Journal of Machine Learning Research 18 (2018) 1-60 Submitted 3/17; Revised 2/18; Published 4/18

On the Behavior of Intrinsically High-Dimensional Spaces: Distances, Direct and Reverse Nearest Neighbors, and Hubness

Fabrizio Angiulli FABRIZIO.ANGIULLI@UNICAL.IT
DIMES – Dept. of Computer, Modeling, Electronics, and Systems Engineering
University of Calabria
87036 Rende (CS), Italy

Editor: Sanjiv Kumar

Abstract

Over the years, different characterizations of the curse of dimensionality have been provided, usually stating the conditions under which, in the limit of the infinite dimensionality, distances become indistinguishable. However, these characterizations almost never address the form of associated distributions in the finite, although high-dimensional, case. This work aims to contribute in this respect by investigating the distribution of distances, and of direct and reverse nearest neighbors, in intrinsically high-dimensional spaces. Indeed, we derive a closed form for the distribution of distances from a given point, for the expected distance from a given point to its kth nearest neighbor, and for the expected size of the approximate set of neighbors of a given point in finite high-dimensional spaces. Additionally, the hubness problem is considered, which is related to the form of the function N_k representing the number of points that have a given point as one of their k nearest neighbors, which is also called the number of k-occurrences. Despite the extensive use of this function, the precise characterization of its form is a longstanding problem. We derive a closed form for the number of k-occurrences associated with a given point in finite high-dimensional spaces, together with the associated limiting probability distribution. By investigating the relationships with the hubness phenomenon emerging in network science, we find that the distribution of node (in-)degrees of some real-life, large-scale networks has connections with the distribution of k-occurrences described herein.

Keywords: high-dimensional data, distance concentration, distribution of distances, nearest neighbors, reverse nearest neighbors, hubness

1. Introduction

Although the size and the dimensionality of collected data are steadily growing, traditional techniques usually slow down exponentially with the number of attributes to be considered and are often overcome by linear scans of the whole data. In particular, the term *curse of dimensionality* (Bellmann, 1961), is used to refer to difficulties arising whenever high-dimensional data has to be taken into account.

One of the main aspects of this curse is known as *distance concentration* (Demartines, 1994), which is the tendency for distances to become almost indiscernible in high-dimensional spaces. This phenomenon may greatly affect the quality and performances of machine learning, data mining, and information-retrieval techniques. This effect results because almost all these techniques rely on the concept of distance, or dissimilarity, among data items in order to retrieve or analyze information. However, whereas low-dimensional spaces show good agreement between geometric proximity and the notion of similarity, as dimensionality increases, different counterintuitive phenomena arise that may be harmful to traditional techniques.

Over time, different characterizations of the curse of dimensionality and related phenomena have been provided (Demartines, 1994; Beyer et al., 1999; Aggarwal et al., 2001; Hinneburg et al., 2000; François et al., 2007). These characterizations usually state conditions under which, according to the limits of infinite dimensionality, distances become indistinguishable. However, almost never do these conditions address the form of associated distributions in finite, albeit high-dimensional, cases.

This work aims to contribute in this area by investigating the distribution of distances and of some related measures in intrinsically high-dimensional data. In particular, the analysis is conducted by applying the central limit theorem to the Euclidean distance random variable to approximate the distance probability distribution between pairs of random vectors, between a random vector and realizations of a random vector, and to obtain the expected distance from a given point to its kth nearest neighbor. It is then shown that an understanding of these distributions can be exploited to gain knowledge of the behavior of high-dimensional spaces, specifically the number of approximate nearest neighbors and the number of reverse nearest neighbors that are also investigated herein.

Nearest neighbors are transversal to many disciplines (Preparata and Shamos, 1985; Dasarathy, 1990; Beyer et al., 1999; Duda et al., 2000; Chávez et al., 2001; Shakhnarovich et al., 2006). In order to try to overcome the difficulty of answering nearest neighbor queries in high-dimensional spaces (Weber et al., 1998; Beyer et al., 1999; Pestov, 2000; Giannella, 2009; Kabán, 2012), the concept of the ϵ-approximate nearest neighbor (Indyk and Motwani, 1998; Arya et al., 1998) has been introduced. The ϵ-neighborhood of a query point is the set of points located at a distance not greater than $(1+\epsilon)$ times the distance separating the query from its true nearest neighbor.

Related to the notion of the ϵ-approximate nearest neighbor is the notion of neighborhood or query instability (Beyer et al., 1999): a query is said to be unstable if the ϵ-neighborhood of the query point consists of most of the data points. Although asymptotic results, such as that reported by Beyer et al. (1999), tell what happens when dimensionality is taken to infinity, nothing is said about the dimensionality at which the nearest neighbors become unstable. Pursuant to this scenario, this paper derives a closed form for the expected size of the ϵ-neighborhood in finite high-dimensional spaces, an expression that is then exploited to determine the critical dimensionality. Also, to quantify the difficulty of (approximate) nearest neighbor search, He et al. (2012) introduced the concept of relative contrast, a measure of separability of the nearest neighbor of the query point from the rest of the data, and provided an estimate which is applicable for finite dimensions. By leveraging the results concerning distance distributions, this paper derives a more accurate estimate for the relative contrast measure.

The number N_k of reverse nearest neighbors, also called the number of k-occurrences or the reverse nearest neighbor count, is the number of data points for which a given point is among their k nearest neighbors. Reverse nearest neighbors are of interest both in the database, information retrieval, and computational geometry literatures (Korn and

Muthukrishnan, 2000; Singh et al., 2003; Tao et al., 2007; Cheong et al., 2011; Yang et al., 2015), with uses having been proposed in the data mining and machine learning fields (Williams et al., 2002; Hautamäki et al., 2004; Radovanovic et al., 2009, 2010; Tomasev et al., 2014; Radovanovic et al., 2015; Tomasev and Buza, 2015), beyond being the objects of study in applied probability and mathematical psychology (Newman et al., 1983; Maloney, 1983; Tversky et al., 1983; Newman and Rinott, 1985; Yao and Simons, 1996).

Despite the usefulness and the extensive use of this construct, the precise characterization of the form of the function N_k both in the finite and infinite dimensional cases is a longstanding problem. What is already known is that for the infinite limit of size and dimension, N_k must converge in its distribution to zero; however, this result and its interpretations seem to be insufficient to characterize its observed behavior in finite samples and dimensions. Consequently, this paper derives a closed form of the number of k-occurrences associated with a given point in finite high-dimensional spaces, together with a generalized closed form of the associated limiting probability distribution that encompasses previous results and provides interpretability of its behavior and of the related hubness phenomenon.

The results, which are first illustrated for independent and identically distributed data, are subsequently extended to independent non-identically distributed data satisfying certain conditions, and then, connections with non-independent real data are depicted. Finally, it is discussed how to potentially extend the approach to Minkowski's metrics and, more generally, to distances satisfying certain conditions of spatial centrality.

Because hubness is a phenomenon of primary importance in network science, we also investigate if the findings relative to the distribution of the reverse nearest neighbors and the emergence of hubs in intrinsically high-dimensional contexts is connected to an analogous phenomenon occurring in the context of networks. The investigation reveals that for some real-life large-scale networks, the distribution of the incoming node degrees is connected to the herein-derived distribution of the infinite-dimensional k-occurrences function, which models the number of reverse k-nearest neighbors in an arbitrarily large feature space of independent dimensions. Hence, the provided distribution also appears to be suitable for modelling node-degree distributions in complex real networks.

The current study can be leveraged in several ways and in different contexts, such as in direct and reverse nearest neighbor searches, density estimation, anomaly and novelty detection, density-based clustering, and network analysis, among others. With regard to its possible applications, we can highlight approximations of measures related to distance distributions, worst-case scenarios for data analysis and retrieval techniques, design strategies that try to mitigate the curse of dimensionality, and models of complex networks. We refer to the concluding section for a more extensive discussion.

The rest of the work is organized as follows. Section 2 discusses related work concerning the concentration of distances, intrinsic dimensionality, and the number of k-occurrences and the associated hubness phenomenon. Section 3 presents the notation used to provide results. Section 4 introduces the main results of the paper. Section 5 discusses relationships between the study of the hubness phenomena occurring in high-dimensional spaces with the analogous phenomena observed in real-life, large-scale, complex networks. Section 6 concludes the work. Finally, the Appendix contains the proofs that are not reported within the main text.

2. Related Work

As already noted, the term *curse of dimensionality* is used to refer to difficulties arising when high-dimensional data must be taken into account, and one of the main aspects of this curse is *distance concentration.* In this regard, Demartines (1994) has shown that the expectation of the Euclidean norm of independent and identically distributed (i.i.d.) random vectors increases as the square root of the dimension, whereas its variance tends toward a constant and, hence, does not depend on the dimensionality. Specifically:

Theorem 1 (Demartines, 1994, cf. Theorem 2.1) *Let $\boldsymbol{X}_d$ be an i.i.d. d-dimensional random vector with common cdf F_X. Then*

$$\mathbf{E}[\|\boldsymbol{X}_d\|_2] = \sqrt{ad-b} + O(1/d) \qquad \textit{and} \qquad \sigma^2(\|\boldsymbol{X}_d\|_2) = b + O(1/\sqrt{d}),$$

where a and b are constants depending on the central moments of F_X up to the fourth order but do not depend on the dimensionality d.

Demartines noticed that, because the Euclidean distance corresponds to the norm of the difference of two vectors, the distance between the i.i.d. random vectors must also exhibit the same behavior. This insightful result explains why high-dimensional vectors appear to be distributed around the surface of a sphere of radius $\mathbf{E}[\|\boldsymbol{X}_d\|]$ and why, because they seem to be normalized, the distances between pairs of high-dimensional random vectors tend to be similar.

The distance concentration phenomenon is usually characterized in the literature by means of a ratio between some measure related to the spread and some measure related to the magnitude of the norm, sometimes presented as the distance from a point located in the origin of the space. In particular, the conclusion is that there is a concentration of distances when the above ratio converges to 0 as the dimensionality tends to infinity.

Some authors have studied the concentration phenomenon by representing a data set as a set of n d-dimensional i.i.d. random vectors $\boldsymbol{X}_d^{(j)}$ ($1 \leq j \leq n$) with not-necessarily common pdfs $f_{X^{(j)}}$. Specifically, the *contrast* is defined as the difference between the largest and the smallest observed norm, or rather the distance from a query point located at the origin, whereas the *relative contrast* is defined as

$$RC_d = \frac{\max_j \|\boldsymbol{X}_d^{(j)}\|_p - \min_j \|\boldsymbol{X}_d^{(j)}\|_p}{\min_j \|\boldsymbol{X}_d^{(j)}\|_p},$$

where $\|\cdot\|_p$ denotes the p-norm $\|\boldsymbol{x}_d\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$, is the contrast normalized with respect to the smallest norm/distance.

Theorem 2 (Adapted from Beyer et al., 1999, cf. Theorem 1) *Let $\boldsymbol{X}_d^{(j)}$ $(1 \leq j \leq n)$ be n d-dimensional random vectors with common cdfs. If*

$$\lim_{d\to\infty} \sigma^2\left(\frac{\|\boldsymbol{X}_d^{(j)}\|_p}{\mathbf{E}[\|\boldsymbol{X}_d^{(j)}\|_p]}\right) = 0, \ \textit{then, for any } \epsilon > 0,\ \lim_{d\to\infty} Pr\left[RC_d \leq \epsilon\right] = 1.$$

If the hypothesis is verified, that is, if the variance of the ratio between the norm of the vectors and their expected value vanishes as the dimensionality goes to infinity, then the relative contrast also becomes smaller and smaller, and all the vectors seem to be located at approximatively the same distance from the reference vector. That is, given a query point $\boldsymbol{Q}_d$, the distance from the nearest and the furthest neighbor become negligible:

$$\lim_{d\to\infty} Pr\left[\max_j \|\boldsymbol{Q}_d - \boldsymbol{X}_d^{(j)}\|_p \leq (1+\epsilon)\min_j \|\boldsymbol{Q}_d - \boldsymbol{X}_d^{(j)}\|_p\right] = 1.$$

In (Beyer et al., 1999), it is shown that i.i.d. random vectors satisfy the above condition.

Other authors have provided characterizations of the concentration phenomenon by providing upper and lower bounds to the relative contrast in the cases of Minkowski and fractional norms (Hinneburg et al., 2000; Aggarwal et al., 2001).

Subsequently, (François et al., 2007) posed the following problem: is the concentration phenomenon a side effect of the *Empty Space Phenomenon* (Bellmann, 1961), just because we consider a finite number of points in a bounded portion of a high-dimensional space? To explore this problem, they studied the concentration phenomenon by taking the same perspective as Demartines, i.e., to refer to a distribution rather than to a finite set of points. The *relative variance*

$$RV_d = \frac{\sigma(\|\boldsymbol{X}_d\|_p)}{\mathbf{E}[\|\boldsymbol{X}_d\|_p]}$$

is a measure of concentration for distributions, corresponding to the ratio between the standard deviation and the expected value of the norm.

Theorem 3 (Adapted from François et al., 2007, cf. Theorem 5) *Let $\boldsymbol{X}_d$ be an i.i.d. d-dimensional random vector. Then*

$$\lim_{d\to\infty} RV_d = 0.$$

From the above result, they conclude that the concentration of the norms in high-dimensional spaces is an intrinsic property of the norms and not a side effect of the finite sample size or of the Empty Space Phenomenon. Because it does not depend on the sample size, this can be regarded as an extension of Demartines' results to all p-norms.

As a consequence of the distance concentration, the separation between the nearest neighbor and the farthest neighbor of a given point tend to become increasingly indistinct as the dimensionality increases.

Related to the analysis of i.i.d. data is the concept of intrinsic dimensionality. Although variables used to identify each datum could not be statistically independent, ultimately, the intrinsic dimensionality of the data is identified as the minimum number of variables needed to represent the data itself (van der Maaten et al., 2009). This corresponds in linear spaces to the number of linearly independent vectors needed to describe each point. As a matter of fact, an extensively used notion of intrinsic dimensionality, the *correlation dimension* (Grassberger and Procaccia, 1983), is based on identifying the dimensionality D at which the Euclidean space is homeomorphic to the manifold containing the support of the data:

$$D = \lim_{\delta\to 0} \frac{\ln F_{\mathrm{dst}}(\delta)}{\ln \delta},$$

where F_{dst} denotes the cumulative distribution function of pairwise distances, which formalizes the notion that in the limit of small length-scales ($\delta \to 0$) upon which the manifold the data lie, the manifold is homeomorphic to the Euclidean space of dimension D.

And, indeed, (Demartines, 1994) mentions that if random vector components are not independent, the concentration phenomenon is still present provided that the actual number D of "degrees of freedom" is sufficiently large. Thus, results derived for i.i.d. data continue to be valid provided that the dimensionality d is replaced with D. Moreover, (Beyer et al., 1999) provided different examples of data presenting concentration, all of which share with the i.i.d. case a sparse correlation structure. (Durrant and Kabán, 2009) noted that it is difficult to identify meaningful workloads that do not exhibit concentration, and showed that for the family of linear latent variable models, a class of data distributions having non-i.i.d. dimensions, the Euclidean distance will not become concentrated as long as the number of relevant dimensions grows no more slowly than the overall data dimensions do. This also confirms that weakly dependent data lead to concentration; however, they also noted that the condition to avoid concentration is not often met in practice.

Another aspect of the curse of dimensionality problem, closely related to the distance concentration and the nearest neighbor relationship, is the so called *hubness* phenomenon. This phenomenon has been previously observed in several applications (Doddington et al., 1998; Singh et al., 2003; Aucouturier and Pachet, 2007), has recently undergone to direct investigation (Radovanovic et al., 2009, 2010; Low et al., 2013), and has been subjected to several different proposed techniques for overcoming the phenomenon (Radovanovic et al., 2015; Tomasev, 2015).

Specifically, consider the number $\mathrm{N}_k(\boldsymbol{x}_d)$ of observed points that have $\boldsymbol{x}_d$ among their k nearest neighbors. N_k is also called *k-occurrences* or the *reverse k-nearest neighbor count*.

It is known that in low dimensional spaces, the distribution of N_k complies with the binomial one and, in particular, for uniformly distributed data in low dimensions, it can be modeled as node in-degrees in the k-nearest neighbor graph, which follows the Erdős-Rényi random graph model, with a binomial degree distribution (Erdős and Rényi, 1959). However, it has been observed that as dimensionality increases, the distribution of N_k becomes skewed to the right, resulting in the emergence of *hubs*, which are points whose reverse k-nearest neighbor counts tend to be meaningfully larger than that associated with any other point.

The distribution of N_k has been explicitly studied in the applied probability and mathematical psychology communities (Newman et al., 1983; Maloney, 1983; Newman and Rinott, 1985; Tversky and Hutchinson, 1986; Yao and Simons, 1996). Almost all the results provided concern a Poisson process that spreads the vectors uniformly over $\mathbb{R}^d$, leading to the conclusion that the limiting distribution of N_k converges to the Poisson distribution with mean k. The case of continuous distributions with i.i.d. components has been considered in (Newman et al., 1983; Newman and Rinott, 1985), where the expression for the infinite-dimensional distribution of N_1 is characterized as follows.

Theorem 4 (Newman et al., 1983, cf. Theorem 7) *Let $\{\boldsymbol{X}_d^{(0)}, \boldsymbol{X}_d^{(1)}, \ldots, \boldsymbol{X}_d^{(n)}\}$ be i.i.d. random vectors with a common continuous cdf having a finite fourth moment. Let $N_1^{n,d}$ denote the number of elements from $\{\boldsymbol{X}_d^{(1)}, \ldots, \boldsymbol{X}_d^{(n)}\}$ whose nearest neighbor with* *respect to the Euclidean distance is* $\boldsymbol{X}_d^{(0)}$*. Then*

$$\lim_{n\to\infty}\lim_{d\to\infty} \mathrm{N}_1^{n,d} \xrightarrow{D} 0 \qquad \textit{and} \qquad \lim_{n\to\infty}\lim_{d\to\infty} \sigma^2(\mathrm{N}_1^{n,d}) = \infty.$$

The interpretation of the above result due to (Tversky et al., 1983) is that if the number of dimensions is large relative to the number of points, a large portion of points will have reverse nearest neighbor count equaling zero, whereas a small fraction (i.e., the hubs) will score large counts.

In order to provide an explanation for hubness, (Radovanovic et al., 2010) noticed that it is expected for points that are closer to the mean of the data distribution to be closer, on average, to all other points. However, empirical evidence indicates that this tendency is amplified by high-dimensionality, making points that reside in the proximity of the datas mean become closer to all other points than their low-dimensional analogues are. This tendency causes high-dimensional points that are closer to the mean to have increased probability of being selected as k-nearest neighbors by other points, even for small values of k.

In order to formalize the above evidence in finite-dimensional spaces, the authors considered the simplified setting of normally distributed i.i.d. d-dimensional random vectors, for which the distribution of Euclidean distances, which are calculated as the square root of the sum of squares of i.i.d. normal variables, corresponds to a chi distribution with d degrees of freedom (Johnson et al., 1994) and the random variable $\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|$, representing the distance from a fixed point $\boldsymbol{x}_d$ to the rest of the data, follows the noncentral chi distribution with d degrees of freedom and noncentrality parameter $\lambda = \|\boldsymbol{x}_d\|$ (Oberto and Pennecchi, 2006).

Theorem 5 (Radovanovic et al., 2010, cf. Theorem 1) *Let* $\lambda_{d,1} = \mu_{\chi(d)} + c_1\sigma_{\chi(d)}$ *and* $\lambda_{d,2} = \mu_{\chi(d)} + c_2\sigma_{\chi(d)}$*, where* $d \in \mathbb{N}^+$*,* $c_1 < c_2 \leq 0$*, and* $\mu_{\chi(d)}$ *and* $\sigma_{\chi(d)}$ *are the mean and standard deviation of the chi distribution with d degrees of freedom, respectively. Define* $\Delta\mu_d(\boldsymbol{x}_{d,1}, \boldsymbol{x}_{d,2}) = \mu_{\chi(d,\lambda_{d,2})} - \mu_{\chi(d,\lambda_{d,1})}$*, where* $\mu_{\chi(d,\lambda)}$ *is the mean of the noncentral chi distribution with d degrees of freedom and noncentrality parameter* λ*. Then, there exists* $d_0 \in \mathbb{N}$ *such that for every* $d > d_0$*,* $\Delta\mu_d(\lambda_{d,1}, \lambda_{d,2}) > 0$*, and* $\Delta\mu_{d+2}(\lambda_{d+2,1}, \lambda_{d+2,2}) > \Delta\mu_d(\lambda_{d,1}, \lambda_{d,2})$*.*

Intuitively, $\lambda_{d,1}$ and $\lambda_{d,2}$ represent two d-dimensional points whose norms are located at c_1 and c_2, resp., standard deviations from the expected value of the norm in the dimensionality d, and for which $\Delta\mu_d(\lambda_{d,1}, \lambda_{d,2})$ is the distance between the expected value of the associated distribution of distances.

As stated by authors, the implication of the above theorem is that hubness is an inherent property of data distributions in high-dimensional space, rather than an artifact of other factors, such as finite sample size. However, Theorem 5 only formalizes the tendency of the difference between the means of the two distance distributions to increase with the dimensionality, and the proof is specific for Gaussian data. No model to predict the number of k-occurrences or to infer the form of the underlying distribution is provided, and the characterization of the distribution probability of N_k remains an open problem.

3. Notation

In the rest of this section, upper case letters, such as X, Y, Z, ..., denote random variables (r.v.) taking values in $\mathbb{R}$. f_X (F_X, resp.) denotes the probability density function (pdf) (probability distribution function (cdf), resp.) associated with X.

Boldface uppercase letters with d as a subscript, such as $\boldsymbol{X}_d$, $\boldsymbol{Y}_d$, $\boldsymbol{Z}_d$, ..., denote d-dimensional random vectors taking values in $\mathbb{R}^d$. The components X_i ($1 \leq i \leq d$) of a random vector $\boldsymbol{X}_d = (X_1, X_2, \ldots, X_d)$ are random variables having pdfs $f_{X_i} = f_i$ (cdf $F_{X_i} = F_i$). A d-dimensional random vector is said to be independent and identically distributed (i.i.d. for short) if its random variables are independent and have common pdf $f_X = f_{X_i}$ (cdf $F_X = F_{X_i}$).

Boldface lowercase letters with d as a subscript, such as $\boldsymbol{x}_d$, $\boldsymbol{y}_d$, $\boldsymbol{z}_d$, ..., denote a specific d-dimensional vector taking value in $\mathbb{R}^d$. The components of a vector $\boldsymbol{x}_d = (x_1, x_2, \ldots, x_d)$, denoted as x_i ($1 \leq i \leq d$), are real scalar values.

Given a random variable X, w.l.o.g. and for simplicity of treatment, sometimes it is assumed that the expected value μ (or μ_X) of f_X is $\mu = 0$. If that is not the case, to satisfy the assumption, it suffices to replace during the analysis the original random variable X with the novel random variable $\hat{X} = X - \mu_X$. Thus, $\hat{\boldsymbol{X}}_d$ denotes the random vector $\boldsymbol{X}_d - \mu_X$.

σ_X or $\sigma(X)$ (or σ alone, whenever X is clear from the context) is the standard deviation of the random variable X. By μ_k ($\hat{\mu}_k$, resp.), or $\mu_{X,k}$ ($\hat{\mu}_{X,k}$, resp.) whenever X is not clear from the context, it is denoted the *k-th moment* (*k-th central moment*, resp.) ($k > 0$) $\mu_k = \mathbf{E}[X^k]$ ($\hat{\mu}_k = \mathbf{E}[(X - \mu_X)^k]$, resp.) of the random variable X, where $\mathbf{E}[X]$ is the expectation of X. Clearly, when $\mu = \mu_1 = 0$, μ_k coincides with $\hat{\mu}_k$ and $\mu_2 = \sigma^2$.

Moments of a pdf f (cdf F, resp.) are those associated with a random variable X having pdf $f_X = f$ (cdf $F_X = f$, resp.). The moments of an i.i.d. random vector $\boldsymbol{X}_d$ are those associated with its cdf F_X.

It is said that a distribution function has *finite* (*central*) *moment* μ_k, if there exists $0 \leq \mu_{top} < \infty$ such that $|\mu_k| \leq \mu_{top}$.

Whenever moments are employed during the proofs, we always assume that they exist and are finite. Moreover, if the random variable associated with a moment employed in a proof is not explicitly stated, we assume that the moment is relative to the common cdf of the random vector(s) occurring in the distribution reported in the statement of the theorem.

Moreover, it is sometimes considered the case that $\mu_3 = 0$, a condition that is referred to as null *skewness*. It is known that odd central moments, provided they exist, are null if the pdf of X is symmetric with respect to the mean (with examples of distributions having null μ_3 value being the Uniform and Normal distributions).

The notation $\mathcal{N}\left(\mu, \sigma^2\right)$ represents the Normal distribution function with mean μ and variance σ^2. By Φ (ϕ, resp.) one denotes the cdf (pdf, resp.) of the standard normal distribution, whereas by Φ_X (ϕ_X, resp.) one denotes the cdf (pdf, resp.) of the normal distribution with mean μ_X and variance σ_X^2.

Let X represent a univariate random variable that is defined in terms of a real-valued function of one or more d-dimensional random vectors. For example, X could be defined as $\|\boldsymbol{X}_d\|^2$. The notation $X \simeq \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ is shorthand to denote the fact that, as $d \to \infty$, the distribution $\widehat{F}_X$ of the standard score $\frac{X - \mu_X}{\sigma_X}$ of X converges to the normal distribution

$\mathcal{N}(0,1)$. Thus, for large values of d, $\mathcal{N}(\mu_X, \sigma_X^2)$ approximates the distribution probability F_X of X, and $Pr[X \leq \delta] \approx \Phi\left(\frac{\delta-\mu_X}{\sigma_X}\right)$.

In the following, $\|\cdot\|$ denotes the L_2 norm, i.e., $\|\boldsymbol{x}_d\| = \sqrt{\sum_{i=1}^d x_i^2}$. Moreover, dist$(P, Q)$ denotes the Euclidean distance $\|P - Q\|$ between (random) vector P and (random) vector Q.

Let $x \in \mathbb{R}$, and let X be a random variable. Then

$$z_{x,X} = \frac{x - \mu_X}{\sigma_X}$$

represents the value x standardized with respect to the mean and the standard deviation of X. For a d-dimensional vector $\boldsymbol{x}_d$, which is the realization of a d-dimensional i.i.d. random vector $\boldsymbol{Y}_d$, the notation $z_{\boldsymbol{x}_d}$ is used as shorthand for $z_{\|\boldsymbol{x}_d\|^2, \|\boldsymbol{Y}_d\|^2}$, i.e.,

$$z_{\boldsymbol{x}_d} = z_{\|\boldsymbol{x}_d\|^2, \|\boldsymbol{Y}_d\|^2} = \frac{\|\boldsymbol{x}_d\|^2 - \mu_{\|\boldsymbol{Y}_d\|^2}}{\sigma_{\|\boldsymbol{Y}_d\|^2}}.$$

Results in the following are derived by considering distributions. However, these results can be applied to a finite set of points by taking into account large samples. In order to deal with a finite set of points, $\{\boldsymbol{Y}_d\}_n$ denotes a set of n random vectors $\{\boldsymbol{Y}_d^{(1)}, \ldots, \boldsymbol{Y}_d^{(n)}\}$, each one distributed as $\boldsymbol{Y}_d$.

Now we recall the Lyapunov Central Limit Theorem (CLT) condition. Consider the sequence $W_1, W_2, W_3, \ldots$ of independent, but not identically distributed, random variables, and let $V_d = \sum_{i=1}^d W_i$. By the Lyapunov CLT condition (Ash and Doléans-Dade, 1999), if for some $\delta > 0$ it holds that

$$\lim_{d\to\infty} \frac{1}{s_d^{2+\delta}} \sum_{i=1}^d \mathbf{E}\left[|W_i - \mathbf{E}[W_i]|^{2+\delta}\right] = 0, \text{ where } s_d^2 = \sum_{i=1}^d \sigma_{W_i}^2, \tag{1}$$

then, as d goes to infinity,

$$\frac{U_d - \mathbf{E}[U_d]}{\sigma(U_d)} = \frac{\sum_{i=1}^d W_i - \sum_{i=1}^d \mathbf{E}[W_i]}{\sqrt{\sum_{i=1}^d \sigma_{W_i}^2}} \to \mathcal{N}(0,1),$$

i.e., the standard score $(V_d - \mathbf{E}[V_d])/\sigma(V_d)$ converges in distribution to a standard normal random variable.

In the following, when a statement involves a d-dimensional vector $\boldsymbol{x}_d$, we will usually assume that $\boldsymbol{x}_d$ is the realization of a specific d-dimensional random vector $\boldsymbol{X}_d$. Moreover, we will say that a result involving the realization $\boldsymbol{x}_d$ of a random vector $\boldsymbol{X}_d$ holds *with high probability* if the statement is true for all the realizations of $\boldsymbol{X}_d$ except for a subset which becomes increasingly less probable as the dimensionality d increases.

Technically, the assumption that $\boldsymbol{x}_d$ is a realization of a random vector $\boldsymbol{X}_d$ is leveraged to attain a proof of convergence in probability. This also means that when we simultaneously account for all the realizations of a random vector $\boldsymbol{X}_d$ (by integrating on all vectors $\boldsymbol{x}_d$ such that $f_X(\boldsymbol{x}_d) > 0$), the existence of such a negligible set does not affect the final result.

4. Results

This section presents the results of the work concerning distribution of distances, nearest neighbors, and reverse nearest neighbors.

Specifically, Section 4.1, concerning the distribution of distances between intrinsically high-dimensional data, derives the expressions for the distance distribution between pairs of random vectors and between a realization of a random vector and a random vector, and analyzes the error associated with expressions.

Section 4.2 takes into account the distribution of distances from nearest neighbors, derives the expected size of the ϵ-neighborhood in high-dimensional spaces, and leverages it to characterize neighborhood instability. The section also derives a novel estimate of the relative contrast measure.

Section 4.3 addresses the problem of determining the number of k-occurrences and determines the closed form of its limiting distribution, showing that it encompasses previous results and provides interpretability of the associated hubness phenomenon.

Section 4.4 generalizes the results derived for the i.i.d. case to independent non-identically distributed data, depicting connections with the behavior in real data.

Section 4.5 discusses relationship to other distances, including Minkowski's metrics and, in general, distances satisfying certain conditions of spatial centrality.

The first three sections deal with i.i.d. random vectors. In these sections, the synthetic data sets considered consist of data generated from a uniform distribution in $[-0.5, +0.5]$, a standard normal distribution, and an exponential distribution with mean 1.

For the proofs that are not reported within the main text, the reader is referred to the Appendix.

4.1 On the Distribution of Distances for i.i.d. Data

First of all, the probability that two d-dimensional i.i.d. random vectors lie at a distance not greater than δ from one another is considered. The expression of Theorem 6 results from the fact that the distribution of the random variable $\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2$ converges towards the normal distribution for large dimensionalities.

Theorem 6 *Let $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ be two d-dimensional i.i.d. random vectors with common cdf F. Then, for large values of d,*

$$Pr\left[\mathrm{dist}(\boldsymbol{X}_d, \boldsymbol{Y}_d) \leq \delta\right] \approx \Phi\left(\frac{\delta^2 - 2d(\mu_2 - \mu^2)}{\sqrt{2d\Big(\mu_4 + \mu_2^2 + 2\mu\big(\mu(2\mu_2 - \mu^2) - 2\mu_3\big)\Big)}}\right).$$

Proof of Theorem 6. The statement follows from the property shown in Lemma 7.

Lemma 7 $\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2 \simeq \mathcal{N}\left(2d(\mu_2 - \mu^2),\ 2d\left(\mu_4 + \mu_2^2 + 2\mu\big(\mu(2\mu_2 - \mu^2) - 2\mu_3\big)\right)\right)$.

Proof of Lemma 7. The squared norm $\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2$ can be written as $\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2 = \|\boldsymbol{X}_d\|^2 + \|\boldsymbol{Y}_d\|^2 - 2\langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle$, where $\|\boldsymbol{X}_d\|^2 \equiv \|\boldsymbol{Y}_d\|^2$, and $\langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle$ are the following random variables

$$\|\boldsymbol{Y}_d\|^2 = \sum_{i=1}^{d} Y_i^2 \text{ and } \langle \boldsymbol{X}_d, \boldsymbol{Y}_d \rangle = \sum_{i=1}^{d} X_i Y_i.$$

The proof proceeds by showing that, as $d \to \infty$, $\|\boldsymbol{X}_d\|^2$, $\|\boldsymbol{Y}_d\|^2$, and $\langle \boldsymbol{X}_d, \boldsymbol{Y}_d \rangle$ are both normally distributed and jointly normally distributed and by determining their covariance, which is accounted for in Propositions 8, 9, 10, and 11, as reported in the following.

Proposition 8 $\|\boldsymbol{Y}_d\|^2 \simeq \mathcal{N}\left(d\mu_2, d(\mu_4 - \mu_2^2)\right)$.

Proof of Proposition 8. Consider the random variable

$$\|\boldsymbol{Y}_d\|^2 = \sum_{i=1}^{d} Y_i^2 = \sum_{i=1}^{d} W_i,$$

where $W_i = Y_i^2$ is a novel random variable. Then, $\mu_W = \mathbf{E}[W_i] = \mathbf{E}[Y_i^2] = \mu_2$, and $\sigma_W^2 = \mathbf{E}[W_i^2] - \mathbf{E}[W_i]^2 = \mathbf{E}[Y_i^4] - \mu_2^2 = \mu_4 - \mu_2^2$.

Consider the sequence $W_1, W_2, W_3, \ldots$ of i.i.d. random variables. By the Central Limit Theorem (CLT for short) (Ash and Doléans-Dade, 1999), the standard score of W_i is such that, as $d \to \infty$,

$$\frac{\sum_{i=1}^{d} W_i - d\mu_W}{\sqrt{d\sigma_W}} = \frac{\sum_{i=1}^{d} Y_i^2 - d\mu_2}{\sqrt{d(\mu_4 - \mu_2^2)}} \to \mathcal{N}(0,1),$$

from which the result follows. ■

Proposition 9 $\langle \boldsymbol{X}_d, \boldsymbol{Y}_d \rangle \simeq \mathcal{N}\big(d\mu^2, d(\mu_2^2 - \mu^4)\big)$.

Proof of Proposition 9. Because $\langle \boldsymbol{X}_d, \boldsymbol{Y}_d \rangle = \sum_{i=1}^{d} X_i Y_i = \sum_{i=1}^{d} W_i$ is the sum of a sequence $W_1, W_2, W_3, \ldots$ of i.i.d. random variables with mean $\mathbf{E}[W_i] = \mathbf{E}[X_i Y_i] = \mathbf{E}[X_i]\mathbf{E}[Y_i] = \mu^2$ and variance $\sigma^2[W_i] = \mathbf{E}[W_i^2] - \mathbf{E}[W_i]^2 = \mathbf{E}[X_i^2 Y_i^2] - (\mu^2)^2 = \mathbf{E}[X_i^2]\mathbf{E}[Y_i^2] - \mu^4 = \mu_2^2 - \mu^4$, from the CLT the result follows. ■

Proposition 10 *As $d \to \infty$, $\|\boldsymbol{X}_d\|^2$, $\|\boldsymbol{Y}_d\|^2$ and $\langle \boldsymbol{X}_d, \boldsymbol{Y}_d \rangle$ are jointly normally distributed.*

Proof of Proposition 10. The statement holds provided that all linear combinations $W = a\|\boldsymbol{X}_d\|^2 + b\|\boldsymbol{Y}_d\|^2 + c\langle \boldsymbol{X}_d, \boldsymbol{Y}_d \rangle$ are normal. Notice that

$$W = a\left(\sum_{i=1}^{d} X_i^2\right) + b\left(\sum_{i=1}^{d} Y_i^2\right) + c\left(\sum_{i=1}^{d} X_i Y_i\right) = \sum_{i=1}^{d}\left(aX_i^2 + bY_i^2 + cX_iY_i\right) = \sum_{i=1}^{d} W_i,$$

where $W_i = aX_i^2 + bY_i^2 + cX_iY_i$ is a novel random variable. Because $W_1, W_2, W_3, \ldots$ is a sequence of i.i.d. random variables, the result follows from the CLT. ■

Proposition 11

$$cov\big(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle\big) = d\mu(\mu_3 - \mu_2\mu)$$
$$\big(\textit{and } cov\big(\|\boldsymbol{X}_d\|^2, \langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle\big) = d\mu(\mu_3 - \mu_2\mu), \textit{ for symmetry}\big)\,.$$

Proof of Proposition 11. See the appendix. ■

Proof of Lemma 7 (continued). Because the random variables $\|\boldsymbol{X}_d\|^2$, $\|\boldsymbol{Y}_d\|^2$, and $\langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle$ are jointly normally distributed (see Proposition 10), their linear combination $\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2 = \|\boldsymbol{X}_d\|^2 + \|\boldsymbol{Y}_d\|^2 - 2\langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle$ is normally distributed with mean $\mu_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2} = \mu_{\|\boldsymbol{X}_d\|^2} + \mu_{\|\boldsymbol{Y}_d\|^2} - 2\mu_{\langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle} = 2d(\mu_2 - \mu^2)$, and variance

$$\begin{aligned}
\sigma^2_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2} &= 2\sigma^2_{\|\boldsymbol{Y}_d\|^2} + (-2)^2\sigma^2_{\langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle} + 4(-2)cov(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle) = \\
&= 2d(\mu_4 - \mu_2^2) + 4d(\mu_2^2 - \mu^4) - 8d\mu(\mu_3 - \mu_2\mu) = \\
&= 2d\Big(\mu_4 + \mu_2^2 + 2\mu\big(\mu(2\mu_2 - \mu^2) - 2\mu_3\big)\Big).
\end{aligned}$$

■

Proof of Theorem 6 (continued). To conclude the proof: $Pr\left[\text{dist}(\boldsymbol{X}_d, \boldsymbol{Y}_d) \le \delta\right] = Pr\left[\text{dist}(\boldsymbol{X}_d, \boldsymbol{Y}_d)^2 \le \delta^2\right] = Pr\left[\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2 \le \delta^2\right] \approx \Phi_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2}(\delta^2)$. ■

Note that, if $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ have a common pdf with null mean ($\mu = 0$), $\|\boldsymbol{Y}_d\|^2$ ($\|\boldsymbol{X}_d\|^2$ equivalently) and $\langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle$ are uncorrelated, and being jointly normal distributed, they are also independent. In such a case, the parameters of the distribution can be expressed in the following simplified form.

Corollary 12 *Let $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ be two d-dimensional i.i.d. random vectors with common cdf F_X having mean μ. Then*

$$\|\hat{\boldsymbol{X}}_d - \hat{\boldsymbol{Y}}_d\|^2 \simeq \mathcal{N}\left(2d\hat{\mu}_2,\ 2d(\hat{\mu}_4 + \hat{\mu}_2^2)\right),$$

where $\hat{\boldsymbol{X}}_d = \boldsymbol{X}_d - \mu$ ($\hat{\boldsymbol{Y}}_d = \boldsymbol{Y}_d - \mu$, resp.) and $\hat{\mu}_k = \mathbf{E}[(X-\mu)^k]$ ($k > 0$) are the central moments of f_X (the moments of $f_{\hat{X}}$, resp.).

Proof of Corollary 12. Immediate from Theorem 7. ■

The notability of the above expression also stems from the following fact.

Proposition 13 $Pr[\text{dist}(\boldsymbol{X}_d, \boldsymbol{Y}_d) \le \delta] = Pr[\text{dist}(\hat{\boldsymbol{X}}_d, \hat{\boldsymbol{Y}}_d) \le \delta]$.

Proof of Proposition 13. Distances are not affected by translation. ■

Until now, it has been assumed that $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ have a common cdf. The following expression takes into account the case of different cdfs.

Corollary 14 *Let $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ be two d-dimensional i.i.d. random vectors with cdfs F_X and F_Y, respectively. Then $\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2 \simeq \mathcal{N}(\mu_{X,Y},\ \sigma^2_{X,Y})$, where*

$$\begin{aligned}
\mu_{X,Y} &= d(\mu_{X,2} + \mu_{Y,2} - 2\mu_X\mu_Y), \text{ and} \\
\sigma^2_{X,Y} &= d\big((\mu_{X,4} - \mu^2_{X,2}) + (\mu_{Y,4} - \mu^2_{Y,2}) + 4\mu_{X,2}\mu_{Y,2} + 4\mu_X\mu_Y\big(\mu_{X,2} + \mu_{Y,2} - \mu_X\mu_Y\big) + \\
&\quad -4\mu_X\mu_{Y,3} - 4\mu_Y\mu_{X,3}\big).
\end{aligned}$$

Proof of Corollary 14. The expression can be obtained by following the same line of reasoning of Theorem 7. ■

To characterize more precisely distance distributions, it is of interest to consider the case in which one of the two vectors is held fixed. With this aim, the following Theorem 15 concerns the probability that a given d-dimensional vector $\boldsymbol{x}_d$ and the realization of a d-dimensional i.i.d. random vector $\boldsymbol{Y}_d$ lie at a distance not greater than δ from one another. The result holds under the condition that $\boldsymbol{x}_d$ itself is the realization of a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$, with the cdf F_X of $\boldsymbol{X}_d$ not necessarily being identical to the cdf F_Y of $\boldsymbol{Y}_d$.

Formally, Theorem 15 holds with high probability because it relies on a proof of convergence in probability exploited in Proposition 17. Although not all the realizations may comply with the condition of Proposition 17 (e.g., consider the case $x_i = c^i$ with $c \neq 1$), it holds anyway for almost all the realizations, except for a set of vanishing measure.

Theorem 15 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$, and let $\boldsymbol{Y}_d$ be a d-dimensional i.i.d. random vector. Then, for large values of d, with high probability*

$$Pr\left[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \leq \delta\right] \approx \Phi\left(\frac{\delta^2 - \|\boldsymbol{x}_d\|^2 - d\mu_2 + 2\mu\sum_{i=1}^d x_i}{\sqrt{d(\mu_4 - \mu_2^2) + 4(\mu_2 - \mu^2)\|\boldsymbol{x}_d\|^2 - 4(\mu_3 - \mu\mu_2)\sum_{i=1}^d x_i}}\right),$$

where moments are relative to the random vector $\boldsymbol{Y}_d$.

Proof of Theorem 15. The proof relies on the result of Lemma 16 considering the distribution of $\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2$.

Lemma 16 *With high probability*

$$\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2 \simeq \mathcal{N}\left(\|\boldsymbol{x}_d\|^2 + d\mu_2 - 2\mu\sum_{i=1}^d x_i,\ d(\mu_4 - \mu_2^2) + 4(\mu_2 - \mu^2)\|\boldsymbol{x}_d\|^2 - 4(\mu_3 - \mu\mu_2)\sum_{i=1}^d x_i\right).$$

Proof of Lemma 16. Consider the equality $\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2 = \|\boldsymbol{x}_d\|^2 + \|\boldsymbol{Y}_d\|^2 - 2\langle\boldsymbol{x}_d, \boldsymbol{Y}_d\rangle$. The proof proceeds by studying the distribution of $\langle\boldsymbol{x}_d, \boldsymbol{Y}_d\rangle$ (see Proposition 18), by showing that $\|\boldsymbol{Y}_d\|^2$ and $\langle\boldsymbol{x}_d, \boldsymbol{Y}_d\rangle$ are jointly normally distributed (see Proposition 19), and by determining their covariance (see Proposition 20). However, a technical result that is leveraged in the sequel of the proof is first needed; this is presented in Proposition 17.

Proposition 17 *Let $\boldsymbol{X}_d$ be a d-dimensional i.i.d. random vector having cdf F_X. Moreover, let p and q be positive integers, and $\beta_0, \beta_1, \ldots, \beta_p$, $\alpha_0, \alpha_1, \ldots, \alpha_q$ be real coefficients such that $\beta_p \neq 0$ and $\alpha_q \neq 0$. Then, for any $\epsilon > 0$,*

$$\lim_{d\to\infty} Pr\left[\left|\frac{\sum_{i=1}^{d}\left(\sum_{j=0}^{p}\beta_j X_i^j\right)}{\left(\sum_{i=1}^{d}\left(\sum_{j=0}^{q}\alpha_j X_i^j\right)\right)^2}\right| \geq \epsilon\right] = 0.$$

Proof of Proposition 17. See the appendix. ■

Proposition 18 *With high probability $\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle \simeq \mathcal{N}\big(\mu\sum_{i=1}^d x_i,\ (\mu_2-\mu^2)\|\boldsymbol{x}_d\|^2\big)$.*

Proof of Proposition 18. Consider the random variable $\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle$:

$$\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle = \sum_{i=1}^{d} x_i Y_i = \sum_{i=1}^{d} W_i,$$

where $W_i = x_i Y_i$ is a novel random variable. Then, $\mu_{W_i} = \mathbf{E}[W_i] = \mathbf{E}[x_i Y_i] = x_i \mathbf{E}[Y_i] = x_i\mu$, and $\sigma^2_{W_i} = \mathbf{E}[W_i^2] - \mathbf{E}[W_i]^2 = \mathbf{E}[x_i^2 Y_i^2] - x_i^2\mu^2 = x_i^2\mu_2 - x_i^2\mu^2 = (\mu_2-\mu^2)x_i^2$.

Consider the sequence $W_1, W_2, W_3, \ldots$ of independent, but not identically distributed, random variables. If the Lyapunov CLT condition reported in Equation (1) holds, the standard score $(\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle - \mu_{\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle})/\sigma_{\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle}$ converges in distribution to a standard normal random variable as d goes to infinity, i.e.,

$$\frac{\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle - \mu_{\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle}}{\sigma_{\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle}} = \frac{\sum_{i=1}^d W_i - \sum_{i=1}^d \mathbf{E}[W_i]}{\sum_{i=1}^d \sigma^2_{W_i}} = \frac{\sum_{i=1}^d x_i Y_i - \mu\sum_{i=1}^d x_i}{\sqrt{\mu_2-\mu^2}\|\boldsymbol{x}_d\|} \to \mathcal{N}(0,1).$$

Thus, consider the Lyapunov condition for $\delta = 2$:

$$\lim_{d\to\infty} \left.\frac{\sum_{i=1}^d \mathbf{E}\left[|W_i - \mathbf{E}[W_i]|^{2+\delta}\right]}{s_d^{2+\delta}}\right|_{\delta=2} = \lim_{d\to\infty}\frac{\sum_{i=1}^d \mathbf{E}\left[|x_i(Y_i-\mu)|^4\right]}{(\mu_2-\mu^2)^2\|\boldsymbol{x}_d\|^4} =$$

$$= \frac{\mu_4 + \mu(6\mu\mu_2 - 4\mu_3 - 3\mu^3)}{(\mu_2-\mu^2)^2}\cdot\lim_{d\to\infty}\frac{\sum_{i=1}^d x_i^4}{\left(\sum_{i=1}^d x_i^2\right)^2} = 0.$$

The above limit converges in probability for the r.v. $\boldsymbol{X}_d$ by Proposition 17. ■

Proposition 19 *As $d\to\infty$, with high probability $\|\boldsymbol{Y}_d\|^2$ and $\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle$ are jointly normally distributed.*

Proof of Proposition 19. See the appendix. ■

Proposition 20 *cov* $\left(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle\right) = (\mu_3 - \mu\mu_2)\sum_{i=1}^{d} x_i$.

Proof of Proposition 20. See the appendix. ■

Proof of Lemma 16 (continued). To conclude the proof of Lemma 16, because the random variables $\|\boldsymbol{Y}_d\|^2$, and $\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle$ are jointly normally distributed, then the random variable $\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2$ is normally distributed with mean

$$\mu_{\|\boldsymbol{x}_d-\boldsymbol{Y}_d\|^2} = \mu_{\|\boldsymbol{x}_d\|^2} + \mu_{\|\boldsymbol{Y}_d\|^2} - 2\mu_{\langle \boldsymbol{x}_d,\boldsymbol{Y}_d\rangle} = \|\boldsymbol{x}_d\|^2 + d\mu_2 - 2\mu\sum_{i=1}^{d} x_i,$$

and variance

$$\begin{aligned}\sigma^2_{\|\boldsymbol{x}_d-\boldsymbol{Y}_d\|^2} &= \sigma^2_{\|\boldsymbol{Y}_d\|^2} + (-2)^2\sigma^2_{\langle \boldsymbol{x}_d,\boldsymbol{Y}_d\rangle} + 2(-2)\ cov(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle). = \\ &= d(\mu_4 - \mu_2^2) + 4(\mu_2 - \mu^2)\|\boldsymbol{x}_d\|^2 - 4(\mu_3 - \mu\mu_2)\left(\sum_{i=1}^{d} x_i\right).\end{aligned}$$

■

Proof of Theorem 15 (continued). To conclude the proof: $Pr\left[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \le \delta\right] = Pr\left[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d)^2 \le \delta^2\right] = Pr\left[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2 \le \delta^2\right] = \Phi_{\|\boldsymbol{x}_d-\boldsymbol{Y}_d\|^2}(\delta^2)$. ■

For distributions having null means, the above expressions can be simplified.

Corollary 21 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$, and let $\boldsymbol{Y}_d$ be a d-dimensional i.i.d. random vector with cdf F_Y having null mean $\mu_Y = 0$. Then, with high probability*

$$\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2 \simeq \mathcal{N}\left(\|\boldsymbol{x}_d\|^2 + d\mu_2,\ d(\mu_4 - \mu_2^2) + 4\mu_2\|\boldsymbol{x}_d\|^2 - 4\mu_3\sum_{i=1}^{d} x_i\right),$$

where the moments are relative to the random vector $\boldsymbol{Y}_d$.

Proof of Corollary 21. The result follows by substituting $\mu = \mu_Y = 0$ in the right-hand side of the statement of Lemma 16. ■

In order to quantify the error associated with the approximations of Theorem 6 and Theorem 15, the Kolmogorov-Smirnov statistic D_n is employed here as an error measure. This statistic is usually used for comparing a theoretical cumulative distribution function F to a given empirical distribution function G_n for n observations, and it is defined as

$$D_n(G_n, F) = \sup_{\delta\in\mathbb{R}} |G_n(\delta) - F(\delta)|\,.$$

In our case, given an empirical distribution function $G_{d,n}$ for n observations and a theoretical distribution function F_d, both related to the dimensionality parameter d, we define the error $err_d(G_{d,n}, F_d)$ as $D_n(G_{d,n}, F_d)$.

As for the approximation of Theorem 6, $F_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2}(\delta) = \Phi\left(\frac{\delta - \mathbf{E}[\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2]}{\sigma(\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2)}\right)$ is employed as theoretical cdf F_d, whereas $F^{emp}_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2,n}(\delta)$, denoting the empirical distribution of the squared distances, is employed as the empirical cdf $G_{d,n}$, and the error measured is $e_d = err_d\left(F^{emp}_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2,n}, F_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2}\right)$.

As the approximation of Theorem 15, given the realization $\boldsymbol{x}_d$ of $\boldsymbol{X}_d$, $F_{\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2}(\delta) = \Phi\left(\frac{\delta - \mathbf{E}[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2]}{\sigma(\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2)}\right)$ is employed as a theoretical cdf, whereas $F^{emp}_{\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2,n}(\delta)$ denotes the empirical cdf. Specifically, we considered three points $\boldsymbol{p}_d^{(i)}$ $(1 \leq i \leq 3)$ as instances of $\boldsymbol{x}_d$. Each point $\boldsymbol{p}_d^{(i)}$ lies k_i (with $k_1 = 0$, $k_2 = 1$, and $k_3 = 5$) standard deviations $\sigma_{\|\boldsymbol{X}_d\|^2}$ away from the mean $\mu_{\|\boldsymbol{X}_d\|^2}$ of the squared norm of $\boldsymbol{X}_d$, i.e., each point $\boldsymbol{p}_d^{(i)}$ is such that $z_{\|\boldsymbol{p}_d^{(i)}\|^2, \|\boldsymbol{X}_d\|^2} = k_i$ (in particular, the generic coordinate of $\boldsymbol{p}_d^{(i)}$ has value $\left((\mu_{\|\boldsymbol{X}_d\|^2} + k_i \cdot \sigma_{\|\boldsymbol{X}_d\|^2})/d\right)^{1/2}$). The error measured for each point is $e_d^{(i)} = err_d\left(F^{emp}_{\|\boldsymbol{p}_d^{(i)} - \boldsymbol{Y}_d\|^2,n}, F_{\|\boldsymbol{p}_d^{(i)} - \boldsymbol{Y}_d\|^2}\right)$.

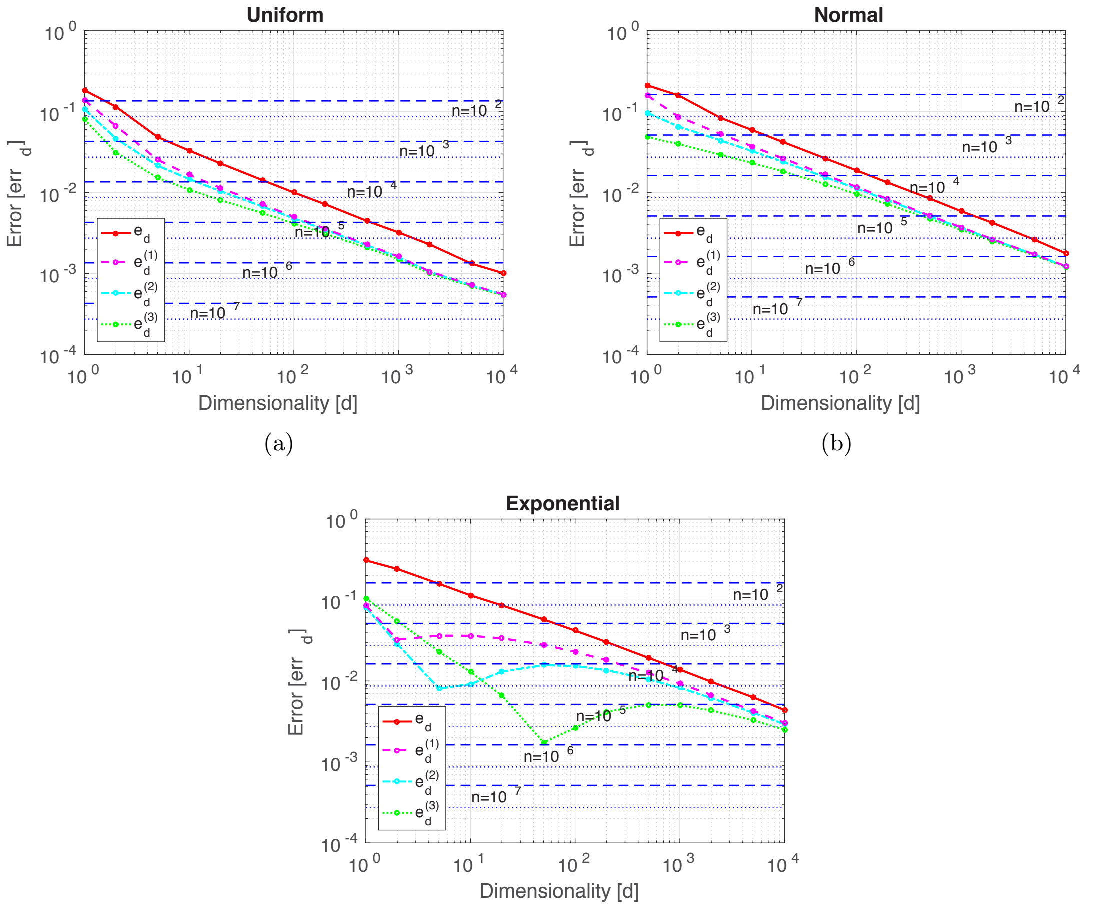
The empirical cdf $F^{emp}_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2,n}$ has been obtained by generating n pairs $(\boldsymbol{x}_d^{(j)}, \boldsymbol{y}_d^{(j)})$ $(1 \leq j \leq n)$ of realizations of the random vectors $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$, respectively, and then by computing $F^{emp}_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2,n}(\delta) = \frac{1}{n}\sum_{j=1}^{n} I_{[0,\delta]}\left(\mathrm{dist}(\boldsymbol{x}_d^{(j)}, \boldsymbol{y}_d^{(j)})\right)$, where I_S denotes the indicator function (with S representing a generic set), such that $I_S(v) = 1$, if $v \in S$, and $I_S(v) = 0$, otherwise. The empirical cdf $F^{emp}_{\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2,n}$, is obtained by generating n realizations $\boldsymbol{y}_d^{(j)}$ $(1 \leq j \leq n)$ of the random vector $\boldsymbol{Y}_d$, and then by computing $F^{emp}_{\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2,n}(\delta) = \frac{1}{n}\sum_{j=1}^{n} I_{[0,\delta]}\left(\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d^{(j)})\right)$.

We note that, for any distance threshold $\delta \geq 0$, the value err_d represents an upper bound to the error committed when the theoretical cdf of Theorem 6 (Theorem 15, resp.) is used to estimate the probability $Pr[\|\boldsymbol{X}_d - \boldsymbol{Y}_d\| \leq \delta]$ ($Pr[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\| \leq \delta]$, resp.).

Figure 1 shows the above defined errors e_d, $e_d^{(1)}$, $e_d^{(2)}$, and $e_d^{(3)}$ (red curves), for distributions $F_X = F_Y$, uniform in $[-0.5, +0.5]$ (Fig. 1a), standard normal (Fig. 1b), and exponential with $\lambda = 1$ (Fig. 1c), respectively.

Before commenting on the results, it must be pointed out that the error err_d depends on the size n of the sample employed to build the empirical distribution. Thus, first we discuss the behavior for unbounded sample sizes n, and then take into account the effect of finite sample sizes. In order to simulate an unbounded sample size, the curves in the figures have been obtained for a very large sample size $n > 1.5 \cdot 10^8$.

From Figures 1a, 1b, and 1c it can be seen that the error err_d decreases with the dimensionality. The trend of the error curves is more regular for the uniform and normal distribution than for the exponential distribution, probably due to the skewness of the exponential distribution. The error e_d associated with the cdf $F_{\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2}$ is greater than the errors $e_d^{(i)}$ associated with the cdf $F_{\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2}$. Intuitively, this can be explained since the degree of uncertainty is reduced if one of the two random vectors is replaced by a fixed point. In general, it holds that $e_d^{(1)} > e_d^{(2)} > e_d^{(3)}$, thus indicating that uncertainty



(a)

(b)

(c)

Figure 1: [Best viewed in color.] Empirical evaluation of the approximation errors of Th. 6 and Th. 15, for dimensionalities $d \in [10^0, 10^4]$ and sample sizes $n \in [10^2, 10^7]$. Error e_d (red solid line) is associated with the expression of Th. 6, whereas errors $e_d^{(1)}$ (magenta dashed line), $e_d^{(2)}$ (cyan dash-dotted line), and $e_d^{(3)}$ (green dotted line) are associated with the expression of Th. 15, for three different points whose squared norm standard scores are 0, 1, and 5, respectively. Horizontal blue lines take into account the sample size n: the dotted line is the expected error for different n values under the hypothesis that the distance distribution is indeed normal; the dashed line is the value under which the hypothesis that the sample is generated according the theoretical distribution can be accepted at the 95% confidence level.

increases towards the most largely populated regions of the space. Moreover, the larger the dimensionality d, the closer the errors $e_d^{(j)}$ to $e_d^{(1)}$.

As anticipated above, the error err_d depends on the size n of the sample employed to build the empirical distribution. Specifically, differently from the case of unbounded n values, for which the error decreases with the dimensionality, for any fixed sample size n, there exists a dimensionality d such that the error converges around a value $\overline{e}_n$. Such a value $\overline{e}_n$ corresponds to the error $D_n(\Phi_n^{emp}, \Phi)$ between the empirical cdf Φ_n^{emp} associated with a random sample of n elements of a (standard) normal distribution and the theoretical cdf Φ of a (standard) normal distribution.

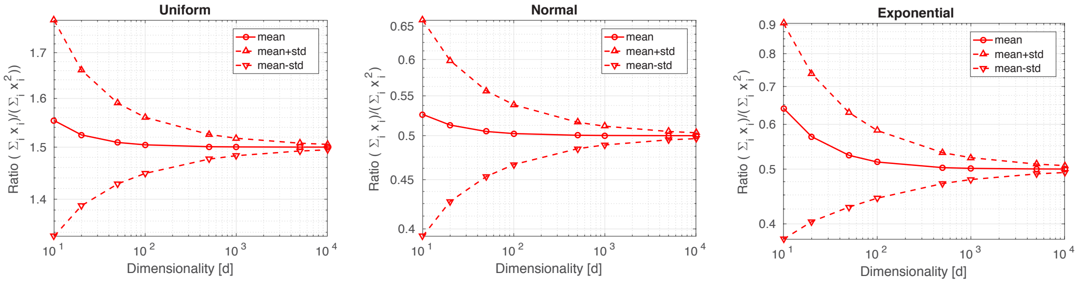
Let K be a random variable having a Kolmogorov distribution. According to the Kolmogorov-Smirnov test, the null hypothesis that the sample of n observations having empirical distribution G_n comes from the hypothesized distribution F is rejected at level $\alpha \in (0,1)$ if the statistic $\sqrt{n} \cdot D_n(G_n, F)$ is greater than the value K_α, where K_α is such that $Pr[K \leq K_\alpha] = 1-\alpha$. It follows from the above that if, for a certain α and sample size n, it holds that err_d is smaller than $e_n^\alpha = K_\alpha \cdot n^{-1/2}$, then the hypothesis that the sample complies with the theoretical distribution can be accepted at the $1-\alpha$ confidence level, e.g., for $\alpha = 0.05$, the value $K_{0.05}$ is 1.3581. Moreover, the expected value $\overline{e}_n$ of $D_n(\Phi_n^{emp}, \Phi)$ approximately corresponds to $K_{\overline{\alpha}} \cdot n^{-1/2}$ with $\overline{\alpha} = 0.44$, i.e., to $\overline{e}_n \approx 0.8673 \cdot n^{-1/2}$.

Horizontal (blue) lines in Figures 1a, 1b, and 1c take into account the effect of the sample size n. Each pair of dashed and dotted lines is associated with a different value of $n \in \{10^2, 10^3, \ldots, 10^7\}$. Dashed lines are associated with the errors $e_n^{0.05}$, whereas dotted lines are associated with the errors $\overline{e}_n$. Let n be the actual sample size, and let d^* be the dimensionality such that the value e_{d^*} of the particular curve e_d is equal to $\overline{e}_n$ (dotted horizontal curve). Then, for $d \geq d^*$, the expected value of e_d tends to $\overline{e}_n$. Thus, for $d < d^*$, the curve of e_d is similar to the one reported in the figure, whereas for $d > d^*$, the curve of e_d tends to be horizontal, with a value close to $\overline{e}_n$. Moreover, if $e_d \leq e_n^{0.05}$ (dashed horizontal curve), then the hypothesis that the sample complies with the hypothesized distribution can be accepted at the 95% confidence level. Informally speaking, this means that in the latter case, the distribution hypothesized in Theorems 6 and 15 is indiscernible from the underlying distribution generating the observed inter-point distances.

In summary, as previously pointed out, because err_d depends on the worst-case threshold value δ, it is an upper bound to the error committed when estimating probabilities by leveraging the results previously presented. The analysis with unbounded sample size highlights that the worst-case error always decreases with the dimensionality. Moreover, let the effective error be defined as the difference between the observed error and the error expected when the data are generated according to the hypothesized distribution. The analysis of finite sample sizes highlights that, in practice, the effective error can become null.

For the distributions F_Y having both a null mean and null skewness ($\mu_3 = 0$), it follows from Propositions 19 and 20 that the random variables $\|\boldsymbol{Y}_d\|^2$ and $\langle \boldsymbol{x}_d, \boldsymbol{Y}_d \rangle$ are independent.

Moreover, the distribution defined in Corollary 21, in Theorem 15 and in Lemma 16, depend only on the squared norm $\|\boldsymbol{x}_d\|^2$, whereas the actual value of $\boldsymbol{x}_d$ does not matter. However, it can be shown that the same property holds also for skewed distributions, since the term $(\sum_i x_i)$ is related to $\|\boldsymbol{x}_d\|^2$, as accounted for in the subsequent result.



(a) (b) (c)

Figure 2: Empirical validation of Proposition 22 on different distributions: (a) uniform ($\mu_1/\mu_2 = 1.5$), (b) normal ($\mu_1/\mu_2 = 0.5$), and (c) exponential ($\mu_1/\mu_2 = 0.5$). The red solid curve represents the expected value μ_W of the ratio $W = (\sum_{i=1}^{d} X_i)/\|\boldsymbol{X}_d\|^2$, whereas the red dashed curves represent the values $\mu_W + \sigma_W$ and $\mu_W - \sigma_W$, measured for $n = 20{,}000$ points and $d \in [10^1, 10^4]$.

Proposition 22 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$ with cdf F_X. Then, for large values of d, with high probability*

$$\frac{\sum_{i=1}^{d} x_i}{\|\boldsymbol{x}_d\|^2} \to \frac{\mu_X}{\mu_{X,2}}.$$

Proof of Proposition 22. See the appendix. ■

Thus, the term $(\sum_i x_i)$ can be approximated by $\frac{\mu_X}{\mu_{X,2}}\|\boldsymbol{x}_d\|^2$.

Notice that the above result also states that for random vectors $\boldsymbol{X}_d$ having null mean, the term $(\sum_i x_i)$ becomes negligible with respect to $\|\boldsymbol{x}_d\|^2$ and, hence, that it can be ignored in the expression reported in Corollary 21, thus removing the dependence from the skewness of the distribution F_Y.

To empirically validate Proposition 22, the mean and the standard deviation of the ratio $W = (\sum_{i=1}^{d} X_i)/\|\boldsymbol{X}_d\|^2$ have been measured on distributions having non-null mean $\mu \neq 0$. Figure 2 reports the result of the experiment for $d \in [10, 10^4]$ and $n = 20{,}000$. Specifically, a uniform distribution with mean $\mu_1 = 0.5$ ($\mu_2 = 0.333$, $\mu_3 = 0.25$, and $\mu_4 = 0.2$) and ratio $\mu_1/\mu_2 = 1.5$ (Fig. 2a), a normal distribution with mean $\mu_1 = 1$ ($\mu_2 = 2$, $\mu_3 = 4$, and $\mu_4 = 10$) and ratio $\mu_1/\mu_2 = 0.5$ (Fig. 2b), and an exponential distribution with mean $\mu_1 = 1$ ($\mu_2 = 2$, $\mu_3 = 6$, and $\mu_4 = 24$) and ratio $\mu_1/\mu_2 = 0.5$ (Fig. 2c), were considered. It can be seen that the expected value $\mathbf{E}[W]$ of the ratio W rapidly converges to the limiting value μ_1/μ_2 and also that the standard deviation $\sigma(W)$ of the ratio W decreases with the dimensionality. Moreover, in all cases, the trend agrees with the prediction of Proposition 22, according to which it holds that $|\mathbf{E}[W] - \mu_1/\mu_2| = O(d^{-1})$ and $\sigma(W) = O(d^{-1/2})$.

4.2 On the Distribution of Nearest Neighbors for i.i.d. Data

Given a real number $\varrho \in [0,1]$, a d-dimensional vector $\boldsymbol{x}_d$ and a d-dimensional random vector $\boldsymbol{Y}_d$, $\mathrm{distnn}_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$ denotes the radius of the smallest neighborhood centered in $\boldsymbol{x}_d$ containing at least the ϱ fraction of the realizations of $\boldsymbol{Y}_d$. Moreover, $nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$, or $nn_\varrho(\boldsymbol{x}_d)$ whenever $\boldsymbol{Y}_d$ is clear from the context, also called *ϱ-th nearest neighbor* of $\boldsymbol{x}_d$ w.r.t. $\boldsymbol{Y}_d$, denotes an element of the set $\{\boldsymbol{y}_d \in \mathbb{R}^d \mid f_Y(\boldsymbol{y}_d) > 0 \text{ and } \mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) = \mathrm{distnn}_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)\}$.[1] $\mathrm{NN}_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$, or $\mathrm{NN}_\varrho(\boldsymbol{x}_d)$ whenever $\boldsymbol{Y}_d$ is clear from the context, denotes the set of points $\{\boldsymbol{y}_d \in \mathbb{R}^d \mid f_Y(\boldsymbol{y}_d) > 0 \text{ and } \mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) \leq \mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d))\}$.

In order to deal with finite sets of n points $\{\boldsymbol{Y}_d\}_n$, the integer parameter $k = \varrho n$ ($k \in \{1, \ldots, n\}$) has to be employed in place of ϱ. Thus, given a positive integer k, $\mathrm{distnn}_k(\boldsymbol{x}_d, \{\boldsymbol{Y}_d\}_n)$ represents the radius of the smallest neighborhood centered in $\boldsymbol{x}_d$ containing at least k points of $\{\boldsymbol{Y}_d\}_n$. Moreover, $nn_k(\boldsymbol{x}_d, \{\boldsymbol{Y}_d\}_n)$ or $nn_k(\boldsymbol{x}_d)$, also called *k-th nearest neighbor* of $\boldsymbol{x}_d$ in $\{\boldsymbol{Y}_d\}_n$, denotes an element of the set $\{\boldsymbol{y}_d \in \{\boldsymbol{Y}_d\}_n \mid \mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) = \mathrm{distnn}_k(\boldsymbol{x}_d, \{\boldsymbol{Y}_d\}_n)\}$.[2] $\mathrm{NN}_k(\boldsymbol{x}_d, \{\boldsymbol{Y}_d\}_n)$, or $\mathrm{NN}_k(\boldsymbol{x}_d, \boldsymbol{Y}_d)$, denotes the set of points $\{\boldsymbol{y}_d \in \{\boldsymbol{Y}_d\}_n \mid \mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) \leq \mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \{\boldsymbol{Y}_d\}_n))\}$.

In the rest of the work, given a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$ with cdf F_X, representing the distribution of the query points, and a d-dimensional i.i.d. random vector $\boldsymbol{Y}_d$ with cdf F_Y, representing the distribution of the data points, we assume w.l.o.g. that F_Y has null mean μ_Y. Indeed, if it is not the case, it is sufficient to replace them with the random vectors $\boldsymbol{X}'_d = \boldsymbol{X}_d - \mu_Y$ and $\boldsymbol{Y}'_d = \hat{\boldsymbol{Y}}_d = \boldsymbol{Y}_d - \mu_Y$ such that $\mu_{Y'} = 0$. Moreover, a realization $\boldsymbol{x}_d$ of $\boldsymbol{X}_d$ can be replaced with $\boldsymbol{x}'_d = \boldsymbol{x}_d - \mu_Y$.

The following result considers the distance separating a vector from its ϱ-th nearest neighbor w.r.t. a d-dimensional i.i.d. random vector.

Lemma 23 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$ having cdf F_X. Consider the ϱ-nearest neighbor $nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$ of $\boldsymbol{x}_d$ w.r.t. a d-dimensional i.i.d. random vector $\boldsymbol{Y}_d$ with cdf F_Y. Assume, w.l.o.g., that F_Y has null mean $\mu_Y = 0$. Then, for large values of d, with high probability*

$$\mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)) \approx \sqrt{\|\boldsymbol{x}_d\|^2 + d\mu_2 + \Phi^{-1}(\varrho)\sqrt{d(\mu_4 - \mu_2^2) + 4\mu_2\|\boldsymbol{x}_d\|^2 - 4\mu_3\sum_{i=1}^d x_i}}.$$

Proof of Lemma 23. By definition, $nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$ is such that

$$Pr\left[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \leq \mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d))\right] = \varrho.$$

By Corollary 21,

$$Pr\left[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \leq \mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d))\right] \approx \Phi\left(\frac{\mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d))^2 - \mathbf{E}[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2]}{\sigma(\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2)}\right).$$

[1] Because our interest is only in the fact that $nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$ satisfies the property $\mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)) = \mathrm{distnn}_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$, it can be assumed that $nn_\varrho(\boldsymbol{x}_d)$ is randomly selected from the above set.

[2] Because our interest is only in the fact that $nn_k(\boldsymbol{x}_d, \boldsymbol{Y}_d)$ satisfies the property $\mathrm{dist}(\boldsymbol{x}_d, nn_k(\boldsymbol{x}_d, \{\boldsymbol{Y}_d)\}_n) = \mathrm{distnn}_k(\boldsymbol{x}_d, \{\boldsymbol{Y}_d\}_n)$, it can be assumed that $nn_k(\boldsymbol{x}_d)$ is randomly selected from the above set.

Hence, $\mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d))^2 \approx \mathbf{E}[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2] + \Phi^{-1}(\varrho)\,\sigma(\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2)$. ■

It has been already pointed out that if F_Y has null skewness ($\mu_3 = 0$), if $F_X = F_Y$, or if F_X has null mean $\mu_X = 0$, the term $4\mu_3(\sum_i x_i)$ can be disregarded.

Due to the difficulty of answering nearest neighbor queries in high-dimensional spaces, different authors have proposed to consider approximate nearest neighbor queries (Indyk and Motwani, 1998; Arya et al., 1998), returning an ϵ-approximate nearest neighbor instead of the exact nearest neighbor: given point $\boldsymbol{x}_d$ and $\epsilon \geq 0$, a point $\boldsymbol{y}_d$ is an ϵ-approximate nearest neighbor of $\boldsymbol{x}_d$ if it holds that $\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) \leq (1+\epsilon)\mathrm{dist}(\boldsymbol{x}_d, nn_1(\boldsymbol{x}_d))$.

Beyer et al. (1999) called a nearest neighbor query *unstable* for a given $\epsilon \geq 0$, if the distance from the query point to most data points is less than $(1+\epsilon)$ times the distance from the query point to its nearest neighbor. Moreover, Beyer et al. (1999) have shown that in many situations, for any fixed $\epsilon > 0$, as dimensionality rises, the probability that a query is unstable converges to 1 (see Theorem 2).

Instability is undesirable because the points that fall in the enlarged query region, also called ϵ-neighborhood, are valid answers to the approximate nearest neighbor problem. Thus, the larger the expected number of data points falling within the ϵ-neighborhoods of the query points, the smaller the meaningfulness of the approximate query scenario.

Definition 24 *Let* $\mathrm{NN}^{\epsilon}_{\varrho}(\boldsymbol{x}_d, \boldsymbol{Y}_d)$ *denote the set of the* ϵ*-approximate* ϱ*-nearest neighbors of* $\boldsymbol{x}_d$*, also called* ϵ*-neighborhood, that are the realizations* $\boldsymbol{y}_d$ *of* $\boldsymbol{Y}_d$ *whose distance from* $\boldsymbol{x}_d$ *is within* $(1+\epsilon)$ *times the distance separating* $\boldsymbol{x}_d$ *from its* ϱ*-th nearest neighbor* $nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$*, i.e.,*

$$\mathrm{NN}^{\epsilon}_{\varrho}(\boldsymbol{x}_d, \boldsymbol{Y}_d) = \{\boldsymbol{y}_d \in \mathbb{R}^d \mid f_Y(\boldsymbol{y}_d) > 0 \text{ and } \mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) \leq (1+\epsilon)\,\mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d))\}.$$

In order to quantify the meaningfulness of ϵ-approximate queries, it is sensible to compute the expected size of the ϵ-neighborhoods associated with query points with respect to the data population, which is the task pursued in the following.

Theorem 25 *Let* $\epsilon \geq 0$*, let* $\boldsymbol{X}_d$ *be a* d*-dimensional i.i.d. random vector with cdf* F_X*, representing the distribution of the query points, and let* $\boldsymbol{Y}_d$ *be a* d*-dimensional i.i.d. random vector with cdf* F_Y *(not necessarily identical to* F_X*), representing the distribution of the data points. Assume, w.l.o.g., that* F_Y *has null mean* $\mu_Y = 0$*. Then, for large values of* d*,*

$$\mathbf{E}[|\mathrm{NN}^{\epsilon}_{k}(\boldsymbol{X}_d, \{\boldsymbol{Y}_d\}_n)|] \approx$$

$$n\Phi\left(\frac{(\epsilon^2+2\epsilon)\,d(\mu_{X,2}+\mu_{Y,2}) + (1+\epsilon)^2\,\Phi^{-1}(\frac{k}{n})\sqrt{d\left(\mu_{Y,4} - \mu_{Y,2}^2 + 4\mu_{Y,2}\mu_{X,2} - 4\mu_{Y,3}\mu_X\right)}}{\sqrt{d\left(\mu_{Y,4} - \mu_{Y,2}^2 + 4\mu_{Y,2}\mu_{X,2} - 4\mu_{Y,3}\mu_X\right) + (\epsilon^2+2\epsilon)^2\,d(\mu_{X,4} - \mu_{X,2}^2)}}\right).$$

Proof of Theorem 25. Consider the probability (exploiting Corollary 21, Proposition 22 and Lemmas 16 and 23)

$$\begin{aligned}
& Pr[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \leq (1+\epsilon)\,\mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d))] = \\
= \;& Pr[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d)^2 \leq (1+\epsilon)^2 \mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d))^2] \approx \\
\approx \;& \Phi\left(\frac{(1+\epsilon)^2\left(\mathbf{E}[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2] + \Phi^{-1}(\varrho)\,\sigma(\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2)\right) - \mathbf{E}[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2]}{\sigma(\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2)} \right) = \\
= \;& \Phi\left(\frac{(\epsilon^2+2\epsilon)\,\mathbf{E}[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2] + (1+\epsilon)^2\,\Phi^{-1}(\varrho)\,\sigma(\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2)}{\sigma(\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2)} \right) = \\
= \;& \Phi\left(\frac{(\epsilon^2+2\epsilon)(\|\boldsymbol{x}_d\|^2 + d\mu_{Y,2}) + (1+\epsilon)^2\,\Phi^{-1}(\varrho)\,\sqrt{d(\mu_{Y,4}-\mu_{Y,2}^2) + 4\mu_{Y,2}\|\boldsymbol{x}_d\|^2 - 4\mu_{Y,3}\frac{\mu_X}{\mu_{X,2}}\|\boldsymbol{x}_d\|^2}}{\sqrt{d(\mu_{Y,4}-\mu_{Y,2}^2) + 4\mu_{Y,2}\|\boldsymbol{x}_d\|^2 - 4\mu_{Y,3}\frac{\mu_X}{\mu_{X,2}}\|\boldsymbol{x}_d\|^2}} \right).
\end{aligned}$$

By taking into account the standard score of $\boldsymbol{x}_d$

$$\|\boldsymbol{x}_d\|^2 = z_{\boldsymbol{x}_d}\sigma_{\|\boldsymbol{X}_d\|^2} + \mu_{\|\boldsymbol{X}_d\|^2} = z_{\boldsymbol{x}_d}\sqrt{d(\mu_{X,4}-\mu_{X,2}^2)} + d\mu_{X,2},$$

and by considering that for α, β, and z finite (note that $\phi(z)$ is practically negligible for $|z| \geq 5$) and d growing, $\sqrt{\alpha d + z\sqrt{\beta d}} \approx \sqrt{\alpha d}$, then the above probability can be approximated with $\Phi(a_{X,Y}^{d,\epsilon,\varrho} + b_{X,Y}^{d,\epsilon,\varrho} z_{\boldsymbol{x}_d})$, where

$$\begin{aligned}
a_{X,Y}^{d,\epsilon,\varrho} &= \frac{(\epsilon^2+2\epsilon)\,(d\mu_{X,2} + d\mu_{Y,2}) + (1+\epsilon)^2\,\Phi^{-1}(\varrho)\,\sqrt{d(\mu_{Y,4}-\mu_{Y,2}^2) + 4d\mu_{Y,2}\mu_{X,2} - 4d\mu_{Y,3}\mu_X}}{\sqrt{d(\mu_{Y,4}-\mu_{Y,2}^2) + 4d\mu_{Y,2}\mu_{X,2} - 4d\mu_{Y,3}\mu_X}}, \\
b_{X,Y}^{d,\epsilon} &= \frac{(\epsilon^2+2\epsilon)\sqrt{d(\mu_{X,4}-\mu_{X,2}^2)}}{\sqrt{d(\mu_{Y,4}-\mu_{Y,2}^2) + 4d\mu_{Y,2}\mu_{X,2} - 4d\mu_{Y,3}\mu_X}}.
\end{aligned}$$

Consider now the expected value

$$\begin{aligned}
\mathbf{E}[|\mathrm{NN}_\varrho^\epsilon(\boldsymbol{X}_d)|] &= \int_{\mathbb{R}^d} Pr[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \leq (1+\epsilon)\,\mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d))] \cdot \Pr[\boldsymbol{X}_d = \boldsymbol{x}_d]\ \mathrm{d}\boldsymbol{x}_d = \\
&= \int_{z_{d,min}}^{z_{d,max}} \Phi\left(a_{X,Y}^{d,\epsilon,\varrho} + b_{X,Y}^{d,\epsilon}\, z_{\boldsymbol{x}_d}\right)\,\phi(z_{\boldsymbol{x}_d})\ \mathrm{d}z_{\boldsymbol{x}_d}.
\end{aligned}$$

The statement then follows by leveraging the following equation (Owen, 1980)

$$\int_{-\infty}^{+\infty} \Phi(a+bz)\,\phi(z)\ \mathrm{d}x = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right), \qquad (2)$$

taking the limits of integration to infinity. Note that the extra domain of integration considered is associated with a negligible probability because $z_{d,min} = (\mu_{\|\boldsymbol{X}_d\|^2} - \inf\|\boldsymbol{X}_d\|^2)/\sigma_{\|\boldsymbol{X}_d\|^2}$ and $z_{d,max} = (\sup\|\boldsymbol{X}_d\|^2 - \mu_{\|\boldsymbol{X}_d\|^2})/\sigma_{\|\boldsymbol{X}_d\|^2}$, are such that both $\phi(z_{d,\mathrm{min}})$ and $\phi(z_{d,\mathrm{max}})$ rapidly approach zero. ■

In order to validate the above result, the expected value $\mathbf{E}[|\mathrm{NN}_k^\epsilon(\boldsymbol{X}_d, \{\boldsymbol{Y}_d\}_n)|]$ is empirically estimated for different values of k, d and $\epsilon \in [0, 0.5]$, by exploiting sets of $n = 10{,}000$

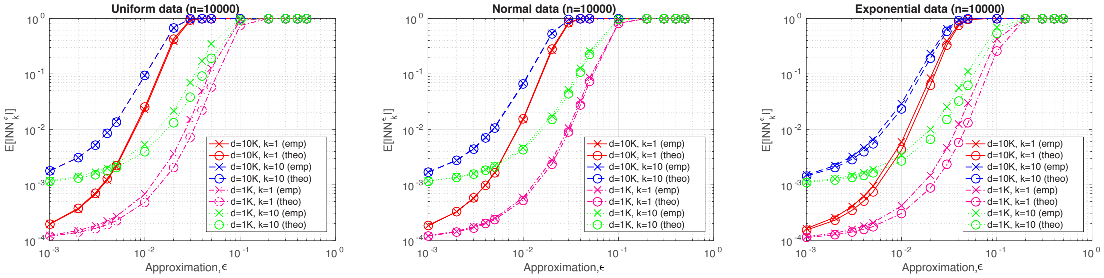


Figure 3: [Best viewed in color.] Comparison between the empirically estimated (x-marked curves) and the predicted by means of Th. 25 (o-marked curves) expected sizes of the ϵ-neighborhood, for $n = 10{,}000$, $d = 1{,}000$ and $k = 1$ (magenta dash-dotted line), $d = 1{,}000$ and $k = 10$ (green dotted line), $d = 10{,}000$ and $k = 1$ (red solid line), and $d = 10{,}000$ and $k = 10$ (blue dashed line).

realizations of the random vector $\boldsymbol{Y}_d$. Results are averaged by considering ten different sets. In the experiment, it is assumed that $F_X = F_Y$ and that each point of the set is used in turn as a query point; thus, the size of the ϵ-neighborhood may vary between k and $n-1$.

Figure 3 reports the results of this experiment for uniform, normal, and exponential i.i.d. data. The value $\mathbf{E}[|\mathrm{NN}_k^\epsilon(\boldsymbol{X}_d, \{\boldsymbol{Y}_d\}_n)|]$ empirically estimated as described above is compared with the value predicted by means of Theorem 25. The curves for the number of neighbors $k \in \{1, 10\}$ and the dimensionalities $d \in \{1{,}000, 10{,}000\}$ are reported. The curves confirm that the prediction follows the trend of the empirical evidence with the error vanishing as the dimensionality increases.

As already stated by Beyer et al. (1999), Theorem 2 only tells us what happens when we take the dimensionality to infinity, but nothing is said about the dimensionality at which do we anticipate nearest neighbors to become unstable, and the issue must be addressed through empirical studies.

The above dimensionality, called the *critical dimensionality*, can, however, be obtained as follows. Let $\theta \in [0, 1]$ represent a fraction of the data elements; the critical dimensionality $d^*_{\varrho,\epsilon,\theta}$ for the parameters ϱ and ϵ at the threshold level θ, also called selectivity, is such that

$$d^*_{\varrho,\epsilon,\theta} = \min\{d \in \mathbb{N}^+ : \mathbf{E}[|\mathrm{NN}_\varrho^\epsilon(\boldsymbol{X}_d, \boldsymbol{Y}_d)|] \geq \theta\},$$

i.e., the dimensionality at which the expected size of the ϵ-neighborhood contains at least the θ fraction of the data points.

Figure 4 reports the critical dimensionality for ϵ varying in $[0.001, 1]$, thresholds $\theta \in \{0.01, 0.1, 0.5\}$, $n = 10{,}000$ and $k = 1$ (i.e. $\varrho = k/(n-1) \approx 0.0001$), obtained by exploiting the expression reported in Theorem 25. For example, for $\theta = 0.1$, the plot says that for dimensionalities below the bottom curve, ϵ-neighborhoods contain on the average 10% of the points (one hundred points for $n = 10{,}000$). Note that analogous predictions can be obtained in a very similar way for any other combination of the parameters ϱ, θ, and ϵ, and distribution function F.

Figure 4 also report the values of the critical dimensionality estimated empirically (dashed lines). The plots highlight that the predicted critical dimensionality tends to the

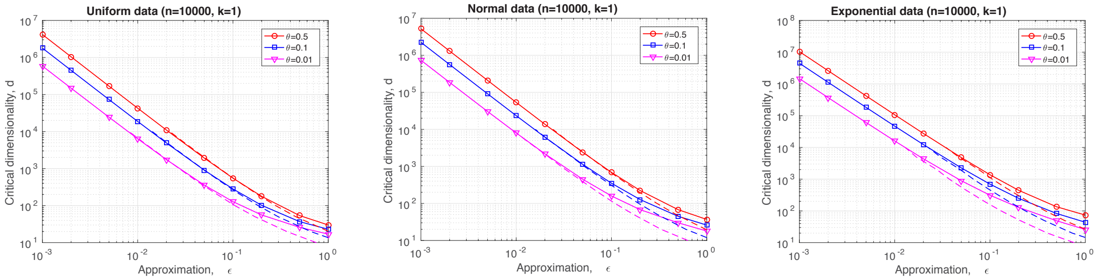


Figure 4: [Best viewed in color.] Critical dimensionality for $\epsilon \in [10^3, 10^0]$, $n = 10{,}000$, $k = 1$, and $\theta = 0.01$ (red solid line), $\theta = 0.1$ (blue solid line), and $\theta = 0.5$ (magenta solid line), predicted by exploiting Th. 25. The dashed curves represent the values of the critical dimensionality estimated empirically.

empirical one for decreasing ϵ and that the rate of convergence is directly proportional to θ. Interestingly, it can be seen that in different cases, the reported critical dimensionality is quite high (e.g., consider $\epsilon = 0.01$). Because approximate nearest neighbors must be associated with small values of θ (e.g., consider $\theta = 0.01$) to be considered meaningful, it can be concluded that the notion of approximate nearest neighbor can be considered meaningful even in high-dimensional spaces provided that the approximation factor ϵ is sufficiently small.

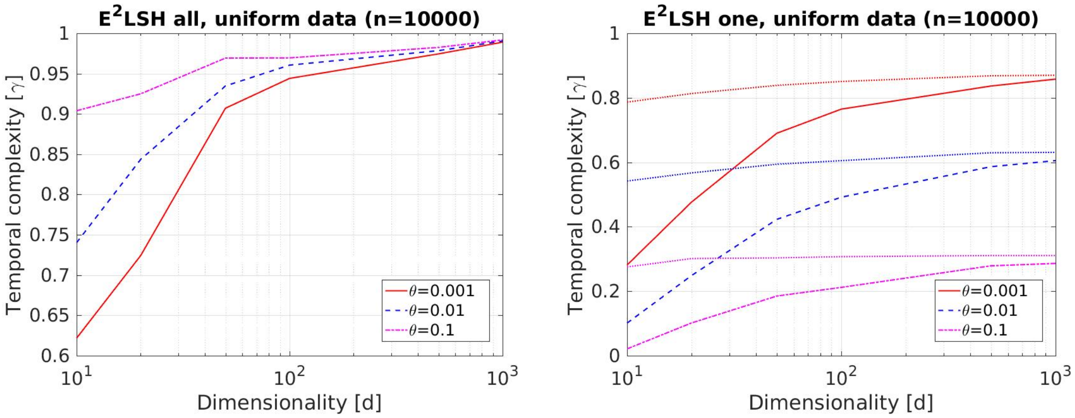
Unfortunately, this does not imply that algorithms perform efficiently in these cases. To illustrate, the researchers have proposed different algorithms for (approximate) nearest neighbor search problems. Most of these algorithms are randomized; that is, they are associated with a failure probability δ. Specifically, the *approximate near(est) neighbor search problem with failure probability* δ is defined as the problem to construct a data structure over a set of points $S \subseteq \mathbb{R}^d$ such that, given any query point $x \in \mathbb{R}^d$, with probability $1 - \delta$ reports:

P1. some $y \in S$ with $\text{dist}(x, y) \leq (1 + \epsilon)r$ (*ϵ-approximate r-near neighbor*);

P2. some $y \in S$ with $\text{dist}(x, y) \leq (1+\epsilon)\cdot\text{dist}(x, nn(x, S))$ (*ϵ-approximate nearest neighbor*);

P3. each point $y \in S$ with $\text{dist}(x, y) \leq r$ (*r-near neighbor reporting*).

The proposed algorithms offer trade-offs between the approximation factor, the space and the query time (Andoni, 2009). From the practical perspective, the space used by an algorithm should be as close to linear as possible. In this case, the best-existing solutions are based on locality-sensitive hashing (LSH) (Indyk and Motwani, 1998; Har-Peled et al., 2012). The idea of the LSH approach is to hash the points in a way that the probability of collision is much higher for points that are close (with the distance r) to each other than for those that are far apart (with distance at least $(1 + \epsilon)r$). Under different assumptions involving the parameters employed (Har-Peled et al., 2012), the LSH algorithm solves the ϵ-approximate r-near neighbor problem using $O(n^{1+\rho_\epsilon})$ extra space, $O(dn^{\rho_\epsilon})$ query time,



(a)

(b)

Figure 5: Temporal complexity of the E^{2}LSH algorithm on uniform data for different selectivity values, namely $\theta = 0.1$ (red solid line), $\theta = 0.01$ (blue dashed line), and $\theta = 0.001$ (magenta dash-dotted line), and dimensions $d \in [10, 10^3]$, estimated by using $n = 10{,}000$ data points and $m = n$ query points. The plot on the left concerns the cost of reporting all the neighbors. The plot on the right concerns the cost of reporting just one neighbor. In the latter plot, dotted curves represent the complexity of sampling until a neighbor is retrieved.

and failure probability $\delta = 1/e + 1/3$.[3] As for the value of the exponent ρ_ϵ, for the Euclidean distance, it is possible to achieve $\rho_\epsilon = 1/(1+\epsilon)^2 + o_\epsilon(1)$ (Andoni and Indyk, 2006), and it is known this bound is tight.

For example, consider that if $\epsilon = 0.01$, then $\rho_\epsilon = 0.980$. Because meaningfulness in intrinsically high-dimensional spaces requires smaller and smaller ϵ values, this means that, if we wish to maintain a pre-defined level of selectivity θ, we expect that the efficiency of LSH-based schemes will diminish with the intrinsic dimensionality of the space.

To empirically illustrate the relationship among selectivity θ, the intrinsic dimensionality d, and the temporal complexity γ of the search algorithm, [4] we analyzed the performances of the E^{2}LSH method as a function of the expected size θ of the r-neighborhood. The E^{2}LSH package solves the randomized r-near neighbor reporting problem exploiting the basic LSH scheme.[5] After preprocessing the data set, E^{2}LSH answers queries, typically in

[3] The failure probability δ can be made arbitrarily small, say $\delta < 1/n$, by running $O(\log(n+m))$ copies of the basic LSH algorithm for *P1*, where n and m denote an upper bound on the number of points in the data structure and on the number of queries performed at any time. Moreover, *P2* can be solved by using as building blocks $O(\log n)$ copies of an algorithm for *P1*, achieving failure probability $O(\delta \log n)$ (Har-Peled et al., 2012). A similar strategy allows solving the nearest neighbor reporting problem (*P3*) by building on different data structures for *P1* associated with increasing values of r (Andoni and Indyk, 2008).

[4] The *temporal complexity* is defined as the exponent $\gamma \geq 0$ such that the total number of distances D computed by the algorithm in order to report its answer is such that $D = n^\gamma$.

[5] The E^{2}LSH package is available for download at `http://www.mit.edu/~andoni/LSH/`.

sub-linear time, with each near neighbor being reported with a certain probability $1 - \delta$ ($= 0.9$ by default). As for the values of the other parameters employed, we used the values determined automatically by the algorithm.

Figure 5 reports the results of the experiments on a family of uniformly distributed data sets composed of $n = 10{,}000$ points with $d \in [10, 10^3]$. We used $m = n$ different query points generated from the same distribution. We also varied the selectivity θ in $\{0.001, 0.01, 0.1\}$ by determining the radius r such that the expected fraction of r-near neighbors of the query points is θ. In Figure 5a, it can be seen that the complexity of the procedure increases with θ, and this can be explained by noting that the total number of points to be reported is directly proportional to θ. However, even if θ is held fixed, in all cases, the complexity of the algorithm for large d values tends to a linear scan of the data or to the cost γ_s of a random sampling procedure.[6] In Figure 5b, the algorithm has been enforced to report at most one near neighbor; hence, it stops the search as soon as it retrieves a near neighbor. It can now be seen that the complexity of the procedure decreases with θ, and this can be explained by noting that the probability of retrieving a neighbor is directly proportional to θ. The dotted curves represent the complexity γ_s of the procedure consisting in randomly selecting points until a r-near neighbor is retrieved. Additionally, in this case, it can be observed that the complexity degrades towards that of the random sampling procedure irrespectively of the selectivity value θ.[7]

The above analysis provides a picture of how much better an approximate search algorithm can perform than the pure random search, as a function of the selectivity and of the intrinsic dimensionality. Although the target neighborhood can be guaranteed to contain not too many points even in very large dimensional spaces, the best search algorithms may fail to perform better than random sampling. This can be explained by the poor separation of distances with the objects that are outside the approximate neighborhood.

In this regard, although the critical dimensionality is a construct with which to attempt to quantify the meaningless of a certain query, the relative contrast C_r (He et al., 2012) is a way to attempt to quantify its difficulty. Given a query point $\boldsymbol{x}_d$, the relative contrast is a measure of separability of the nearest neighbor of $\boldsymbol{x}_d$ from the rest of the data set points.

Definition 26 (Adapted from He et al., 2012) *Let DS be a data set consisting of n realizations of a random vector $\boldsymbol{Y}_d$. The relative contrast for the data set DS for a query $\boldsymbol{x}_d$, being the realization of a random vector $\boldsymbol{X}_d$, is defined as $C_r^k(\boldsymbol{x}_d) = \frac{\mathbf{E}[\mathrm{dist}(\boldsymbol{x}_d, DS)]}{\mathbf{E}[\mathrm{distnn}_k(\boldsymbol{x}_d, DS)]}$. Taking expectations with respect to queries, the relative contrast for the data set DS is*

$$C_r^k = \frac{\mathbf{E}[\mathrm{dist}(\boldsymbol{X}_d, DS)]}{\mathbf{E}[\mathrm{distnn}_k(\boldsymbol{X}_d, DS)]}.$$

He et al. (2012) provided an estimate of the relative contrast for a data set valid for independent dimensions and, moreover, provided bounds on the cost of LSH-based nearest neighbor search algorithms taking into account the relative contrast. They also noted that

[6] Indeed, the expected number n^{γ_s} of points to be randomly picked in order to retrieve the $1 - \delta$ fraction of the $n\theta$ data points that are r-near neighbors of the query point is $n^{\gamma_s} = n(1 - \delta)$ and $\gamma_s = 1 + \log(1 - \delta)/\log(n)$. E.g., for $1 - \delta = 0.9$, $\gamma_s = 0.9886$.

[7] Note that, for a query having selectivity θ, the expected number of points to be randomly picked in order to retrieve exactly one r-near neighbor is $n^{\gamma_s} = 1/\theta$ and, hence, $\gamma_s = -\log(\theta)/\log(n)$.

the analysis of Beyer et al. (1999) and François et al. (2007) agree with the asymptotic behavior of the relative contrast. We refer to (He et al., 2012) for the details.

Here, we show that by exploiting the previous results, we can derive an approximation for the relative contrast C_r^k of a data set that results to be more accurate than the estimate provided by He et al. (2012). In addition, we can derive a closed form for the relative contrast $C_r^k(\boldsymbol{x}_d)$ of an individual query point.

Theorem 27 *Let $\boldsymbol{X}_d$ be a d-dimensional i.i.d. random vector with cdf F_X and let $\boldsymbol{Y}_d$ be a d-dimensional i.i.d. random vector with cdf F_Y. Assume, w.l.o.g., that F_Y has null mean $\mu_Y = 0$. Then, for large values of d,*

$$C_r^k \approx \sqrt{\frac{\sqrt{d}(\mu_{Y,2}+\mu_{X,2})}{\sqrt{d}(\mu_{Y,2}+\mu_{X,2}) + \Phi^{-1}\left(\frac{k}{n}\right)\sqrt{\mu_{Y,4}-\mu_{Y,2}^2+4\mu_{Y,2}\mu_{X,2}-4\mu_{Y,3}\mu_X}}}.$$

Proof of Theorem 27. Consider the expected value of the squared distance separating a query point $\boldsymbol{X}_d$ from $nn_\varrho(\boldsymbol{X}_d, \boldsymbol{Y}_d)$ (leveraging Proposition 8 and Lemma 23):

$$\begin{aligned}
\mathbf{E}[\|\boldsymbol{X}_d - nn_\varrho(\boldsymbol{X}_d,\boldsymbol{Y}_d)\|^2] &= \mathbf{E}[\mathrm{distnn}_\varrho(\boldsymbol{X}_d,\boldsymbol{Y}_d)^2] = \mathbf{E}[\mathrm{dist}(\boldsymbol{X}_d, nn_\varrho(\boldsymbol{X}_d,\boldsymbol{Y}_d))^2] = \\
&= \int_{\mathbb{R}^d} Pr[\boldsymbol{X}_d = \boldsymbol{x}_d]\cdot \mathrm{dist}(\boldsymbol{x}_d, nn_\varrho(\boldsymbol{x}_d,\boldsymbol{Y}_d))^2 \,\mathrm{d}\boldsymbol{x}_d = \\
&= \int_0^{+\infty} \phi_{\|\boldsymbol{X}_d\|^2}(R)\cdot\left(\mu_{\|\boldsymbol{x}_d-\boldsymbol{Y}_d\|^2} + \Phi^{-1}(\varrho)\,\sigma_{\|\boldsymbol{x}_d-\boldsymbol{Y}_d\|^2}\right)\Big|_{\|\boldsymbol{x}_d\|^2=R} \mathrm{d}R = \\
&= \int_0^{+\infty} \phi_{\|\boldsymbol{X}_d\|^2}(R)\cdot\left(R + d\mu_{Y,2} + \Phi^{-1}(\varrho)\sqrt{d(\mu_{Y,4}-\mu_{Y,2}^2) + 4\mu_{Y,2}R - 4\mu_{Y,3}\frac{\mu_X}{\mu_{X,2}}R}\right)\mathrm{d}R.
\end{aligned}$$

After approximating the R under the square root with the expected value $\mu_{\|\boldsymbol{X}_d\|^2} = d\mu_{X,2}$ of $\boldsymbol{X}_d$:

$$\begin{aligned}
\mathbf{E}[\mathrm{dist}(\boldsymbol{X}_d, nn_\varrho(\boldsymbol{X}_d,\boldsymbol{Y}_d))^2] &= \int_0^{+\infty} R\cdot\phi_{\|\boldsymbol{X}_d\|^2}(R)\,\mathrm{d}R + \\
&+ \left(d\mu_{Y,2} + \Phi^{-1}(\varrho)\sqrt{d(\mu_{Y,4}-\mu_{Y,2}^2) + 4d\mu_{Y,2}\mu_{X,2} - 4d\mu_{Y,3}\mu_X}\right)\cdot\int_0^{+\infty}\phi_{\|\boldsymbol{X}_d\|^2}(R)\,\mathrm{d}R = \\
&= d(\mu_{X,2}+\mu_{Y,2}) + \Phi^{-1}(\varrho)\sqrt{d(\mu_{Y,4}-\mu_{Y,2}^2+4\mu_{X,2}\mu_{Y,2}-4\mu_{Y,3}\mu_X)}.
\end{aligned}$$

Indeed, the left hand integral above corresponds to the expected value $\mu_{\|\boldsymbol{X}_d\|^2} = d\mu_{X,2}$ of of the random variable $\|\boldsymbol{X}_d\|^2$, whereas the right hand integral evaluates to one.

According to the Jensen inequality (Johnson et al., 1994), if g is a concave function, then $\mathbf{E}[g(X)] \leq g(\mathbf{E}[X])$; moreover, the larger the relative variance σ_X/μ_X of X, the closer the two above values, i.e., $\mathbf{E}[g(X)] \approx g(\mathbf{E}[X])$. Specifically, $\mathbf{E}[\|\boldsymbol{X}_d\|] = \mathbf{E}[\sqrt{\sum_i X_i^2}] \leq \sqrt{\mathbf{E}[\sum_i X_i^2]} = \sqrt{\mathbf{E}[\|\boldsymbol{X}_d\|^2]}$ and, because $\sigma_{\|\boldsymbol{X}_d\|^2}/\mu_{\|\boldsymbol{X}_d\|^2} = O(d^{-1/2})$, for large values of d, $\mathbf{E}[\mathrm{dist}(\boldsymbol{X}_d,\boldsymbol{Y}_d)] \approx \sqrt{\mathbf{E}[\mathrm{dist}(\boldsymbol{X}_d,\boldsymbol{Y}_d)^2]}$, and $\mathbf{E}[\mathrm{distnn}_\varrho(\boldsymbol{X}_d,\boldsymbol{Y}_d)] \approx \sqrt{\mathbf{E}[\mathrm{distnn}_\varrho(\boldsymbol{X}_d,\boldsymbol{Y}_d)^2]}$. ■

We can also provide the relative contrast $C_r^k(\boldsymbol{x}_d)$ of an individual query point $\boldsymbol{x}_d$.

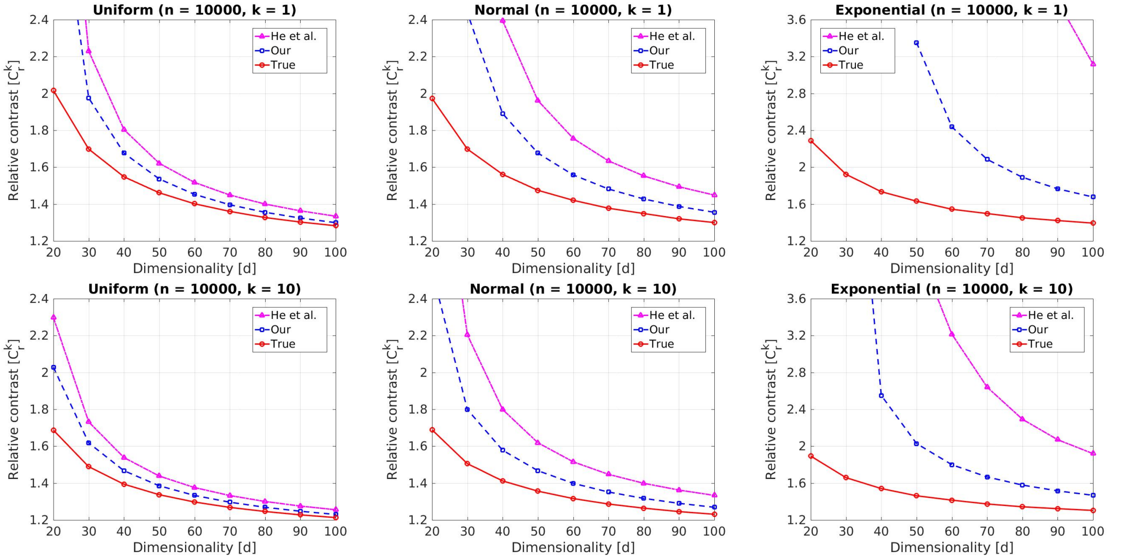


Figure 6: [Best viewed in color.] Comparison between the estimate of the relative contrast C_r provided in Theorem 27 (blue dashed line) and the estimate provided by He et al. (2012) (magenta dash-dotted line). The red solid line represents the value of the relative contrast estimated empirically.

Corollary 28 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector, and let $\boldsymbol{Y}_d$ be a d-dimensional i.i.d. random vector with cdf F_Y. Then, for large values of d, with high probability*

$$C_r^k(\boldsymbol{x}_d) \approx \sqrt{\frac{\|\boldsymbol{x}_d\|^2 + d\mu_2 - 2\mu \sum_{i=1}^d x_i}{\|\boldsymbol{x}_d\|^2 + d\mu_2 - 2\mu \sum_{i=1}^d x_i + \Phi^{-1}\left(\frac{k}{n}\right)\sqrt{d(\mu_4 - \mu_2^2) + 4\mu_2\|\boldsymbol{x}_d\|^2 - 4(\mu_3 - \mu\mu_2)\sum_{i=1}^d x_i}}}.$$

Proof of Corollary 28. Following the same line of reasoning of Theorem 27, $C_r^k(\boldsymbol{x}_d) \approx \sqrt{\frac{\mathbf{E}[\mathrm{dist}(\boldsymbol{x}_d,\boldsymbol{Y}_d)^2]}{\mathbf{E}[\mathrm{distnn}_\varrho(\boldsymbol{x}_d,\boldsymbol{Y}_d)^2]}}$, and the statement follows by leveraging Theorem 15 and Lemma 23. ■

Figure 6 compares the approximation of the relative contrast provided in Theorem 27 to the approximation provided by He et al. (2012). In all the cases, the approximation of Theorem 27 is the more accurate. This can be understood by noting that He et al. (2012) estimated the relative contrast by considering the differences $X_i - Y_i$ between the components of a query point $\boldsymbol{X}_d$ and of a data point $\boldsymbol{Y}_d$ as novel random variables, and then by determining their expectations and standard deviations. This corresponds to ignoring the form of the distribution of distances from each individual query point and all the data points, a relationship that is conversely taken into account in Theorem 27, due to the leveraging of Theorem 16, Proposition 8, and Lemma 23.

4.3 On the Distribution of Reverse Nearest Neighbors for i.i.d. Data

Given a real number $\varrho \in [0,1]$, a d-dimensional random vector $\boldsymbol{Y}_d$, and a realization $\boldsymbol{x}_d$ of $\boldsymbol{Y}_d$, it is said that a realization $\boldsymbol{y}_d$ of $\boldsymbol{Y}_d$ is a *ϱ reverse nearest neighbor* of $\boldsymbol{x}_d$ w.r.t. $\boldsymbol{Y}_d$ if $\boldsymbol{x}_d \in \mathrm{NN}_\varrho(\boldsymbol{y}_d, \boldsymbol{Y}_d)$.

The size $\mathrm{N}_\varrho(\boldsymbol{x}_d, \boldsymbol{Y}_d)$, or $\mathrm{N}_\varrho(\boldsymbol{x}_d)$ whenever $\boldsymbol{Y}_d$ is clear from the context, of the *ϱ reverse nearest neighborhood* of $\boldsymbol{x}_d$ w.r.t. $\boldsymbol{Y}_d$, also called *reverse ϱ nearest neighbor count* or *ϱ-occurrences*, is the fraction of realizations $\boldsymbol{y}_d$ of $\boldsymbol{Y}_d$ such that $\boldsymbol{x}_d \in \mathrm{NN}_\varrho(\boldsymbol{y}_d, \boldsymbol{Y}_d)$.

As in the previous sections, in order to deal with finite sets of n points $\{\boldsymbol{Y}_d\}_n$, the integer parameter $k = \varrho n$ ($k \in \{1, \ldots, n\}$) must be employed in place of ϱ. In such a case, we speak of *k reverse nearest neighborhood*, *reverse k nearest neighbor count*, or *k-occurrences*.

Before going into the main results, the following expression provides the probability that a realization, having norm value R, of a d-dimensional i.i.d. random vector $\boldsymbol{Y}_d$ lies at distance not greater than δ from a given d-dimensional vector $\boldsymbol{x}_d$.

Lemma 29 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$ with cdf F_X, and let $\boldsymbol{Y}_d$ be a d-dimensional i.i.d. random vector with cdf F_Y. Assume, w.l.o.g., that F_Y has null mean $\mu_Y = 0$. Then, for large values of d, with high probability*

$$Pr\left[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \le \delta \mid \|\boldsymbol{Y}_d\| = R\right] \approx \Phi\left(\frac{\delta^2 - R^2 - \|\boldsymbol{x}_d\|^2}{2\|\boldsymbol{x}_d\|\sqrt{\mu_2}}\right),$$

where moments are relative to the constrained random vector $\boldsymbol{Y}_d$.

Proof of Lemma 29. See the appendix. ■

The noteworthiness of the above expression lies in the fact that, by combining it with Proposition 8, it is possible in some cases to replace multi-dimensional integrations involving the full event space $\mathbb{R}^d$ with one-dimensional integrations over the domain $\mathbb{R}_0^+$ of the squared-norm values. Specifically, it is essential to the proof of the following result.

Theorem 30 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector $\boldsymbol{Y}_d$, with cdf F_Y having, w.l.o.g., null mean $\mu = 0$. Consider the reverse k nearest neighbor count $N_\varrho(\boldsymbol{x}_d)$ of $\boldsymbol{x}_d$ w.r.t. $\boldsymbol{Y}_d$. Then, for large values of d, with high probability*

$$\mathrm{N}_\varrho(\boldsymbol{x}_d) \approx \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{\mu_4 + 3\mu_2^2} - z_{\boldsymbol{x}_d}\sqrt{\mu_4 - \mu_2^2}}{2\mu_2}\right).$$

Proof of Theorem 30. First of all, note that $\mathrm{N}_\varrho(\boldsymbol{x}_d) = Pr[\boldsymbol{x}_d \in \mathrm{NN}_\varrho(\boldsymbol{Y}_d)]$. Consider the probability

$$\begin{aligned} Pr[\boldsymbol{x}_d \in \mathrm{NN}_\varrho(\boldsymbol{Y}_d)] &= \int_{\mathbb{R}^d} Pr\big[\boldsymbol{x}_d \in \mathrm{NN}_\varrho(\boldsymbol{y}_d)\big] \cdot Pr[\boldsymbol{Y}_d = \boldsymbol{y}_d]\ \mathrm{d}\boldsymbol{y}_d = \\ &= \int_{\mathbb{R}^d} Pr\big[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) \le \mathrm{dist}(\boldsymbol{y}_d, nn_\varrho(\boldsymbol{y}_d))\big] \cdot Pr[\boldsymbol{Y}_d = \boldsymbol{y}_d]\ \mathrm{d}\boldsymbol{y}_d = \\ &= \int_0^{+\infty} Pr\big[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \le \mathrm{dist}(\boldsymbol{Y}_d, nn_\varrho(\boldsymbol{Y}_d)) \mid \|\boldsymbol{Y}_d\|^2 = R\big] \cdot Pr\big[\|\boldsymbol{Y}_d\|^2 = R\big]\ \mathrm{d}R. \end{aligned}$$

By Lemma 23 and Proposition 22, for $\|\boldsymbol{Y}_d\|^2 = R$,

$$\mathrm{dist}(\boldsymbol{Y}_d, nn_\varrho(\boldsymbol{Y}_d))^2 = R + d\mu_2 + \Phi^{-1}(\varrho)\sqrt{d(\mu_4 - \mu_2^2) + 4\mu_2 R},$$

while by Lemma 29,

$$\begin{aligned} Pr\big[\mathrm{dist}(\boldsymbol{x}_d, \boldsymbol{Y}_d) \leq \mathrm{dist}(\boldsymbol{Y}_d, nn_\varrho(\boldsymbol{Y}_d)) \mid \|\boldsymbol{Y}_d\|^2 = R\big] \approx \\ \approx \Phi\left(\frac{\mathrm{dist}(\boldsymbol{Y}_d, nn_\varrho(\boldsymbol{Y}_d))^2 - R - \|\boldsymbol{x}_d\|^2}{2\|\boldsymbol{x}_d\|\sqrt{\mu_2}}\right) = \\ = \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{d(\mu_4 - \mu_2^2) + 4\mu_2 R} + d\mu_2 - \|\boldsymbol{x}_d\|^2}{2\|\boldsymbol{x}_d\|\sqrt{\mu_2}}\right), \end{aligned}$$

from which it follows that

$$Pr[\boldsymbol{x}_d \in nn_\varrho(\boldsymbol{Y}_d)] \approx \int_0^{+\infty} \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{d(\mu_4 - \mu_2^2) + 4\mu_2 R} + d\mu_2 - \|\boldsymbol{x}_d\|^2}{2\|\boldsymbol{x}_d\|\sqrt{\mu_2}}\right) \phi_{\|\boldsymbol{Y}_d\|^2}(R)\ \mathrm{d}R,$$

where moments are relative to the constrained random vector $\boldsymbol{Y}_d$.

The proof proceeds by expressing $\|\boldsymbol{x}_d\|^2$ and R in terms of their standard scores with respect to the random variable $\|\boldsymbol{Y}_d\|^2$, i.e.,

$$\|\boldsymbol{x}_d\|^2 = \mu_{\|\boldsymbol{Y}_d\|^2} + z_{\boldsymbol{x}_d} \cdot \sigma_{\|\boldsymbol{Y}_d\|^2} \quad \text{and} \quad R = \mu_{\|\boldsymbol{Y}_d\|^2} + z_{R,\|\boldsymbol{Y}_d\|^2} \cdot \sigma_{\|\boldsymbol{Y}_d\|^2}.$$

By substituting in the left-hand side above, and by considering that for α and β being finite and d growing, $\sqrt{\alpha d} \approx \sqrt{\alpha d + \sqrt{\beta d}}$,

$$\begin{aligned} \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{d(\mu_4 - \mu_2^2) + 4d\mu_2^2 + z_{R,\|\boldsymbol{Y}_d\|^2} 4\mu_2\sqrt{d(\mu_4 - \mu_2^2)}} + d\mu_2 - d\mu_2 - z_{\boldsymbol{x}_d}\sqrt{d(\mu_4 - \mu_2^2)}}{2\sqrt{\mu_2}\sqrt{d\mu_2 + z_{\boldsymbol{x}_d}\sqrt{d(\mu_4 - \mu_2^2)}}}\right) \approx \\ \approx \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{d(\mu_4 + 3\mu_2^2)} - z_{\boldsymbol{x}_d}\sqrt{d(\mu_4 - \mu_2^2)}}{2\mu_2\sqrt{d}}\right) = \\ = \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{\mu_4 + 3\mu_2^2} - z_{\boldsymbol{x}_d}\sqrt{\mu_4 - \mu_2^2}}{2\mu_2}\right) = C(z_{\boldsymbol{x}_d}, \varrho). \end{aligned}$$

Since, for large values of d, moments tend to their unconstrained values (see proof of Lemma 29), the last expression depends on $z_{\boldsymbol{x}_d}$ and on ϱ, but not on R. Thus

$$Pr[\boldsymbol{x}_d \in \mathrm{NN}_\varrho(\boldsymbol{Y}_d)] \quad \approx \quad C(z_{\boldsymbol{x}_d}, \varrho) \int_0^{+\infty} \phi_{\|\boldsymbol{Y}_d\|^2}(R)\ \mathrm{d}R = C(z_{\boldsymbol{x}_d}, \varrho).$$

■

As for the expression reported in the statement of Theorem 30, it does not explicitly depend on the dimensionality and on the exact position of the point $\boldsymbol{x}_d$ but only on the relative position of the squared norm of the point with respect to the expected value. Thus, the following definition naturally arises.

Definition 31 *Let z denote the standard score of the squared norm. Then, the infinite dimensional k-occurrences function $\mathrm{N}_\varrho^\infty : \mathbb{R} \mapsto [0,1]$, defined as*

$$\mathrm{N}_\varrho^\infty(z) \quad = \quad \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{\mu_4 + 3\mu_2^2} - z\sqrt{\mu_4 - \mu_2^2}}{2\mu_2}\right), \tag{3}$$

represents the fraction of points having a point with squared norm standard score z among their ϱ nearest neighbors.

An alternative expression can be provided by leveraging the *kurtosis* $\kappa = \frac{\mu_4}{\mu_2^2}$, a well known measure of tailedness of a probability distribution that is

$$\mathrm{N}_\varrho^\infty(z) \quad = \quad \Phi\left(\Phi^{-1}(\varrho)\frac{\sqrt{\kappa+3}}{2} - z\frac{\sqrt{\kappa-1}}{2}\right). \tag{4}$$

In particular, it holds from the development of Theorem 30 that

$$\lim_{d\to\infty} Pr[\mathrm{N}_\varrho(\boldsymbol{X}_d) = \mathrm{N}_\varrho^\infty(z)] = \phi(z), \tag{5}$$

from which it can be seen that the point $z_0 \to -\infty$ is such that $\mathrm{N}_\varrho^\infty(z_0) \to 1$ and $\phi(z_0) \to 0$. This point precisely corresponds with the expected value of $\boldsymbol{X}_d$ (the origin of the space for variables with null mean) and is the point most likely to be selected as nearest neighbor by any other point. At the same time, it is the point least likely to be observed as a realization of $\boldsymbol{X}_d$ among those having norms smaller than the expected value.

As for the point $z_\infty \to \infty$, it is such that $\mathrm{N}_\varrho^\infty(z_\infty) \to 0$ and $\phi(z_\infty) \to 0$. Hence, it is the less likely to be observed as a realization of $\boldsymbol{X}_d$, but it is also the less likely to be selected as a nearest neighbor. This point is the furthest from the mean, and it is located on the boundary of a bounded region or ad infinitum.

Figure 7 shows the curves of $\mathrm{N}_\varrho^\infty$ (red lines) for i.i.d. data coming from different distributions, together with the empirical N_k/n values (black dots), with $k = \varrho n$, in a random sample of $n = 10{,}000$ points. Theoretical $\mathrm{N}_\varrho^\infty$ curves represent the picture of what happens ad infinitum, because they provide the values to which the k-occurrences counts converge for large dimensions. We observed that in most cases, the convergence arises very soon, often a few tens of dimensions suffice. Indeed, empirical values tend to distribute along the associated curves. While the first two distributions have null skewness, the same behavior is exhibited by the third one, which instead has non-null skewness, even if, in this case, convergence appears to be slower. In any case, it appears that the value of k-occurrences predicted by means of the function $\mathrm{N}_\varrho^\infty$ usually is in good agreement with the empirical evidence, even for the smallest dimension considered in the figure. For similar plots on real data sets, we refer to Figures 10, 11, and 12, reported in the following.

It is now of interest to obtain the cdf and pdf of $\mathrm{N}_\varrho(\boldsymbol{X}_d)$ for large values, together with the associated variance and expected value.

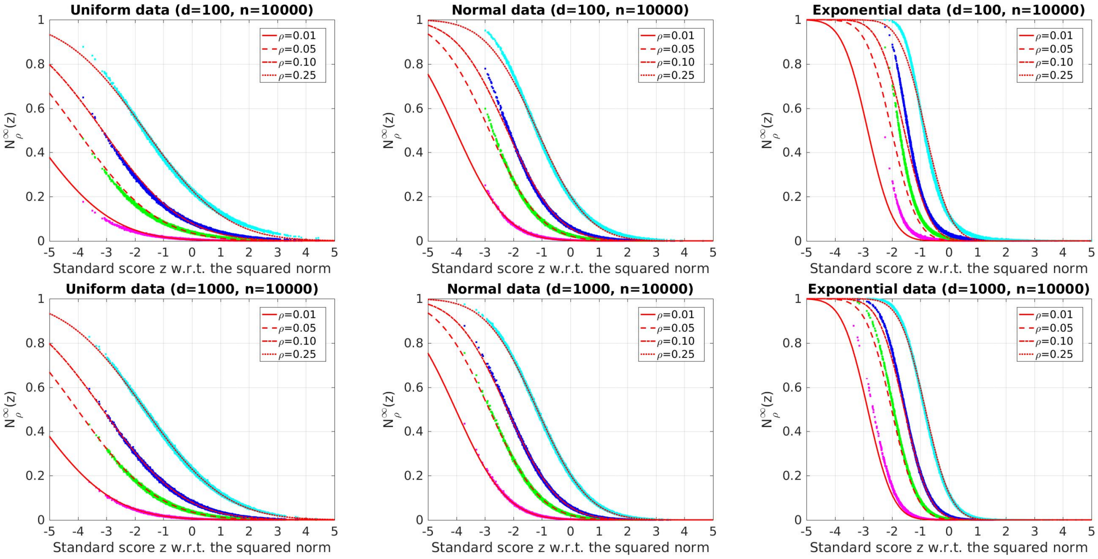


Figure 7: [Best viewed in color.] Comparison between the empirical values of the relative number of k-occurrences (N_k/n), estimated in a random sample of $n = 10{,}000$ points with $d \in \{100, 1{,}000\}$, and the values predicted by means of the function $\mathrm{N}^{\infty}_{\varrho}$ (red curves), for $\varrho = 0.01$ (magenta dots and solid red line), $\varrho = 0.01$ (green dots and dashed red line), $\varrho = 0.01$ (blue dots and dash-dotted red line), and $\varrho = 0.01$ (cyan dots and dotted red line).

Theorem 32

$$(i) \qquad \lim_{d\to\infty} Pr[\mathrm{N}_{\varrho}(\boldsymbol{X}_d) \leq \theta] = \Phi\left(\frac{\Phi^{-1}(\theta)2\mu_2 - \Phi^{-1}(\varrho)\sqrt{\mu_4 + 3\mu_2^2}}{\sqrt{\mu_4 - \mu_2^2}}\right),$$

$$(ii) \qquad \lim_{d\to\infty} Pr[\mathrm{N}_{\varrho}(\boldsymbol{X}_d) = \theta] = \frac{2\mu_2}{\sqrt{\mu_4 - \mu_2^2}} \cdot \frac{1}{\phi(\Phi^{-1}(\theta))} \cdot \phi\left(\frac{\Phi^{-1}(\theta)2\mu_2 - \Phi^{-1}(\varrho)\sqrt{\mu_4 + 3\mu_2^2}}{\sqrt{\mu_4 - \mu_2^2}}\right),$$

$$(iii) \qquad \lim_{d\to\infty} \sigma^2(\mathrm{N}_{\varrho}(\boldsymbol{X}_d)) = \varrho(1-\varrho) - 2T\left(\Phi^{-1}(\varrho), \frac{2\mu_2}{\sqrt{2(\mu_4 + \mu_2^2)}}\right), \textit{ and}$$

$$(iv) \qquad \lim_{d\to\infty} \mathbf{E}[\mathrm{N}_{\varrho}(\boldsymbol{X}_d)] = \varrho,$$

$$\textit{where} \qquad T(h,a) = \phi(h)\int_0^a \frac{\phi(hx)}{1+x^2}\,\mathrm{d}x \qquad \textit{is the Owen's T function.}$$

Proof of Theorem 32. See the appendix. ■

Figure 8 shows the cdfs and the pdfs of the limiting distributions of the k-occurrences for uniform, normal, and exponential distributions. As expected, the probability of observing large values of N_{ϱ} increases with ϱ. Moreover, it can be observed from the pdf functions that for the exponential distribution, the probability of observing $\mathrm{N}_{\varrho} \approx 1$ is not negligible

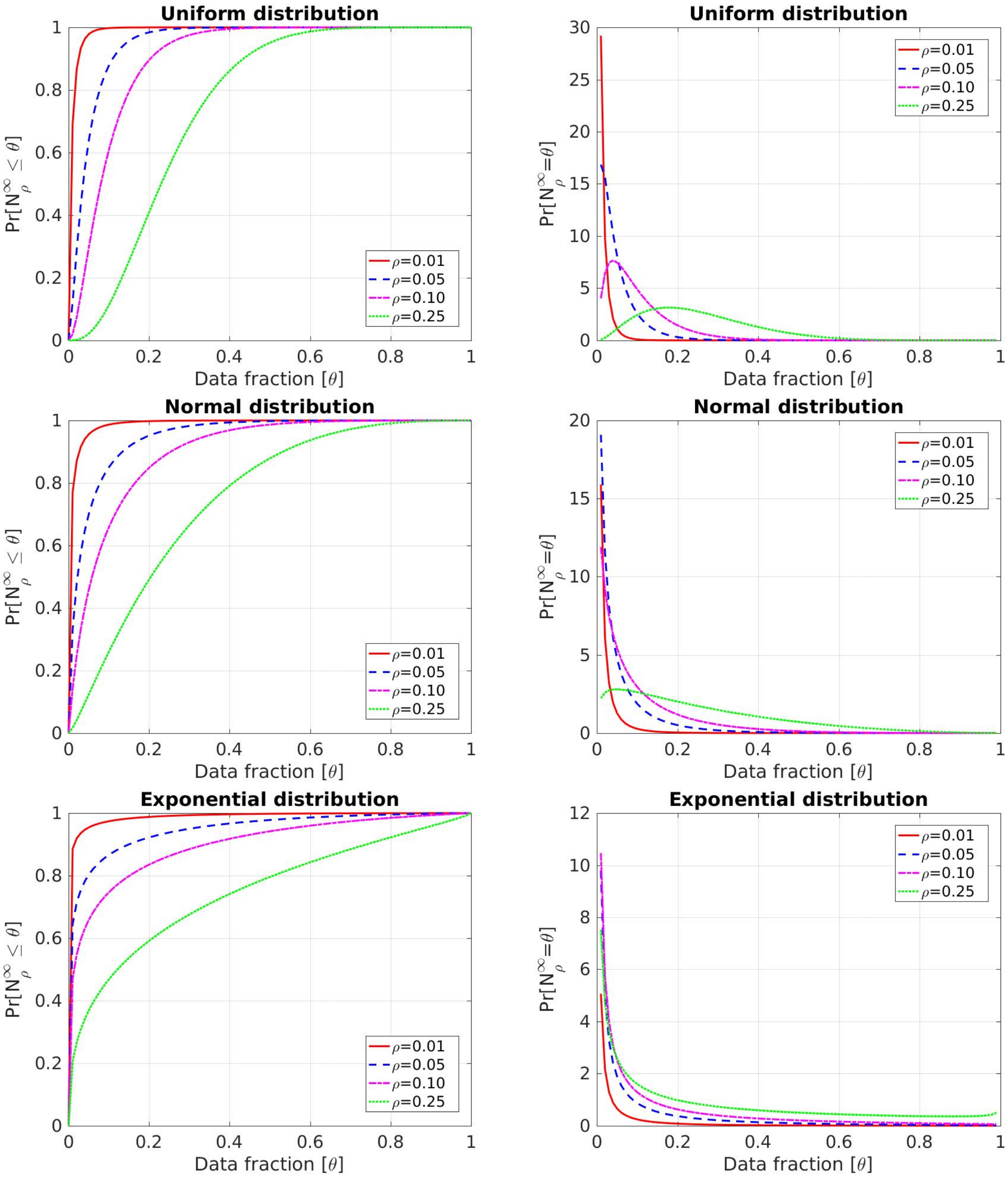


Figure 8: [Best viewed in color.] Cumulative distribution function (left column) and probability density function (right column) of the limiting distribution of the number of k-occurrences for i.i.d. random vectors (see Th. 32) according to different families of distributions, for $\varrho = 0.01$ (red solid line), $\varrho = 0.05$ (blue dashed line), $\varrho = 0.10$ (magenta dash-dotted line), and $\varrho = 0.25$ (green dotted line).

even for moderately large values of ϱ. This behavior can be better understood by looking at Figure 7, where theoretical curves of the exponential are approaching 1 earlier.

Corollary 33 *Let $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ be two d-dimensional i.i.d. random vectors with common cdf F_Y having, w.l.o.g., null mean $\mu_Y = 0$. Then, for large values of d,*

$$Pr[\mathrm{N}_k(\boldsymbol{X}_d, \{\boldsymbol{Y}_d\}_n) \leq h] \approx \Phi\left(\frac{\Phi^{-1}(\frac{h}{n})2\mu_2 - \Phi^{-1}(\frac{k}{n})\sqrt{\mu_4 + 3\mu_2^2}}{\sqrt{\mu_4 - \mu_2^2}}\right).$$

Proof of Corollary 33. The statement follows immediately from Theorem 32. ■

In order to compare the solution here derived to the large dimensional limits of the function $\mathrm{N}_k^{n,d}$ provided by Newman et al. (1983), the same limits are derived next.

Corollary 34 *Let k be a fixed positive integer. Then*

$$(i) \lim_{n\to\infty}\lim_{d\to\infty} \mathrm{N}_k^{n,d} \xrightarrow{D} 0, \quad (ii) \lim_{n\to\infty}\lim_{d\to\infty} \sigma^2(\mathrm{N}_k^{n,d}) = \infty, \quad \textit{and } (iii) \lim_{n\to\infty}\lim_{d\to\infty} \mathbf{E}[\mathrm{N}_k^{n,d}] = k.$$

Proof of Corollary 34. See the appendix. ■

Results provided by Newman et al. (1983), correspond to points (i) and (ii) above for the case $k = 1$. The point (iii) is reported only as a check, because k is expected by definition of the k-occurrences.

The interpretation of the above result provided by Tversky et al. (1983), which is typically how it is reported in the related literature, is that if the number of dimensions is large relative to the number of points, one may expect to have a large proportion of points with reverse nearest neighbor counts equaling 0, and a small proportion of points with high count values. However, according to the previous findings, the convergence in distribution to zero does not have to be motivated by the excess of the dimensions with respect to the sample size, but rather by the use of a fixed-size neighborhood parameter k in the presence of large samples. As a matter of fact, the curves reported in Figure 8 tend to the zero distribution only for $\varrho = k/n \to 0$. Moreover, large counts can also be achieved in the case of small samples and large dimensionalities, as shown in Figures 7 and 8. E.g., from Equation (5), the expected number of points such that $z \leq -1$ ($z \leq -2$, resp.) is about the 15.9% (2.3%, resp.) for any sample size n. All of this suggests that hubness is definitely not an artifact of a finite sample.

4.4 The Distribution of Independent Non-Identically Distributed Data

Previous results can be extended to independent non-identically distributed random vectors. With this aim, we consider the following proposition.

Given a sequence $W_1, W_2, \ldots, W_d$ of independent non-identically distributed random variables having non-null variances[8] and finite central moments $\hat{\mu}_{i,k}$ up to the eighth mo-

[8] Clearly, variables having constant domain, hence null variance, can be disregarded because they do not alter distances.

ment, we say that the sequence has *comparable* central moments if there exist positive constants $\hat{\mu}_{\max} \geq \max_{i,k}\{|\hat{\mu}_{i,k}|\}$ and $\hat{\mu}_{\min} \leq \min_i\{|\hat{\mu}_{i,k}| : \hat{\mu}_{i,k} \neq 0\}$. Intuitively, this guarantees that the ratio between the greatest and the smallest non-null moment remains limited.[9]

Proposition 35 *Let $U_d = \sum_{i=1}^{d} W_i$ be a random variable defined as the summation of a sequence of independent, but not identically distributed, random variables W_i having comparable central moments. Then*

$$U_d \simeq \mathcal{N}\left(\sum_{i=1}^{d} \mu_{W_i},\ \sum_{i=1}^{d} \sigma^2_{W_i}\right) = \mathcal{N}\left(d \cdot \overline{\mu}_W,\ d \cdot \overline{\sigma^2}_W\right),$$

where $\overline{\mu}_W = (1/d)\sum_{i=1}^{d} \mu_{W_i}$ and $\overline{\sigma^2}_W = (1/d)\sum_{i=1}^{d} \sigma^2_{W_i}$.

Proof of Proposition 35. See the appendix. ■

François et al. (2007, cf. Proposition 2) affirmed that Theorem 3 holds for independent non-identically distributed variables provided that they are normalized. Authors justify this result by noting that norms will concentrate because normalization prevents variables from having too little effect on the distance values. According to this interpretation, normalization is essential for having comparable variances. (Recall that the variance is the second order central moment.)

Definition 36 *Let $\boldsymbol{Y}_d = (Y_1, Y_2, \ldots, Y_d)$ be an independent non-identically distributed d-dimensional random vector with cdfs $\boldsymbol{F}_Y = (F_{Y_1}, F_{Y_2}, \ldots, F_{Y_d})$ having k-th moments $\boldsymbol{\mu_k} = (\mu_{Y_{1,k}}, \mu_{Y_{2,k}}, \ldots, \mu_{Y_{d,k}}) = (\mu_{1,k}, \mu_{2,k}, \ldots, \mu_{d,k})$. Moreover, given a positive integer $h, k \geq 1$, let $\tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{h}}$ denote the average h-th degree of the k-th central moments of $\boldsymbol{Y}_d$, also referred to as average central moment for simplicity, defined as*

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{h}} = \frac{1}{d}\sum_{i=1}^{d} \hat{\mu}_{i,k}^{h} = \frac{1}{d}\sum_{i=1}^{d} \mathbf{E}[(Y_i - \mu_{Y_i})^k]^h,$$

where $\hat{\mu}_{i,k}$ denotes the k-th central moment of Y_i.

Because considering random variables having null means simplifies expressions, for the sake of simplicity, we next consider the case of independent non-identically distributed random vectors having common cdfs, but we note that a similar result also holds in the more general case $\boldsymbol{F_X} \neq \boldsymbol{F_Y}$.

Theorem 37 *Let $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ be two independent non-identically distributed d-dimensional random vectors with common cdfs $\boldsymbol{F}$ having means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$, non null variances,*

[9] This definition fits the Lyapunov condition. In general, given a sequence $W_1, W_2, \ldots, W_d$ of independent non-identically distributed random variables having non-null finite variances, then their standardized sum converges in distribution to a standard normal random variable if and only if the Feller-Lindeberg condition holds (Feller, 1971). According to this condition, the variance $\sigma(W_i)$ of any individual term never dominates their sum s_d; hence, $\lim_{d\to\infty} \max_{i=1}^{d} \sigma_i^2(W_i)/s_d^2 = 0$. Because this both necessary and sufficient for the CLT to hold, the Feller-Lindeberg condition implies the Lyapunov condition.

and comparable central moments, and let $\boldsymbol{x}_d$ denote a realization of $\boldsymbol{X}_d$. The results of Sections 4.1, 4.2 and 4.3 can be applied to $\boldsymbol{X}_d$, $\boldsymbol{Y}_d$, and $\boldsymbol{x}_d$ by taking into account the average central moments of $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ and the realization $\boldsymbol{x}_d - \boldsymbol{\mu}$.

Proof of Theorem 37. See the appendix for details. ■

To illustrate the above results, real data sets having dimensionality varying at some order of magnitude are considered. The data sets are briefly described next. The *Statlog* (Landsat Satellite) data set[10] consists of multi-spectral values of pixels in 3 × 3 neighborhoods in a satellite image ($d = 36$, $n = 6{,}435$). The SIFT data set[11] consists of the base vectors of the `ANN_SIFT10` evaluation set used to evaluate the quality of approximate nearest neighbors search algorithms and consists of SIFT image descriptors ($d = 128$, $n = 10{,}000$). The MNIST data set[12] consists of handwritten digits which have been size-normalized and centered in a 28 × 28 image. The test examples have been employed ($d = 784$, $n = 10{,}000$). The *Sports* data set[13] consists of time series representing sensor measurements associated with activities performed by eight subjects for 5 minutes ($d = 5{,}625$, $n = 9{,}120$). The *NIPS* textual data set[14] consists of counts associated with words appearing in 5,812 NIPS conference papers published between 1987 and 2015 ($d = 11{,}463$, $n = 5{,}812$). The *RNA-Seq* data set[15] consists of gene expressions levels, measured by a illumina HiSeq sequencing platform, of patients having different types of tumors ($d = 20{,}531$, $n = 801$).

In the following, we also consider the *shuffled* version of the original data set. Specifically, the shuffled version of a given data set is obtained by randomly permuting the elements within every attribute. As already noted by François et al. (2007), the shuffled data set is marginally distributed as the original one, but because all the relationships between variables are destroyed, its components are independent, and its intrinsic dimension is equal to its embedding dimension.

Figure 9 reports the empirical cdf of the squared distance (solid line) associated with each data set, together with the theoretical cdf (dashed line) associated with independent but not identically data having the same average central moments of the original data, as reported in Theorem 37. The latter curve has been obtained by using the average central moments of the data according to Theorem 37. The empirical cdf of the shuffled data is also reported (dotted line).

From the linearity of the expected value, for any pair of d-dimensional random vectors, it follows that

$$\mathbf{E}[\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2] = d(\tilde{\boldsymbol{\mu}}_{\boldsymbol{X},\mathbf{2}} + \tilde{\boldsymbol{\mu}}_{\boldsymbol{Y},\mathbf{2}}) \quad \text{and} \quad \mathbf{E}[\|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2] = \|\boldsymbol{x}_d - \boldsymbol{\mu}_{\boldsymbol{Y}}\|^2 + d\tilde{\boldsymbol{\mu}}_{\boldsymbol{Y},\mathbf{2}},$$

where the equality holds also for dependent and non-identically distributed random vectors. Hence, the expected value of the pairwise distances between data set points is identical to

[10] Data available at `https://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29`.

[11] Data available at `http://corpus-texmex.irisa.fr/`.

[12] Data available at `http://yann.lecun.com/exdb/mnist/`.

[13] Data available at `https://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities`.

[14] Data available at `https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015`.

[15] Data available at `https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq`.

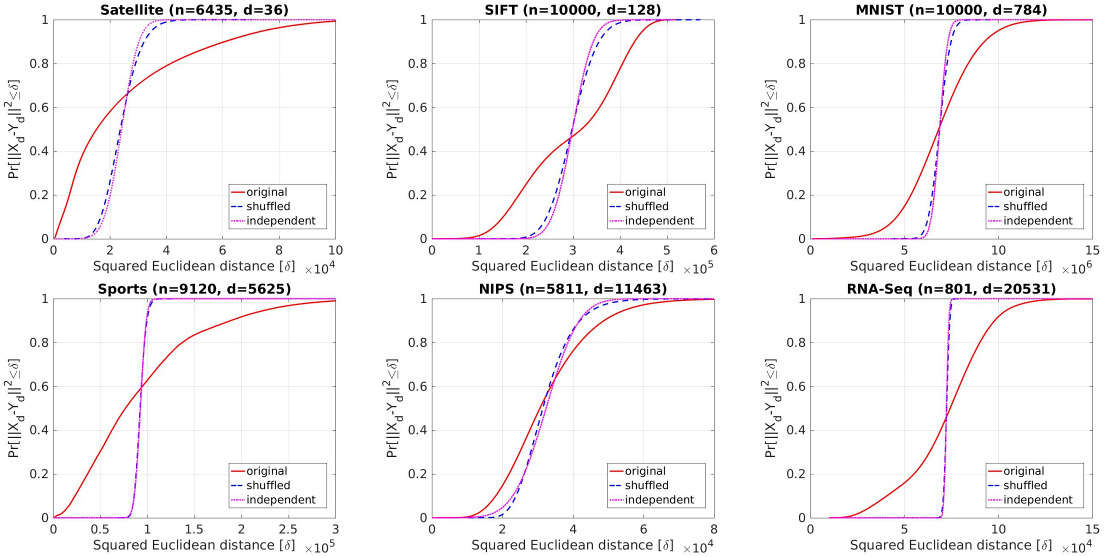


Figure 9: [Best viewed in color.] Pairwise distance distributions for real data sets, including the original data (red solid line), the shuffled data (blue dashed line), and the equivalent independent data (magenta dotted line). The last curve corresponds to the theoretical cdf associated with independent but not identically data having the same average central moments of the original data, as reported in Theorem 37.

the expected value of the theoretical distribution derived under the i.i.d. hypothesis, and also to that of the shuffled data.

The difference between the curves of the original and of the independent data suggests that the intrinsic dimensionality of the data at hand is smaller than that of the embedding space, because dependencies evidently exist between the attributes. Moreover, it can be seen that the empirical cdf of the shuffled data is very similar to that of the theoretical cdf.

These result confirm the correctness of Theorem 37, whose prediction agrees with the empirical observation on real independent data. Moreover, its approximation is accurate even in moderately large spaces, because the correspondence is good even for the smallest data set considered ($d = 36$). Moreover, these experiments testify to the meaningfulness of the analysis here accomplished as a worst-case analysis scenario, corresponding to the case in which relationships between variables are absent.

In order to verify if the data sets satisfy the requirements for the CLT to be applied, the value of the finite Lyapunov CLT condition (see Equation 1, for $\delta = 2$) has been determined on the data at hand (with the variable W_i taking value over all the terms $(x_{j,i} - x_{k,i})^2$ that can be formed with distinct pairs of data set points x_j and x_k, $1 \leq j < k \leq n$). Table 1 reports the value of the above condition (indicated as LC) together with the Relative Variance (RV) of the norm of data set points, for both the original data set (note that the shuffled data presents the same LC value as the original one) and its normalized version (obtained by substituting each attribute X_i with $(X_i - \mu_i)/\sigma_i$):

Data set	Original		Normalized	
	LC	RV	LC	RV
Satellite	0.012675	0.463934	0.014629	0.427250
SIFT	0.003382	0.063242	0.049320	0.090368
MNIST	0.000403	0.133234	25.163236	0.502937
Sports	0.072869	0.432149	0.546147	0.361839
NIPS	0.745970	0.257738	0.307322	0.241819
RNA-Seq	0.000412	0.131759	0.096444	0.175561

Table 1: Comparison between the values of the Lyapunov CLT Condition (LC) and the Relative Variance (RV) against those of the real data sets.

A small LC value (say, less than 1) indicates that the normal approximation is correct. This condition is met for all the configurations except for the normalized *MNIST* data set. Indeed, normalizing this data set is not meaningful, because attributes (corresponding to pixels) are already homogeneous (their domain consists of 256 gray levels encoded as byte values) and because the normalization has only the negative effect of exaggerating the range of variation of pixels whose domain is formed almost entirely of zeros. As a result, a few attributes dominate the distance, and this deteriorates convergence to normality. The relative variance has been reported for comparison, because it is a measure of the concentration of the data. (Note that the relative variance of the shuffled data is not coincident with that of the original one.)

Figure 10 reports the relative number of k-occurrences associated with data set points represented in terms of their squared norm standard score z. Different values for the parameter ϱ have been employed, Specifically, $\varrho \in \{0.01, 0.05, 0.10, 0.25\}$ (the color of points in the figure is magenta for $\varrho = 0.01$, green for $\varrho = 0.05$, blue for $\varrho = 0.10$ and cyan for $\varrho = 0.25$). The theoretical curves of $\mathrm{N}_{\varrho}^{\infty}(z)$, based on average central moments of the data, are also reported for comparison.

Interestingly, the real distributions of k-occurrences follow the trends of the theoretical curves; however, in contrast to the independent case, counts have much larger variability. The interpretation is that variability is associated with a lower intrinsic dimensionality and dependencies among variables, because independent data are much more widely distributed along the theoretical trend, as can be seen in Figure 7. Indeed, such behavior is also observed when considering the shuffled data set (see Figure 11). Moreover, in this case, the empirical evidence matches the behavior predicted by Theorem 37 for the independent case. In some cases, the trend appears to be different albeit generally analogous. It appears that the degree of agreement between the empirical evidence and the the theoretical prediction is directly proportional to the value of the LC condition reported in Table 1.

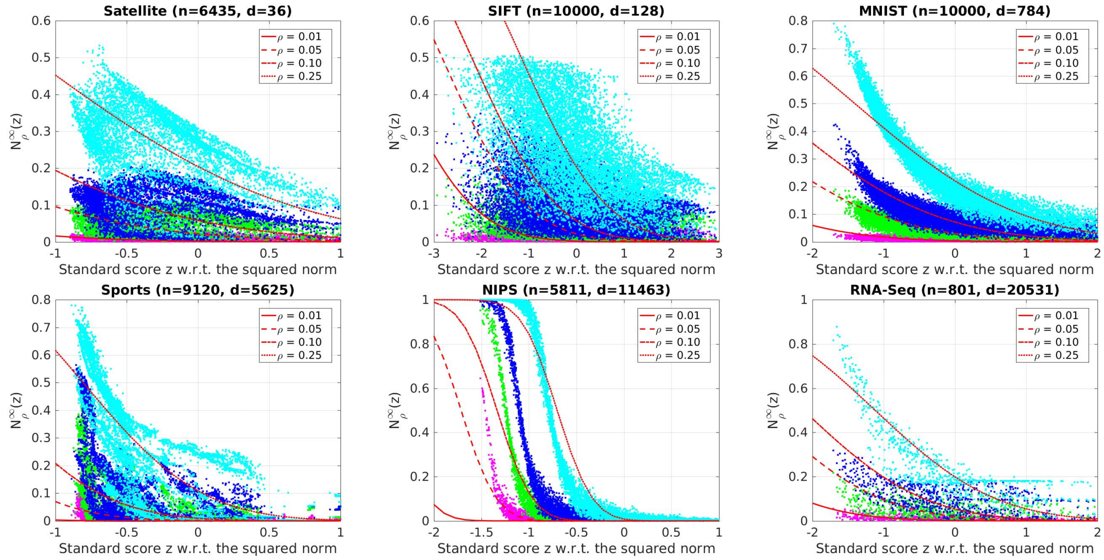


Figure 10: [Best viewed in color.] Relative number of k-occurrences associated with data set points, represented in terms of their squared norm standard score z (different colors), and the theoretical prediction according to the infinite dimensional k-occurrences function reported in Equation (4.3) (red lines), for the following values of $\varrho = k/n$: $\varrho_1 = 0.01$ (magenta-colored points and solid red line), $\varrho_2 = 0.05$ (green-colored points and dashed red line), $\varrho_3 = 0.10$ (blue-colored points and dash-dotted red line), and $\varrho_4 = 0.25$ (cyan-colored points and dotted red line).

4.5 Extension to Other Distances

In general, one may attempt to extend some of the properties discussed above to distances having the general form

$$\operatorname{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) = h\left(\sum_{i=1}^{d} g(x_i, y_i)\right), \tag{6}$$

with $g : \mathbb{R}^2 \mapsto \mathbb{R}$ being commutative and not identically constant, and $h : \mathbb{R} \mapsto \mathbb{R}$ strictly monotonic and, hence, invertible. Indeed, let $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ be d-dimensional i.i.d. random vectors, and consider the random variable

$$h^{-1}\left(\operatorname{dist}(\boldsymbol{X}_d, \boldsymbol{Y}_d)\right) = \sum_{i=1}^{d} g(X_i, Y_i) = \sum_{i=1}^{d} W_i.$$

Because $W_1, W_2, W_3, \ldots$ is a sequence of i.i.d. random variables, by the CLT, it can be said that $h^{-1}\left(\operatorname{dist}(\boldsymbol{X}_d, \boldsymbol{Y}_d)\right) \simeq \Phi\big(d \cdot \mathbf{E}[g(X_i, Y_i)],\ d \cdot \sigma^2(g(X_i, Y_i))\big)$ and, for large values of d,

$$Pr\left[\operatorname{dist}(\boldsymbol{X}_d, \boldsymbol{Y}_d) \le \delta\right] \approx \Phi\left(\frac{h(\delta) - d \cdot \mathbf{E}[g(X_i, Y_i)]}{\sqrt{d} \cdot \sigma(g(X_i, Y_i))}\right).$$

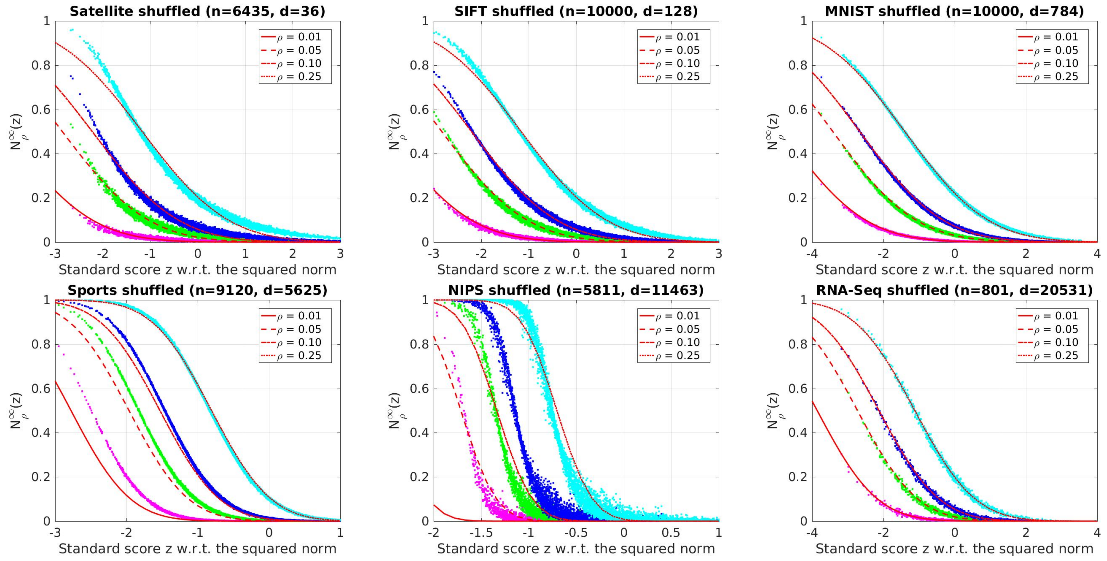


Figure 11: [Best viewed in color.] Relative number of k-occurrences associated with data set points (the shuffled version of each data set here being considered, which is obtained by randomly permuting the elements within every attribute), represented in terms of their squared norm standard score z (different colors), and theoretical prediction according to the infinite dimensional k-occurrences function reported in Equation (4.3), for the following values of $\varrho = k/n$: $\varrho_1 = 0.01$ (magenta-colored points and solid red line), $\varrho_2 = 0.05$ (green-colored points and dashed red line), $\varrho_3 = 0.10$ (blue-colored points and dash-dotted red line), and $\varrho_4 = 0.25$ (cyan-colored points and dotted red line).

As an example, consider the Minkowski norm L_p, $\|\boldsymbol{x}_d\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}$, with p a positive integer. Let, for the sake of simplicity, p be even, then

$$\begin{aligned}
\mathbf{E}[\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|_p^p] &= d\sum_{j=0}^{p}(-1)^j \binom{p}{j} \mu_{X,p-j}\mu_{Y,j}, \text{ and} \\
\sigma^2(\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|_p^p) &= d\sum_{j=0}^{p} \binom{p}{j}^2 \sigma^2(X_i^{p-j}Y_i^j) + \\
&\quad +d\sum_{j=0}^{p}\sum_{j\neq k=0}^{p} (-1)^{j+k}\binom{p}{j}\binom{p}{k} cov(X_i^{p-j}Y_i^j, X_i^{p-k}Y_i^k).
\end{aligned}$$

Newman and Rinott (1985) reported a generalized version of Theorem 4, in which distances of the form used in Equation (6) are considered.

Theorem 38 (Adapted from Newman and Rinott, 1985, cf. Theorem 3) *Consider the generalized distance function reported in Equation* (6)*. Let* $\beta_g = corr\big(g(X,Y), g(X,Z)\big)$ *be the correlation between* $g(X,Y)$ *and* $g(X,Z)$*, where* X*,* Y*,* Z *are i.i.d. random variables* *with common distribution F, and let $0 < \sigma^2(g(X,Y)) < \infty$. If $\beta_g > 0$, then Theorem 4 holds even if the generalized distance is employed instead of the Euclidean distance.*

Both the Euclidean distance and all Minkowski's metrics with $p \neq 0$ respect the condition $\beta_g > 0$. This condition implies that the location of vector components plays a role when computing pairwise distances, because when it holds, the closer the coordinate value x_i of $\boldsymbol{x}_d$ to the the expected position of X_i, the more likely it is that the vector $\boldsymbol{x}_d$ will be close to the other realizations of the same random vector.

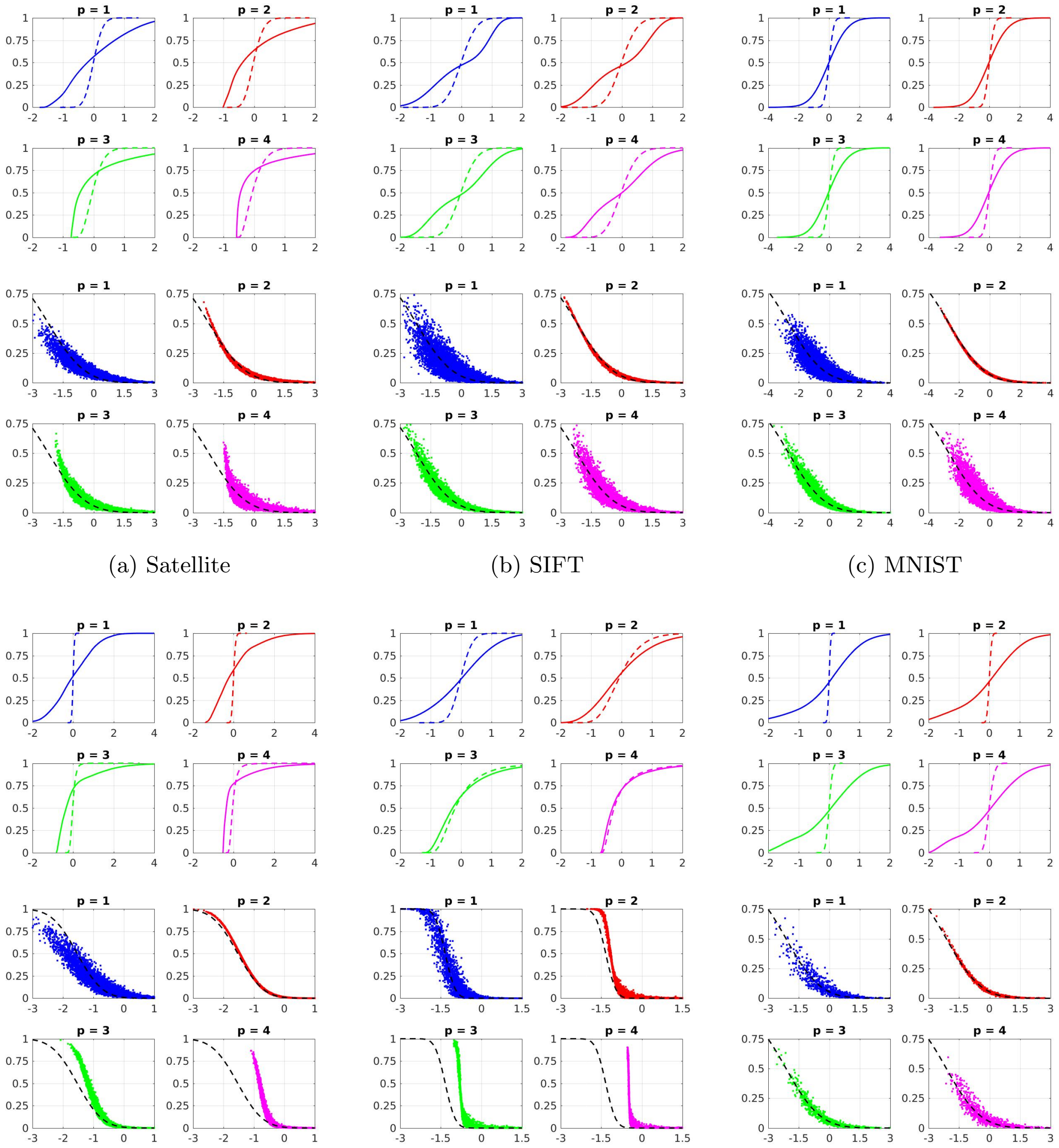
In contrast, the case $\beta_g = 0$ means that no vector occupies a special position. This condition is valid, for example, for Poisson process, which spread the vectors uniformly over $\mathbb{R}^d$. In this case, all positions within the space become equivalent and, hence, no concentration of distances is exhibited. As already noted by Radovanovic et al. (2010), the absence of spatial centrality can be intuitively used to explain the absence of hubness for cosine distance, since in this setting, no vector is more spatially central than any other, and the observation of the emergence of hubness for distance measures such as the L_p norm, Bray-Curtis, normalized Euclidean, and Canberra.

Because Theorems 16, 25 and 30, as well as related ones, are based on spatial centrality in that (the standard score of) the squared norm plays a special role in their formulation, it is conceivable that by following the same line here presented, similarly behaving closed forms can be obtained for any other distance presenting spatial centrality (namely, such that $\beta_g > 0$), expressed in terms of the standard score (or other degree) of some measure $\ell(\boldsymbol{x}_d)$ of centrality of $\boldsymbol{x}_d$. For example, for L_p norms, the natural measure of centrality is $\|\boldsymbol{x}_d\|_p^p$.

Figure 12 reports the distribution of pairwise distances and the relative number of k-occurrences for different Minkowski's metrics p. Namely, $p \in \{1,2,3,4\}$ (colors employed are blue for $p=1$, red for $p=2$, green for $p=3$, and magenta for $p=4$), on the real data sets considered in the previous section.

To facilitate the comparison of results involving different metrics, both distance values and norm values have been normalized. Specifically, let $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ be two independent and not identically distributed random vectors whose components X_i and Y_i have the same central moments of the i-th attribute of the data set, and let $\boldsymbol{x}_d$ and $\boldsymbol{y}_d$ be two data set points. As for pairwise distance distributions, on the x-axis, the value $z_{dist}(\boldsymbol{x}_d, \boldsymbol{y}_d) = \frac{\|\boldsymbol{x}_d-\boldsymbol{y}_d\|_p^p-\mathbf{E}[\|\boldsymbol{X}_d-\boldsymbol{Y}_d\|_p^p]}{\sigma(\|\boldsymbol{X}_d-\boldsymbol{Y}_d\|_p^p)}$ is reported, whereas for the relative number of k-occurrences, the value $z_{norm}(\boldsymbol{x}_d) = \frac{\|\boldsymbol{x}_d\|_p^p-\mathbf{E}[\|\boldsymbol{X}_d\|_p^p]}{\sigma(\|\boldsymbol{X}_d\|_p^p)}$ is reported on the x-axis.

In Figure 12, plots concerning the pairwise distance distribution (that are, for each data set, the four plots on the top), report the cdf associated with the original data set (solid line) and the cdf associated with the shuffled data (dashed line). The curve of the cdf associated with the equivalent independent data—that is, the cdf of the normal distribution having mean $\mathbf{E}[\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|_p^p]$ and standard deviation $\sigma(\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|_p^p)$—is not reported for clarity, because its curve overlaps with that of the shuffled data. In Figure 12, plots concerning k-occurrences (which are, for each data set, the four plots on the bottom) report the relative number of k-occurrences (for $k = \varrho n$ and $\varrho = 0.1$) associated with the points of the shuffled data set (color varying with p) together with the value $\mathrm{N}_k^\infty(z_{norm})$ of the infinite dimensional k-occurrences function reported in Equation (4.3) evaluated in $z_{norm}(\boldsymbol{x}_d)$ (black dashed



(a) Satellite (b) SIFT (c) MNIST

(d) Sports (e) NIPS (f) RNA-Seq

Figure 12: [Best viewed in color.] Experimental results on real data for different Minkowski's metrics p (blue for $p = 1$, red for $p = 2$, green for $p = 3$, and magenta for $p = 4$). For each data set, the 4 plots on the top show the cdfs of pairwise distances for the original (solid line) and shuffled data (dashed line). Moreover, for each data set, the 4 plots on the bottom show the relative number of k-occurrences ($k = \varrho n$ with $\varrho = 0.1$) for the points of the shuffled data set (colored dots) and the value of the function $\mathrm{N}_{\varrho}^{\infty}$ (black dashed line) reported in Equation (4.3). The values on the x-axis are standard scores using $\|\cdot\|_p^p$ as measure of centrality.

line). Interestingly, as previously hypothesized, by representing the data in terms of the standard score of the measure of centrality $\|\cdot\|_p^p$, a similar behavior can also be observed for different Minkowski's metrics p. In general, the results for $p = 1$ are very similar to those for $p = 2$, whereas for $p > 2$, it seems that the degree of agreement is related to the value of the LC condition reported in the first column of Table 1.

5. Relationship with Hubness in Network Science

Because hubness is a phenomenon of primary importance in network science, we wondered if the findings relative to the distribution of the reverse nearest neighbors and the emergence of hubs in intrinsically high-dimensions have connections with the analogous phenomenon occurring in the context of networks.

It is well understood that complex networks (Barabási and Pósfai, 2016) arise from different natural and human-made systems, e.g., the Internet, the world-wide web, citation networks and some social networks. These networks exhibit as a major property a few nodes, called *hubs*, with unusually high degree as compared to the other nodes of the network.

In most cases, it has been observed that networks are approximatively scale-free that is, they have approximate power-law degree distributions. Specifically, in most cases, the approximation is true only for the tails of the node degree distribution. Tails are associated with the larger node degrees, which are also the less probable observations, whereas most of the probability mass is associated with the smallest degree values.

Early well-known random graph models, such as the Erdős and Rényi (1959) model, do not exhibit power laws. Thus, other models for generating scale-free networks have been proposed. The Bianconi and Barabási (2001) model generates scale-free networks based on three important concepts that have been observed in real networks: growth, preferential attachment, and fitness. Preferential attachment means that the more connected a node is, the more likely it is to receive new links. Fitness is an intrinsic value associated with each node, defined as the ability to attract new links.

At least two analogies between the study herein conducted and what was depicted above can be identified by regarding nodes as point in a high-dimensional space.

First, in some cases, the behavior of the theoretical pdf of the function N_k^∞ exhibits a transition (on the basis of the value $\varrho = k/n$) between the two aforementioned families of networks (see Theorem 32 (*ii*) and the bottom left plot in Figure 8 for uniform data); namely, the behavior is binomial-like for high ϱ values and skewed to the right with the emergence of hubs for small ϱ values.

Second, the squared norm standard score z for points can be conceptualized as a value of fitness that is assigned to nodes/points according to a certain probability ($\phi(z)$, that is the standard Normal pdf, in the case of points); consequently, $\mathrm{N}_k^\infty(z)$ represents the relative expected number of times that a certain node/point with fitness z will be referred by any other node of the network/data set. The more central the node (the closer the point to the mean), the higher its fitness and the higher is its probability of being selected as a neighbor by the rest of the points.

To ascertain whether the above analogies adhere to empirical evidence, we examined the node (in-)degree distribution of real networks. Given a directed network, consisting of

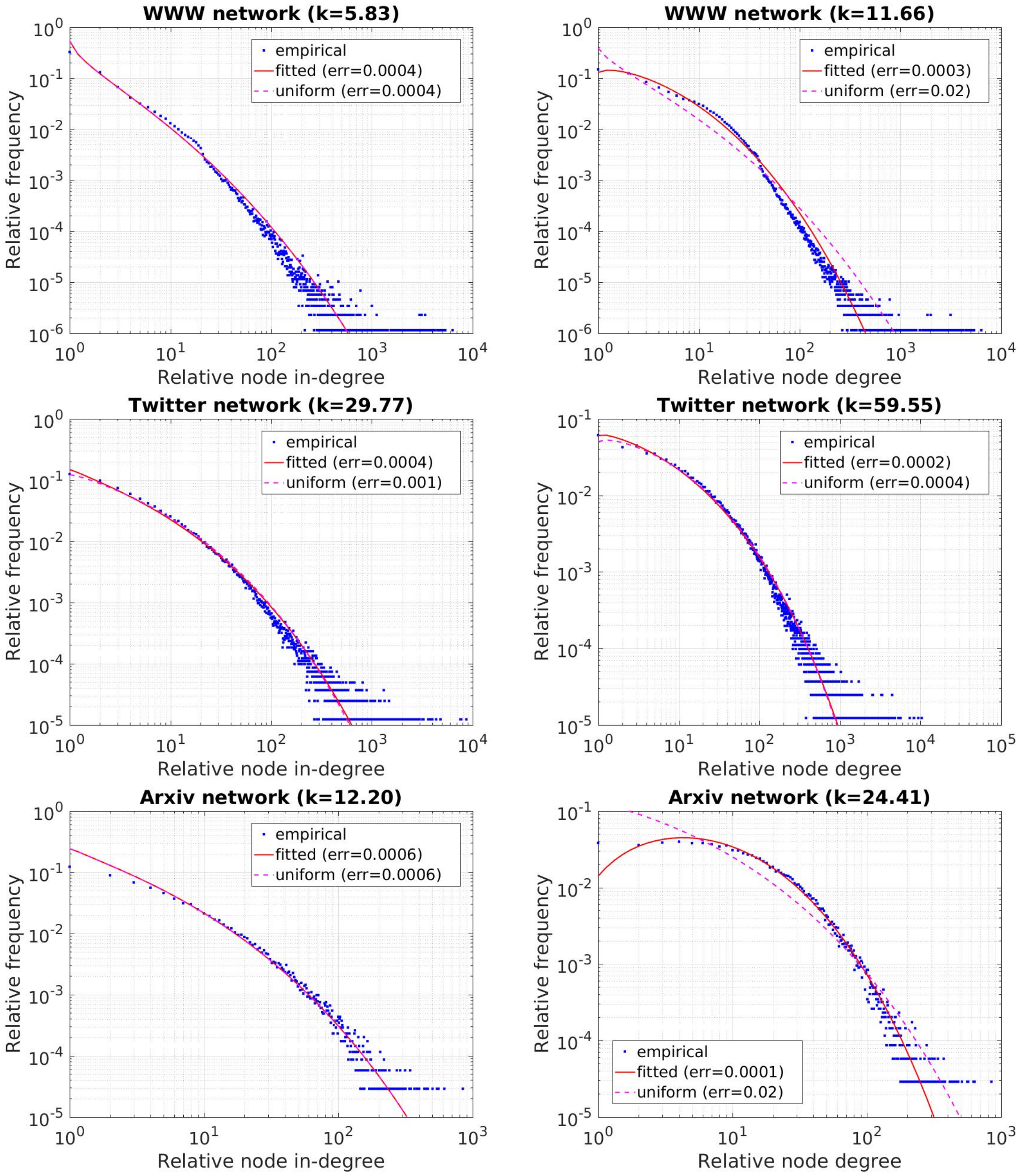


Figure 13: [Best viewed in color.] Relative node (in-)degrees (horizontal axis) and associated relative frequency (vertical axis) in log-log scale for different real-life complex networks (blue dots). Lines are associated with the probability of observing a point which has a certain number of reverse k-nearest neighbors in a set of n other points. Here, the average (in-)degree number is used as value for k. The magenta dashed line is associated with a uniform distribution in $[-0.5, +0.5]$. The red solid line is obtained using the moments minimizing the Cramér-von Mises distance between the theoretical and empirical cdfs.

n nodes and m arcs, let e_i (e_i^{in}, resp.) denote the number of arcs connected to (coming into, resp.) node n_i ($1 \leq i \leq n$).

Let $F_k(h)$ denote the discrete version of the cdf associated with the infinite dimensional k-occurrences function N_k^{∞} (see Corollary 33). Then, $p_k(h) = F_k(h) - F_k(h-1)$ denotes the probability of observing a point which has exactly h reverse k-nearest neighbors in a set of n other points. As for the value of k, we used the average number of (incoming) arcs, that is $\overline{k} = \frac{1}{n}\sum_{i=1} e_i$ and $\overline{k}^{in} = \frac{1}{n}\sum_{i=1} e_i^{in}$, which is in general a rational number. This is consistent at least for incoming edges with Theorem 32 (*iv*), for which the expected value of N_k is indeed k.

We considered directed networks from the *Stanford Large Network Dataset Collection* of the Stanford Network Analysis Project (SNAP) (Leskovec and Krevl, 2014). We obtained similar results in different cases. Figure 13 reports the results concerning the following three directed networks: *WWW* (n = 875,713 nodes and m = 5,105,039 arcs), the `web-Google` web graph from Google; *Twitter* (n = 81,306 and m = 1,768,149), the `ego-Twitter` social circles from Twitter; and *Arxiv* (n = 34,546 and m = 421,578), the `cit-HepPh` arXiv high energy physics paper citation network.

Plots in Figure 13 are in a log-log scale and report on the horizontal axis the node degree and on the vertical axis the associated relative frequency. Blue dots represent the empirical values associated with each network. The magenta dashed line represents the function $p_k^u = p_k$ associated with a random variable uniformly distributed in $[-0.5, +0.5]$ ($\mu_2 = 0.08\overline{3}$, $\mu_4 = 0.0125$, and $\kappa = 1.8$). The red solid line represents the function $p_k^* = p_k$ using the values $\mu_2 = \mu_2^*$ and $\mu_4 = \mu_4^*$ (or $\kappa^* = \mu_4^*/(\mu_2^*)^2$) that minimize the Cramér-von Mises distance, also called *err* in the plots, between the cdf F_k and the empirical cdf of the node (in-)degree distribution. The Cramér-von Mises criterion corresponds to the integral of the squared difference between the empirical and the estimated distribution functions and is used to judge the goodness of fit of a cdf compared with an empirical cdf. This criterion depends on the entire cdf and gives more importance to the most probable observations. Alternative curves can be obtained by using other criteria.

Interestingly, we observed in different cases a good agreement between the empirical distribution of the number of incoming edges and the expected number of k-occurrences associated with the function p^u. In particular, the distance *err* for uniform data is in most cases similar to the distance for p^* associated with the best values for the parameters according to the Cramér-von mises statistics, and this is true especially for node in-degrees.

This suggests that for some real networks, the distribution of the incoming node degrees has connections with the herein-derived infinite-dimensional k-occurrences function N_k^{∞}, which models the number of reverse k-nearest neighbors in an arbitrarily large feature space of independent dimensions. Moreover, the function N_k^{∞} appears to be suitable to be leveraged as a model for node-degrees distributions in complex real networks. We are currently investigating to what extent the above observations can be generalized and the use of these findings for the generation of realistic synthetic networks.

6. Concluding Discussion

This work investigated the distribution of distances in intrinsically high-dimensional spaces and leveraged this analysis to gain knowledge of phenomena related to the so-called dimensionality curse.

The study has been focused on independent data, because it is usually assumed that the number of independent dimensions dictates the intrinsic dimensionality of the data. By applying the central limit theorem to the Euclidean distance random variable, we obtained an approximation of the distance probability distribution between a given realization of a random vector and a random vector. The analysis of the error associated with the approximation highlights that, whereas the worst-case error always decreases with the dimensionality, there are configurations of n and d for which the hypothesized distance distribution can be considered equivalent, in terms of the expected empirical error, to the underlying distribution generating the observed inter-point distances.

With reference to the distribution of the nearest neighbors, we derived the expected distance of a point from its k-th nearest neighbor and the expected size of the ϵ-neighborhood in finite high-dimensional spaces, that is, the average number of points which emerge as ϵ-approximate neighbors of any other point, and then exploited it to determine the intrinsic dimensionality at which the neighborhood is expected to become unstable, called critical dimensionality. Also, a better estimate for the relative contrast for quantifying query difficulty has been obtained.

Moreover, the function N_k, or number of k-occurrences, representing the number of points which have a given point as one of their k nearest neighbors, has been investigated. Despite the extensive use of this function in many fields, including, among others, applied statistics and mathematical psychology, data mining and machine learning, information retrieval and computational geometry, the precise characterization of its form has been a longstanding problem. The limiting probability distribution of the function N_k has been derived, thereby providing full interpretability of the associated hubness problem.

It is well understood that complex networks arising from different natural conditions exhibit a few nodes, called *hubs*, of unusually high degree as compared to the other nodes of the network. Thus, we investigated if the findings relative to the distribution of the reverse nearest neighbors and the emergence of hubs in intrinsically high-dimensions is associated with the analogous phenomenon occurring in the context of networks. We concluded that for some real-life large-scale networks the distribution of the incoming node degrees is closely related to the herein-derived infinite-dimensional k-occurrences function N_k^∞ associated with uniform data and that this function is suitable to be leveraged as a model for node-degrees distribution in complex real networks.

We believe that the current study can be leveraged in several ways and in different contexts, such as direct and reverse nearest neighbor search, density estimation, anomaly and novelty detection, density-based clustering, and network analysis, as well as others, because almost all of them are based on the concepts of direct and reverse nearest neighbors. As for the study's possible applications, one is to obtain approximations of measures that are related to distance distributions. Another is to exploit the distributions for independent data as a worst-case scenario for data analysis and retrieval techniques in order to understand their behavior and limitations, in terms of meaningfulness or computational cost, as dimensionality increases. Moreover, a deeper understanding of the behavior of intrinsically high dimensional spaces is fundamental to the design strategies that seek to mitigate the curse of dimensionality. For example, a line of research seeks to alleviate the problem by designing dissimilarity functions that suffer less on i.i.d. uniformly distributed features (François et al., 2007; Hsu and Chen, 2009). Moreover, from the discussion of Section 5, different applications within geometric models of complex networks can be devised.

To illustrate, the approximation of relative contrasts described in Section 4.2 results in estimates more accurate than those already provided, because the approach leverages a refined characterization of the distance distribution herein provided, which takes into account the relationship between the norm of the query point and its expected distance to the data points.

Additionally, the *Concentration Free Outlier Factor* (CFOF) recently introduced by Angiulli (2017) is a measure that aims to overcome the concentration problem in density estimation and outlier detection, whose behavior emerges from that of the k-occurrences function. Specifically, for a given parameter $\varrho \in [0,1]$ representing a fraction of the data population, the CFOF score of point $\boldsymbol{x}_d$ is $\mathrm{CFOF}(\boldsymbol{x}_d) = \min\{k/n : \mathrm{N}_k(\boldsymbol{x}_d) \geq n\varrho\}$, which is the smallest value for neighborhood parameter k (normalized on n) for which $\boldsymbol{x}_d$ presents a reverse neighborhood with a size of at least $n\varrho$. The intuition is that isolated points will require larger values of k than inliers in order to be selected as neighbors by an equal-sized fraction of the data population. In contrast to almost all known outlier detection measures, CFOF scores do not exhibit concentration. By leveraging the closed form of the function N_k it is possible to formally see that CFOF outliers are few in number and separated from inliers even in intrinsically high-dimensional spaces, whereas the direct use of the number of k-occurrences for outlier detection is prone to false positives. For further details, we refer to (Angiulli, 2017).

The understanding of properties characterizing high-dimensional spaces is also fundamental for enhancing intrinsic dimensionality estimation techniques. For example, Granata and Carnevale (2016), due to the difficulty of correctly working with the distance probability density function at small-length scales, propose to reconstruct that pdf at intermediate scales and then to compare it with a known pdf of a uniform distribution on a d-dimensional support.

Acknowledgments

The author would like to thank the anonymous reviewers for their meticulous work in reviewing the manuscript and for providing valuable suggestions that have helped to improve the presentation of the results.

Appendix A. Proofs

Proposition 11

$$\mathit{cov}\big(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle\big) = d\mu(\mu_3 - \mu_2\mu)$$
$$\big(\mathit{and}\ \mathit{cov}\big(\|\boldsymbol{X}_d\|^2, \langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle\big) = d\mu(\mu_3 - \mu_2\mu),\ \mathit{for\ symmetry}\big)\,.$$

Proof of Proposition 11. Consider the covariance

$$\begin{aligned}
cov\left(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle\right) &= \mathbf{E}\big[\|\boldsymbol{Y}_d\|^2 \cdot \langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle\big] - \mathbf{E}\big[\|\boldsymbol{Y}_d\|^2\big] \cdot \mathbf{E}\big[\langle \boldsymbol{X}_d, \boldsymbol{Y}_d\rangle\big] = \\
&= \mathbf{E}\left[\left(\sum_{i=1}^{d} Y_i^2\right) \cdot \left(\sum_{j=1}^{d} X_j Y_j\right)\right] - d\mu_2 \cdot d\mu^2 = \\
&= \mathbf{E}\left[\sum_{i=1}^{d} Y_i^3 X_i + \sum_{i=1}^{d} \sum_{i \neq j=1}^{d} X_j Y_j Y_i^2\right] - d^2\mu_2\mu^2 = \\
&= d\mu_3\mu + d(d-1)\mu_2\mu^2 - d^2\mu_2\mu^2 = d\mu(\mu_3 - \mu_2\mu).
\end{aligned}$$

■

Proposition 17 *Let $\boldsymbol{X}_d$ be a d-dimensional i.i.d. random vector having cdf F_X. Moreover, let p and q be positive integers, and $\beta_0, \beta_1, \ldots, \beta_p$, $\alpha_0, \alpha_1, \ldots, \alpha_q$ be real coefficients such that $\beta_p \neq 0$ and $\alpha_q \neq 0$. Then, for any $\epsilon > 0$,*

$$\lim_{d\to\infty} Pr\left[\left|\frac{\sum_{i=1}^{d}\left(\sum_{j=0}^{p} \beta_j X_i^j\right)}{\left(\sum_{i=1}^{d}\left(\sum_{j=0}^{q} \alpha_j X_i^j\right)\right)^2}\right| \geq \epsilon\right] = 0.$$

Proof of Proposition 17. Let $U_i = \left(\sum_{j=0}^{p} \beta_j X_i^j\right)$ and $V_i = \left(\sum_{j=0}^{q} \alpha_j X_i^j\right)$ $(1 \leq i \leq d)$. Moreover, let U be $\sum_{i=1}^{d} U_i$ and let V be $\sum_{i=1}^{d} V_i$. Now it is shown that, for all $\epsilon > 0$,

$$\lim_{d\to\infty} Pr\left[\left|\frac{U}{V^2}\right| \geq \epsilon\right] = 0.$$

The mean and variance of V_i are as follows (mean and variance of U_i are similar):

$$\begin{aligned}
\mathbf{E}[V_i] &= \mathbf{E}\left[\sum_{j=1}^{q} \alpha_j X_i^j\right] = \sum_{j=1}^{q} \alpha_j \mu_j, \\
\sigma^2(V_i) &= \mathbf{E}[V_i^2] - \mathbf{E}[V_i]^2 = \mathbf{E}\left[\left(\sum_{j=1}^{q} \alpha_j X_i^j\right)^2\right] - \left(\sum_{j=1}^{q} \alpha_j \mu_j\right)^2 = \\
&= \mathbf{E}\left[\sum_{j=1}^{q} \alpha_j^2 X_i^{2j} + \sum_{j=1}^{q} \sum_{k\neq j}^{q} \alpha_j \alpha_k X_i^j X_i^k\right] - \left(\sum_{j=1}^{q} \alpha_j^2 \mu_j^2 + \sum_{j=1}^{q} \sum_{k\neq j}^{q} \alpha_j \alpha_k \mu_j \mu_k\right) = \\
&= \sum_{j=1}^{q} \alpha_j^2 \mu_{2j} + \sum_{j=1}^{q} \sum_{k\neq j}^{q} \alpha_j \alpha_k \mu_j \mu_k - \sum_{j=1}^{q} \alpha_j^2 \mu_j^2 - \sum_{j=1}^{q} \sum_{k\neq j}^{q} \alpha_j \alpha_k \mu_j \mu_k = \\
&= \sum_{j=1}^{q} \alpha_j^2 \left(\mu_{2j} - \mu_j^2\right).
\end{aligned}$$

We assume that moments up to $2\max\{p,q\}$ exist finite. Since both U and V are the sum of d independent identically distributed random variables, by the CLT, as $d\to\infty$,

$$U \approx \mathcal{N}\left(d\sum_{j=1}^{p}\alpha_j\mu_j,\ d\sum_{j=1}^{p}\alpha_j^2\left(\mu_{2j}-\mu_j^2\right)\right) \text{ and } V \approx \mathcal{N}\left(d\sum_{j=1}^{p}\beta_j\mu_j,\ d\sum_{j=1}^{p}\beta_j^2\left(\mu_{2j}-\mu_j^2\right)\right).$$

Consider now the random variable V^2, having mean

$$\mathbf{E}[V^2] \;=\; \mathbf{E}[V]^2+\sigma^2(V) = d^2\mu_{V_i}^2 + d\sigma_{V_i}^2 = O(d^2).$$

Now it is essential to show that $\sigma^2(V^2)=O(d^3)$ and $cov(U,V^2)=O(d^2)$.

As for the variance $\sigma^2(V^2)=\mathbf{E}[V^4]-\mathbf{E}[V^2]^2$ of V^2, notice that the term of higher order in $\mathbf{E}[V^4]$ derives from the summation

$$\mathbf{E}\left[\sum_{i\neq j\neq k\neq h} V_iV_jV_kV_h\right] = d(d-1)(d-2)(d-3)\mu_{V_i}^4 = (d^4-6d^3+11d^2-6d)\mu_{V_i}^4,$$

and that all the other terms in $\mathbf{E}[V^4]$ are $O(d^3)$. As for $\mathbf{E}[V^2]^2=(d^2\mu_{V_i}^2+d\sigma_{V_i}^2)^2 = d^4\mu_{V_i}^4+2d^3\mu_{V_i}^2\sigma_{V_i}^2+d^2\sigma_{V_i}^4$. Since both $\mathbf{E}[V^4]$ and $\mathbf{E}[V^2]^2$ contain as term of higher order $d^4\mu_{V_i}^4$, it then follows that $\sigma^2(V^2)=O(d^3)$.

As for the covariance $cov(U,V^2)=\mathbf{E}[U\cdot V^2]-\mathbf{E}[U]\mathbf{E}[V^2]$, similar considerations hold. Indeed, notice that the term of higher order in $\mathbf{E}[U\cdot V^2]$ derives from the summation:

$$\mathbf{E}\left[\sum_{i\neq j\neq k} U_iV_jV_k\right] = d(d-1)(d-2)\mu_{U_i}\mu_{V_i}^2$$

and that all the other terms in $\mathbf{E}[U\cdot V^2]$ are $O(d^2)$. As for $\mathbf{E}[U]\mathbf{E}[V^2]=d\mu_{U_i}(d^2\mu_{V_i}^2+d\sigma_{V_i}^2) = d^3\mu_{U_i}\mu_{V_i}^2+d^2\mu_{U_i}\sigma_{V_i}^2$. Since both $\mathbf{E}[U\cdot V^2]$ and $\mathbf{E}[U]\mathbf{E}[V^2]$ contain as term of higher order $d^3\mu_{U_i}\mu_{V_i}^2$, it then follows that $cov(U,V^2)=O(d^2)$.

By exploiting Taylor series, it can be written:

$$\begin{aligned}\mathbf{E}\left[\frac{U}{V^2}\right] &\approx \frac{\mathbf{E}[U]}{\mathbf{E}[V^2]}-\frac{cov(U,V^2)}{\mathbf{E}[V^2]^2}+\frac{\mathbf{E}[U]}{\mathbf{E}[V^2]^3}\sigma^2(V^2), \text{ and}\\ \sigma^2\left(\frac{U}{V^2}\right) &\approx \left(\frac{\mathbf{E}[U]}{\mathbf{E}[V^2]}\right)^2\left(\frac{\sigma^2[U]}{\mathbf{E}[U]^2}+\frac{\sigma^2(V^2)}{\mathbf{E}[V^2]^2}-\frac{2cov(U,V^2)}{\mathbf{E}[U]\mathbf{E}[V^2]}\right).\end{aligned}$$

Then

$$\begin{aligned}\lim_{d\to\infty}\mathbf{E}\left[\frac{U}{V^2}\right] &= \lim_{d\to\infty}\left(\frac{O(d)}{O(d^2)}-\frac{O(d^2)}{O(d^4)}+\frac{O(d)}{O(d^6)}O(d^3)\right) = \lim_{d\to\infty}O(d^{-1})=0, \text{ and}\\ \lim_{d\to\infty}\sigma^2\left(\frac{U}{V^2}\right) &= \lim_{d\to\infty}\left(\frac{O(d)}{O(d^2)}\right)^2\left(\frac{O(d)}{O(d^2)}+\frac{O(d^3)}{O(d^4)}-\frac{O(d^2)}{O(d^3)}\right) = \lim_{d\to\infty}O(d^{-3})=0.\end{aligned}$$

Since both $\mathbf{E}[U/V^2]$ and $\sigma^2(U/V^2)$ are vanishing as $d\to\infty$, by exploiting the Chebicheff theorem it can be proved that U/V^2 converges in probability to 0. Let $W_d=U/V^2$ then,

$\{W_d\}$ converges in probability towards the value zero, if for all $\epsilon > 0$ the following limit evaluates to 0:

$$\lim_{d\to\infty} Pr\left[|W_d| \geq \epsilon\right] = \lim_{d\to\infty} Pr\left[|W_d - \mathbf{E}[W_d]| \geq \epsilon\right] \leq \lim_{d\to\infty} \frac{\sigma^2(W_d)}{\epsilon^2} = \lim_{d\to\infty} \frac{1}{\epsilon^2 O(d^3)} = 0.$$

■

Proposition 19 *As $d \to \infty$, with high probability $\|\boldsymbol{Y}_d\|^2$ and $\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle$ are jointly normally distributed.*

Proof of Proposition 19. The proof follows a line similar to that exploited in Propositions 10 and 18. It must be shown that all linear combinations

$$Z = a\|\boldsymbol{Y}_d\|^2 + b\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle = a\left(\sum_{i=1}^{d} Y_i^2\right) + b\left(\sum_{i=1}^{d} x_i Y_i\right) = \sum_{i=1}^{d}(aY_i^2 + bx_iY_i) = \sum_{i=1}^{d} W_i$$

are normally distributed, where $W_1, W_2, W_3, \ldots$ form a sequence of independent, but not identically distributed, random variables. The proof is completed by noticing that

$$\begin{aligned}
\mathbf{E}\left[|W_i - \mathbf{E}[W_i]|^4\right] &= \mathbf{E}\left[W_i^4 - 4W_i^3\mathbf{E}[W_i] + 6W_i^2\mathbf{E}[W_i]^2 - 4W_i\mathbf{E}[W_i]^3 + \mathbf{E}[W_i]^4\right] = \\
&= \mathbf{E}[W_i^4] - 4\mathbf{E}[W_i^3]\mathbf{E}[W_i] + 6\mathbf{E}[W_i^2]\mathbf{E}[W_i]^2 - 3\mathbf{E}[W_i]^4 = \sum_{j=0}^{4} \alpha_j x^j,
\end{aligned}$$

and

$$\sigma^2_{W_i} = (b^2\mu_2)x_i^2 + (2ab\mu_3)x_i + (a^2\mu_4 - a\mu_2^2) = \beta_2 x_i^2 + \beta_1 x_i + \beta_0,$$

from which the Lyapunov CLT condition (see Equation 1) for $\delta = 2$:

$$\lim_{d\to\infty} \left.\frac{\mathbf{E}\left[|W_i - \mathbf{E}[W_i]|^{2+\delta}\right]}{s_d^{2+\delta}}\right|_{\delta=2} = \lim_{d\to\infty} \frac{\sum_{i=1}^{d}\left(\sum_{j=0}^{4}\alpha_j x_i^j\right)}{\left(\sum_{i=1}^{d}\left(\sum_{j=0}^{2}\beta_j x_i^j\right)\right)^2} = 0.$$

The above limit converges in probability to zero for the r.v. $\boldsymbol{X}_d$ by Proposition 17. ■

Proposition 20 $cov\left(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle\right) = (\mu_3 - \mu\mu_2)\sum_{i=1}^{d} x_i.$

Proof of Proposition 20. Consider the covariance

$$\begin{aligned}
cov\left(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle\right) &= \mathbf{E}\big[\|\boldsymbol{Y}_d\|^2 \cdot \langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle\big] - \mathbf{E}\big[\|\boldsymbol{Y}_d\|^2\big] \cdot \mathbf{E}\big[\langle \boldsymbol{x}_d, \boldsymbol{Y}_d\rangle\big] = \\
&= \mathbf{E}\left[\left(\sum_{i=1}^{d} Y_i^2\right) \cdot \left(\sum_{j=1}^{d} x_j Y_j\right)\right] - d\mu_2 \cdot \mu \sum_{i=1}^{d} x_i = \\
&= \mathbf{E}\left[\sum_{i=1}^{d}\sum_{j=1}^{d} x_j Y_j Y_i^2\right] - d\mu\mu_2 \sum_{i=1}^{d} x_i = \\
&= \sum_{i=1}^{d} x_i \mathbf{E}[Y_i^3] + \sum_{i=1}^{d-1}\sum_{i\neq j=1}^{d} x_j \mathbf{E}[Y_j]\mathbf{E}[Y_i^2] - d\mu\mu_2 \sum_{i=1}^{d} x_i = \\
&= \mu_3 \sum_{i=1}^{d} x_i + (d-1)\mu\mu_2 \sum_{i=1}^{d} x_i - d\mu\mu_2 \sum_{i=1}^{d} x_i = (\mu_3 - \mu\mu_2)\sum_{i=1}^{d} x_i.
\end{aligned}$$

■

Proposition 22 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$ with cdf F_X. Then, for large values of d, with high probability*

$$\frac{\sum_{i=1}^{d} x_i}{\|\boldsymbol{x}_d\|^2} \to \frac{\mu_X}{\mu_{X,2}}.$$

Proof of Proposition 22. Assume that in general $\mu \in \mathbb{R}$. By the CLT, following the same line of reasoning of Proposition 8, it can be seen that the random variable

$$U = \sum_{i=1}^{d} X_i \approx \mathcal{N}\left(d\mu, d(\mu_2 - \mu^2)\right).$$

Let $V = \|\boldsymbol{X}_d\|^2 = \sum_{i=1}^{d} X_i^2$. By Proposition 8, $V \approx \mathcal{N}\big(d\mu_2, d(\mu_4 - \mu_2^2)\big)$. As for the covariance $cov(U, V)$, it is $d(\mu_3 - \mu\mu_2)$. Consider the ratio U/V. Now it is shown that

$$\lim_{d\to\infty} Pr\left[\left|\frac{U}{V} - \frac{\mu}{\mu_2}\right| \geq \epsilon\right] = 0.$$

By exploiting Taylor series, the mean of U/V is:

$$\begin{aligned}
\mathbf{E}\left[\frac{U}{V}\right] &\approx \frac{\mathbf{E}[U]}{\mathbf{E}[V]} - \frac{cov(U,V)}{\mathbf{E}[V]^2} + \frac{\mathbf{E}[U]}{\mathbf{E}[V]^3}\sigma^2(V) = \frac{d\mu}{d\mu_2} - \frac{d(\mu_3 - \mu\mu_2)}{d^2\mu_2^2} + \frac{d\mu}{d^3\mu_2^3} d(\mu_4 - \mu_2^2) = \\
&= \frac{\mu}{\mu_2} - \frac{\mu_3 - \mu\mu_2}{d\mu_2^2} + \frac{\mu}{d\mu_2^3}(\mu_4 - \mu_2^2) = \frac{\mu}{\mu_2} + \frac{1}{d} \cdot \frac{\mu(\mu_4 - \mu_2^2) - \mu_2(\mu_3 - \mu\mu_2)}{\mu_2^3},
\end{aligned}$$

while the variance of U/V is:

$$\begin{aligned}
\sigma^2\left(\frac{U}{V}\right) &\approx \left(\frac{\mathbf{E}[U]}{\mathbf{E}[V]}\right)^2\left(\frac{\sigma^2[U]}{\mathbf{E}[U]^2}+\frac{\sigma^2(V)}{\mathbf{E}[V]^2}-\frac{2cov(U,V)}{\mathbf{E}[U]\mathbf{E}[V]}\right)= \\
&= \left(\frac{d\mu}{d\mu_2}\right)^2\left(\frac{d(\mu_2-\mu^2)}{d^2\mu^2}+\frac{d(\mu_4-\mu_2^2)}{d^2\mu_2^2}-\frac{2d(\mu_3-\mu\mu_2)}{d^2\mu\mu_2}\right)= \\
&= \frac{1}{d}\cdot\frac{\mu_2^2(\mu_2-\mu^2)+\mu(\mu_4-\mu_2^2)-2\mu\mu_2(\mu_3-\mu\mu_2)}{\mu_2^4}.
\end{aligned}$$

The statement then follows by applying the Chebicheff theorem to show that U/V converges in probability to μ/μ_2.

$$\lim_{d\to\infty} Pr\left[\left|\frac{U}{V}-\frac{\mu}{\mu_2}\right|\geq\epsilon\right]=\lim_{d\to\infty} Pr\left[\left|\frac{U}{V}-\mathbf{E}\left[\frac{U}{V}\right]\right|\geq\epsilon\right]\leq\lim_{d\to\infty}\frac{\sigma^2\left(\frac{U}{V}\right)}{\epsilon^2}=\lim_{d\to\infty}\frac{1}{\epsilon^2 O(d)}=0.$$

■

Lemma 29 *Let $\boldsymbol{x}_d$ denote a realization of a d-dimensional i.i.d. random vector $\boldsymbol{X}_d$ with cdf F_X and let $\boldsymbol{Y}_d$ be a d-dimensional i.i.d. random vector with cdf F_Y. Assume, w.l.o.g., that F_Y has null mean $\mu_Y=0$. Then, for large values of d, with high probability*

$$Pr\left[\text{dist}(\boldsymbol{x}_d,\boldsymbol{Y}_d)\leq\delta \mid \|\boldsymbol{Y}_d\|=R\right]\approx\Phi\left(\frac{\delta^2-R^2-\|\boldsymbol{x}_d\|^2}{2\|\boldsymbol{x}_d\|\sqrt{\mu_2}}\right),$$

where moments are relative to the random vector $\boldsymbol{Y}_d$.

Proof of Lemma 29. By Proposition 19, $\|\boldsymbol{Y}_d\|^2$ and $\langle\boldsymbol{x}_d,\boldsymbol{Y}_d\rangle$ are jointly normally distributed. Then,

$$\begin{aligned}
Pr[\text{dist}(\boldsymbol{x}_d,\boldsymbol{Y}_d)\leq\delta \mid \|\boldsymbol{Y}_d\|=R] &= Pr[\|\boldsymbol{x}_d-\boldsymbol{Y}_d\|\leq\delta \mid \|\boldsymbol{Y}_d\|=R]= \\
&= Pr[\|\boldsymbol{x}_d-\boldsymbol{Y}_d\|^2\leq\delta^2 \mid \|\boldsymbol{Y}_d\|^2=R^2]= \\
&= Pr[\|\boldsymbol{x}_d\|^2+\|\boldsymbol{Y}_d\|^2-2\langle\boldsymbol{x}_d,\boldsymbol{Y}_d\rangle\leq\delta^2 \mid \|\boldsymbol{Y}_d\|^2=R^2].
\end{aligned}$$

Notice that, for distributions F_Y having null skewness ($\mu_{Y,3}=0$), by Proposition 20, $\|\boldsymbol{Y}_d\|^2$ and $\langle\boldsymbol{x}_d,\boldsymbol{Y}_d\rangle$ are both uncorrelated and independent, and it can be written that

$$Pr[\|\boldsymbol{x}_d\|^2+\|\boldsymbol{Y}_d\|^2-2\langle\boldsymbol{x}_d,\boldsymbol{Y}_d\rangle\leq\delta^2 \mid \|\boldsymbol{Y}_d\|^2=R^2]=Pr[\|\boldsymbol{x}_d\|^2+R^2-2\langle\boldsymbol{x}_d,\boldsymbol{Y}_d\rangle\leq\delta^2],$$

from which the statement follows by exploiting Proposition 18.

More in general ($\mu_{Y,3}\neq 0$) by leveraging properties within Theorem 15 and by Proposition 22 the distribution of the squared norm $\|\boldsymbol{x}_d-\boldsymbol{Y}_d\|^2$ subject to the constraint that $\|\boldsymbol{Y}_d\|^2=R^2$, tends to the normal distribution with mean

$$\mu=\|\boldsymbol{x}_d\|^2+R^2$$

and variance (by the assumption that F_X has null mean $\mu_X=0$)

$$\sigma^2=4\mu_{Y,2}\|\boldsymbol{x}_d\|^2-4\mu_{Y,3}\frac{\mu_X}{\mu_{X,2}}\|\boldsymbol{x}_d\|^2=4\mu_{Y,2}\|\boldsymbol{x}_d\|^2,$$

from which the above expression again follows.

The above holds under the assumption that R itself is selected with high probability, that is $R^2 \approx \mathbf{E}[\|\boldsymbol{Y}_d\|^2]$, as also assumed in Theorem 30. For a generic R, the moments μ_2 and μ conditioned on $\|\boldsymbol{Y}_d\| = R$ must be used. As for $\mu_2 = \mathbf{E}[Y^2 \mid \|\boldsymbol{Y}_d\|^2 = R^2] = \mathbf{E}[Y^2 \mid d\mathbf{E}[Y^2] = R^2] = R^2/d$. As for μ conditioned on $\|\boldsymbol{Y}_d\|^2 = R^2$, it is $\mu = \mu_Y = 0$ for symmetric distributions, since for each $\boldsymbol{y}_d$ such that $\|\boldsymbol{y}_d\|^2 = R^2$, it holds that $\|-\boldsymbol{y}_d\|^2 = R^2$ and $f_Y(\boldsymbol{y}_d) = f_Y(-\boldsymbol{y}_d)$.

Moreover, the closer R^2 to $\mathbf{E}[\|\boldsymbol{Y}_d\|^2] = \mu_{\|\boldsymbol{Y}_d\|^2} = d\mu_{Y,2}$, the closer the moments to their unconditioned values. Indeed, for $k \geq 1$

$$\begin{aligned}\mathbf{E}[Y^k \mid \|\boldsymbol{Y}_d\|^2 = R^2] = \tfrac{1}{Pr[\|\boldsymbol{Y}_d\|^2 = R^2]} \left(\int_{\mathbb{R}} y^k\, f_Y(y)\, Pr\left[\left(\sum_{j>1}^d Y_j^2\right) = R^2 - y^2\right] \mathrm{d}y\right) \approx \\ \approx \tfrac{1}{\phi_{\|\boldsymbol{Y}_d\|^2}(R^2)} \left(\int_{\mathbb{R}} y^k\, f_Y(y)\, \phi_{\|\boldsymbol{Y}_d\|^2}\left(R^2 - y^2\right) \mathrm{d}y\right).\end{aligned}$$

Hence, since μ_Y exists finite, as $d \to \infty$

$$\begin{aligned}\mathbf{E}\left[Y^k \mid \|\boldsymbol{Y}_d\|^2 = \mu_{\|\boldsymbol{Y}_d\|^2}\right] \approx \tfrac{1}{\phi_{\|\boldsymbol{Y}_d\|^2}\left(\mu_{\|\boldsymbol{Y}_d\|^2}\right)} \left(\int_{\mathbb{R}} y^k\, f_Y(y)\, \phi_{\|\boldsymbol{Y}_d\|^2}\left(\mu_{\|\boldsymbol{Y}_d\|^2} - y^2\right) \mathrm{d}y\right) = \\ = \tfrac{1}{\phi(0)} \left(\int_{\mathbb{R}} y^k\, f_Y(y)\, \phi\left(-y^2/\sigma_{\|\boldsymbol{Y}_d\|^2}\right) \mathrm{d}y\right) \approx \tfrac{\phi(0)}{\phi(0)} \int_{\mathbb{R}} y^k\, f_Y(y)\, \mathrm{d}y = \mu_{Y,k}.\end{aligned}$$

■

Theorem 32

$$(i) \qquad \lim_{d\to\infty} Pr[\mathrm{N}_\varrho(\boldsymbol{X}_d) \leq \theta] = \Phi\left(\frac{\Phi^{-1}(\theta)2\mu_2 - \Phi^{-1}(\varrho)\sqrt{\mu_4 + 3\mu_2^2}}{\sqrt{\mu_4 - \mu_2^2}}\right),$$

$$(ii) \qquad \lim_{d\to\infty} Pr[\mathrm{N}_\varrho(\boldsymbol{X}_d) = \theta] = \frac{2\mu_2}{\sqrt{\mu_4 - \mu_2^2}} \cdot \frac{1}{\phi(\Phi^{-1}(\theta))} \cdot \phi\left(\frac{\Phi^{-1}(\theta)2\mu_2 - \Phi^{-1}(\varrho)\sqrt{\mu_4 + 3\mu_2^2}}{\sqrt{\mu_4 - \mu_2^2}}\right),$$

$$(iii) \qquad \lim_{d\to\infty} \sigma^2(\mathrm{N}_\varrho(\boldsymbol{X}_d)) = \varrho(1-\varrho) - 2T\left(\Phi^{-1}(\varrho), \frac{2\mu_2}{\sqrt{2(\mu_4 + \mu_2^2)}}\right), \textit{ and}$$

$$(iv) \qquad \lim_{d\to\infty} \mathbf{E}[\mathrm{N}_\varrho(\boldsymbol{X}_d)] = \varrho,$$

$$\textit{where} \qquad T(h,a) = \phi(h)\int_0^a \frac{\phi(hx)}{1+x^2}\,\mathrm{d}x \qquad \textit{is the Owen's } T \textit{ function.}$$

Proof of Theorem 32. Since

$$\mathrm{N}_\varrho^\infty(Z_\theta) \leq \theta \implies Z_\theta \leq \frac{\Phi^{-1}(\theta)2\mu_2 - \Phi^{-1}(\varrho)\sqrt{\mu_4 + 3\mu_2^2}}{\sqrt{\mu_4 - \mu_2^2}},$$

point (i) corresponds to $\Phi(Z_\theta)$ and point (ii) to $\frac{\mathrm{d}}{\mathrm{d}\theta}\Phi(Z_\theta)$.

As for points (iii) and (iv), consider the Owen's Gaussian-type integrals reported in Equation (2) and in the following equation due to Owen (1980):

$$\int_{-\infty}^{+\infty} \Phi(a+bx)^2\phi(x)\,\mathrm{d}x = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) - 2T\left(\frac{a}{\sqrt{1+b^2}}, \frac{1}{\sqrt{1+2b^2}}\right). \tag{7}$$

Let us consider first point (*iv*). Since

$$\lim_{d\to\infty} \mathbf{E}[\mathrm{N}_\varrho(\boldsymbol{X}_d)] \quad = \quad \int_{-\infty}^{+\infty} \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{\mu_4+3\mu_2^2} - z\sqrt{\mu_4-\mu_2^2}}{2\mu_2}\right) \phi(z)\ \mathrm{d}z,$$

by substituting $a = \frac{\Phi^{-1}(\varrho)\sqrt{\mu_4+3\mu_2^2}}{2\mu_2}$ and $b = -\frac{\sqrt{\mu_4-\mu_2^2}}{2\mu_2}$ in Equation (2):

$$\lim_{d\to\infty} \mathbf{E}[\mathrm{N}_\varrho(\boldsymbol{X}_d)] = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{\mu_4+3\mu_2^2}}{\sqrt{\mu_4+3\mu_2^2}}\right) = \Phi(\Phi^{-1}(\varrho)) = \varrho.$$

As for point (*iii*),

$$\lim_{d\to\infty} \sigma^2(\mathrm{N}_\varrho(\boldsymbol{X}_d)) = \lim_{d\to\infty} \mathbf{E}[\mathrm{N}_\varrho(\boldsymbol{X}_d)^2] - \mathbf{E}[\mathrm{N}_\varrho(\boldsymbol{X}_d)]^2,$$

and by substituting a and b as above in the first (see Equation 2) and second (see Equation 7) Owen's formula the result is obtained. ■

Corollary 34 *Let k be a fixed positive integer. Then*

$$(i) \lim_{n\to\infty}\lim_{d\to\infty} \mathrm{N}_k^{n,d} \xrightarrow{D} 0, \quad (ii) \lim_{n\to\infty}\lim_{d\to\infty} \sigma^2(\mathrm{N}_k^{n,d}) = \infty, \ \textit{and}\ (iii) \lim_{n\to\infty}\lim_{d\to\infty} \mathbf{E}[\mathrm{N}_k^{n,d}] = k.$$

Proof of Corollary 34. All the points derive from Theorem 32. As for point (1) it suffices to show that $F_{N_k^{\infty,\infty}}(h) = Pr[N_k^{\infty,\infty} \leq h] = 1$ for $h > 0$, since $h = 0$ is not a continuity point for the cdf:

$$\begin{aligned}
\lim_{n\to\infty}\lim_{d\to\infty} Pr[\mathrm{N}_k^{n,d} > 0] \quad &= \quad 1 - \lim_{n\to\infty}\lim_{d\to\infty} Pr[\mathrm{N}_k^{n,d} \leq 0] = 1 - \lim_{n\to\infty} Pr[\mathrm{N}_k^{n,\infty} \leq 0] = \\
&= \quad 1 - \lim_{n\to\infty} \Phi\left(\frac{\Phi^{-1}(0)2\mu_2 - \Phi^{-1}(\frac{k}{n})\sqrt{\mu_4+3\mu_2^2}}{\sqrt{\mu_4-\mu_2^2}}\right) = \\
&= \quad 1 - \Phi\left(\Phi^{-1}(0)\frac{2\mu_2 - \sqrt{\mu_4+3\mu_2^2}}{\sqrt{\mu_4-\mu_2^2}}\right) = 1 - \Phi(-\infty) = 1.
\end{aligned}$$

As for point (2),

$$\begin{aligned}
\lim_{n\to\infty}\lim_{d\to\infty} \sigma^2(\mathrm{N}_k^{n,d}) \quad &= \quad \lim_{n\to\infty} n^2\left(\frac{k}{n} - \frac{k}{n^2} - 2T\left(\Phi^{-1}\left(\frac{k}{n}\right), \frac{2\mu_2}{\sqrt{2(\mu_4+\mu_2^2)}}\right)\right) = \\
&= \quad \lim_{n\to\infty} nk - k - 0 = \infty.
\end{aligned}$$

As for point (3),

$$\lim_{n\to\infty}\lim_{d\to\infty} \mathbf{E}[\mathrm{N}_k^{n,d}] \quad = \quad \lim_{d\to\infty} n\left(\frac{k}{n}\right) = k.$$

■

Proposition 35 *Let $U_d = \sum_{i=1}^d W_i$ be a random variable defined as the summation of a sequence of independent, but not identically distributed, random variables W_i having comparable central moments. Then*

$$U_d \simeq \mathcal{N}\left(\sum_{i=1}^d \mu_{W_i},\ \sum_{i=1}^d \sigma^2_{W_i}\right) = \mathcal{N}\left(d \cdot \overline{\mu}_W,\ d \cdot \overline{\sigma^2}_W\right),$$

where $\overline{\mu}_W = (1/d)\sum_{i=1}^d \mu_{W_i}$ and $\overline{\sigma^2}_W = (1/d)\sum_{i=1}^d \sigma^2_{W_i}$.

Proof of Proposition 35. For variables W_i having comparable central moments the Lyapunov CLT condition holds:

$$\lim_{d\to\infty} \frac{\sum_{i=1}^d \mathbf{E}\left[(W_i - \mathbf{E}[W_i])^4\right]}{\left(\sum_{i=1}^d \sigma^2(W_i)\right)^2} = \lim_{d\to\infty} \frac{\sum_{i=1}^d \hat{\mu}_{i,4}}{\left(\sum_{i=1}^d \hat{\mu}_{i,2}\right)^2} \leq \lim_{d\to\infty} \frac{d\hat{\mu}_{max}}{(d\hat{\mu}_{min})^2} = \lim_{d\to\infty} \frac{\hat{\mu}_{max}}{d\hat{\mu}^2_{min}} = 0.$$

■

Theorem 37. *Let $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ be two independent non-identically distributed d-dimensional random vectors with common cdfs $\boldsymbol{F}$ having means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$, non null variances, and comparable central moments, and let $\boldsymbol{x}_d$ denote a realization of $\boldsymbol{X}_d$. The results of Sections 4.1, 4.2 and 4.3 can be applied to $\boldsymbol{X}_d$, $\boldsymbol{Y}_d$, and $\boldsymbol{x}_d$ by taking into account the average central moments of $\boldsymbol{X}_d$ and $\boldsymbol{Y}_d$ and the realization $\boldsymbol{x}_d - \boldsymbol{\mu}$.*

Proof of Theorem 37. W.l.o.g. assume that $\boldsymbol{\mu} = (0, \ldots, 0)$, for otherwise it is sufficient to replace vector $\boldsymbol{X}_d$ with $\hat{\boldsymbol{X}}_d = \boldsymbol{X}_d - \boldsymbol{\mu}$ and vector $\boldsymbol{Y}_d$ with $\hat{\boldsymbol{Y}}_d = \boldsymbol{Y}_d - \boldsymbol{\mu}$. Thus, from now $\mu_i = \mathbf{E}[X_i] = \mathbf{E}[Y_i] = 0$.

Let $\boldsymbol{\mu_k} = (\mu_{k,1}, \ldots, \mu_{k,d})$ denote the k-th moments of $\boldsymbol{X}_d$ ($\boldsymbol{Y}_d$, resp.). The result follows by taking into account that the variables X_i and Y_i ($1 \leq i \leq d$) are independent but not identically distributed and, hence, by exploiting the average moments to formulate expressions.

Let $\boldsymbol{p} = (p_1, \ldots, p_d)$ and $\boldsymbol{q}_d = (q_1, \ldots, q_d)$ two d-dimensional vectors, and let h be a positive integer. In the following we denote by $\boldsymbol{p^k}$ the vector $\boldsymbol{p^k} = (p_1^k, \ldots, p_d^h)$ and by $\boldsymbol{p} \cdot \boldsymbol{q}$ the scalar product $\langle \boldsymbol{p}, \boldsymbol{q} \rangle = \sum_{i=1}^d p_i q_i$ of $\boldsymbol{p}$ and $\boldsymbol{q}$.

Thus, as for the results of Section 4.1, we obtain the following expressions:

(P37.1) $\|\boldsymbol{Y}_d\|^2 \simeq \mathcal{N}\left(d\tilde{\boldsymbol{\mu}}_{\mathbf{2}},\ d(\tilde{\boldsymbol{\mu}}_{\mathbf{4}} - \tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}})\right)$;

(P37.2) $\langle \boldsymbol{X}_d, \boldsymbol{Y}_d \rangle \simeq \mathcal{N}\left(0,\ \tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}\right)$;

(P37.3) $cov(\|\boldsymbol{Y}_d\|^2, \langle \boldsymbol{X}_d, \boldsymbol{Y}_d \rangle) = 0$;

(P37.4) $\|\boldsymbol{X}_d - \boldsymbol{Y}_d\|^2 \simeq \mathcal{N}\left(2d\tilde{\boldsymbol{\mu}}_{\mathbf{2}},\ 2d(\tilde{\boldsymbol{\mu}}_{\mathbf{4}} - \tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}})\right)$;

(P37.5) $\langle \boldsymbol{x}_d, \boldsymbol{Y}_d \rangle \simeq \mathcal{N}\left(0,\ \boldsymbol{\mu}_{\mathbf{2}} \cdot \boldsymbol{x}^{\mathbf{2}}\right)$;

(P37.6) $cov(\boldsymbol{Y}_d, \langle \boldsymbol{x}_d, \boldsymbol{Y}_d \rangle) = \boldsymbol{\mu}_{\mathbf{3}} \cdot \boldsymbol{x}_{\boldsymbol{d}}$;

$$(\mathbf{P37.7})\ \|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2 \simeq \mathcal{N}\left(\|\boldsymbol{x}_d\|^2 + d\tilde{\boldsymbol{\mu}}_{\mathbf{2}},\ d(\tilde{\boldsymbol{\mu}}_{\mathbf{4}} - \tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}) + 4\boldsymbol{\mu}_{\mathbf{2}} \cdot \boldsymbol{x}_{\boldsymbol{d}}^{\mathbf{2}} - 4\boldsymbol{\mu}_{\mathbf{3}} \cdot \boldsymbol{x}_{\boldsymbol{d}}\right).$$

Expression (**P37.1**) can be derived by exploiting $\mathbf{E}[\|\boldsymbol{Y}_d\|^2]$ and $\sigma(\|\boldsymbol{Y}_d\|^2)$ in terms of the average central moments of the random vectors, as illustrated next:

$$\begin{aligned}
\mathbf{E}[\|\boldsymbol{Y}_d\|^2] &= \mathbf{E}\left[\textstyle\sum_i Y_i^2\right] = \sum_i \mathbf{E}[Y_i^2] = \sum_i \mu_{i,2} = d\tilde{\boldsymbol{\mu}}_{\mathbf{2}},\\
\mathbf{E}[\|\boldsymbol{Y}_d\|^4] &= \mathbf{E}\left[\left(\textstyle\sum_i Y_i^2\right)^2\right] = \mathbf{E}\left[\textstyle\sum_i Y_i^4 + \sum_{i\neq j} Y_i^2 Y_j^2\right] = \sum_{i=1}^d \mu_{i,4} + \sum_{i\neq j}\mu_{i,2}\mu_{j,2}, \text{ and}\\
\sigma^2\left(\|\boldsymbol{Y}_d\|^2\right) &= \mathbf{E}[\|\boldsymbol{Y}_d\|^4] - \mathbf{E}[\|\boldsymbol{Y}_d\|^2]^2 = \mathbf{E}[\|\boldsymbol{Y}_d\|^4] - (\textstyle\sum_i \mu_{i,2})^2 =\\
&= \mathbf{E}[\|\boldsymbol{Y}_d\|^4] - \left(\textstyle\sum_i \mu_{i,2}^2 + \sum_{i\neq j}\mu_{i,2}\mu_{j,2}\right) = \sum_i \mu_{i,4} + \sum_i \mu_{i,2}^2 = d(\tilde{\boldsymbol{\mu}}_{\mathbf{4}} + \tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}).
\end{aligned}$$

The other expressions can be obtained in an analogous manner.

Moreover, by using the same line of reasoning of Proposition 22 it can be shown that:

$$(\mathbf{P37.8})\ \frac{\boldsymbol{\mu}_{\mathbf{3}} \cdot \boldsymbol{x}_{\boldsymbol{d}}}{\|\boldsymbol{x}_d\|^2} \to \frac{\boldsymbol{\mu}_{\mathbf{3}} \cdot \boldsymbol{\mu}}{\boldsymbol{\mu}_{\mathbf{2}}} = 0, \text{ and } (\mathbf{P37.9})\ \frac{\boldsymbol{\mu}_{\mathbf{2}} \cdot \boldsymbol{x}_{\boldsymbol{d}}^{\mathbf{2}}}{\|\boldsymbol{x}_d\|^2} \to \frac{\tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}}{\tilde{\boldsymbol{\mu}}_{\mathbf{2}}},$$

and, hence, (**P37.7'**) can be reformulated only in terms of the squared norm of $\boldsymbol{x}_d$:

$$(\mathbf{P37.7'})\ \|\boldsymbol{x}_d - \boldsymbol{Y}_d\|^2 \simeq \mathcal{N}\left(\|\boldsymbol{x}_d\|^2 + d\tilde{\boldsymbol{\mu}}_{\mathbf{2}},\ d(\tilde{\boldsymbol{\mu}}_{\mathbf{4}} - \tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}) + 4\tfrac{\tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}}{\tilde{\boldsymbol{\mu}}_{\mathbf{2}}}\|\boldsymbol{x}_d\|^2\right).$$

Expressions of Sections 4.2 and 4.3 can be obtained by exploiting the above ones and by following the same line of reasoning. For completeness, we report the final expression of the number of k-occurrences:

$$(\mathbf{P37.10})\ N_{\varrho}^{\infty}(z) = \Phi\left(\frac{\Phi^{-1}(\varrho)\sqrt{\tilde{\boldsymbol{\mu}}_{\mathbf{4}} + 3\tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}} - z\sqrt{\tilde{\boldsymbol{\mu}}_{\mathbf{4}} - \tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}}}{2\sqrt{\tilde{\boldsymbol{\mu}}_{\mathbf{2}}^{\mathbf{2}}}}\right).$$

■

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In *Proceedings of the International Conference on Database Theory (ICDT)*, pages 420–434, London, UK, 4-6 January 2001.

Alexandr Andoni. *NN search : the old, the new, and the impossible.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.

Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 459–468, Berkeley, California, USA, 21-24 October 2006.

Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117–122, 2008.

Fabrizio Angiulli. Concentration free outlier detection. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 3–19, Skopje, Macedonia, 18-22 September 2017.

Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

Robert B. Ash and Catherine A. Doléans-Dade. *Probability & Measure Theory*. Academic Press, New York, NY, USA, 1999.

Jean-Julien Aucouturier and Francois Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern Recognition*, 41(1):272–284, 2007.

Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge University Press, England, 2016.

Richard E. Bellmann. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, New Jersey, USA, 1961.

Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Proceedings of the International Conference on Database Theory (ICDT)*, pages 217–235, Jerusalem, Israel, 10-12 January 1999.

Ginestra Bianconi and Albert-László Barabási. Competition and multiscaling in evolving networks. *Europhysics Letters*, 54, 2001.

Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

Otfried Cheong, Antoine Vigneron, and Juyoung Yon. Reverse nearest neighbor queries in fixed dimension. *Internatinal Journal of Computational Geometry & Applications*, 21(2): 179–188, 2011.

Belur V. Dasarathy. *Nearest neighbor (NN) norms: NN pattern classification techniques*. IEEE Computer Society Press, Los Alamitos, CA, USA, 1990.

Pierre Demartines. *Analyse de Données par Réseaux de Neurones Auto-Organisés*. PhD thesis, Institut National Polytechnique de Grenoble, France, 1994.

George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 1351–1354, 1998.

Richar O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, New York, NY, USA, 2000.

Robert J. Durrant and Ata Kabán. When is 'nearest neighbour' meaningful: A converse theorem and implications. *Journal of Complexity*, 25(4):385–397, 2009.

Paul Erdős and Alfrèd Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York, 3rd edition, 1971.

Damien François, Vincent Wertz, and Michel Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007.

Chris Giannella. New instability results for high-dimensional nearest neighbor search. *Information Processing Letters*, 109(19):1109–1113, 2009.

Daniele Granata and Vincenzo Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific Reports*, 6, 2016.

Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D Nonlinear Phenomena*, 9:189–208, 1983.

Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012.

Ville Hautamäki, Ismo Kärkkäinen, and Pasi Fränti. Outlier detection using k-nearest neighbour graph. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 430–433, Cambridge, UK, 23-26 August 2004.

Junfeng He, Sanjiv Kumar, and Shih-Fu Chang. On the difficulty of nearest neighbor search. In *Proceedings of the International Conference on Machine Learning, (ICML)*, Edinburgh, Scotland, 26 June-1 July 2012.

Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. What is the nearest neighbor in high dimensional spaces? In *Proceedings of International Conference on Very Large Data Bases (VLDB)*, pages 506–515, Cairo, Egypt, 10-14 September 2000.

Chih-Ming Hsu and Ming-Syan Chen. On the design and applicability of distance functions in high-dimensional data space. *IEEE Transactions on Knowledge and Data Engineering*, 21(4):523–536, 2009.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the ACM Symposium on the Theory of Computing (STOC)*, pages 604–613, Dallas, Texas, USA, 23-26 May 1998.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Wiley, 1994.

Ata Kabán. Non-parametric detection of meaningless distances in high dimensional data. *Statistics and Computing*, 22(2):375–385, 2012.

Flip Korn and S. Muthukrishnan. Influence sets based on reverse nearest neighbor queries. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 201–212, Dallas, Texas, USA, 16-18 May 2000.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

Thomas Low, Christian Borgelt, Sebastian Stober, and Andreas Nürnberger. The hubness phenomenon: Fact or artifact? In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, pages 267–278. Springer, Berlin, Heidelberg, 2013.

Laurence T. Maloney. Nearest neighbor analysis of point processes: Simulations and evaluations. *Journal of Mathematical Psychology*, 27(3):251–260, 1983.

Charles M. Newman and Yosef Rinott. Nearest neighbors and Voronoi volumes in high-dimensional point processes with various distance functions. *Advances in Applied Probability*, 17(4):794–809, 1985.

Charles M. Newman, Yosef Rinott, and Amos Tversky. Nearest neighbors and Voronoi regions in certain point processes. *Advances in Applied Probability*, 15(4):726–751, 1983.

Luca Oberto and Francesca Pennecchi. Estimation of the modulus of a complex-valued quantity. *Metrologia*, 43(6):531–538, 2006.

Donald B. Owen. A table of normal integrals. *Communications in Statistics: Simulation and Computation*, 89:389–419, 1980.

Vladimir Pestov. On the geometry of similarity search: Dimensionality curse and concentration of measure. *Information Processing Letters*, 73(1-2):47–51, 2000.

Franco P. Preparata and Michael I. Shamos. *Computational Geometry: An Introduction.* Springer-Verlag New York, Inc., New York, NY, USA, 1985.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Nearest neighbors in high-dimensional data: the emergence and influence of hubs. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 865–872, Montreal, Quebec, Canada, 14-18 June 2009.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11: 2487–2531, 2010.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Reverse nearest neighbors in unsupervised distance-based outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1369–1382, 2015.

Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk, editors. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice.* MIT Press, March 2006.

Amit Singh, Hakan Ferhatosmanoglu, and Ali Saman Tosun. High dimensional reverse nearest neighbor queries. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 91–98, New Orleans, Louisiana, USA, 2-8 November 2003.

Yufei Tao, Dimitris Papadias, Xiang Lian, and Xiaokui Xiao. Multidimensional reverse k NN search. *The VLDB Journal*, 16(3):293–316, 2007.

Nenad Tomasev. Taming the empirical hubness risk in many dimensions. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 891–899, Vancouver, BC, Canada, 30 April-2 May 2015.

Nenad Tomasev and Krisztian Buza. Hubness-aware kNN classification of high-dimensional data in presence of label noise. *Neurocomputing*, 160:157–172, 2015.

Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic. The role of hubness in clustering high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):739–751, 2014.

Amos Tversky and John Wesley Hutchinson. Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(1):3–22, 1986.

Amos Tversky, Yosef Rinott, and Charles M. Newman. Nearest neighbor analysis of point processes: Applications to multidimensional scaling. *Journal of Mathematical Psychology*, 27(3):235–250, 1983.

Laurens van der Maaten, Eric Postma, and Jaapvan den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg University, The Netherlands, 2009.

Roger Weber, Hans-Jörg Schek, and Stephen Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 194–205, New York, NY, USA, 24-27 August 1998.

Graham J. Williams, Rohan A. Baxter, Hongxing He, Simon Hawkins, and Lifang Gu. A comparative study of RNN for outlier detection in data mining. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 709–712, Maebashi City, Japan, 9-12 December 2002.

Shiyu Yang, Muhammad A. Cheema, Xuemin Lin, and Wei Wang. Reverse k nearest neighbors query processing: Experiments and analysis. *Proceedings of the VLDB Endowment (PVLDB)*, 8(5):605–616, 2015.

Yi-Ching Yao and Gordon Simons. A large-dimensional independent and identically distributed property for nearest neighbor counts in poisson processes. *Annals of Applied Probability*, 6(2):561–571, 1996.