Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2015, Article ID 467195, 17 pages
http://dx.doi.org/10.1155/2015/467195

Research Article

Determination of System Dimensionality from Observing Near-Normal Distributions

Shahid Razzaq and Shehzad Khalid

Department of Computer Engineering, Bahria University, Islamabad 44000, Pakistan

Correspondence should be addressed to Shehzad Khalid; shehzad_khalid@hotmail.com

Received 17 June 2014; Revised 31 August 2014; Accepted 4 September 2014

Academic Editor: Sher Afzal Khan

This paper identifies a previously undiscovered behavior of uniformly distributed data points or vectors in high dimensional ellipsoidal models. Such models give near normal distributions for each of its dimensions. Converse of this may also be true; that is, for a normal-like distribution of an observed variable, it is possible that the distribution is a result of uniform distribution of data points in a high dimensional ellipsoidal model, to which the observed variable belongs. Given the currently held notion of normal distributions, this new behavior raises many interesting questions. This paper also attempts to answer some of those questions. We cover both volume based (filled) and surface based (shell) ellipsoidal models. The phenomenon is demonstrated using statistical as well as mathematical approaches. We also show that the dimensionality of the latent model, that is, the number of hidden variables in a system, can be calculated from the observed distribution. We call the new distribution "*Tanazur*" and show through experiments that it is at least observed in one real world scenario, that of the motion of particles in an ideal gas. We show that the Maxwell-Boltzmann distribution of particle speeds can be explained on the basis of Tanazur distributions.

1. Introduction

Probability theory has acquired a special status in statistics as it is essential to many real life applications involving quantitative analysis of large sets of data. Probability density function (pdf) is a function that describes the probability of a random variable taking certain values. Certain pdf occurs frequently in statistics as they can model many natural or physical processes and hence has acquired significant importance in probability theory. Some of these prominent continuous probability distributions include uniform, Laplacian, normal, gamma, and beta distributions.

Uniform distribution is a rectangular distribution where each observation has equal probability of occurrence. It is majorly used in generation of pseudorandom numbers in various simulation experiments. Laplace distribution is a double exponential distribution and is computed in terms of absolute distance of observation from mean instead of squared distance as in the case of normal distribution. The normal or Gaussian distribution is considered to be the most widely observable and prominent distribution in statistic that is used in variety of disciplines including social sciences, statistics, machine learning, data mining, simulation and modeling, and natural sciences. According to [1], this prominence of normal distribution is due to two reasons. First, it is very easy to analytically control the normal distribution as substantial results involving normal distribution can be derived in explicit form. Secondly, the normal distribution has its basis in central limit theorem, which states that, under mild conditions, the sum of a large number of random variables drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution.

There is an interesting relationship between distribution of the data points and their vector components. It has been observed that the uniform distribution of data point vectors in high dimensional ellipsoidal models give near normal distribution for the vector components. As the number of vector components increases, the generated density distribution would get flatter and the observed distribution becomes a complete uniform distribution in the presence of actual number of vector components. This phenomenon gives rise to an important question of whether the reverse of this phenomenon is possible. This implies that, given a normal distribution of data points with certain number of observable dimensions, we can predict the dimensionality of the parent model with the assumption that the parent model with the assumption that the parent model is ellipsoidal and exhibits uniform data point distribution. This paper attempts to answer this question for volume based and surface based ellipsoidal models. We further target to show that the dimensionality of the latent model, that is, the number of hidden variables in a system, can be calculated from the observed distribution.

The remainder of this paper is organized as follows. Section 2 presents a brief review of frequently occurring probability density function (pdf) and their application in modeling various physical/natural processes. In Section 3, Monte Carlo method is used to show how uniform distributions in ellipsoidal models give near normal distributions for single variables. Furthermore, the mathematical basis of the new distributions is discussed. Section 4 presents a method for determining the dimensionality of a latent (uniformly distributed) ellipsoidal model from any observed near normal distribution. Section 5 demonstrates via experiments that the new distribution is observed in real world scenarios and that other distributions can be explained on the basis of this new distribution. The last section summarizes the finding and gives direction for the future work.

2. Background

The subject of probability theory has gained significant importance as it is the foundation on which all the statistics are generated. It becomes a basis of modeling anything that can be considered as a random process. The variety of commonly occurring probability density distributions exists in literature. The difference between two i.i.d. exponential random variables is governed by a Laplace distribution. Various applications of Laplace distributions include signal processing [[2], speech recognition [3]], credit risks in finance engineering [4], and Kalman Filter [5–11]. The Laplacian of Gaussian distribution has been applied in spectral theory [12, 13], eigenspace decomposition [14], and so forth.

Normal distribution is a very commonly observable distribution which can be perceived as a function that tells the probability of data point falling between any two real limits. The observation from a normal distribution tends to pile up around a particular value, referred to as mean, instead of spreading uniformly in the state space R, thus having a symmetric distribution about its mean. The normal distribution is usually denoted by $N(\mu, \sigma^2)$ [15] where μ and σ are the mean and standard deviation, respectively. It has an attractive capacity of generating simple models for complex real life phenomena to a relatively good degree of accuracy. Normal distribution has been applied in variety of fields. In data mining, normal distribution has been excessively used for clustering, modeling, classification, and novelty detection. Multivariate Gaussian [16–18] and Gaussian Mixture Models [13, 19–22] are well-known statistical models for modeling and classification of variety of data. Normal distribution has also been widely used for novelty detection [16, 22–24]. One of the important reasons of dominance of normal distribution is its basis in central limit theorem which explains the ubiquitous occurrence of the normal distribution in nature. A central limit theorem is based on any theory from a set of weak-convergence theories [25]. They all express the fact that a sum of many independent and identically distributed (i.i.d.) random variables having finite variance will tend to be distributed according to normal distribution. Central limit theorem and in turn normal distribution has its wide application in sampling. Other applications and characterization of normal distribution have been discussed in detail by [26, 27].

The analysis of low dimensional projections of higher dimensional distributions was done by Sudakov [28]. It was observed that uniform distribution of data points in high dimensional convex bodies gives near normal distributions in lower dimensions [29–31]. Building on their work, we experiment with the reverse, that is, determination of dimensionality of the original (uniform) distributed model, from the observation of its projections in lower dimensions. As a case study, we apply the concept to ideal gasses, that of determining the number of particles in the system by observing the speed distribution of its particles.

In this paper, we present a new kind of distribution as an alternative to the Normal Distribution. The advantages of the new distribution are

(i) for a given system representing the bounds of the observed variable. Unlike normal distribution, the new distribution restricts the range of observed variable(s) according to system's model,

(ii) using the interdependence of the model variables to explain the formation of observed distributions,

(iii) allowing the number of hidden system variables (dimensions) to be determined from the observed distributions,

(iv) having backward compatibility with the normal distribution (for medium to high model dimensionality) in characteristics other than those mentioned above.

3. Uniform Distribution in Ellipsoidal Models

Before a formal discussion can be carried out on the subject, some terms need to be identified. Model here refers to mathematical descriptions of systems. Systems can range anywhere from physical system as in the ones found in physics, biology, metrology, and so forth to computational ones as in computer science and simulations. Ellipsoidal models refer to systems which can be modeled by mathematical equations that describe an ellipse. The equation below represents an n-dimensional ellipse:

$$\sum_{i=1}^{n} \frac{x_i^2}{a_i^2} = 1. \tag{1}$$

Each of the x_i variables in the system takes up a dimension in the ellipsoidal model. The maximum value of a dimension is bounded by the value of radius a_i of that dimension. A special case of the ellipsoidal model is the spherical model where all dimensions of the system have the same maximum value (i.e., radius). This spherical model is represented below with r as the said radius:

$$\sum_{i=1}^{n} \frac{x_i^2}{r^2} = 1 \quad \text{OR} \quad \sum_{i=1}^{n} x_i^2 = r^2. \tag{2}$$

The given ellipsoidal models, whether sphere of ellipse in shape, represent a system. For the n dimensions or variables in the system, any combination which satisfies the given equation is said to constitute a data point vector in the ellipsoidal model. The system is therefore defined by the set of variables and the limiting condition (in this case the radius).

The ellipsoidal models represented above are one of two types of models possible. The one presented above, owing to the fact that they are represented by a mathematical equation (as opposed to an inequality), is a surface based (shell) ellipsoidal model; that is, all data points take up positions on the outer surface of the n-dimensional ellipsoid. A second type of ellipsoidal model is the volumetric model, which constitutes regions both on the surface of the ellipsoid and on those inside. Therefore the whole volume of the ellipsoidal model is covered by the model inequality formula, which is given as

$$\sum_{i=1}^{n} \frac{x_i^2}{a_i^2} \le 1. \tag{3}$$

Uniform distribution in any of these models refers to the distribution, in n dimensions, of the data points such that they are distributed evenly across the model, with no preference for any particular region of the model.

What follows are an empirical analysis for discrete uniform distribution and a mathematical/theoretical analysis for continuous uniform distribution of data points in the model. Empirical analysis primarily consists of generating statistically uniform n-dimensional model data using random variables. This method is also known as the Monte Carlo method. This empirical method is supplemented with another approach, that is, that of generating discrete evenly spaced data points. For both cases, the observed data distribution in lower dimensions is compared to the original (uniform) one. The theoretical analysis of the observed behavior consists of using mathematical equations for n-dimensional volumetric as well as surface slices. After the logical analysis of the observed distributions, these new distribution curves are named.

3.1. Uniform Distribution with Monte Carlo Method. Uniform distribution of data points in an n-dimensional model can be approximated using Monte Carlo method. Generally, the number of possible combinations of variables in a model increases exponentially with the number of dimensions of the model. Monte Carlo method has the advantage of being saleable for higher dimensions while roughly preserving uniformity of data point distribution. Using this method, random values are generated for each of the n dimensions such that the combined set of values satisfies the conditions of the model. For a volumetric ellipsoidal model, this would be data points within the volume defined by the ellipsoid. Whereas, for surface based models, this would be data points on the surface only, with no data point on the inside. Generation of data points in this manner does not guarantee a completely uniform distribution, but, for the purpose of statistical inference, it is sufficient and, more importantly, scalable to higher dimensions.

We use Monte Carlo method on volumetric ellipsoidal models. Both the general case of ellipse models and the special case of sphere models are used. 100,000 data points are generated randomly for models of different dimensionality. Total of 11 different n-ball (n-dimensional spheres) and n-ellipse models are used. Of the n variables or dimensions of the model, the distribution of any one variable is observed in across the range of possible values for the dimension. The choice of variable out of the n possible options is arbitrary.

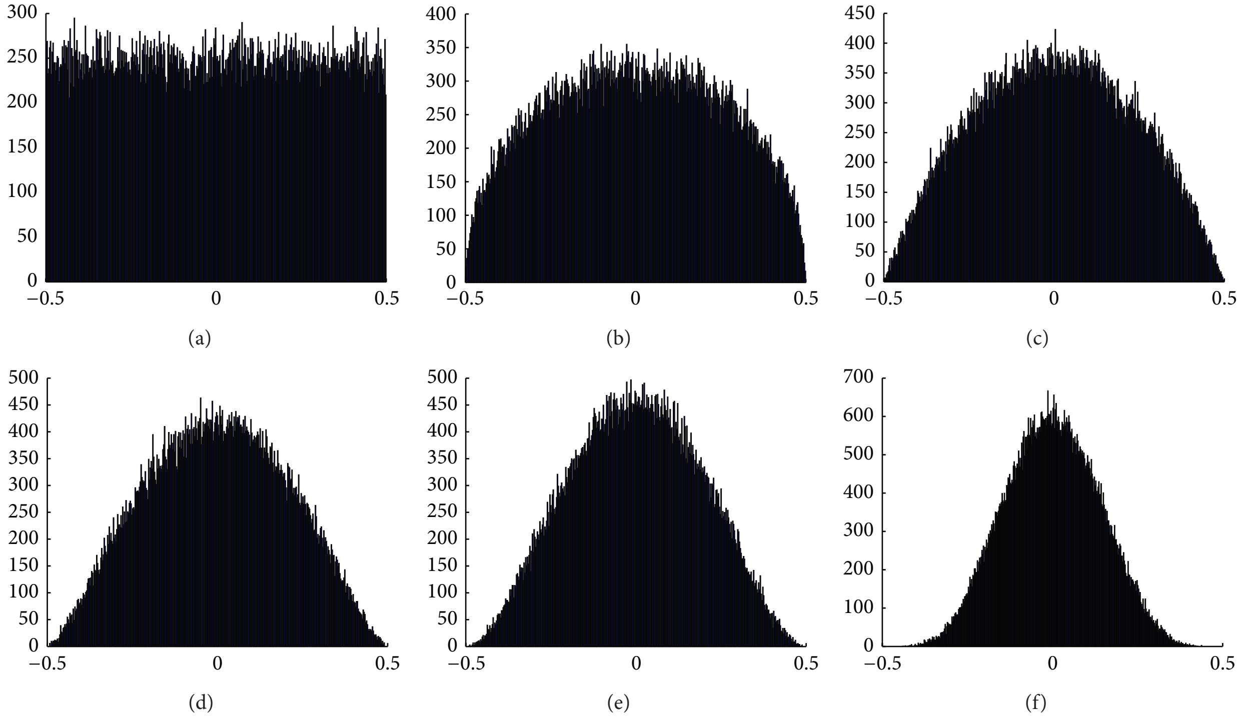
Equation for volumetric n-ball is

$$\sum_{i=1}^{n} x_i^2 \le r^2. \tag{4}$$

Equation for volumetric n-ellipse is specified in (3). Each data point of the models is an n-dimensional vector, with each component representing a dimension. The value of each vector component is generated from a computer based pseudorandom number generator. To avoid any bias in the generation and selection of vectors, values of all the constituent components are filled for a vector before it is tested for compliance using model constraints.

If the models are visualized in all n-dimensions, the distribution of vectors would be somewhat uniform across the model, with no particular concentration of vectors in any region of the model. If small pockets of greater vector density are found, that would be purely coincidental as the model constraints do not impose any such bias. On the other hand, if a subset of the dimensions is observed, the distribution of vectors does not stay uniform. For our case, we show the observation of distribution of a single vector component, that is, only one dimension, and the same dimension is observed across the 100,000 vectors (data points). This is analogous in the real world to observing a single variable, where the information of the other related variables is not known. These distributions are shown for different n-ball models in Figure 1. Similar process is repeated for n-dimensional ellipse models and the results shown in Figure 2. The radius a1 of the first dimension is kept at 0.5 to allow comparison with n-ball histograms, while the radius of the other dimension is varied as described in the description of Figure 2.

The first distribution, of the set of distributions in Figures 1 and 2, shows a more or less uniform distribution. There are spikes which is the concentration of data points in certain bins, but they are more or less arbitrary, with no particular bias towards any side or region. This is because all of the vector components of the 1-dimensional model (i.e., 1 of 1 vector component) are being observed. This is, however, not true for other distributions, where only 1 out of the n-dimensions of the model is being observed. When



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1: Distribution of a vector component for different n-ball models. The vector component can take a range of values between −0.5 (min) and +0.5 (max), and therefore the n-ball model has radius $r = 0.5$. The distributions represent the count of data points that lie in any of 400 bins dividing the [−0.5,0.5] range. ((a), (b), (c)) 1-ball, 2-ball, and 3-ball. ((d), (e), (f)) 4-ball, 5-ball, and 11-ball.

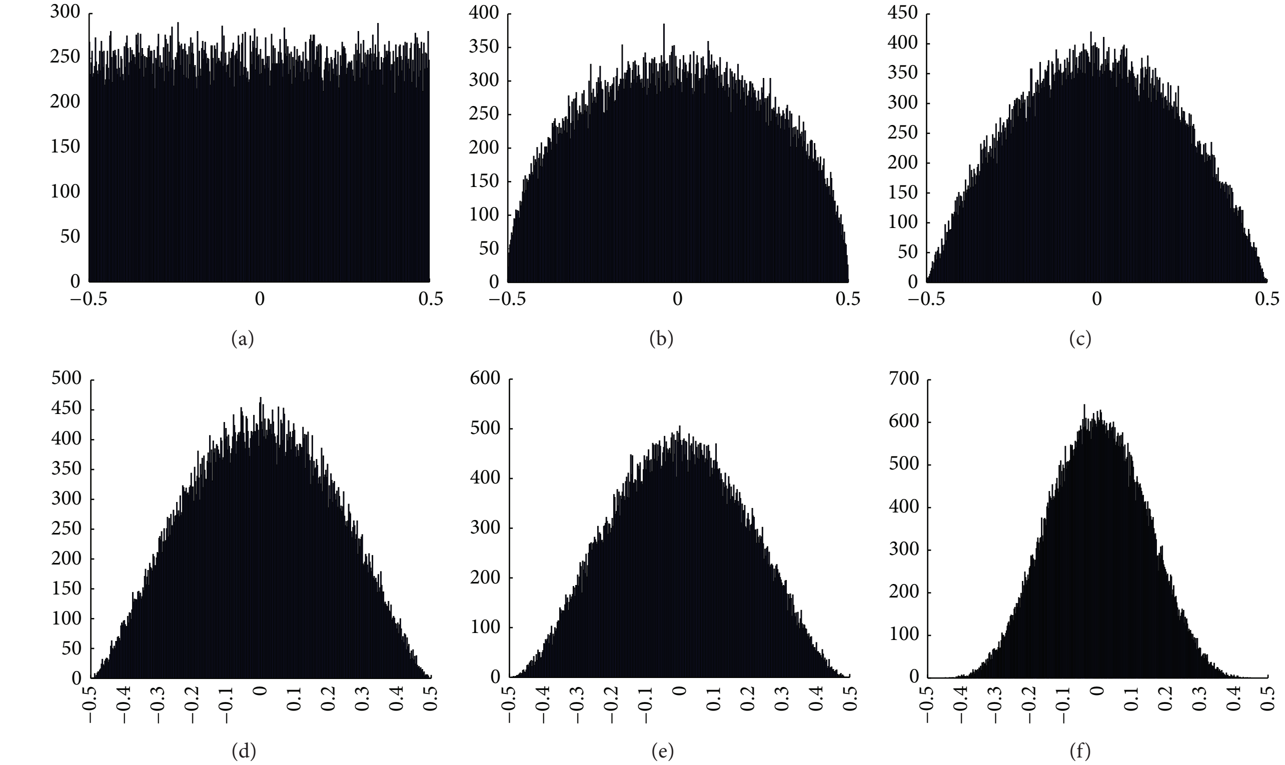
the difference of the actual versus observed dimensions, that is, the value of $(n - 1)$, is low, the observed distribution is not particularly interesting. But, for large values of $n - 1$, the distribution starts to have a striking resemblance with the normal distribution. As the dimensionality of the data models is increased, fewer data samples are recorded at the extremities of dimensions owing to the geometrical restriction of the model, as compared to the center. Therefore, few data points are found closer to −0.5 and +0.5, as compared to the center at 0.0, and result in a relatively high histogram bin count at the center, as compared to the two ends. The geometrical restrictions of the sphere and ellipse based model have more constricted at the extremities as compared to the center and result in the curve distributions as seen in Figures 1 and 2.

This phenomenon, however, interesting, is not exhibited for every model that is constricted at the extremities. N-cross polytope is a model which has decreased volume away from the center and is represented by the equation:

$$\sum_{i=1}^{n} |x_i| \le R. \tag{5}$$

Here, the sum of the values of the vector components is capped at a constant R. Figure 3 shows the distribution of vector components for n-cross-polytope models. It is clear that the observed distribution is different from the ones observed in Figures 1 and 2. In fact, the observed distribution of Figure 3 resembles that of the Laplacian distribution. It is also not surprising that the power of the Euler's constant in the equations representing the Laplacian distribution is linear, very much like the equation of the n-cross polytope. Nevertheless, this discussion is outside the scope of the current paper and the rest of the paper focuses on the comparison of distributions of vector components in ellipsoidal models with the normal distribution.

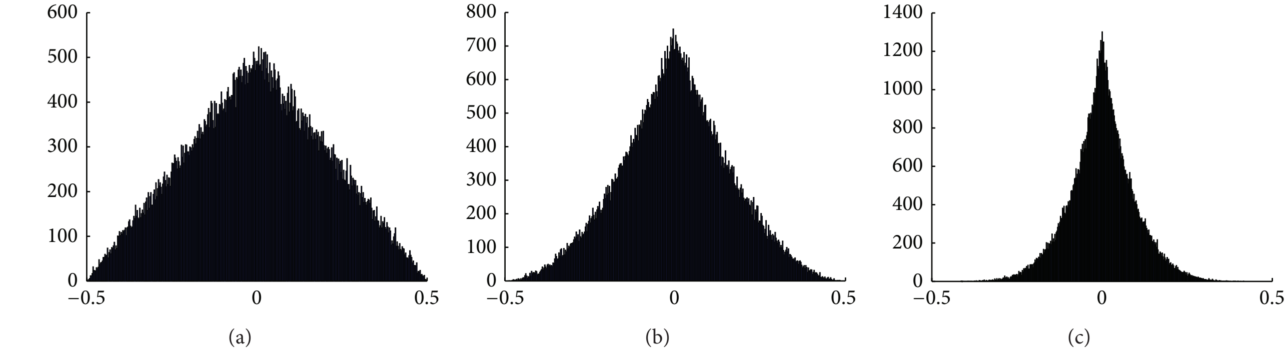
A visualization of the higher dimensions of the ellipsoidal models is not possible, but the same phenomenon can be observed visually in lower dimensions. Figure 4 helps explain the behavior in 2-dimensional sphere and ellipse models. Here, the model visualization shows uniform distribution when considering all (2) dimensions of the model but shows nonuniform distribution when observing a single dimension. The fact that the shape of the observed distribution is the same for both spheres and ellipses can appear nonintuitive but can be explained if scaling is considered. For the vertically elongated ellipsoidal model of Figure 4, the number of data points at the center is certainly higher than that of the corresponding sphere, but the ratio of the neighboring bin count remains the same for both sphere and ellipses. Therefore, for a 2D ellipse, with vertical radius twice that of the horizontal one, the bin count at the center is twice that at the center of sphere, but scaling vertically by 0.5 shows that there is no actual difference between the distributions. In other words, the ratio of the decrease in the number of data points for both models, as we move away from the center, is the same. Both models exhibit the same kind of distribution curve. This phenomenon is observed in higher dimensional models as



(a) (b) (c)

(d) (e) (f)

Figure 2: Distribution of a vector component for different n-ellipse models. The radius of the observed dimension (vector component) is $a_1 = 0.5$, as with n-ball models, again 400 bin histograms are used. The radii for the different dimensions are $a_1 = 0.5$, $a_2 = 1.0$, $a_3 = 0.25$, $a_4 = 1.5$, $a_5 = 0.16$, $a_6 = 2.0$, $a_7 = 0.13$, $a_8 = 2.5$, $a_9 = 0.1$, $a_{10} = 3.0$, and $a_{11} = 0.08$. ((a)–(f)) 1-ellipse, 2-ellipse, 3-ellipse, 4-ellipse, 5-ellipse, and 11-ellipse.



(a) (b) (c)

Figure 3: Distribution of a vector component for different n-cross polytope models. The value of constant R is set at 0.5 again using 400 bin histograms. The distribution is particularly different from the normal distribution. ((a) to (c)) 2-cross-polytope, 3-cross-polytope, and 6-cross-polytope.

well and results in the similarity of the distributions for sphere and ellipse based models.

Statistical inferences can be drawn from pseudouniform distributions generated in Monte Carlo methods. The plausibility of the inference, however, is dependent on the random number generator used for creating pseudouniform distributions. For a more robust analysis of the phenomenon, discrete uniform distributions need to be created instead. The brute force variable permutations give more uniform distribution in the discrete variable space but have the disadvantage of dimensionality explosion for high dimensional models. The number of data points generated in high dimensions grows exponentially and the task becomes intractable. Nevertheless, an analysis of manually generated discrete uniform distribution for relatively lower dimensions is given next.

3.2. *Discrete Uniform Distribution in Ellipsoidal Models.*

In order to obviate any statistical bias due to Monte Carlo method, particularly the pseudorandom number generator, discrete uniform distributions have been generated for the ellipsoidal model. This ensures that all regions in every dimension have the exact same data point density. The

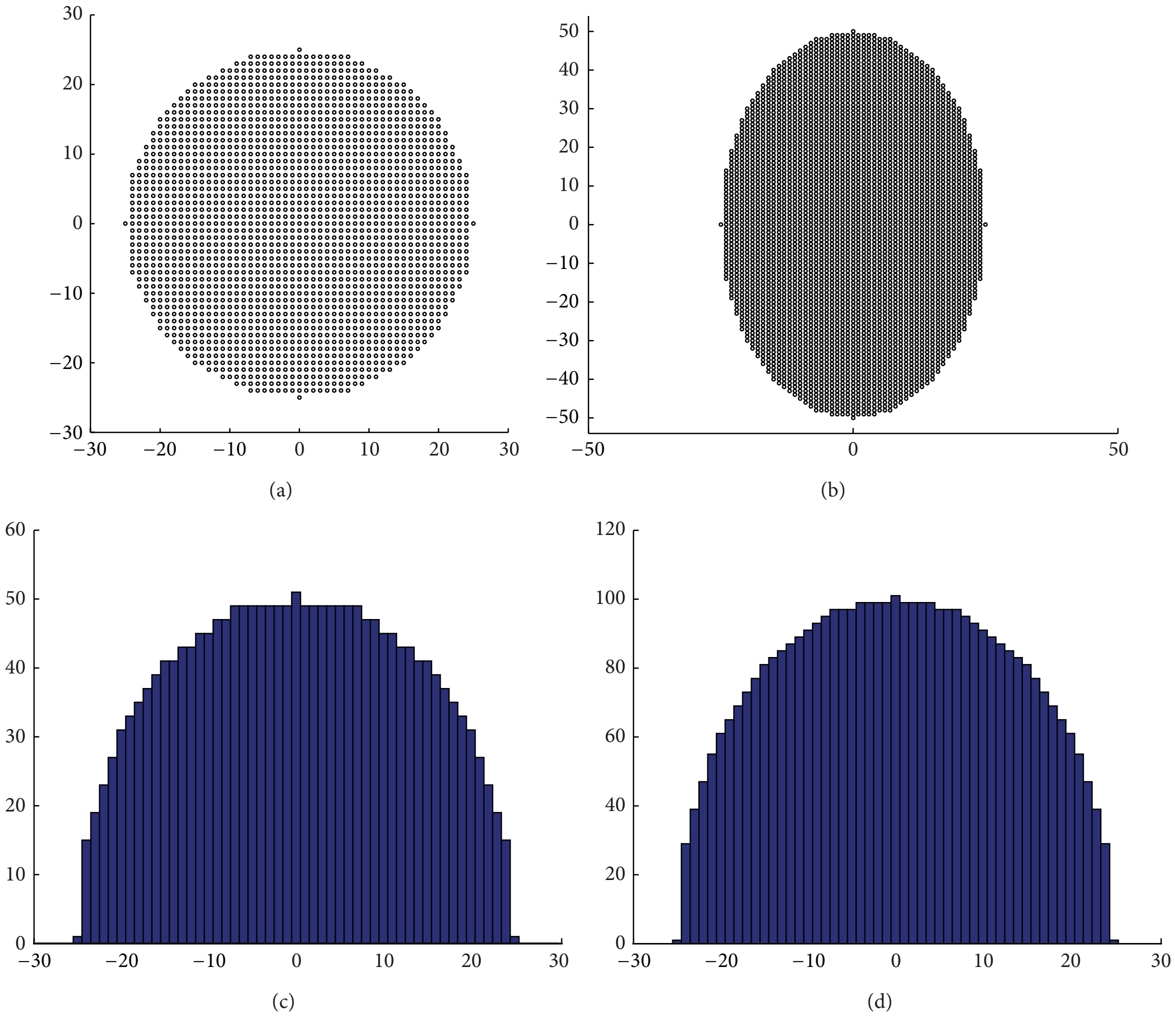


Figure 4: Two-dimensional plot of data points in a 2-dimensional sphere (a) and ellipse model (b). The distribution is of uniform type when observing all dimensions but becomes skewed when observing fewer (one) dimensions. The histogram of sphere (c) and ellipse (d) data points shows greater data point count at the center, that is, where the value of dimension is 0, as compared to the extremities. Histogram bins maintain the same relative ratio for both sphere and ellipse.

method of generating such data points is very simple. All permutations of the discrete values of the dimensions are tested for model conformity. The generated model is similar to the one seen in Figure 4.

Discrete uniform distributions were created for sphere and ellipse based models. Figure 5 shows the distribution curve when observing a single dimension. Both spherical and elliptical models in 1 dimension are represented by a line and can be seen as the uniform distribution (horizontal red line). The distribution curves of Figure 5 have been normalized in the vertical axis and therefore the range of values of the curve in the vertical axis is in the range [0,1.0]. This normalization allows for comparison between distribution curves generated for models of different dimensionality. The next curve after the uniform red line of the 1-dimensional model is the green curve of the 2D models (a 2D circle and a 2D ellipse). The distribution curve is no longer uniform, as the original uniform distribution was created in higher dimensions. The distribution curve for 3D sphere and 3D ellipse comes next in blue color. The curve at this point is still convex shaped, with no change in the sign of the 2nd differential of the curve. Distribution curves for the 4-dimensional spherical and elliptical models are the first to show sign of concavity. Direction of gradient change reverses twice in the curve and once in each half of the distribution. Moreover, the distribution curve for higher dimensions is narrower as compared to that for the lower dimensions. For subsequent models, from 5 to 7 dimensional spheres and ellipses, the resultant distribution curves are still narrower and increasingly give appearance of a normal distribution.

Let the number of dimensions, for which the data point distribution was being observed, be represented by k. For Figure 5, the value of $k = 1$ as only 1 of n total dimensions of the model is being observed. The choice of dimension being viewed is arbitrary and does not affect the shape of the distribution. The distribution curves record the density

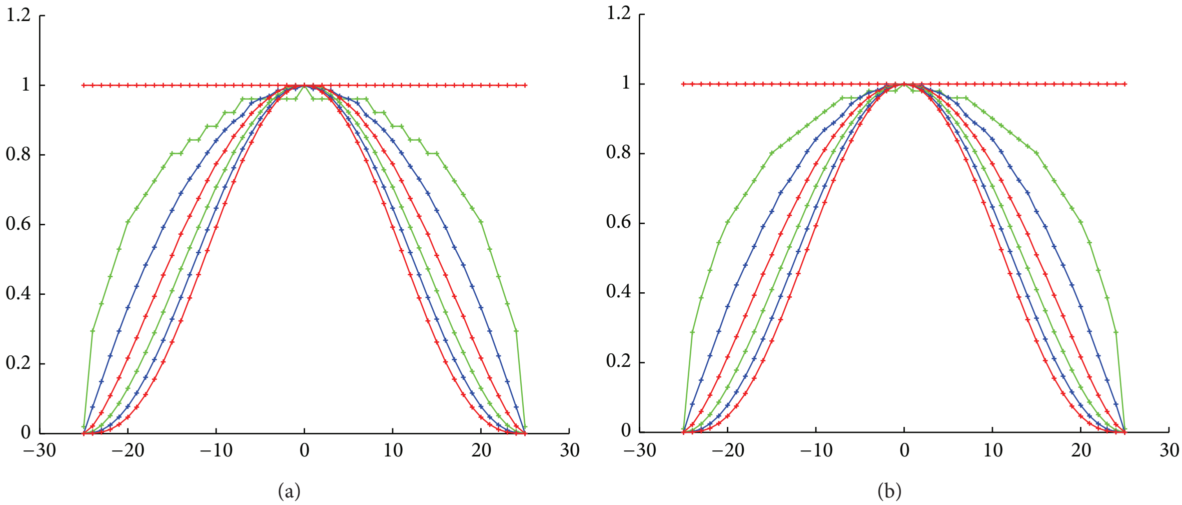


Figure 5: Distribution curves for one vector component ($k = 1$) in spherical (a) and elliptical (b) models. (a) Seven spherical models of different dimensionality, ranging from 1-ball to 7-ball, with radius $r = 25$. As the dimensionality of the model increases, the original uniform distribution of vectors is seen more like the normal distribution for its vector components. (b) Seven ellipsoidal models ranging from 1-ellipse to 7-ellipse. Here radii of the ellipse are $a_1 = 25$, $a_2 = 50$, $a_3 = 12.5$, $a_4 = 75$, $a_5 = 8.33$, $a_6 = 100$, $a_7 = 6.25$, $a_8 = 125$, $a_9 = 5$, $a_{10} = 150$, and $a_{11} = 4.16$.

of the data points as observed with respect to k of n total dimensions. If the value of k was gradually increased from 1 to n, the generated density distribution would become flatter and spread out further into the k dimensional distribution space. At $k = n$, the observed distribution would be a uniform distribution in k dimensions.

As discussed earlier, one drawback of the use of discrete uniform distribution is the dimensionality explosion for higher dimensions. For the given radius of the spheres and ellipses ($r = 25$ for Figure 5), the process of data point generation beyond 7 dimensions quickly becomes intractable. Moreover, with discrete variable values, the intermediate vector component values are not being modeled. This is handled next using mathematical analysis for continuous uniform distribution.

3.3. Analysis of Continuous Uniform Distribution in Ellipsoidal Models. For the analysis of the ellipsoidal modal distributions in the continuous domain, we have to use mathematical tools. The two types of models discussed earlier, namely, the surface based models and volume based models, rely on the concept of space for the data points. The greater the surface area or volume, the greater the number of distinct data points that can fit the space will be. In other words, the number of vectors, having their vector component value in a certain range (of the observed component), is directly proportional to the volumetric slice of the entire model in that range of the vector component. To get the complete distribution curve, consisting of different ranges of the observed vector component, we can integrate volumetric slices of the model over the range of a scalar component.

Table 1: Volume and surface area of n-ball spherical models as a function of the model dimensionality.

Model	1-ball (1D line)	2-ball (circle)	3-ball (sphere)	n-ball (nD sphere)
Volume	$2r$	πr^2	$4/3\pi r^3$	$C_{vn} r^n$
Surface area	—	$2\pi r$	$4\pi r^2$	$C_{sn} r^{n-1}$

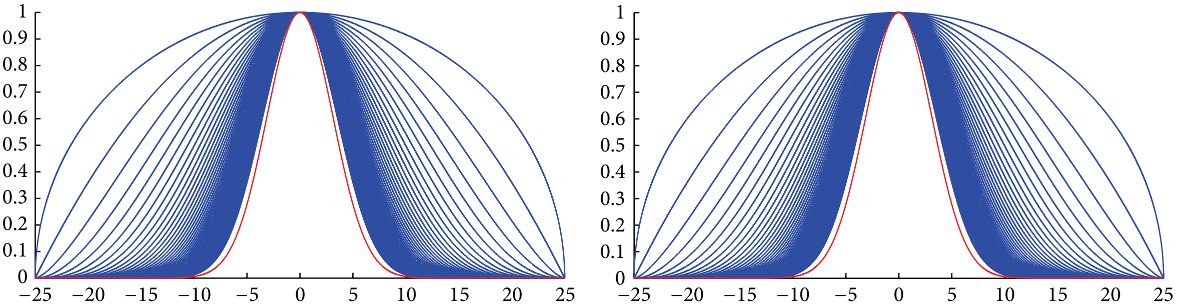
For a spherical model, the volume is proportional to the power of the radius and seen in Table 1. For simplification, we consider an n-ball model of unit radius ($r = 1$). The integration of volumetric slice over the scalar component is given as

$$\text{Volume} = \int_{-r}^{r} \text{Volumetric Slice} * dx. \tag{6}$$

Here, "x" is the observed vector component. The generic formula for a unit n-ball is

$$V_n = C_{vn} \cdot \int_{-1}^{1} \left(1 - x^2\right)^{(n-1)/2} dx. \tag{7}$$

The integrand ($(1 - x^2)^{(n-1)/2}$) represents the volumetric slice for an n-dimensional model, when it is integrated across one dimension x. The constant C_{vn} has different values for different n-balls, as seen in Table 1. As discussed earlier, we are interested in the volumetric ratios of different ranges and not in the absolute value of the volume. Also the distribution curves are normalized afterwards, which renders constants like C_{vn} irrelevant. Figure 6(a) shows the plot of



(a)

(b)

Figure 6: (a) Plot of the vertically normalized volumetric curve of 50 models in the range of 2-ball to 51-ball. Blue curves show the integrand $(1 - x^2)^{(n-1)/2}$ as a function of the vector component x and represent the distribution of data points as a function of the observed vector component x. (b) Plot of the vertically normalized surface area curve of 50 models in the range of 2-ball to 51-ball. Blue curves show the integrand $(1 - x^2)^{(n-2)/2}$ as a function of the vector component x. Notice the presence of flat line at $y = 1$.

the normalized volumetric curve or integrand $((1-x^2)^{(n-1)/2})$ as a function of the vector component x. The curves have been scaled across (along the dimension x) here by a factor of 25 to allow for curve comparisons with discrete uniform distribution curves of Figure 5.

The above integrands are for volume based ellipsoidal models, that is, models that allow data point vectors to be generated anywhere within the volume of the ellipsoid. Distribution curves for surface based ellipsoidal models need to be calculated as well. For a spherical model, the surface area is proportional to the power of the radius as seen in Table 1. The integration of surface slice over the scalar component is given as

$$\text{Surface Area } d = \int_{-r}^{r} \text{Surface Slice} * dx,$$
$$S_n = C_{sn} \cdot \int_{-1}^{1} \left(1 - x^2\right)^{(n-2)/2} dx. \tag{8}$$

The constant C_{sn} also has different values at different dimensions as seen in Table 1. Once again, C_{sn} can be ignored as we are interested in the ratio of the surface slices of spherical models. Moreover, vertical normalization of the surface area curve makes the distribution agnostic of multiplication with constants. Figure 6(b) shows the plot of the normalized surface area curve or integrand $(1-x^2)^{(n-2)/2}$ as a function of the vector component x. Again the curves have been scaled up horizontally by a factor of 25 for the purpose of comparison.

The distribution curves of the volumetric and surface based models look identical, except for a one dimensional shift. The outermost (semicircular) curve of the volumetric model corresponds to 2-ball, whereas the one for surface based models corresponds to 3-ball. The curve for 2-ball in surface based model is a flat line at $y = 1$. The shift of dimension is due to the difference in the exponent of the radius between the formulas representing volume and surface area. The curves are otherwise the same.

The above discussion and derivations are for a spherical model of unit radius. For spherical models having radius values other than 1, the corresponding distribution can be reached by multiplying the horizontal dimension by the said radius. This is equivalent to scaling the distribution curve horizontally. Similarly, instead of deriving separate equations for elliptical models, the same equations can be used along with horizontal scaling factors equal to the radius of the observed dimension.

3.4. Naming the New Distribution. Before proceeding, we give formal nomenclature for the observed distribution curves. The distributions observed above for the ellipsoidal models are a result of the projection of a higher dimensional uniform distribution, over a lower dimensional space. The distributions are projections of the actual (uniform) distribution and hence we give it a name "Tanazur" (pronounced "te-na-zer"), which is Urdu for perspective. Unlike the normal distributions, the Tanazur distributions are a set of distinct distribution curves. With normal distributions, scaling along the variable's axis gives curves for other standard deviation values. On the contrary, no scaling can equate Tanazur distributions of different dimensionality. This has been seen earlier with the inflection point positions.

For now, we represent univariate Tanazur Distribution as

$$X \sim T(r, n). \tag{9}$$

Here, n is the dimensionality of the ellipsoidal model and r is the radius along the observed dimension. A more formal description and representation of the distribution will be given in Section 4.

4. Model Dimensionality Determination from Vector Component's Distribution Curve

As we have seen, uniform distribution of data point vectors in n-dimensional ellipsoidal models (both volumetric and

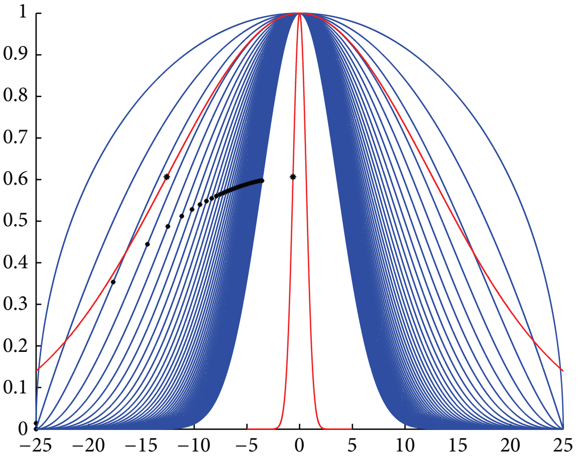


Figure 7: Inflection points of the curves shown with a black asterisk. Blue curves represent the Tanazur distributions whereas the red curves represent the normal distribution. Outermost Tanazur distribution is $T(25, 2)$ and innermost is $T(25, 51)$.

surface base) gives near normal distributions for the vector components. What we determine now is whether the reverse is possible; that is, given a normal looking distribution on a variable, can the dimensionality of the parent model be found, if the parent model is assumed to be ellipsoidal and exhibits uniform data point distribution. As an example, if the innermost blue curve of the two plots in Figure 6 was observed, the intended method should indicate that the dimensionality of the parent model is 51.

Before a discussion can be carried out for the dimensionality determination, some features of the distribution curves of Figure 6 need to be mentioned. All the distribution curves inside each plot of Figure 6 have their own characteristic shape. No two curves of the plot are the same, regardless of any scaling transform that is applied in the horizontal axis. The relative ratios of different sections of the curves are maintained even after scaling. In other words, changing a dimensional radius of a spherical model to form an elliptical model produces a distribution curve which can be scaled down along the observed dimension to yield the distribution of the spherical model. This is similar to the normal distribution where curves of different standard deviation are only (horizontally) scaled versions of each other. But, with Tanazur distribution, the characteristics of the curves corresponding to models of different dimensionality are different. No scaling can equate such curves. A geometrical measure is required for identification of these curves, which are shown in Figure 6 plots.

A geometrical feature, called the inflection point, is capable of identifying the different Tanazur distribution curves and is shown in Figure 7. Inflection points on a curve have the characteristic property that the second derivative of the curve at their particular location reaches 0. More accurately, the direction of gradient change switches directions (from +ive to −ive or vice versa). The vertical position of the inflection points of normalized distribution curves is always scale-invariant. In other words, the vertical position, as a percentage of the vertical length, does not change for a given curve regardless of scaling along the horizontal axis (i.e., change in dimensional radii of the model). The same can be said for the normal distribution. Figure 7 shows two different normal distribution curves corresponding to two different standard deviation values. As can be seen, the height of the inflection point of these red curves remains the same. The characteristic position of the inflection point of the normal distribution is approximately 60.65%, upwards from the horizontal axis. For a normal distribution centered on the origin, this point is calculated below:

$$\begin{gathered} f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \\ f''(x) = \frac{\left(e^{-x^2/2\sigma^2}\right) x^2 - \sigma^2}{\sigma^4\sqrt{2\pi}}. \end{gathered} \tag{10}$$

Value of x for which the normal distribution has 2nd derivative of 0 is

$$\begin{gathered} 0 = \frac{\left(e^{-x^2/2\sigma^2}\right) x^2 - \sigma^2}{\sigma^4\sqrt{2\pi}}, \\ 0 = \left(x^2 - \sigma^2\right), \\ x = \pm\sigma. \end{gathered} \tag{11}$$

Therefore, the inflection point of the normal distribution is at the 1st standard deviation from the mean. The value of the inflection point in the vertical axis is given by $f(\sigma)$. Consider

$$f(\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\sigma^2/2\sigma^2} = \frac{1}{\sigma\sqrt{2\pi e}}. \tag{12}$$

In order to arrive at the normalized vertical position of the inflection point, that is, as a percentage of the maximum

Table 2: I_y of $T(r, n)$ for different dimensionality (n) of the parent volume based ellipsoidal model.

n	1–3	4	5	6	7	8	9	10	100	1,000	10,000	100,000	1×10^8
$I_y = t(\sigma')/t(0)$	N/A	0.35355	0.44444	0.48714	0.51200	0.52828	0.53978	0.54832	0.60188	0.60608	0.60649	0.60653*	0.60653**

*First six decimal points of the inflection point vertical are the same as those for normal distribution. **First eight decimal points of the inflection point vertical are the same as those for normal distribution.

height of the normal distribution, the peak value of $f(x)$ is required. This peak value of the normal distribution is at the mean; that is, $x = 0$:

$$f(0) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0^2/2\sigma^2} = \frac{1}{\sigma\sqrt{2\pi}}. \tag{13}$$

For the normal distribution, vertical position of the inflection point, as a percentage of the vertical range, is given as

$$\begin{aligned} \frac{f(\sigma)}{f(0)} &= \frac{1/\sigma\sqrt{2\pi e}}{1/\sigma\sqrt{2\pi}}, \frac{f(\sigma)}{f(0)} \\ &= \frac{1}{\sqrt{e}} = 0.60653065971263342360379953499118 \\ &\approx 60.6531\%. \end{aligned} \tag{14}$$

Considering volume based ellipsoidal models, the Tanazur distributions are represented as

$$\begin{gathered} t(x) = \left(1 - x^2\right)^{(n-2)/2} \\ t''(x) = (n-1)\left(1 - x^2\right)^{(n-5)/2}\left((n-2)x^2 - 1\right), \\ 0 = (n-1)\left(1 - x^2\right)^{(n-5)/2}\left((n-2)x^2 - 1\right). \end{gathered} \tag{15}$$

Value of x for which the Tanazur distributions have 2nd derivative of 0 is

$$\begin{gathered} 0 = \left((n-2)x^2 - 1\right), \\ x = \pm\sqrt{\frac{1}{(n-2)}}. \end{gathered} \tag{16}$$

Let this value of x be denoted as σ'. Consider

$$x = \pm\sigma'. \tag{17}$$

Again, to get the vertical position of the inflection point as a percentage of the vertical range, we need to divide $t(x)$ by $t(0)$:

$$\begin{gathered} t\left(\pm\sigma'\right) = \left(1 - \left(\pm\sqrt{\frac{1}{(n-2)}}\right)^2\right)^{(n-1)/2}, \\ t\left(\sigma'\right) = \left(1 - \frac{1}{(n-2)}\right)^{(n-1)/2} \end{gathered} \tag{18}$$

and $t(0)$ is given as

$$t(0) = \left(1 - 0^2\right)^{(n-1)/2} = 1. \tag{19}$$

If "I" represents the inflection point of $t(x)$, then I_y is given as

$$\begin{aligned} \frac{t\left(\sigma'\right)}{t(0)} = I_y &= \frac{(1 - 1/(n-2))^{((n-1)/2)}}{1} \\ &= \left(1 - \frac{1}{(n-2)}\right)^{((n-1)/2)}. \end{aligned} \tag{20}$$

Table 2 shows the normalized vertical position I_y of the inflection points Tanazur distributions and is visualized in Figure 7. The value of I_y for $T(r, n)$ appears to converge to that of the normal distribution (≈60.6531%) as the value of n is increased. This is discussed in Section 4.1.

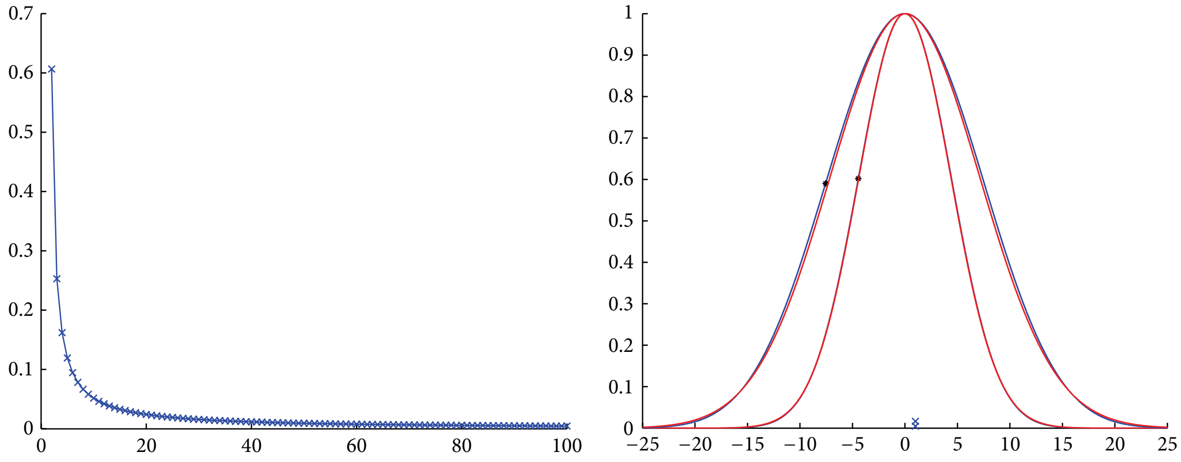
Using the framework of inflection point positions, we can handle dimensionality determination of a parent ellipsoidal model from the near normal distribution of any of its vector components. If I_y is known, the dimensionality of the parent model can be determined by solving for n in (20). Value of I_y can also be compared to a table, like the one given above, to get the correct model dimensionality.

The calculations of I_y for different $T(r, n)$ have been done while assuming volumetric ellipsoidal models. As discussed earlier, the Tanazur distribution for both volume based and surface based models is the same, except for a difference of one dimension. For surface based ellipsoidal models, the dimensionality of the parent model is one more than the value calculated for volume based models.

Inflection point is one of the features of $T(r, n)$ (and also $N(\mu, \sigma^2)$) that can be used for dimensionality determination. Another feature that can be used is the length of the tail in the distribution. This, however, is not discussed here and we base our discussion on the inflection point vertical (I_y).

4.1. Similarity between Tanazur and Normal Distributions.

We have already seen from Figure 7 that as the dimensionality of the parent ellipsoidal model is increased, the inflection point vertical for $T(r, n)$ appears to converge to that of $N(\mu, \sigma^2)$. The difference between I_y of $T(r, n)$ and $N(\mu, \sigma^2)$



(a)

(b)

FIGURE 8: (a) Plot of the difference between I_y of $T(r,n)$ and $N(\mu,\sigma^2)$. Horizontal axis represents number of dimensions n while the vertical axis shows delta in I_y. (b) Comparison of the blue $T(r,n)$ curve and the red $N(\mu,\sigma^2)$ curve when the value of $n = 30$ (outer curve) and $n = 100$ (inner curve). Tanazur and Normal distribution curves have been superimposed where red curves represent the normal distribution ($N(0, 90.25)$ and $N(0, 34.03)$) and the background blue curves represent Tanazur distribution ($T(40, 30)$ and $T(44, 100)$).

is shown in Figure 8. The limit of $I_y = t(\sigma')/t(0)$, as n approaches infinity, is given as

$$\begin{aligned} I_y &= \frac{t(\sigma')}{t(0)} = \lim_{n\to\infty}\left(1-\frac{1}{(n-2)}\right)^{((n-1)/2)} \\ &= \lim_{n\to\infty}\left(\frac{n-3}{n-2}\right)^{((n-1)/2)}, \\ I_y &= \frac{t(\sigma')}{t(0)} = \frac{1}{\sqrt{e}} = \frac{f(\sigma)}{f(0)} \\ &= 0.606530659712633423603 \approx 60.6531\%. \end{aligned} \tag{21}$$

Given infinite dimensions of an ellipsoidal model, the inflection point of the model's distribution curve, when observing any of the dimensions, will tend towards the inflection point of the normal distribution. The distribution curve for a high dimensional model ($n = 100$) is shown in Figure 8(b) and the superimposed curves show that there is minimal difference between Tanazur and Normal distributions.

Even though it is clear from Figure 8 that $T(r,n)$ and $N(\mu,\sigma^2)$ are very similar at higher dimensions and that their inflection point verticals (I_y) converge at the limit, the curves themselves cannot be equal. This is because the upper limit of x is unbounded in case of $N(\mu,\sigma^2)$, whereas it is limited in case of $T(r,n)$ by the dimensional radius r. In other words, normal distribution has an infinitely long tail whereas Tanazur distribution does not.

4.2. Formulation of Tanazur Distribution. Like other distributions, Tanazur distribution is also of two kinds: univariate and multivariate. The notation $X \sim T(r,n)$ is for the univariate Tanazur distribution. It assumes that only 1 dimension of the n-dimensional ellipsoidal model is being observed. Here r refers to the radius of the observed dimension, in the ellipsoidal model.

Multivariate Tanazur distribution occurs when more than one dimension of the model is observed, leading to a multidimensional Tanazur distribution. If the dimensionality of the uniformly distributed ellipsoidal model is n and the Tanazur Distribution is being observed in n' dimensions, then the multivariate Tanazur Distribution is denoted as

$$X \sim T(r, n, n'). \tag{22}$$

Here r is a vector of length n', representing the radii of the n' dimensions in the ellipsoidal model. The resultant vector X has n' dimensions.

Probability distribution function requires that the area under the curve be equal to 1. The probability density function (pdf) for the univariate Tanazur Distribution can be formulated as

$$t(x; r, n) = \frac{(r^2 - x^2)^{((n-1)/2)}}{\int_{-r}^{r} (r^2 - x^2)^{((n-1)/2)}\, dx}. \tag{23}$$

The factor in the denominator ensures that the area under the curve is equal to 1. The function returns the probability of observing a value x for a dimension of radius r in an n-dimensional volume based ellipsoidal model. For a surface based ellipsoidal model, the function returns the probability of observing a value x for a dimension of radius r in an $n + 1$ dimensional model. The cumulative distribution function (cdf) is given as

$$t(x; r, n) = \frac{\int_{-r}^{x} (r^2 - x^2)^{((n-1)/2)}\, dx}{\int_{-r}^{r} (r^2 - x^2)^{((n-1)/2)}\, dx}. \tag{24}$$

So far we have used integral forms of the probability density function (pdf) for the univariate Tanazur Distribution. For easy mathematical analysis, we demonstrate calculation of the closed form for a given model dimensionality. The probability of a variable taking up certain value in the filled 3D spherical model is equal to the ratio of the area of sphere slice at that value and the total sphere volume. This can be written as

$$t(x;r,3) = \pi \times \left(\frac{r^2 - x^2}{(4/3)\pi r^3}\right),$$
$$t(x;r,3) = \frac{3}{4r} \times \left(1 - \frac{x^2}{r^2}\right). \tag{25}$$

The discussion so far has implicitly assumed that the distribution is spread symmetrically around the value 0 in the observed dimension. This is generally not the case in many practical scenarios. The mean value of the distribution can be handled outside of the formula above, where the translation along the observed dimension can happen after scaling for radii. When generating Tanazur distributions, the process is

(i) generate distribution for unit spherical model;

(ii) scale distribution for each observed dimension according to the radius of the model in the said dimension;

(iii) translate the distribution in each dimension according to the value of the mean for the given dimension.

5. Experimentation

The relevance of Tanazur distributions in real world scenarios is demonstrated using an experiment. The physical phenomenon of motion of molecules in an ideal gas is governed by the kinetic theory of gasses which states that the kinetic energy of molecules is conserved. The gas molecules under the conditions of standard temperature and pressure exhibit motion primarily on the basis of particle collisions, with minimal effects from other intermolecular forces. These particle collisions happen in such a way that the kinetic energy, of the particles involved in the collision, is conserved. This conservation of kinetic energy can be modeled as a surface based ellipsoidal model.

Distribution of particle speeds in an ideal gas has been described by Maxwell-Boltzmann Distribution [32] which is given as

$$f(v) = \sqrt{\frac{2}{\pi}\left(\frac{m}{kT}\right)^3} v^2 e^{-mv^2/2kT}. \tag{26}$$

Numerous experiments [33–36] have involved the measurement of the speed distribution of such particles under different condition.

Direction observation of gas molecules has not been possible so far and therefore indirect means of measuring approximate particle speed have been used. Speed of particles is calculated based on the motion of tiny dust or pollen particles suspended in the gas. Our experimentation setup, however, consists of a computer based simulation of a 2-dimensional gas. Two-dimensional gas is a known experimental setup [33, 34] for ideal gases and restricts the motion of gas particles to a single plane, thereby simplifying the experiment. This simplification does not compromise the behavior of gas particles. Motion of particles of an ideal gas is governed by the law of conservation of kinetic energy, which is given as

$$\sum_{k=1}^{n} \frac{m_k v_{kt}^2}{2} = \sum_{k=1}^{n} \frac{m_k v_{kt'}^2}{2}. \tag{27}$$

Here v_{kt} represents velocity of particle k at time t and $v_{kt'}$ represents velocity at another time t'. The above equation can be rewritten in terms of the total kinetic energy E of the system:

$$E = \sum_{k=1}^{n} \frac{m_k v_k^2}{2}. \tag{28}$$

The subscript "t" has been removed as the gasses in thermodynamic equilibrium maintain the total kinetic energy E over time. With constant kinetic energy of the system, the conservation of kinetic energy can be modeled as a surface-ellipsoidal model. The above equation can be simplified further by assuming a homogeneous ideal gas. A homogeneous gas has the same type of particles and hence the same value of mass throughout:

$$E = \frac{m}{2}\sum_{k=1}^{n} v_k^2 \Longrightarrow \frac{2E}{m} = \sum_{k=1}^{n} v_k^2. \tag{29}$$

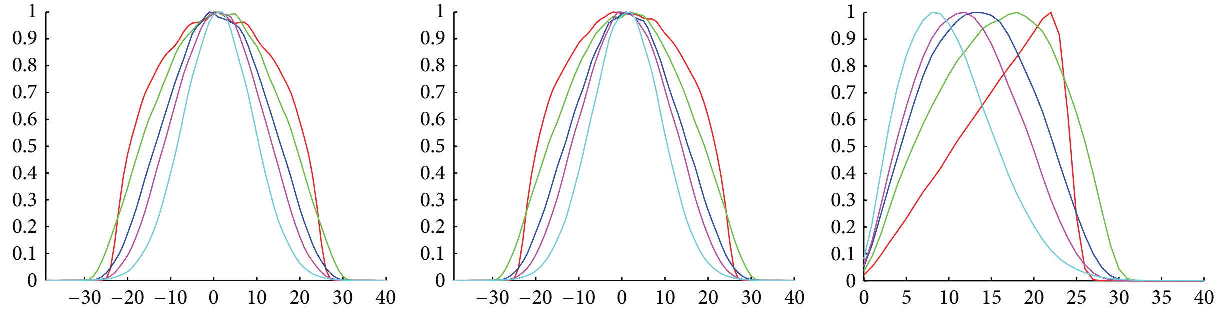
For a 2D gas, the velocity component can be split up into its component velocities in the 2 dimensions:

$$\frac{2E}{m} = \sum_{k=1}^{n} \left(v_{xk}^2 + v_{yk}^2\right). \tag{30}$$

This equation is of the same form as that of the equation of a spherical model, specified in (2).

Therefore the conservation of kinetic energy in a 2D ideal gas scenario can be represented by a surface based spherical model with $2n$ dimensions. The radius of the spherical model is equal to $\sqrt{2E/m}$ which is also the maximum velocity attainable by a single particle in such system.

Computer based simulation of 2-dimensional ideal gas was performed. Kinetic energy of the system of particles was conserved in the elastic collisions of particles with each other as well as with the walls of the container. Only 2-way particle collisions were considered; that is, simultaneous collision of 3 or more particles was avoided to reduce computational complexity. Different simulation scenarios were performed which consisted of varying number of gas particles in the system. As the velocity of each particle is represented by 2 dimensions in the 2D gas, the dimensionality of the spherical model is twice that of the number of particles in the system. Figure 9 shows the results of the experiments that were performed.



(a)

(b)

(c)

FIGURE 9: Distributions generated though computer simulations of different particle scenarios: red = 2 particles, green = 3 particles, blue = 4 particles, magenta = 5 particles, and cyan = 100 particles. ((a) to (c)) Distribution of V_x, V_y, and $|V|$.

It can be seen from Figure 9 that the velocity components of a randomly selected particle exhibit Tanazur Distribution. When the dimensionality of the system is low, that is, few particles in the system, the distribution curve is similar to the one seen in lower order Tanazur distribution. As the number of particles in the gas is increased, the resulting distribution appears more like the normal distribution. Inflection point of the distribution curve can be used to find the dimensionality of the parent ellipsoidal model. What this implies is that, by observing the distribution of velocity components of a single particle, the total number of particles in the system can be calculated.

The above mentioned surface based spherical model was for a homogeneous gas. A heterogeneous gas on the other hand is modeled by a surface based elliptical model:

$$1 = \sum_{k=1}^{n} \frac{m_k v_k^2}{2E} = \sum_{k=1}^{n} \frac{v_{xk}^2 + v_{yk}^2}{2E/m_k}. \tag{31}$$

The term m_k represents the mass of particle k. The above equation can be written as

$$1 = \sum_{k=1}^{n} \frac{v_{xk}^2}{\left(\sqrt{2E/m_k}\right)^2} + \frac{v_{yk}^2}{\left(\sqrt{2E/m_k}\right)^2}. \tag{32}$$

The equation is the same form as that of the equation of an ellipse, as presented in (1). The radius of the ellipse for a given dimension is equal to $\sqrt{2E/m_k}$; that is, the radius of the dimension is related to the mass of the particle represented by the dimension.

As mentioned earlier, this experimental setup uses two dimensions to represent the 2 velocity components for every particle. The two velocity components V_x and V_y, together, give the velocity V of a particle. The speed of the particle is given as

$$|V| = \sqrt{v_{xk}^2 + v_{yk}^2}. \tag{33}$$

The distribution of particle speed in gasses as described by the Maxwell-Boltzmann distribution [32] is shown in Figure 10. This distribution has the characteristic property of being skewed, with a long tail at higher velocities. Similar skewed speed distributions were observed in our 2D ideal gas simulation and can be seen in Figure 9. This skewed distribution of particle speed $|V|$ can also be explained on the basis of Tanazur Distributions. More specifically, we show that this phenomenon is an outcome of multivariate Tanazur distributions.

Particle velocity components V_x and V_y together make up the velocity vector of a particle. The magnitude of this velocity vector is the speed of the particle and is never negative. If the velocities components of a particle are uniformly distributed in the 2D V_xV_y 2-ball (circular) model, then the distribution in 2 dimensions would look similar to Figure 4(a). Each data point in the model represents an observation on the combination of velocity components of the particle. The surface integral for the 2D V_xV_y 2-ball model gives the area of the circle:

$$\begin{aligned} \text{Area} &= \int_0^r \text{Ring Circumfrance} * dx, \\ \text{Area} &= \int_0^r 2\pi x * dx = 2\pi \int_0^r x * dx. \end{aligned} \tag{34}$$

As the number of particles in each speed range is equal to the area of the circular ring defined by that speed range, the uniform distribution in the V_xV_y circular model gives a skewed distribution in $|V|$, with more data point in the higher speed ranges as compared to the lower ones. This distribution is shown in Figure 11 and is equal to the distribution of the integrand in (35). If the distribution of data points in V_xV_y plane was to change from a uniform distribution to a more skewed distribution, this skewed data point density in the circular model will also be reflected on the distribution of $|V|$. If the scalar value of data point density in the 2D model is given by a probability density function $f(x)$, then the resultant surface integral over scalar field $(f(x))$ gives

$$\text{Area} = 2\int_0^r x * f(x) * dx. \tag{35}$$

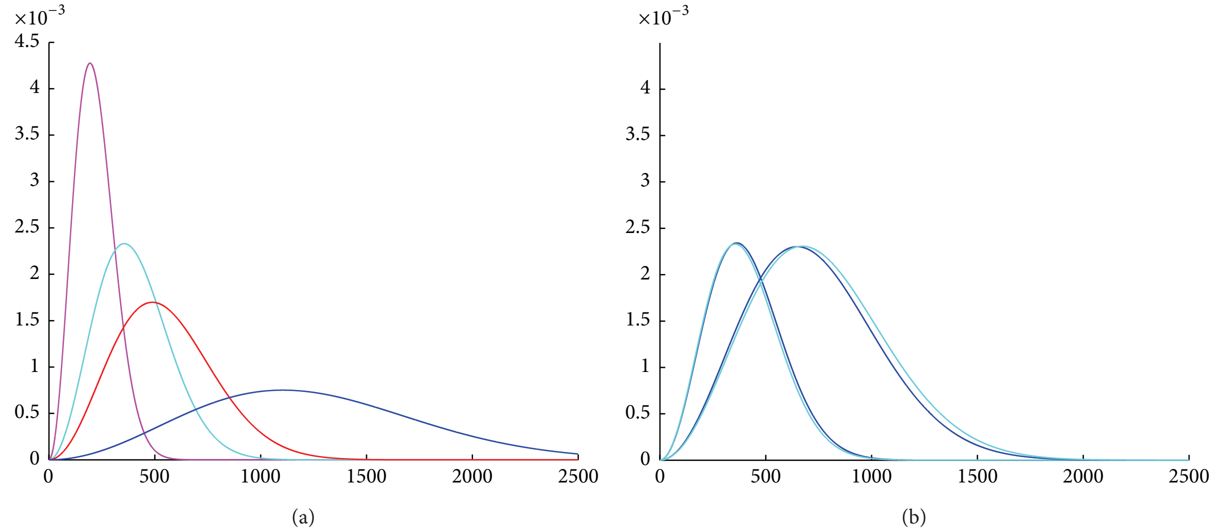


Figure 10: (a) Maxwell-Boltzmann distribution for speed of particles in noble gasses at 0 degree C (273.15 K). (blue: Helium, red: Argon, cyan: Argon, and magenta: Xenon). (b) Tanazur Distribution (black) compared to distribution for one noble gas (Argon). Argon distributions are scaled differently for comparison here, with more accurate distribution corresponding to $T(3800, 253, 3)$ and less accurate distribution corresponding to $T(3400, 53, 3)$.

If $f(x) = t(x; r, n, n')$, then the resultant density is

$$\text{Area} = 2\int_0^r x * t\left(x; r, n, n'\right) * dx. \tag{36}$$

In other words, we are considering the distribution of data points in the 2D plane to be a multivariate Tanazur distribution. For $t(x; [25, 25], 4, 2)$, the resultant 2D Tanazur distribution of the 4D spherical model is shown in Figure 11. We also observe that the distribution of $|V|$ for data point distribution described by $t(x; [25, 25], 4, 2)$ looks changed from the original form. This distribution of a 4D spherical model corresponds to a system containing 2 particles. If the number of particles is increased, the corresponding distributions take the form shown in Figure 12. This is similar to the distribution of $|V|$ observed in the experiments and shown in Figure 9.

The results of the 2D gas simulation suggest that the data points in the ellipsoidal model may be distributed uniformly at the model level, that is, when observing all dimensions of the model. What this also suggests is that there is no bias in the individual velocity components of the particles (as observed by the symmetric distributions of the velocity components) or in the combined state of the particles (as assumed in the uniform distribution on $2 * n$ dimensional vectors in the model). The skewed distribution of the particle speed is due to the (skewed) Tanazur distribution from higher dimensions on the 2D V_xV_y plane of the particles. For 3D gas scenarios with 3 degrees of freedom, the calculated distribution of $|V|$ is

$$\begin{aligned} |V| &= \sqrt{v_{xk}^2 + v_{yk}^2 + v_{zk}^2}, \\ \text{Volume} &= \int_0^r 4\pi x^2 * dx = 4\pi \int_0^r x^2 * dx. \end{aligned} \tag{37}$$

If Tanazur distributions from higher dimensions are mapped onto this 3D volume, the distribution of $|V|$ can be calculated as

$$\begin{aligned} \text{Volume} &= 4\int_0^r x^2 * f(x) * dx \\ &= 4\int_0^r x^2 * t\left(x; r, n, n'\right) * dx. \end{aligned} \tag{38}$$

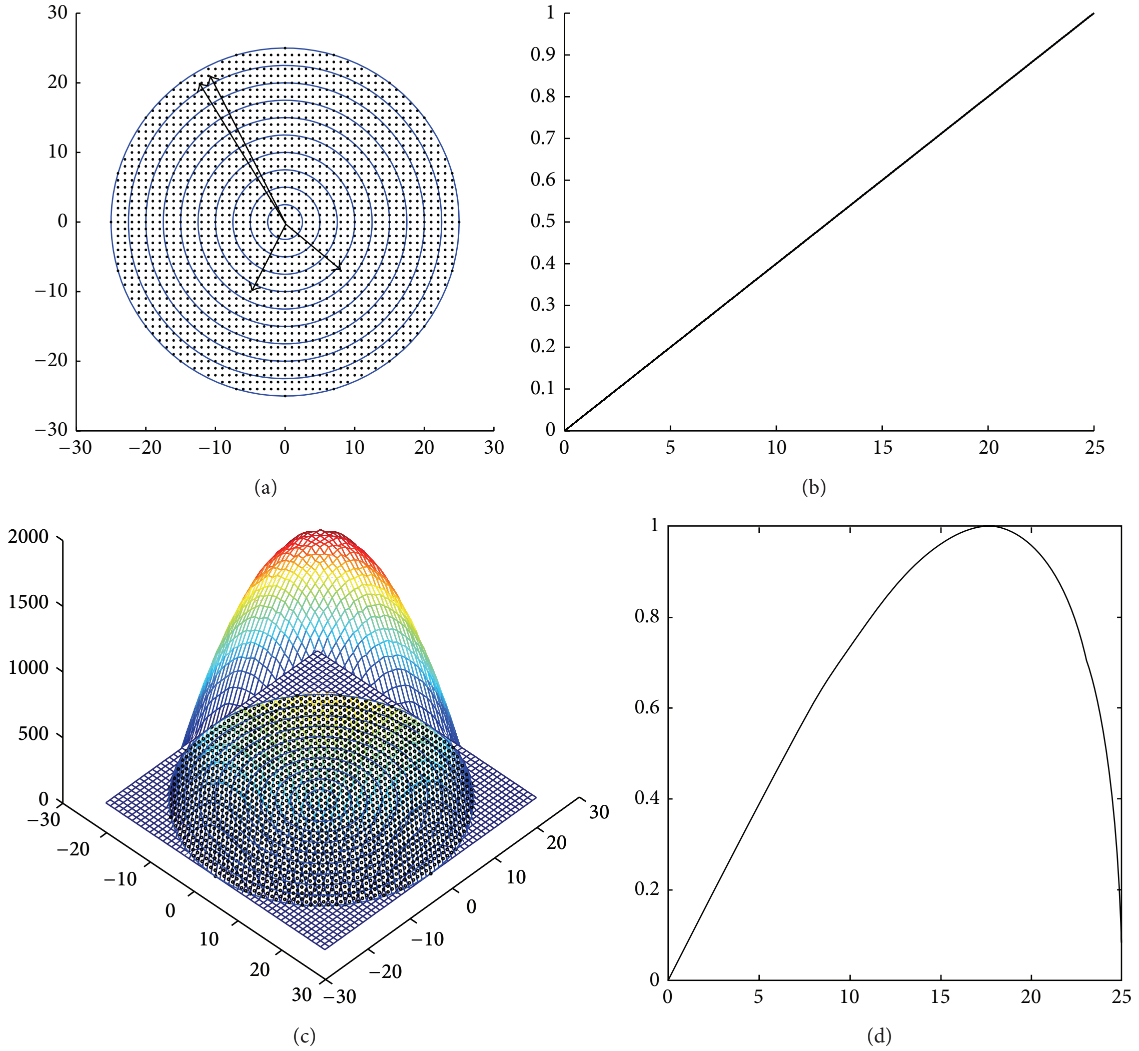
The distributions for $|V|$ can thus be calculated on the basis of the Tanazur distribution, which is projected onto a 3D sphere. The distribution of—V—generated using Maxwell-Boltzmann distribution's equation is compared with that generated by Tanazur distribution equation and is shown in Figure 10(b). As the dimensionality of the model is increased, the $|V|$ Tanazur distributions look very much identical to Maxwell-Boltzmann's. In other words, the shape of the $|V|$ distribution depends on the number of particles in the experiment. Tanazur distribution therefore predicts that as the number of particles in the system is decreased, the tail of the speed distribution $|V|$ (i.e., at higher speed range) will get smaller.

The above experiment was based on a surface based ellipsoidal model. An example of a volume based ellipsoidal model can be that of a system of particles in which the upper bound of the average particle speed is the speed constant c:

$$c > \sum_{i=1}^{n} \frac{v_i}{n}. \tag{39}$$

For a homogeneous gas, this can be written like the equation of spherical models:

$$\left(\sqrt{cn}\right)^2 > \sum_{i=1}^{n} \left(\sqrt{v_{xi}^2 + v_{yi}^2 + v_{zi}^2}\right)^2. \tag{40}$$



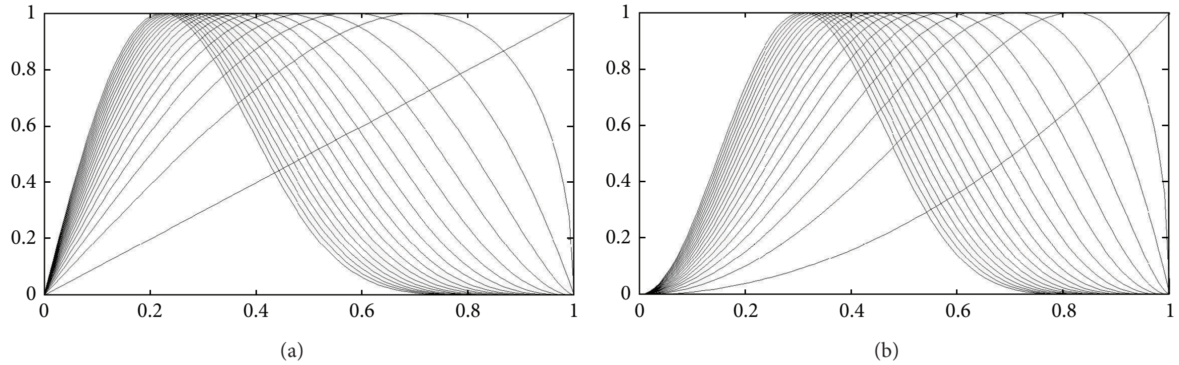
(a)

(b)

(c)

(d)

Figure 11: (a) $T([25, 25], 1)$ for a 2-ball (1 particle) with the two dimensions corresponding to particles V_x and V_y. (c) $T([25, 25], 3)$ for a 4-ball (2 particles) with the observed dimensions corresponding to V_x and V_y of first particle. (b) Speed distribution $|V| = \sqrt{v_{xk}^2 + v_{yk}^2}$ for the Tanazur Distribution shown in top left image. (d) Speed distribution corresponding to the Tanazur Distribution shown in (c).



(a)

(b)

Figure 12: Calculated distributions of $|V|$ for multivariate Tanazur distributions corresponding to different model dimensionalities. (a) Counter-clockwise from the horizontal (at the origin) for a 2D gas scenario, distributions of $|V|$ for $T(25, 2, 2)$, $T(25, 3, 2)$, $T(25, 4, 2)$, and $T(25, 21, 2)$. (b) Counter-clockwise from the horizontal (at the origin) for a 3D gas (3 velocity components) scenario, distributions of $|V|$ for $T(25, 3, 3)$, $T(25, 4, 3)$, $T(25, 5, 3)$, and $T(25, 22, 3)$.

Therefore, the radius of such a system is $\sqrt{cn}$ and each dimension of the spherical model represents square root of the speed component of a particle. If we assume that the distribution of the model vectors in this n-dimensional spherical model is uniform, then Tanazur distribution should be observed.

The proposed method performs a statistical analysis of the system and therefore does not give a deterministic solution. But when compared to deterministic methods of particle count determination like the Ideal Gas Law ($pV = nRT$), it offers an advantage. With sensor errors aside, the Ideal Gas Law observations can alter the state of the system by changing the energy state of the system. For example, observing the pressure of the gas system can alter the temperature and vice versa. The proposed statistical method does not suffer from this dilemma as the observed near normal curve merely shifts along an axis, while maintaining its distinct distribution curve. This is because a change in quantities like pressure or temperature does not change the underlying variable upon which the curve shape depends, that is, the particle count of the gas.

6. Conclusion

By empirical and mathematical methods, we have shown that uniform distributions in higher dimensional ellipsoidal models can be observed as "near normal" distributions in lower dimensions. For an n-dimensional ellipsoidal model, as the dimensionality of a system increases, the observed distribution of any variable tends towards the normal distribution. Conversely, by observing the "near normal" distribution of a variable, it may be possible to predict the number of variables in the system.

Many of the phenomena observed in nature can be modeled as surface or volume based ellipsoidal models. The experimental section shows one such scenario. The apparent flexibility of the ellipse equation, with variable dimensional radii, allows many real world scenarios and processes to be represented as ellipsoidal models.

There has been a long held belief in the scientific community about the random nature of the observed variables in a normal distribution. The findings of this paper offer an alternate explanation for such observations. It also suggests that what ultimately appears to be a bias in the states a variable can take up can in fact be a result of an unbiased (or uniform) distribution of the variable states in the state space. Perhaps nature restricts the limits of the sandbox for the play of the variables but does not intervene in the act of play itself. The extent to which the model presented here can explain other real world observations is something which will only be known in the days to come. But for now Tanazur distributions give a new perspective on old observations. The choice of the distribution's name "*Tanazur*", Urdu for perspective, reflects this.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] G. Casella and L. Berger, *Statistical Inference*, Duxbury, 2nd edition, 2001.

[2] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate Laplace distribution," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 300–303, 2006.

[3] A. Mohammadi, F. Almasganj, and A. Taherkhani, "Missing Feature reconstruction with Multivariate Laplace distribution (MLD) for noise robust phoneme recognition," in *Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP '08)*, pp. 836–840, 2008.

[4] K. Giesecke and S. Zhu, "Transform analysis for point processes and applications in credit risk," *Mathematical Finance*, vol. 23, no. 4, pp. 742–762, 2013.

[5] R. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME Journal of Basic Engineering*, vol. 82, no. 1, pp. 33–45, 1960.

[6] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[7] G. Welch and G. Bishop, "An introduction to the Kalman filter," Tech. Rep. TR 95-041, 2000.

[8] Y. Chen, T. Huang, and Y. Rui, "Parametric contour tracking using unscented Kalman filter," in *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 613–616, June 2002.

[9] S. Wachter and H.-H. Nagel, "Tracking persons in monocular image sequences," *Computer Vision and Image Understanding*, vol. 74, no. 3, pp. 174–192, 1999.

[10] N. Peterfreund, "Robust tracking of position and velocity with Kaiman snakes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 564–569, 1999.

[11] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[12] Y. N. Andrew, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Advances in Neural Information and Processing Systems*, vol. 14, 2001.

[13] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, "Object trajectory-based activity classification and recognition using hidden Markov models," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[14] F. Porikli and T. Haga, "Event detection by eigenvector decomposition using object and frame features," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '04)*, 2004.

[15] D. Williams, *Probability with Martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, UK, 1991.

[16] A. Naftel and S. Khalid, "Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space," *Multimedia Systems*, vol. 12, no. 3, pp. 227–238, 2006.

[17] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1450–1464, 2006.

[18] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1168–1181, 2007.

[19] F. Bashir, A. Khokhar, and D. Schonfeld, "Automatic object trajectory-based motion recognition using Gaussian mixture models," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 1532–1535, Amsterdam, The Netherlands, July 2005.

[20] J. Yu, "Localized Fisher discriminant analysis based complex chemical process monitoring," *AIChE Journal*, vol. 57, no. 7, pp. 1817–1828, 2011.

[21] T. Brotherton, T. Johnson, and G. Chadderdon, "Classification and novelty detection using linear models and a class dependent elliptical basis function neural network," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 2, pp. 876–879, Anchorage, Alaska, USA, 1998.

[22] S. Roberts and L. Tarassenko, "A probabilistic resource allocating network for novelty detection," *Neural Computation*, vol. 6, pp. 270–284, 1994.

[23] T. Odin and D. Addison, "Novelty detection using neural network technology," in *Proceedings of the International Congress on Condition Monitoring and Diagnostic Engineering (COMADEM '00)*, 2000.

[24] D.-Y. Yeung and C. Chow, "Parzen-window network intrusion detectors," in *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 4, pp. 385–388, 2002.

[25] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York, NY, USA, 1999.

[26] W. Bryc, *Normal Distribution Characterizations with Applications*, vol. 100 of *Lecture Notes in Statistics*, 2005.

[27] A. M. Kagan, Y. V. Linnik, and C. R. Rao, *Characterization Problems of Mathematical Statistics*, John Wiley & Sons, New York, NY, USA, 1973.

[28] V. Sudakov, "Typical distributions of linear functionals on spaces of high dimension," *Soviet Mathematics—Doklady*, vol. 92, no. 6, pp. 1578–1582, 1978.

[29] D. Williams, *Probability with Martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, UK, 1991.

[30] H. von Weizsäcker, "Sudakov's typical marginals, random linear functionals and a conditional central limit theorem," *Probability Theory and Related Fields*, vol. 107, no. 3, pp. 313–324, 1997.

[31] M. Anttila, K. Ball, and I. Perissinaki, "The central limit problem for convex bodies," *Transactions of the American Mathematical Society*, vol. 355, no. 12, pp. 4723–4735, 2003.

[32] R. C. Dunbar, "Deriving the Maxwell distribution," *Journal of Chemical Education*, vol. 59, no. 1, pp. 22–23, 1982.

[33] R. P. Bonomo and F. Riggi, "The evolution of the speed distribution for a two−dimensional ideal gas: a computer simulation," *The American Journal of Physics*, vol. 52, no. 54, 1984.

[34] J. M. Montanero, A. Santos, and V. Garzó, "Distribution function for large velocities of a two-dimensional gas under shear flow," *Journal of Statistical Physics*, vol. 88, no. 5-6, pp. 1165–1181, 1997.

[35] A. Barrat, T. Biben, Z. Rácz, E. Trizac, and F. van Wijland, "On the velocity distributions of the one-dimensional inelastic gas," *Journal of Physics A: Mathematical and General*, vol. 35, no. 3, pp. 463–480, 2002.

[36] A. Santos, "Nonlinear viscosity and velocity distribution function in a simple longitudinal flow," *Physical Review E*, vol. 62, no. 5, pp. 6597–6607, 2000.