even} \\ \sqrt{\pi} \left(\frac{d!!}{2^{(d+1)/2}}\right) & \text{if } d \text{ is odd} \end{cases} \quad (3.16)$$

Plugging in the values of d in (3.10) gives us the volume of the hypersphere in different dimensions. For example, the volume of the 1-dimensional hyper-sphere/ball with radius r , which is simply the closed interval $[-r, r]$ is $2r$. The volume of the 2-dimensional hyper-sphere/ball, which is simply a circle with radius r , is πr^2 . The volume of the usual 3-dimensional ball or sphere is $\frac{4}{3}\pi r^3$.

Surface Area We can also compute the *surface area* of the hypersphere, which is given as:

$$\mathbb{A}(S_d(r)) = \frac{d}{dr} \mathbb{V}(S_d(r)) = \left(\frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)}\right) dr^{d-1} = \left(\frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}\right) r^{d-1} \quad (3.17)$$

We can quickly verify that for 2D, the surface area of a circle is given as $2\pi r$, and in 3D the surface area of sphere is given as $4\pi r^2$.

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 $r = 1$,

$$\lim_{d \rightarrow \infty} \mathbb{V}(S_d(1)) = \lim_{d \rightarrow \infty} \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)} \rightarrow 0 \quad (3.18)$$

In fact (3.10) achieves its highest volume for $d = 5$ when $\mathbb{V}(S_5(1)) = 5.257$.

3.3 Hyperspheres Inscribed within Hypercubes

We next look at the space enclosed within the largest hypersphere that can be accommodated within a hypercube (which represents the data-space). 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.

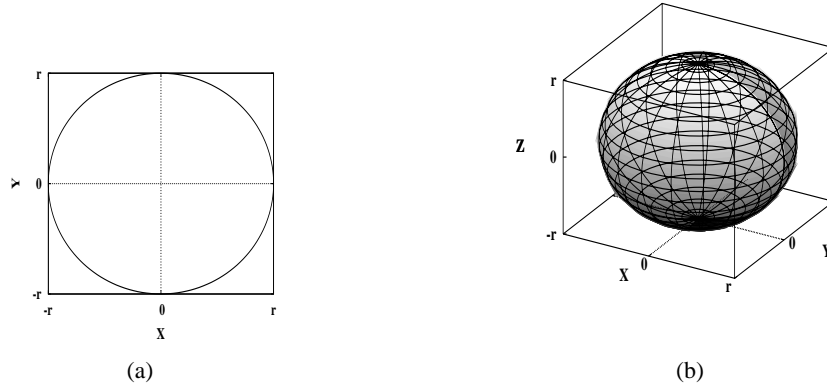


Figure 3.1: Hypersphere inscribed inside a hypercube in (a) 2D, (b) 3D

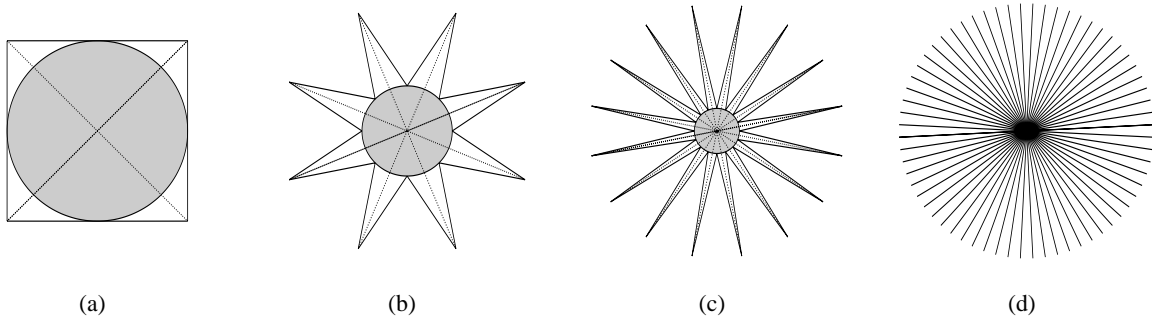


Figure 3.2: Conceptual View of High Dimensional Space in (a) 2D, (b) 3D, (c) 4D, and (d) Higher Dimensions. In d dimensions there are 2^d “corners” and 2^{d-1} diagonals.

In 2 Dimensions

$$\frac{\mathbb{V}(S_2(r))}{\mathbb{V}(H_2(2r))} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} \approx 75\% \quad (3.19)$$

Thus an inscribed circle occupies $\frac{\pi}{4}$ of the volume of its enclosing square, as illustrated in Figure 3.1(a).

In 3 Dimensions

$$\frac{\mathbb{V}(S_3(r))}{\mathbb{V}(H_3(2r))} = \frac{\frac{4}{3}\pi r^3}{8r^3} = \frac{\pi}{6} \approx 50\% \quad (3.20)$$

An inscribed sphere takes up only $\frac{\pi}{6}$ of the volume of its enclosing cube, as shown in Figure 3.1, which is quite a sharp decrease over the 2D case.

In d Dimensions As the dimensionality d increases asymptotically, we get:

$$\boxed{\lim_{d \rightarrow \infty} \frac{\mathbb{V}(S_d(r))}{\mathbb{V}(H_d(2r))} = \lim_{d \rightarrow \infty} \frac{\pi^{d/2}}{2^d \Gamma(\frac{d}{2} + 1)} \rightarrow 0} \quad (3.21)$$

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 emerges is that high-dimensional space looks like a rolled-up porcupine, as illustrated in Figure 3.2.

3.4 Volume of Thin Hypersphere Shell

Let us now consider the volume of a thin hypersphere shell of width ϵ along the surface of a hypersphere of radius r . In other words, the volume of the thin shell is given as the difference between the hyperspheres of radius r and $r - \epsilon$, as illustrated in Figure 3.3.

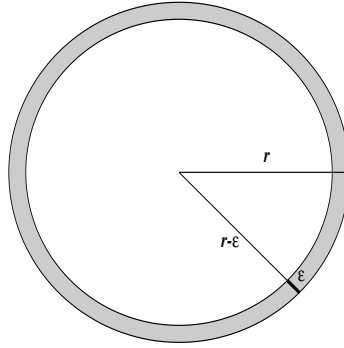


Figure 3.3: Volume of the Thin Shell: Comparing $\mathbb{V}(S_d(r - \epsilon))$ to $\mathbb{V}(S_d(r))$ for small $\epsilon > 0$

Let $S_d(r, \epsilon)$ denote the thin hypershell of width ϵ within the radius r . It’s volume is then given as

$$\boxed{\mathbb{V}(S_d(r, \epsilon)) = \mathbb{V}(S_d(r)) - \mathbb{V}(S_d(r - \epsilon)) = K_d r^d - K_d (r - \epsilon)^d.} \quad (3.22)$$

where $K_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)}$ from (3.10).

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

$$\frac{\mathbb{V}(S_d(r, \epsilon))}{\mathbb{V}(S_d(r))} = \frac{K_d r^d - K_d (r - \epsilon)^d}{K_d r^d} = 1 - \left(1 - \frac{\epsilon}{r}\right)^d \quad (3.23)$$

For example, for a circle in two-dimensions, with $r = 1$ and $\epsilon = 0.01$ the volume of the thin shell is $1 - (0.99)^2 = 0.0199 \approx 2\%$. 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 $1 - (0.99)^3 = 0.0297 \approx 3\%$, which is still a relatively small fraction. However, as d increases, in the limit we obtain:

$$\boxed{\lim_{d \rightarrow \infty} \frac{\mathbb{V}(S_d(r, \epsilon))}{\mathbb{V}(S_d(r))} = \lim_{d \rightarrow \infty} 1 - \left(1 - \frac{\epsilon}{r}\right)^d \rightarrow 1} \quad (3.24)$$

This means that in high dimensional spaces most of the volume is concentrated around the surface (within ϵ) of the hypersphere, and the center is essentially void. One consequence of this phenomena is that for any point p , if we assume that the space is centered on p , almost all other points will appear to be equally far from it, since they are all scattered in the thin surface shell.

3.5 Diagonals in Hyperspace

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

Consider the angle θ_d between the diagonal vector $\mathbf{1}$ and the first axis \mathbf{e} in d dimensions. In two dimensions, we have $\cos(\theta_2) = \frac{1}{\sqrt{2}}$ whereas in three dimensions, we have $\cos(\theta_3) = \frac{1}{\sqrt{3}}$. See Figure 3.4 for an illustration of the diagonal vector and θ_d for $d = 2$ and $d = 3$.

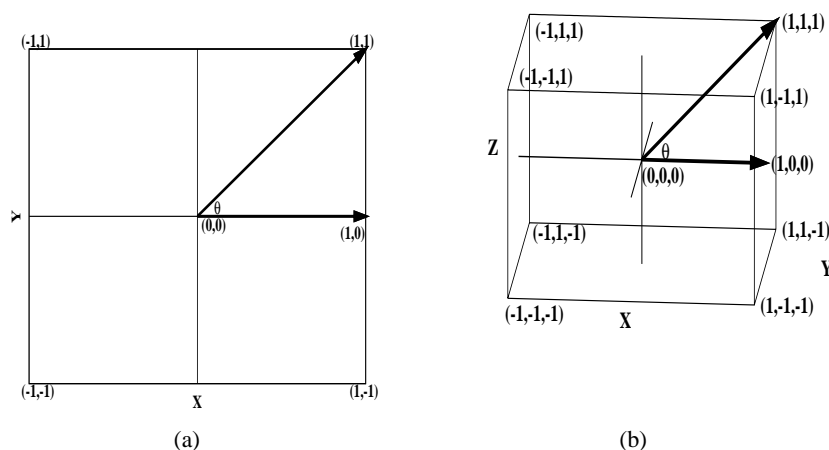


Figure 3.4: Angle between diagonal $\mathbf{1}_d$ vector and \mathbf{e}_d^1 in (a) 2D, (b) 3D

In general in d dimensions, for the angle between the d -dimensional diagonal vector $\mathbf{1}$ and the first axis \mathbf{e} is given as

$$\cos(\theta_d) = \frac{\mathbf{e}^T \mathbf{1}}{\|\mathbf{e}\| \|\mathbf{1}\|} = \frac{\mathbf{e}^T \mathbf{1}}{\sqrt{\mathbf{e}^T \mathbf{e}} \sqrt{\mathbf{1}^T \mathbf{1}}} = \frac{1}{\sqrt{1} \sqrt{d}} = \frac{1}{\sqrt{d}} \quad (3.25)$$

As d increases, we get

$$\lim_{d \rightarrow \infty} \cos(\theta_d) = \lim_{d \rightarrow \infty} \frac{1}{\sqrt{d}} \rightarrow 0 \implies \lim_{d \rightarrow \infty} \theta_d \rightarrow \frac{\pi}{2} = 90^\circ \quad (3.26)$$

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. This implies that in high dimensions all of the diagonal vectors are almost perpendicular (or orthogonal) to all the axes! Since there are 2^d corners in a d -dimensional hyperspace, there are 2^d diagonal vectors from the origin to each of the corners. Since 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!

3.6 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 ratio $\alpha > 0$, of the density at the mean.

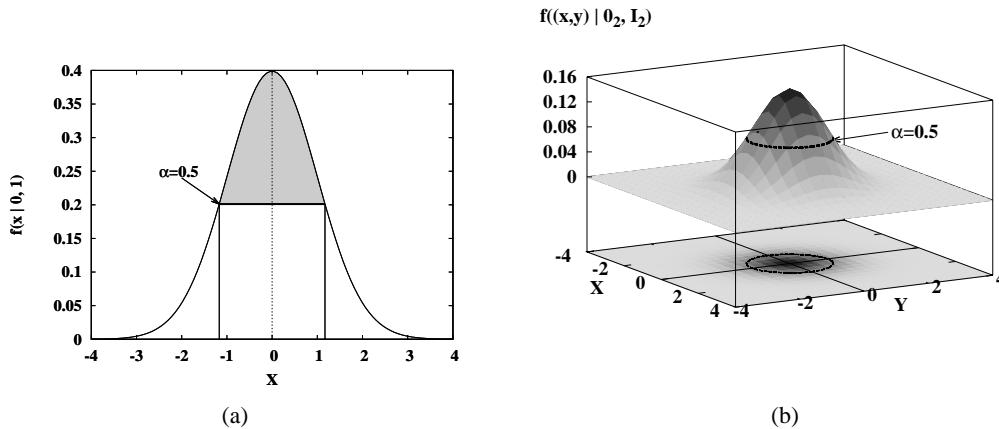


Figure 3.5: Probability of α Fraction of the Density at Mean in (a) 1D, (b) 2D

From (1.20), when $\boldsymbol{\mu} = \mathbf{0}_d$ (the d dimensional zero vector), and when $\boldsymbol{\Sigma} = \mathbf{I}_d$ (the $d \times d$ identity matrix), we have:

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{\mathbf{x}^T \mathbf{x}}{2}} \quad (3.27)$$

At the mean $\boldsymbol{\mu} = \mathbf{0}_d$, the density is $f(\mathbf{0}_d) = \frac{1}{(\sqrt{2\pi})^d}$. We want to consider the value \mathbf{x} where the density is α times the density at the mean, i.e., we want to solve for \mathbf{x} , where $\frac{f(\mathbf{x})}{f(\mathbf{0})} = \alpha$. From (3.27) we have

$$\frac{f(\mathbf{x})}{f(\mathbf{0})} = e^{-\frac{\mathbf{x}^T \mathbf{x}}{2}}$$

which gives us

$$\ln\left(\frac{f(\mathbf{x})}{f(\mathbf{0})}\right) = -\frac{\mathbf{x}^T \mathbf{x}}{2}$$

which in turn implies that

$$-2 \ln\left(\frac{f(\mathbf{x})}{f(\mathbf{0})}\right) = \mathbf{x}^T \mathbf{x} = \sum_{i=1}^d (x^i)^2 \quad (3.28)$$

$$-2 \ln(\alpha) = \sum_{i=1}^d (x^i)^2 \quad (3.29)$$

THEOREM: If the random variables $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k$ are independent and identically distributed and if each variable has a standard normal distribution, then their squared sum $\mathbf{y}_1^2 + \mathbf{y}_2^2 + \dots + \mathbf{y}_k^2$ has a χ^2 distribution with k degrees of freedom.

The projection of the standard multivariate normal onto any attribute \mathbf{x}^i is a standard univariate normal. Based on the above theorem, we immediately have that $\sum_{i=1}^d (\mathbf{x}^i)^2$ follows a χ^2 distribution with d degrees of freedom. In other words, the probability that a point is within the α times the density at the mean, given as $P\left(\frac{f(\mathbf{x})}{f(\mathbf{0})} \geq \alpha\right)$, is the same as the probability $P(\chi^2(d) \leq -2\ln(\alpha))$ using (3.29). Thus we have:

$$P\left(\frac{f(\mathbf{x})}{f(\mathbf{0})} \geq \alpha\right) = P(\chi^2(d) \leq -2\ln(\alpha)) = \int_0^{-2\ln(\alpha)} f(\chi^2|q=d) \quad (3.30)$$

where $f(\chi^2|q=d)$ is the chi-squared distribution with d degrees of freedom, given in (2.8).

Let us consider the probability of a point being within 1% of the density at the mean, i.e., when $\alpha = 0.01$. From (3.30) we have

$$P\left(\frac{f(\mathbf{x})}{f(\mathbf{0})} \geq 0.01\right) = P(\chi^2(d) \leq -2\ln(0.01)) = P(\chi^2(d) \leq 9.21)$$

We can now get the probability of a point being within 1% of the mean density by evaluating the χ^2 distribution for different degrees of freedom (the number of dimensions). For $d = 1$, we find that the probability is $P(\chi^2(1) \leq 9.21) = 99.7593\%$. For $d = 2$ the probability decreases slightly to $P(\chi^2(2) \leq 9.21) = 98.999\%$. Figure 3.6 show the probability for higher dimensions.

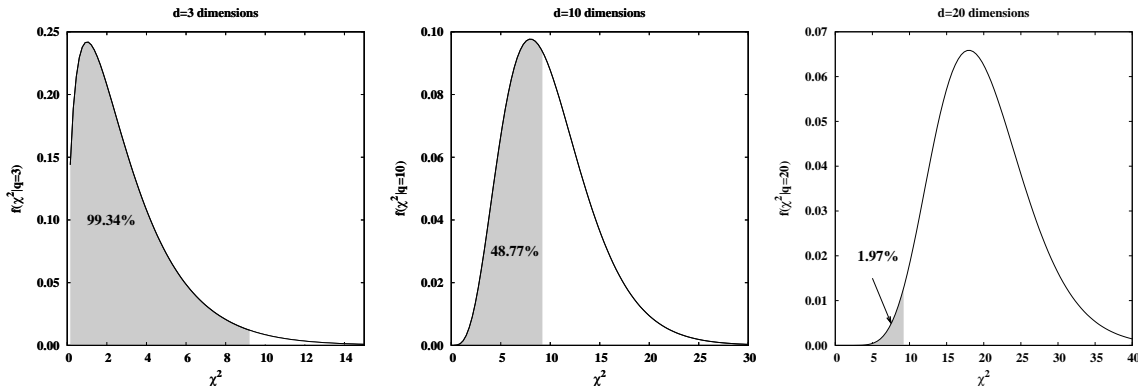


Figure 3.6: Probability that $\chi^2 < 9.21$ for Increasing Dimensions

We can observe that as dimensionality increases, this probability decreases sharply, and eventually tends to zero,

$$\lim_{d \rightarrow \infty} P(\chi^2(d) \leq -2\ln(\alpha)) \rightarrow 0 \quad (3.31)$$

In other words in higher dimensions the probability mass around the mean decreases significantly, as opposed to the 1D, 2D or 3D cases.