Journal of Machine Learning Research 11 (2010) 2487-2531 Submitted 12/09; Revised 6/10; Published 9/10

Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data*

Miloš Radovanović RADACHA@DMI.UNS.AC.RS
Department of Mathematics and Informatics
University of Novi Sad
Trg D. Obradovića 4, 21000 Novi Sad, Serbia

Alexandros Nanopoulos NANOPOULOS@ISMLL.DE
Institute of Computer Science
University of Hildesheim
Marienburger Platz 22, D-31141 Hildesheim, Germany

Mirjana Ivanović MIRA@DMI.UNS.AC.RS
Department of Mathematics and Informatics
University of Novi Sad
Trg D. Obradovića 4, 21000 Novi Sad, Serbia

Editor: Ulrike von Luxburg

Abstract

Different aspects of the curse of dimensionality are known to present serious challenges to various machine-learning methods and tasks. This paper explores a new aspect of the dimensionality curse, referred to as *hubness*, that affects the distribution of k-occurrences: the number of times a point appears among the k nearest neighbors of other points in a data set. Through theoretical and empirical analysis involving synthetic and real data sets we show that under commonly used assumptions this distribution becomes considerably skewed as dimensionality increases, causing the emergence of *hubs*, that is, points with very high k-occurrences which effectively represent "popular" nearest neighbors. We examine the origins of this phenomenon, showing that it is an inherent property of data distributions in high-dimensional vector space, discuss its interaction with dimensionality reduction, and explore its influence on a wide range of machine-learning tasks directly or indirectly based on measuring distances, belonging to supervised, semi-supervised, and unsupervised learning families.

Keywords: nearest neighbors, curse of dimensionality, classification, semi-supervised learning, clustering

1. Introduction

The curse of dimensionality, a term originally introduced by Bellman (1961), is nowadays commonly used in many fields to refer to challenges posed by high dimensionality of data space. In the field of machine learning, affected methods and tasks include Bayesian modeling (Bishop, 2006), nearest-neighbor prediction (Hastie et al., 2009) and search (Korn et al., 2001), neural networks (Bishop, 1996), and many others. One aspect of the dimensionality curse is *distance concentration*, which denotes the tendency of distances between all pairs of points in high-dimensional data to become almost equal. Distance concentration and the meaningfulness of nearest neighbors in high

*. A preliminary version of this paper appeared in the Proceedings of the 26th International Conference on Machine Learning (Radovanović et al., 2009).

dimensions has been thoroughly explored (Beyer et al., 1999; Hinneburg et al., 2000; Aggarwal et al., 2001; François et al., 2007). The effect of the phenomenon on machine learning was demonstrated, for example, in studies of the behavior of kernels in the context of support vector machines, lazy learning, and radial basis function networks (Evangelista et al., 2006; François, 2007).

There exists another aspect of the curse of dimensionality that is related to nearest neighbors (NNs), which we will refer to as *hubness*. Let $D \subset \mathbb{R}^d$ be a set of d-dimensional points and $N_k(\mathbf{x})$ the number of *k-occurrences* of each point $\mathbf{x} \in D$, that is, the number of times $\mathbf{x}$ occurs among the k nearest neighbors of all other points in D, according to some distance measure. Under widely applicable conditions, as dimensionality increases, the distribution of N_k becomes considerably skewed to the right, resulting in the emergence of *hubs*, that is, points which appear in many more k-NN lists than other points, effectively making them "popular" nearest neighbors. Unlike distance concentration, hubness and its influence on machine learning have not been explored in depth. In this paper we study the causes and implications of this aspect of the dimensionality curse.

As will be described in Section 4, the phenomena of distance concentration and hubness are related, but distinct. Traditionally, distance concentration is studied through asymptotic behavior of norms, that is, distances to the origin, with increasing dimensionality. The obtained results trivially extend to reference points other than the origin, and to pairwise distances between all points. However, the asymptotic tendencies of distances of all points to different reference points do not necessarily occur at the same *speed*, which will be shown for normally distributed data by our main theoretical result outlined in Section 4.2, and given with full details in Section 5.1. The main consequence of the analysis, which is further discussed in Section 5.2 and supported by theoretical results by Newman et al. (1983) and Newman and Rinott (1985), is that the hubness phenomenon is an inherent property of data distributions in high-dimensional space under widely used assumptions, and not an artefact of a finite sample or specific properties of a particular data set.

The above result is relevant to machine learning because many families of ML algorithms, regardless of whether they are supervised, semi-supervised, or unsupervised, directly or indirectly make use of distances between data points (and, with them, k-NN graphs) in the process of building a model. Moreover, the hubness phenomenon recently started to be observed in application fields like music retrieval (Aucouturier and Pachet, 2007), speech recognition (Doddington et al., 1998), and fingerprint identification (Hicklin et al., 2005), where it is described as a problematic situation, but little or no insight is offered into the origins of the phenomenon. In this paper we present a unifying view of the hubness phenomenon through theoretical analysis of data distributions, and empirical investigation including numerous synthetic and real data sets, explaining the origins of the phenomenon and the mechanism through which hubs emerge, discussing the role of *antihubs* (points which appear in very few, if any, k-NN lists of other points), and studying the effects on common supervised, semi-supervised, and unsupervised machine-learning algorithms.

After discussing related work in the next section, we make the following contributions. First, we demonstrate the emergence of hubness on synthetic and real data in Section 3. The following section provides a comprehensive explanation of the origins of the phenomenon, through empirical and theoretical analysis of artificial data distributions, as well as observations on a large collection of real data sets, linking hubness with the *intrinsic* dimensionality of data. Section 5 presents the details of our main theoretical result which describes the mechanism through which hubs emerge as dimensionality increases, and provides discussion and further illustration of the behavior of nearest-neighbor relations in high dimensions, connecting our findings with existing theoretical results. The role of dimensionality reduction is discussed in Section 6, adding further support to the previ-

ously established link between intrinsic dimensionality and hubness, and demonstrating that dimensionality reduction may not constitute an easy mitigation of the phenomenon. Section 7 explores the impact of hubness on common supervised, semi-supervised, and unsupervised machine-learning algorithms, showing that the information provided by hubness can be used to significantly affect the success of the generated models. Finally, Section 8 concludes the paper, and provides guidelines for future work.

2. Related Work

The hubness phenomenon has been recently observed in several application areas involving sound and image data (Aucouturier and Pachet, 2007; Doddington et al., 1998; Hicklin et al., 2005). Also, Jebara et al. (2009) briefly mention hubness in the context of graph construction for semi-supervised learning. In addition, there have been attempts to avoid the influence of hubs in 1-NN time-series classification, apparently without clear awareness about the existence of the phenomenon (Islam et al., 2008), and to account for possible skewness of the distribution of N_1 in reverse nearest-neighbor search (Singh et al., 2003),[1] where $N_k(\mathbf{x})$ denotes the number of times point $\mathbf{x}$ occurs among the k nearest neighbors of all other points in the data set. None of the mentioned papers, however, successfully analyze the causes of hubness or generalize it to other applications. One recent work that makes the connection between hubness and dimensionality is the thesis by Berenzweig (2007), who observed the hubness phenomenon in the application area of music retrieval and identified high dimensionality as a cause, but did not provide practical or theoretical support that would explain the mechanism through which high dimensionality causes hubness in music data.

The distribution of N_1 has been explicitly studied in the applied probability community (Newman et al., 1983; Maloney, 1983; Newman and Rinott, 1985; Yao and Simons, 1996), and by mathematical psychologists (Tversky et al., 1983; Tversky and Hutchinson, 1986). In the vast majority of studied settings (for example, Poisson process, d-dimensional torus), coupled with Euclidean distance, it was shown that the distribution of N_1 converges to the Poisson distribution with mean 1, as the number of points n and dimensionality d go to infinity. Moreover, from the results by Yao and Simons (1996) it immediately follows that, in the Poisson process case, the distribution of N_k converges to the Poisson distribution with mean k, for any $k \geq 1$. All these results imply that no hubness is to be expected within the settings in question. On the other hand, in the case of continuous distributions with i.i.d. components, for the following specific order of limits it was shown that $\lim_{n\to\infty}\lim_{d\to\infty}\mathrm{Var}(N_1) = \infty$, while $\lim_{n\to\infty}\lim_{d\to\infty} N_1 = 0$, in distribution (Newman et al., 1983, p. 730, Theorem 7), with a more general result provided by Newman and Rinott (1985). According to the interpretation by Tversky et al. (1983), this suggests that if the number of dimensions is large relative to the number of points one may expect a small proportion of points to become hubs. However, the importance of this finding was downplayed to a certain extent (Tversky et al., 1983; Newman and Rinott, 1985), citing empirically observed slow convergence (Maloney, 1983), with the attention of the authors shifting more towards similarity measurements obtained directly from psychological and cognitive experiments (Tversky et al., 1983; Tversky and Hutchinson, 1986) that do not involve vector-space data. In Section 5.2 we will discuss the above results in more detail, as well as their relations with our theoretical and empirical findings.

It is worth noting that in ε-neighborhood graphs, that is, graphs where two points are connected if the *distance* between them is less than a given limit ε, the hubness phenomenon does not occur.

1. Reverse nearest-neighbor queries retrieve data points that have the query point $\mathbf{q}$ as their nearest neighbor.

Settings involving randomly generated points forming ε-neighborhood graphs are typically referred to as random geometric graphs, and are discussed in detail by Penrose (2003).

Concentration of distances, a phenomenon related to hubness, was studied for general distance measures (Beyer et al., 1999; Durrant and Kabán, 2009) and specifically for Minkowski and fractional distances (Demartines, 1994; Hinneburg et al., 2000; Aggarwal et al., 2001; François et al., 2007; François, 2007; Hsu and Chen, 2009). Concentration of cosine similarity was explored by Nanopoulos et al. (2009).

In our recent work (Radovanović et al., 2009), we performed an empirical analysis of hubness, its causes, and effects on techniques for classification, clustering, and information retrieval. In this paper we extend our findings with additional theoretical and empirical insight, offering a unified view of the origins and mechanics of the phenomenon, and its significance to various machine-learning applications.

3. The Hubness Phenomenon

In Section 1 we gave a simple set-based deterministic definition of N_k. To complement this definition and introduce N_k into a probabilistic setting that will also be considered in this paper, let $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n$, be $n+1$ random vectors drawn from the same continuous probability distribution with support $\mathcal{S} \subseteq \mathbb{R}^d$, $d \in \{1, 2, \ldots\}$, and let *dist* be a distance function defined on $\mathbb{R}^d$ (not necessarily a metric). Let functions $p_{i,k}$, where $i, k \in \{1, 2, \ldots, n\}$, be defined as

$$p_{i,k}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \text{ is among the } k \text{ nearest neighbors of } \mathbf{x}_i, \text{ according to } dist, \\ 0, & \text{otherwise.} \end{cases}$$

In this setting, we define $N_k(\mathbf{x}) = \sum_{i=1}^{n} p_{i,k}(\mathbf{x})$, that is, $N_k(\mathbf{x})$ is the random number of vectors from $\mathbb{R}^d$ that have $\mathbf{x}$ included in their list of k nearest neighbors. In this section we will empirically demonstrate the emergence of hubness through increasing skewness of the distribution of N_k on synthetic (Section 3.1) and real data (Section 3.2), relating the increase of skewness with the dimensionality of data sets, and motivating the subsequent study into the origins of the phenomenon in Section 4.

3.1 A Motivating Example

We start with an illustrative experiment which demonstrates the changes in the distribution of N_k with varying dimensionality. Let us consider a random data set consisting of 10000 d-dimensional points, whose components are independently drawn from the uniform distribution in range $[0, 1]$, and the following distance functions: Euclidean (l_2), fractional $l_{0.5}$ (proposed for high-dimensional data by Aggarwal et al., 2001), and cosine. Figure 1(a–c) shows the empirically observed distributions of N_k, with $k = 5$, for (a) $d = 3$, (b) $d = 20$, and (c) $d = 100$. In the same way, Figure 1(d–f) depicts the empirically observed N_k for points randomly drawn from the i.i.d. normal distribution.

For $d = 3$ the empirical distributions of N_5 for the three distance functions (Figure 1(a, d)) are consistent with the binomial distribution. This is expected by considering k-occurrences as node in-degrees in the k-nearest neighbor digraph. For randomly distributed points in low dimensions, the degree distributions of the digraphs closely resemble the degree distribution of the Erdős-Rényi (ER) random graph model, which is is binomial and Poisson in the limit (Erdős and Rényi, 1959).

As dimensionality increases, the observed distributions of N_5 depart from the random graph model and become more skewed to the right (Figure 1(b, c), and Figure 1(e, f) for l_2 and $l_{0.5}$).

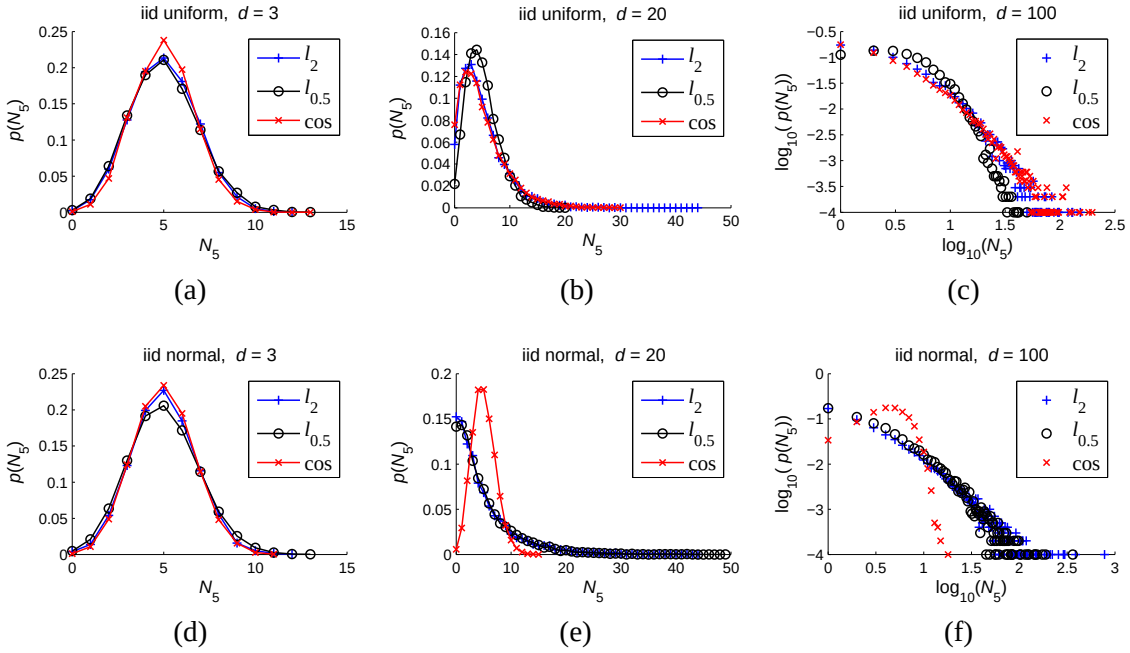


Figure 1: Empirical distribution of N_5 for Euclidean, $l_{0.5}$, and cosine distances on (a–c) i.i.d. uniform, and (d–f) i.i.d. normal random data sets with $n = 10000$ points and dimensionality (a, d) $d = 3$, (b, e) $d = 20$, and (c, f) $d = 100$ (log-log plot).

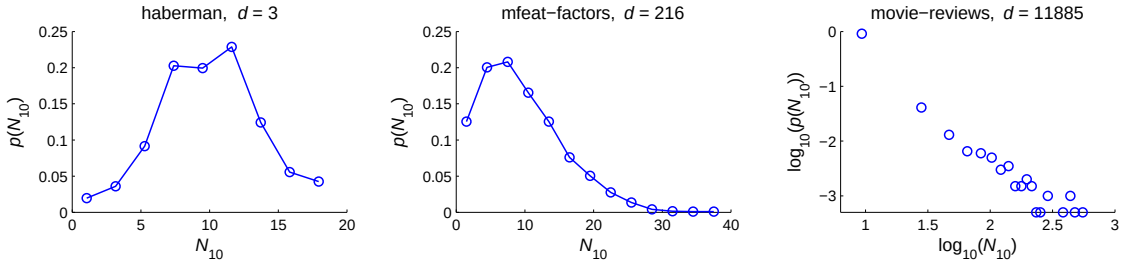
We verified this by being able to fit major right portions (that is, tails) of the observed distributions with the log-normal distribution, which is highly skewed.[2] We made similar observations with various k values (generally focusing on the common case $k \ll n$, where n is the number of points in a data set), distance measures (l_p-norm distances for both $p \geq 1$ and $0 < p < 1$, Bray-Curtis, normalized Euclidean, and Canberra), and data distributions. In virtually all these cases, skewness exists and produces hubs, that is, points with high k-occurrences. One exception visible in Figure 1 is the combination of cosine distance and normally distributed data. In most practical settings, however, such situations are not expected, and a thorough discussion of the necessary conditions for hubness to occur in high dimensions will be given in Section 5.2.

3.2 Hubness in Real Data

To illustrate the hubness phenomenon on real data, let us consider the empirical distribution of N_k ($k = 10$) for three real data sets, given in Figure 2. As in the previous section, a considerable increase in the skewness of the distributions can be observed with increasing dimensionality.

In all, we examined 50 real data sets from well known sources, belonging to three categories: UCI multidimensional data, gene expression microarray data, and textual data in the bag-of-words

2. Fits were supported by the χ^2 goodness-of-fit test at 0.05 significance level, where bins represent the number of observations of individual N_k values. These empirical distributions were compared with the expected output of a (discretized) log-normal distribution, making sure that counts in the bins do not fall below 5 by pooling the rightmost bins together.



(a) (b) (c)

Figure 2: Empirical distribution of N_{10} for three real data sets of different dimensionalities.

representation,[3] listed in Table 1. The table includes columns that describe data-set sources (2nd column), basic statistics (data transformation (3rd column): whether standardization was applied, or for textual data the used bag-of-words document representation; the number of points (n, 4th column); dimensionality (d, 5th column); the number of classes (7th column)), and the distance measure used (Euclidean or cosine, 8th column). We took care to ensure that the choice of distance measure and preprocessing (transformation) corresponds to a realistic scenario for the particular data set.

To characterize the asymmetry of N_k we use the standardized third moment of the distribution of k-occurrences,

$$S_{N_k} = \frac{\mathrm{E}(N_k - \mu_{N_k})^3}{\sigma_{N_k}^3},$$

where μ_{N_k} and σ_{N_k} are the mean and standard deviation of N_k, respectively. The corresponding (9th) column of Table 1, which shows the empirical $S_{N_{10}}$ values for the real data sets, indicates that the distributions of N_{10} for most examined data sets are skewed to the right.[4] The value of k is fixed at 10, but analogous observations can be made with other values of k.

It can be observed that some S_{N_k} values in Table 1 are quite high, indicating strong hubness in the corresponding data sets.[5] Moreover, computing the Spearman correlation between d and S_{N_k} over all 50 data sets reveals it to be strong (0.62), signifying that the relationship between dimensionality and hubness extends from synthetic to real data in general. On the other hand, careful scrutiny of the charts in Figure 2 and S_{N_k} values in Table 1 reveals that for real data the impact of dimensionality on hubness may not be as strong as could be expected after viewing hubness on synthetic data in Figure 1. Explanations for this observation will be given in the next section.

3. We used the movie review polarity data set v2.0 initially introduced by Pang and Lee (2004), while the computers and sports data sets were first used by Radovanović and Ivanović (2006). Preprocessing of all text data sets (except dexter, which is already preprocessed) involved stop-word removal and stemming using the Porter stemmer. Documents were transformed into the bag-of-words representation with word weights being either term frequencies (tf), or term frequencies multiplied by inverse document frequencies (tf-idf), with the choice based on independent experiments involving several classifiers. All term frequency vectors were normalized to average document length.
4. If $S_{N_k} = 0$ there is no skewness, positive (negative) values signify skewness to the right (left).
5. For comparison, sample skewness values for i.i.d. uniform data and Euclidean distance, shown in Figure 1(a–c), are 0.121, 1.541, and 5.445 for dimensionalities 3, 20, and 100, respectively. The values for i.i.d. normal data from Figure 1(d–f) are 0.118, 2.055, and 19.210.

Name	Src.	Trans.	n	d	d_{mle}	Cls.	Dist.	$S_{N_{10}}$	$S^S_{N_{10}}$	Clu.	$C^{N_{10}}_{\text{dm}}$	$C^{N_{10}}_{\text{cm}}$	$\widetilde{BN}_{10}$	$C^{N_{10}}_{BN_{10}}$	CAV
abalone	UCI	stan	4177	8	5.39	29	l_2	0.277	0.235	62	−0.047	−0.526	0.804	0.934	0.806
arcene	UCI	stan	100	10000	22.85	2	l_2	0.634	2.639	2	−0.559	−0.684	0.367	0.810	0.455
arrhythmia	UCI	stan	452	279	21.63	16	l_2	1.984	6.769	8	−0.867	−0.892	0.479	0.898	0.524
breast-w	UCI	stan	699	9	5.97	2	l_2	1.020	0.667	7	−0.062	−0.240	0.052	0.021	0.048
diabetes	UCI	stan	768	8	6.00	2	l_2	0.555	0.486	15	−0.479	−0.727	0.322	0.494	0.337
dorothea	UCI	none	800	100000	201.11	2	cos	2.355	1.016	19	−0.632	−0.672	0.108	0.236	0.092
echocardiogram	UCI	stan	131	7	4.92	2	l_2	0.735	0.438	5	−0.722	−0.811	0.372	0.623	0.337
ecoli	UCI	stan	336	7	4.13	8	l_2	0.116	0.208	8	−0.396	−0.792	0.223	0.245	0.193
gisette	UCI	none	6000	5000	149.35	2	cos	1.967	4.671	76	−0.667	−0.854	0.045	0.367	0.241
glass	UCI	stan	214	9	4.37	7	l_2	0.154	0.853	11	−0.430	−0.622	0.414	0.542	0.462
haberman	UCI	stan	306	3	2.89	2	l_2	0.087	−0.316	11	−0.330	−0.573	0.348	0.305	0.360
ionosphere	UCI	stan	351	34	13.57	2	l_2	1.717	2.051	18	−0.639	−0.832	0.185	0.464	0.259
iris	UCI	stan	150	4	2.96	3	l_2	0.126	−0.068	4	−0.275	−0.681	0.087	0.127	0.147
isolet1	UCI	stan	1560	617	13.72	26	l_2	1.125	6.483	38	−0.306	−0.760	0.283	0.463	0.352
mfeat-factors	UCI	stan	2000	216	8.47	10	l_2	0.826	5.493	44	−0.113	−0.688	0.063	0.001	0.145
mfeat-fourier	UCI	stan	2000	76	11.48	10	l_2	1.277	4.001	44	−0.350	−0.596	0.272	0.436	0.415
mfeat-karhunen	UCI	stan	2000	64	11.82	10	l_2	1.250	8.671	40	−0.436	−0.788	0.098	0.325	0.205
mfeat-morph	UCI	stan	2000	6	3.22	10	l_2	−0.153	0.010	44	−0.039	−0.424	0.324	0.306	0.397
mfeat-pixel	UCI	stan	2000	240	11.83	10	l_2	1.035	3.125	44	−0.210	−0.738	0.049	0.085	0.107
mfeat-zernike	UCI	stan	2000	47	7.66	10	l_2	0.933	3.389	44	−0.185	−0.657	0.235	0.252	0.400
musk1	UCI	stan	476	166	6.74	2	l_2	1.327	3.845	17	−0.376	−0.752	0.237	0.621	0.474
optdigits	UCI	stan	5620	64	9.62	10	l_2	1.095	3.789	74	−0.223	−0.601	0.044	0.097	0.168
ozone-eighthr	UCI	stan	2534	72	12.92	2	l_2	2.251	4.443	49	−0.216	−0.655	0.086	0.300	0.138
ozone-onehr	UCI	stan	2536	72	12.92	2	l_2	2.260	5.798	49	−0.215	−0.651	0.046	0.238	0.070
page-blocks	UCI	stan	5473	10	3.73	5	l_2	−0.014	0.470	72	−0.063	−0.289	0.049	−0.046	0.068
parkinsons	UCI	stan	195	22	4.36	2	l_2	0.729	1.964	8	−0.414	−0.649	0.166	0.321	0.256
pendigits	UCI	stan	10992	16	5.93	10	l_2	0.435	0.982	104	−0.062	−0.513	0.014	−0.030	0.156
segment	UCI	stan	2310	19	3.93	7	l_2	0.313	1.111	48	−0.077	−0.453	0.089	0.074	0.332
sonar	UCI	stan	208	60	9.67	2	l_2	1.354	3.053	8	−0.550	−0.771	0.286	0.632	0.461
spambase	UCI	stan	4601	57	11.45	2	l_2	1.916	2.292	49	−0.376	−0.448	0.139	0.401	0.271
spectf	UCI	stan	267	44	13.83	2	l_2	1.895	2.098	11	−0.616	−0.729	0.300	0.595	0.366
spectrometer	UCI	stan	531	100	8.04	10	l_2	0.591	3.123	17	−0.269	−0.670	0.200	0.225	0.242
vehicle	UCI	stan	846	18	5.61	4	l_2	0.603	1.625	25	−0.162	−0.643	0.358	0.435	0.586
vowel	UCI	stan	990	10	2.39	11	l_2	0.766	0.935	27	−0.252	−0.605	0.313	0.691	0.598
wdbc	UCI	stan	569	30	8.26	2	l_2	0.815	3.101	16	−0.449	−0.708	0.065	0.170	0.129
wine	UCI	stan	178	13	6.69	3	l_2	0.630	1.319	3	−0.589	−0.874	0.076	0.182	0.084
wpbc	UCI	stan	198	33	8.69	2	l_2	0.863	2.603	6	−0.688	−0.878	0.340	0.675	0.360
yeast	UCI	stan	1484	8	5.42	10	l_2	0.228	0.105	34	−0.421	−0.715	0.527	0.650	0.570
AMLALL	KR	none	72	7129	31.92	2	l_2	1.166	1.578	2	−0.868	−0.927	0.171	0.635	0.098
colonTumor	KR	none	62	2000	11.22	2	l_2	1.055	1.869	3	−0.815	−0.781	0.305	0.779	0.359
DLBCL	KR	none	47	4026	16.11	2	l_2	1.007	1.531	2	−0.942	−0.947	0.338	0.895	0.375
lungCancer	KR	none	181	12533	59.66	2	l_2	1.248	3.073	6	−0.537	−0.673	0.052	0.262	0.136
MLL	KR	none	72	12582	28.42	3	l_2	0.697	1.802	2	−0.794	−0.924	0.211	0.533	0.148
ovarian-61902	KR	none	253	15154	9.58	2	l_2	0.760	3.771	10	−0.559	−0.773	0.164	0.467	0.399
computers	dmoz	tf	697	1168	190.33	2	cos	2.061	2.267	26	−0.566	−0.731	0.312	0.699	0.415
dexter	UCI	none	300	20000	160.78	2	cos	3.977	4.639	13	−0.760	−0.781	0.301	0.688	0.423
mini-newsgroups	UCI	tf-idf	1999	7827	3226.43	20	cos	1.980	1.765	44	−0.422	−0.704	0.524	0.701	0.526
movie-reviews	PaBo	tf	2000	11885	54.95	2	cos	8.796	7.247	44	−0.604	−0.739	0.398	0.790	0.481
reuters-transcribed	UCI	tf-idf	201	3029	234.68	10	cos	1.165	1.693	3	−0.781	−0.763	0.642	0.871	0.595
sports	dmoz	tf	752	1185	250.24	2	cos	1.629	2.543	27	−0.584	−0.736	0.260	0.604	0.373

Table 1: Real data sets. Data sources are the University of California, Irvine (UCI) Machine Learning Repository, Kent Ridge (KR) Bio-Medical Data Set Repository, dmoz Open Directory, and www.cs.cornell.edu/People/pabo/movie-review-data/ (PaBo).

4. The Origins of Hubness

This section moves on to exploring the causes of hubness and the mechanisms through which hubs emerge. Section 4.1 investigates the relationship between the position of a point in data space and hubness. Next, Section 4.2 explains the mechanism through which hubs emerge as points in high-dimensional space that become closer to other points than their low-dimensional counterparts, outlining our main theoretical result. The emergence of hubness in real data is studied in Section 4.3, while Section 4.4 discusses hubs and their opposites—antihubs—and the relationships between hubs, antihubs, and different notions of outliers.

4.1 The Position of Hubs

Let us consider again the i.i.d. uniform and i.i.d. normal random data examined in the previous section. We will demonstrate that the position of a point in data space has a significant effect on its k-occurrences value, by observing the sample mean of the data distribution as a point of reference. Figure 3 plots, for each point $\mathbf{x}$, its $N_5(\mathbf{x})$ against its Euclidean distance from the empirical data mean, for $d = 3, 20, 100$. As dimensionality increases, stronger correlation emerges, implying that points closer to the mean tend to become hubs. We made analogous observations with other values of k, and combinations of data distributions and distance measures for which hubness occurs. It is important to note that proximity to one global data-set mean correlates with hubness in high dimensions when the underlying data distribution is *unimodal*. For multimodal data distributions, for example those obtained through a mixture of unimodal distributions, hubs tend to appear close to the means of individual component distributions of the mixture. In the discussion that follows in Section 4.2 we will assume a unimodal data distribution, and defer the analysis of multimodal distributions until Section 4.3, which studies real data.

4.2 Mechanisms Behind Hubness

Although one may expect that some random points are closer to the data-set mean than others, in order to explain the mechanism behind hub formation we need to (1) understand the geometrical and distributional setting in which some points tend to be closer to the mean than others, and then (2) understand why such points become hubs in higher dimensions.[6]

Hubness appears to be related to the phenomenon of distance concentration, which is usually expressed as the ratio between some measure of spread (for example, standard deviation) and some measure of magnitude (for example, the mean) of distances of all points in a data set to some arbitrary reference point (Aggarwal et al., 2001; François et al., 2007). If this ratio converges to 0 as dimensionality goes to infinity, it is said that the distances concentrate. Based on existing theoretical results discussing distance concentration (Beyer et al., 1999; Aggarwal et al., 2001), high-dimensional points are approximately lying on a hypersphere centered at the data-set mean. Moreover, the results by Demartines (1994) and François et al. (2007) specify that the distribution of distances to the data-set mean has a non-negligible variance for any finite d.[7] Hence, the existence of a non-negligible number of points closer to the data-set mean is *expected* in high dimensions.

6. We will assume that random points originate from a unimodal data distribution. In the multimodal case, it can be said that the observations which follow are applicable around one of the "peaks" in the pdf of the data distribution.

7. These results apply to l_p-norm distances, but our numerical simulations suggest that other distance functions mentioned in Section 3.1 behave similarly. Moreover, any point can be used as a reference instead of the data mean, but we observe the data mean since it plays a special role with respect to hubness.

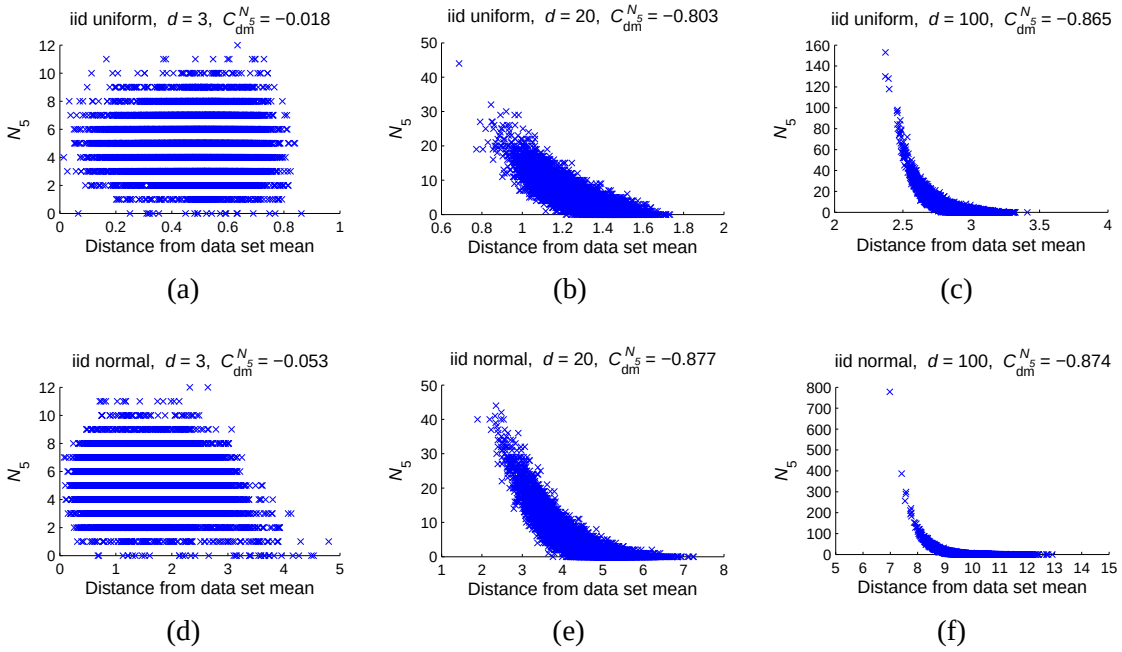


Figure 3: Scatter plots and Spearman correlation of $N_5(\mathbf{x})$ against the Euclidean distance of point $\mathbf{x}$ to the sample data-set mean for (a–c) i.i.d. uniform and (d–f) i.i.d. normal random data sets with (a, d) $d = 3$, (b, e) $d = 20$, and (c, f) $d = 100$.

To illustrate the above discussion, Figure 4 depicts, for i.i.d. normal data, the distribution of Euclidean distances of all points to the true data mean (the origin) for several d values. By definition, the distribution of distances is actually the Chi distribution with d degrees of freedom (as the square root of the sum of squares of i.i.d. normal variables, Johnson et al., 1994).[8] In this setting, distance concentration refers to the fact that the standard deviation of distance distributions is asymptotically constant with respect to increasing d, while the means of the distance distributions asymptotically behave like $\sqrt{d}$ (a direct consequence of the results by François et al., 2007, discussed further in Section 5.1). On the other hand, for l_p-norm distances with $p > 2$, the standard deviation would tend to 0 (François et al., 2007). However, for any finite d, existing variation in the values of random coordinates causes some points to become closer to the distribution mean than others. This happens despite the fact that all distance values, in general, may be increasing together with d.

To understand why points closer to the data mean become hubs in high dimensions, let us consider the following example. We observe, within the i.i.d. normal setting, two points drawn from the data, but at specific positions with respect to the origin: point $\mathbf{b}_d$ which is at the expected distance from the origin, and point $\mathbf{a}_d$ which is two standard deviations closer. In light of the above, the distances of $\mathbf{a}_d$ and $\mathbf{b}_d$ from the origin change with increasing d, and it could be said that different $\mathbf{a}_d$-s (and $\mathbf{b}_d$-s) occupy analogous positions in the data spaces, with respect to changing d. The distances of $\mathbf{a}_d$ (and $\mathbf{b}_d$) to all other points, again following directly from the definition (Oberto and Pennecchi, 2006), are distributed according to *noncentral* Chi distributions with d degrees of

8. For this reason, in Figure 4 we plot the known pdf, not the empirically obtained distribution.

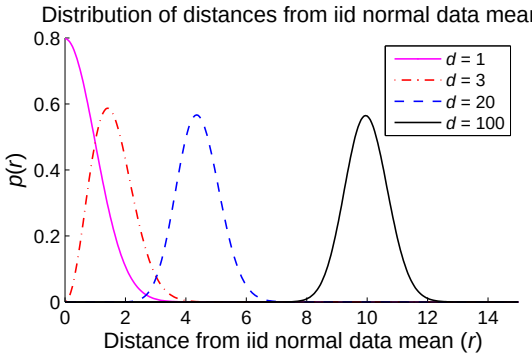
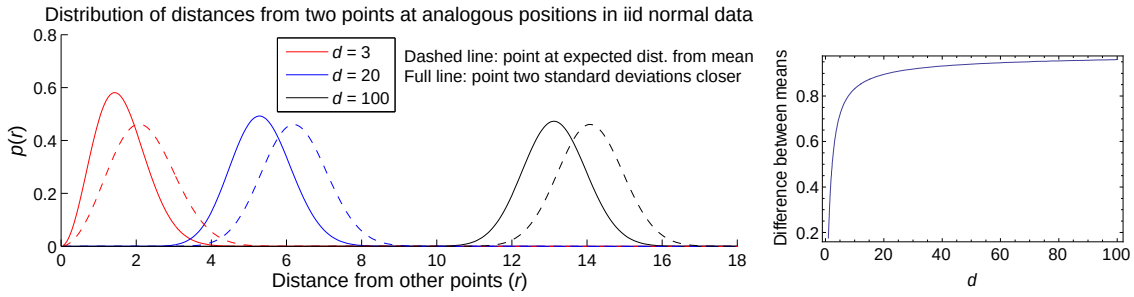


Figure 4: Probability density function of observing a point at distance r from the mean of a multivariate d-dimensional normal distribution, for $d = 1, 3, 20, 100$.



(a) (b)

Figure 5: (a) Distribution of distances to other points from i.i.d. normal random data for a point at the expected distance from the origin (dashed line), and a point two standard deviations closer (full line). (b) Difference between the means of the two distributions, with respect to increasing d.

freedom and noncentrality parameter λ equaling the distance of $\mathbf{a}_d$ ($\mathbf{b}_d$) to the origin. Figure 5(a) plots the probability density functions of these distributions for several values of d. It can be seen that, as d increases, the distance distributions for $\mathbf{a}_d$ and $\mathbf{b}_d$ move away from each other. This tendency is depicted more clearly in Figure 5(b) which plots the difference between the means of the two distributions with respect to d.

It is known, and expected, for points that are closer to the mean of the data distribution to be closer, on average, to all other points, for any value of d. However, the above analysis indicates that this tendency is amplified by high dimensionality, making points that reside in the proximity of the data mean become closer (in relative terms) 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 inclusion probability into k-NN lists of other points, even for small values of k. We will discuss this relationship further in Section 5.2.

In terms of the notion of node *centrality* typically used in network analysis (Scott, 2000), the above discussion indicates that high dimensionality amplifies what we will call the *spatial centrality*

of a point (by increasing its proximity to other points), which, in turn, affects the *degree centrality* of the corresponding node in the k-NN graph (by increasing its degree, that is, N_k). Other notions of node centrality, and the structure of the k-NN graph in general, will be studied in more detail in Section 5.2.1. The rest of this section will focus on describing the mechanism of the observed spatial centrality amplification.

In the preceding discussion we selected two points from i.i.d. normal data with specific distances from the origin expressed in terms of expected distance and deviation from it, and tracked the analogues of the two points for increasing values of dimensionality d. Generally, we can express the distances of the two points to the origin in terms of "offsets" from the expected distance measured by standard deviation, which in the case of i.i.d. normal random data would be $\lambda_{d,1} = \mu_{\chi(d)} + c_1\sigma_{\chi(d)}$ and $\lambda_{d,2} = \mu_{\chi(d)} + c_2\sigma_{\chi(d)}$, where $\lambda_{d,1}$ and $\lambda_{d,2}$ are the distances of the first and second point to the origin, $\mu_{\chi(d)}$ and $\sigma_{\chi(d)}$ are the mean and standard deviation of the Chi distribution with d degrees of freedom, and c_1 and c_2 are selected constants (the offsets). In the preceding example involving points $\mathbf{a}_d$ and $\mathbf{b}_d$, we set $c_1 = -2$ and $c_2 = 0$, respectively. However, analogous behavior can be observed with arbitrary two points whose distance to the data mean is below the expected distance, that is, for $c_1, c_2 \leq 0$. We describe this behavior by introducing the following notation: $\Delta\mu_d(\lambda_{d,1}, \lambda_{d,2}) = |\mu_{\chi(d,\lambda_{d,2})} - \mu_{\chi(d,\lambda_{d,1})}|$, where $\mu_{\chi(d,\lambda_{d,i})}$ is the mean of the noncentral Chi distribution with d degrees of freedom and noncentrality parameter $\lambda_{d,i}$ ($i \in \{1,2\}$). In the following theorem, which we prove in Section 5.1, we show that $\Delta\mu_d(\lambda_{d,1}, \lambda_{d,2})$ increases with increasing values of d.

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 $d \in \mathbb{N}^+$, $c_1, c_2 \leq 0$, $c_1 < c_2$, 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*

$$\Delta\mu_d(\lambda_{d,1}, \lambda_{d,2}) = \mu_{\chi(d,\lambda_{d,2})} - \mu_{\chi(d,\lambda_{d,1})}\,,$$

where $\mu_{\chi(d,\lambda_{d,i})}$ is the mean of the noncentral Chi distribution with d degrees of freedom and noncentrality parameter $\lambda_{d,i}$ ($i \in \{1,2\}$).

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

$$\Delta\mu_{d+2}(\lambda_{d+2,1}, \lambda_{d+2,2}) > \Delta\mu_d(\lambda_{d,1}, \lambda_{d,2})\,. \tag{1}$$

The main statement of the theorem is given by Equation 1, which expresses the tendency of the difference between the means of the two distance distributions to increase with increasing dimensionality d. It is important to note that this tendency is obtained through analysis of *distributions* of data and distances, implying that the behavior is an inherent property of data distributions in high-dimensional space, rather than an artefact of other factors, such as finite sample size, etc. Through simulation involving randomly generated points we verified the behavior for i.i.d. normal data by replicating very closely the chart shown in Figure 5(b). Furthermore, simulations suggest that the same behavior emerges in i.i.d. uniform data,[9] as well as numerous other unimodal random data distributions, producing charts of the same shape as in Figure 5(b). Real data, on the other hand,

9. The uniform cube setting will be discussed in more detail in Section 5.2, in the context of results from related work (Newman and Rinott, 1985).

tends to be clustered, and can be viewed as originating from a *mixture* of distributions resulting in a multimodal distribution of data. In this case, the behavior described by Theorem 1, and illustrated in Figure 5(b), is manifested primarily on the individual component distributions of the mixture, that is, on clusters of data points. The next section takes a closer look at the hubness phenomenon in real data sets.

4.3 Hubness in Real Data

Results describing the origins of hubness given in the previous sections were obtained by examining data sets that follow specific distributions and generated as i.i.d. samples from these distributions. To extend these results to real data, we need to take into account two additional factors: (1) real data sets usually contain dependent attributes, and (2) real data sets are usually clustered, that is, points are organized into groups produced by a mixture of distributions instead of originating from a single (unimodal) distribution.

To examine the first factor (dependent attributes), we adopt the approach of François et al. (2007) used in the context of distance concentration. For each data set we randomly permute the elements within every attribute. This way, attributes preserve their individual distributions, but the dependencies between them are lost and the *intrinsic dimensionality* of data sets increases, becoming equal to their embedding dimensionality d (François et al., 2007). In Table 1 (10th column) we give the empirical skewness, denoted as $S^S_{N_k}$, of the shuffled data. For the vast majority of high-dimensional data sets, $S^S_{N_k}$ is considerably higher than S_{N_k}, indicating that hubness actually depends on the intrinsic rather than embedding dimensionality. This provides an explanation for the apparent weaker influence of d on hubness in real data than in synthetic data sets, which was observed in Section 3.2.

To examine the second factor (many groups), for every data set we measured: (i) the Spearman correlation, denoted as $C^{N_{10}}_{\mathrm{dm}}$ (12th column), of the observed N_k and distance from the data-set mean, and (ii) the correlation, denoted as $C^{N_{10}}_{\mathrm{cm}}$ (13th column), of the observed N_k and distance to the closest group mean. Groups are determined with K-means clustering, where the number of clusters for each data set, given in column 11 of Table 1, was determined by exhaustive search of values between 2 and $\lfloor\sqrt{n}\rfloor$, to maximize $C^{N_{10}}_{\mathrm{cm}}$.[10] In most cases, $C^{N_{10}}_{\mathrm{cm}}$ is considerably stronger than $C^{N_{10}}_{\mathrm{dm}}$. Consequently, in real data, hubs tend to be closer than other points to their respective cluster centers (which we verified by examining the individual scatter plots).

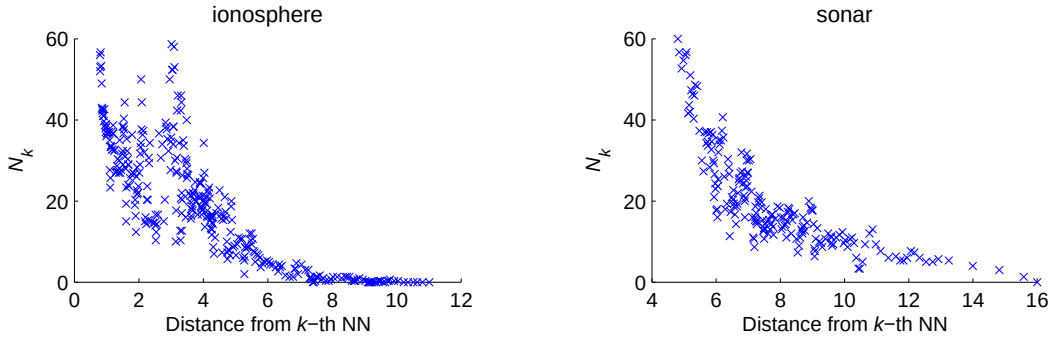
To further support the above findings, we include the 6th column (d_{mle}) to Table 1, corresponding to intrinsic dimensionality measured by the maximum likelihood estimator (Levina and Bickel, 2005). Next, we compute Spearman correlations between various measurements from Table 1 over all 50 examined data sets, given in Table 2. The observed skewness of N_k, besides being strongly correlated with d, is even more strongly correlated with the intrinsic dimensionality d_{mle}. Moreover, intrinsic dimensionality positively affects the correlations between N_k and the distance to the data-set mean / closest cluster mean, implying that in higher (intrinsic) dimensions the positions of hubs become increasingly localized to the proximity of centers.

Section 6, which discusses the interaction of hubness with dimensionality reduction, will provide even more support to the observation that hubness depends on intrinsic, rather than embedding dimensionality.

10. We report averages of $C^{N_{10}}_{\mathrm{cm}}$ over 10 runs of K-means clustering with different random seeding, in order to reduce the effects of chance.

	d	d_{mle}	$S_{N_{10}}$	$C_{\text{dm}}^{N_{10}}$	$C_{\text{cm}}^{N_{10}}$	$\widetilde{BN}_{10}$	$C_{BN_{10}}^{N_{10}}$
d_{mle}	0.87						
$S_{N_{10}}$	0.62	0.80					
$C_{\text{dm}}^{N_{10}}$	−0.52	−0.60	−0.42				
$C_{\text{cm}}^{N_{10}}$	−0.43	−0.48	−0.31	0.82			
$\widetilde{BN}_{10}$	−0.05	0.03	−0.08	−0.32	−0.18		
$C_{BN_{10}}^{N_{10}}$	0.32	0.39	0.29	−0.61	−0.46	0.82	
CAV	0.03	0.03	0.03	−0.14	−0.05	0.85	0.76

Table 2: Spearman correlations over 50 real data sets.



(a) (b)

Figure 6: Correlation between low N_k and outlier score ($k = 20$).

4.4 Hubs and Outliers

The non-negligible variance of the distribution of distances to the data mean described in Section 4.2 has an additional "side": we also expect points farther from the mean and, therefore, with much lower observed N_k than the rest.[11] Such points correspond to the bottom-right parts of Figure 3(b, c, e, f), and will be referred to as *antihubs*. Since antihubs are far away from all other points, in high dimensions they can be regarded as distance-based *outliers* (Tan et al., 2005).

To further support the connection between antihubs and distance-based outliers, let us consider a common outlier score of a point as the distance from its kth nearest neighbor (Tan et al., 2005). Low N_k values and high outlier scores are correlated as exemplified in Figure 6(a, b) (in their lower-right parts) for two data sets from Table 1.

Next, let us recall the i.i.d. normal random data setting, and the probability density function corresponding to observing a point at a specified distance from the mean, plotted in Figure 4. An analogous chart for real data is given in Figure 7, which shows the empirical distributions of distances from the closest cluster mean for three real data sets, as described in Section 4.3. In both figures it can be seen that in low dimensions the probability of observing a point near a center is quite high, while as dimensionality increases it becomes close to zero. If we now consider a *probabilistic* definition of an outlier as a point with a low probability of occurrence (Tan et al., 2005), in high

11. Assuming the presence of hubs, the existence of points with low N_k is implied by the constant-sum property of N_k: for any data set D, $\sum_{\mathbf{x}\in D} N_k(\mathbf{x}) = k|D|$ (Aucouturier and Pachet, 2007).

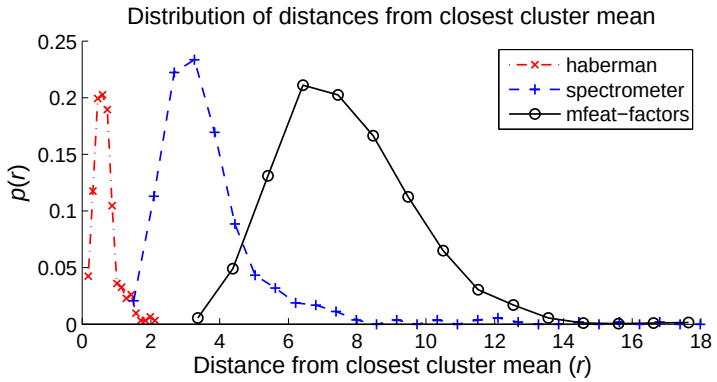


Figure 7: Probability density function of observing a point at distance r from the closest cluster mean for three real data sets.

dimensions hubs actually *are* outliers, as points closer to the distribution (or cluster) mean than the majority of other points. Therefore, somewhat counterintuitively, it can be said that hubs are points that reside in low-density regions of the high-dimensional data space, and are at the same time close to many other points. On the other hand, distance-based outliers correspond to probabilistic outliers that are farther away from the center(s). It follows that the definitions of distance-based and probabilistic outliers significantly diverge from one another in high dimensions, with distance-based outliers only partially corresponding to probabilistic outliers. To prevent confusion in the rest of the paper, we shall continue to refer to "hubs" and "outliers" in the distance-based sense. Outliers will be analyzed further in Section 7.3.2.

5. Proofs and Discussion

This section is predominantly devoted to the theoretical aspects of the behavior of distances in high-dimensional space and the hubness phenomenon. Section 5.1 provides a step-by-step proof of Theorem 1 which was introduced in Section 4.2, while Section 5.2 discusses additional aspects of the phenomenon in the context of the geometry of high-dimensional spaces and the properties of data distributions which occupy them.

5.1 Proof of Theorem 1

In this section we analyze the behavior of distances that provides the mechanism for the formation of hubs, introduced in Section 4.2, culminating with the proof of Theorem 1. Section 5.1.1 reviews distance concentration results by François et al. (2007), while Section 5.1.2 discusses distance distributions in i.i.d. normal random data and extended interpretations of the results by François et al. (2007) in this setting. The notion of asymptotic equivalence that will be used in the proof of Theorem 1 is the subject of Section 5.1.3. The expectation of the noncentral Chi distribution, which is a key feature in the analysis of distance distributions in i.i.d. normal random data, is defined in Section 5.1.4, together with the generalized Laguerre function on which it relies. Properties of the generalized Laguerre function that will be used in the proof of Theorem 1 are presented in Section 5.1.5. Finally, the proof of Theorem 1 is given in Section 5.1.6.

5.1.1 DISTANCE CONCENTRATION RESULTS

We begin by reviewing the main results of François et al. (2007) regarding distance concentration. Let $\mathbf{X}_d = (X_1, X_2, \ldots, X_d)$ be a random d-dimensional variable with i.i.d. components: $X_i \sim \mathcal{F}$, and let $\|\mathbf{X}_d\|$ denote its Euclidean norm. For random variables $|X_i|^2$, let $\mu_{|\mathcal{F}|^2}$ and $\sigma^2_{|\mathcal{F}|^2}$ signify their mean and variance, respectively. François et al. (2007) prove the following two lemmas.[12]

Lemma 2 (François et al., 2007, Equation 17, adapted)

$$\lim_{d\to\infty} \frac{\mathrm{E}(\|\mathbf{X}_d\|)}{\sqrt{d}} = \mu_{|\mathcal{F}|^2}\,.$$

Lemma 3 (François et al., 2007, Equation 21, adapted)

$$\lim_{d\to\infty} \mathrm{Var}(\|\mathbf{X}_d\|) = \frac{\sigma^2_{|\mathcal{F}|^2}}{4\mu_{|\mathcal{F}|^2}}\,.$$

The above lemmas imply that, for i.i.d. random data, the expectation of the distribution of Euclidean distances to the origin (Euclidean norms) asymptotically behaves like $\sqrt{d}$, while the standard deviation is asymptotically constant. From now on, we will denote the mean and variance of random variables that are distributed according to some distribution $\mathcal{F}$ by $\mu_{\mathcal{F}}$ and $\sigma^2_{\mathcal{F}}$, respectively.

5.1.2 DISTANCES IN I.I.D. NORMAL DATA

We now observe more closely the behavior of distances in i.i.d. normal random data. Let $\mathbf{Z}_d = (Z_1, Z_2, \ldots, Z_d)$ be a random d-dimensional vector whose components independently follow the standard normal distribution: $Z_i \sim \mathcal{N}(0;1)$, for every $i \in \{1, 2, \ldots, d\}$. Then, by definition, random variable $\|\mathbf{Z}_d\|$ follows the Chi distribution with d degrees of freedom: $\|\mathbf{Z}_d\| \sim \chi(d)$. In other words, $\chi(d)$ is the distribution of Euclidean distances of vectors drawn from $\mathbf{Z}_d$ to the origin. If one were to fix another reference vector $\mathbf{x}_d$ instead of the origin, the distribution of distances of vectors drawn from $\mathbf{Z}_d$ to $\mathbf{x}_d$ would be completely determined by $\|\mathbf{x}_d\|$ since, again by definition, random variable $\|\mathbf{Z}_d - \mathbf{x}_d\|$ follows the noncentral Chi distribution with d degrees of freedom and noncentrality parameter $\lambda = \|\mathbf{x}_d\|$: $\|\mathbf{Z}_d - \mathbf{x}_d\| \sim \chi(d, \|\mathbf{x}_d\|)$.

In light of the above, let us observe two points, $\mathbf{x}_{d,1}$ and $\mathbf{x}_{d,2}$, drawn from $\mathbf{Z}_d$, and express their distances from the origin in terms of offsets from the expected distance, with the offsets described using standard deviations: $\|\mathbf{x}_{d,1}\| = \mu_{\chi(d)} + c_1\sigma_{\chi(d)}$ and $\|\mathbf{x}_{d,2}\| = \mu_{\chi(d)} + c_2\sigma_{\chi(d)}$, where $c_1, c_2 \leq 0$. We will assume $c_1 < c_2$, that is, $\mathbf{x}_{d,1}$ is closer to the data distribution mean (the origin) than $\mathbf{x}_{d,2}$. By treating c_1 and c_2 as constants, and varying d, we observe analogues of two points in spaces of different dimensionalities (roughly speaking, points $\mathbf{x}_{d,1}$ have identical "probability" of occurrence at the specified distance from the origin for every d, and the same holds for $\mathbf{x}_{d,2}$). Let $\lambda_{d,1} = \|\mathbf{x}_{d,1}\|$ and $\lambda_{d,2} = \|\mathbf{x}_{d,2}\|$. Then, the distributions of distances of points $\mathbf{x}_{d,1}$ and $\mathbf{x}_{d,2}$ to all points from the data distribution $\mathbf{Z}_d$ (that is, the distributions of random variables $\|\mathbf{Z}_d - \mathbf{x}_{d,1}\|$ and $\|\mathbf{Z}_d - \mathbf{x}_{d,2}\|$) are noncentral Chi distributions $\chi(d, \lambda_{d,1})$ and $\chi(d, \lambda_{d,2})$, respectively. We study the behavior of these two distributions with increasing values of d.

Lemmas 2 and 3, taking $\mathcal{F}$ to be the standard normal distribution $\mathcal{N}(0;1)$, and translating the space so that $\mathbf{x}_{d,1}$ or $\mathbf{x}_{d,2}$ become the origin, imply that both $\mu_{\chi(d,\lambda_{d,1})}$ and $\mu_{\chi(d,\lambda_{d,2})}$ asymptotically

12. François et al. (2007) provide a more general result for l_p norms with arbitrary $p > 0$.

behave like $\sqrt{d}$ as $d \to \infty$, while $\sigma^2_{\chi(d,\lambda_{d,1})}$ and $\sigma^2_{\chi(d,\lambda_{d,2})}$ are both asymptotically constant.[13] However, for $\mathbf{x}_{d,1}$ and $\mathbf{x}_{d,2}$ placed at different distances form the origin ($\lambda_{d,1} \neq \lambda_{d,2}$, that is, $c_1 \neq c_2$), these asymptotic tendencies do not occur at the same *speed*. In particular, we will show that as d increases, the difference between $\mu_{\chi(d,\lambda_{d,1})}$ and $\mu_{\chi(d,\lambda_{d,2})}$ actually *increases*. If we take $\mathbf{x}_{d,1}$ to be closer to the origin than $\mathbf{x}_{d,2}$ ($c_1 < c_2$), this means that $\mathbf{x}_{d,1}$ becomes closer to all other points from the data distribution $\mathbf{Z}_d$ than $\mathbf{x}_{d,2}$, simply by virtue of increasing dimensionality, since for different values of d we place the two points at analogous positions in the data space with regards to the distance from the origin.

5.1.3 ASYMPTOTIC EQUIVALENCE

Before describing our main theoretical result, we present several definitions and lemmas, beginning with the notion of asymptotic equivalence that will be relied upon.

Definition 4 *Two real-valued functions $f(x)$ and $g(x)$ are* asymptotically equal, *$f(x) \approx g(x)$, iff for every $\varepsilon > 0$ there exists $x_0 \in \mathbb{R}$ such that for every $x > x_0$, $|f(x) - g(x)| < \varepsilon$.*

Equivalently, $f(x) \approx g(x)$ iff $\lim_{x\to\infty} |f(x) - g(x)| = 0$. Note that the $\approx$ relation is different from the divisive notion of asymptotic equivalence, where $f(x) \sim g(x)$ iff $\lim_{x\to\infty} f(x)/g(x) = 1$.

The following two lemmas describe approximations that will be used in the proof of Theorem 1, based on the $\approx$ relation.

Lemma 5 *For any constant $c \in \mathbb{R}$, let $f(d) = \sqrt{d+c}$, and $g(d) = \sqrt{d}$. Then, $f(d) \approx g(d)$.*

Proof

$$\lim_{d\to\infty} \left|\sqrt{d+c} - \sqrt{d}\right| = \lim_{d\to\infty} \left|\left(\sqrt{d+c} - \sqrt{d}\right)\frac{\sqrt{d+c}+\sqrt{d}}{\sqrt{d+c}+\sqrt{d}}\right| = \lim_{d\to\infty} \left|\frac{c}{\sqrt{d+c}+\sqrt{d}}\right| = 0. \qquad \blacksquare$$

Lemma 6 *$\mu_{\chi(d)} \approx \sqrt{d}$, and $\sigma_{\chi(d)} \approx 1/\sqrt{2}$.*

Proof Observe the expression for the mean of the $\chi(d)$ distribution,

$$\mu_{\chi(d)} = \sqrt{2}\,\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}.$$

The equality $x\Gamma(x) = \Gamma(x+1)$ and the convexity of $\log\Gamma(x)$ yield (Haagerup, 1982, p. 237):

$$\Gamma\left(\frac{d+1}{2}\right)^2 \leq \Gamma\left(\frac{d}{2}\right)\Gamma\left(\frac{d+2}{2}\right) = \frac{d}{2}\Gamma\left(\frac{d}{2}\right)^2,$$

13. More precisely, the lemmas can be applied only for points $\mathbf{x}'_{d,i}$ that have equal values of all components, since after translation data components need to be i.i.d. Because of the symmetry of the Gaussian distribution, the same expectations and variances of distance distributions are obtained, for every d, with any $\mathbf{x}_{d,i}$ that has the same norm as $\mathbf{x}'_{d,i}$, thereby producing identical asymptotic results.

and

$$\Gamma\left(\frac{d+1}{2}\right)^2 = \frac{d-1}{2}\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{d+1}{2}\right) \geq \frac{d-1}{2}\Gamma\left(\frac{d}{2}\right)^2,$$

from which we have

$$\sqrt{d-1} \leq \sqrt{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \leq \sqrt{d}.$$

From Lemma 5 it now follows that $\mu_{\chi(d)} \approx \sqrt{d}$.

Regarding the standard deviation of the $\chi(d)$ distribution, from Lemma 3, taking $\mathcal{F}$ to be the standard normal distribution, we obtain

$$\lim_{d\to\infty} \sigma^2_{\chi(d)} = \frac{\sigma^2_{\chi^2(1)}}{4\mu_{\chi^2(1)}} = \frac{1}{2},$$

since the square of a standard normal random variable follows the chi-square distribution with one degree of freedom, $\chi^2(1)$, whose mean and variance are known: $\mu_{\chi^2(1)} = 1$, $\sigma^2_{\chi^2(1)} = 2$. It now directly follows that $\sigma_{\chi(d)} \approx 1/\sqrt{2}$. ■

5.1.4 Expectation of the Noncentral Chi Distribution

The central notion in Theorem 1 is the noncentral Chi distribution. To express the expectation of the noncentral Chi distribution, the following two definitions are needed, introducing the Kummer confluent hypergeometric function ${}_1F_1$, and the generalized Laguerre function.

Definition 7 (Itô, 1993, p. 1799, Appendix A, Table 19.I)
For $a, b, z \in \mathbb{R}$, the Kummer confluent hypergeometric function ${}_1F_1(a;b;z)$ is given by

$${}_1F_1(a;b;z) = \sum_{k=0}^{\infty} \frac{(a)_k}{(b)_k} \cdot \frac{z^k}{\Gamma(k+1)},$$

where $(\cdot)_k$ is the Pochhammer symbol, $(x)_k = \frac{\Gamma(x+k)}{\Gamma(x)}$.

Definition 8 (Itô, 1993, p. 1811, Appendix A, Table 20.VI)
For $\nu, \alpha, z \in \mathbb{R}$, the generalized Laguerre function $L_\nu^{(\alpha)}(z)$ is given by

$$L_\nu^{(\alpha)}(z) = \frac{\Gamma(\nu+\alpha+1)}{\Gamma(\nu+1)} \cdot \frac{{}_1F_1(-\nu;\alpha+1;z)}{\Gamma(\alpha+1)}.$$

The expectation of the noncentral Chi distribution with d degrees of freedom and noncentrality parameter λ, denoted by $\mu_{\chi(d,\lambda)}$, can now be expressed via the generalized Laguerre function (Oberto and Pennecchi, 2006):

$$\mu_{\chi(d,\lambda)} = \sqrt{\frac{\pi}{2}} L_{1/2}^{(d/2-1)}\left(-\frac{\lambda^2}{2}\right). \tag{2}$$

5.1.5 PROPERTIES OF THE GENERALIZED LAGUERRE FUNCTION

The proof of Theorem 1 will rely on several properties of the generalized Laguerre function. In this section we will review two known properties and prove several additional ones as lemmas.

An important property of the generalized Laguerre function is its infinite differentiability in z, with the result of differentiation again being a generalized Laguerre function:

$$\frac{\partial}{\partial z} L_{\nu}^{(\alpha)}(z) = -L_{\nu-1}^{(\alpha+1)}(z)\,. \tag{3}$$

Another useful property is the following recurrence relation:

$$L_{\nu}^{(\alpha)}(z) = L_{\nu-1}^{(\alpha)}(z) + L_{\nu}^{(\alpha-1)}(z)\,. \tag{4}$$

Lemma 9 *For $\alpha > 0$ and $z < 0$:*
(a) $L_{-1/2}^{(\alpha)}(z)$ *is a positive monotonically increasing function in z, while*
(b) $L_{1/2}^{(\alpha)}(z)$ *is a positive monotonically decreasing concave function in z.*

Proof (a) From Definition 8,

$$L_{-1/2}^{(\alpha)}(z) = \frac{\Gamma(\alpha+1/2)}{\Gamma(1/2)} \cdot \frac{{}_1F_1(1/2;\alpha+1;z)}{\Gamma(\alpha+1)}\,.$$

Since $\alpha > 0$ all three terms involving the Gamma function are positive. We transform the remaining term using the equality (Itô, 1993, p. 1799, Appendix A, Table 19.I):

$${}_1F_1(a;b;z) = e^z\,{}_1F_1(b-a;b;-z)\,, \tag{5}$$

which holds arbitrary $a, b, z \in \mathbb{R}$, obtaining

$${}_1F_1(1/2;\alpha+1;z) = e^z\,{}_1F_1(\alpha+1/2;\alpha+1;-z)\,.$$

From Definition 7 it now directly follows that $L_{-1/2}^{(\alpha)}(z)$ is positive for $\alpha > 0$ and $z < 0$.

To show that $L_{-1/2}^{(\alpha)}(z)$ is monotonically increasing in z, from Equation 3 and Definition 8 we have

$$\frac{\partial}{\partial z} L_{-1/2}^{(\alpha)}(z) = -L_{-3/2}^{(\alpha+1)}(z) = -\frac{\Gamma(\alpha+1/2)}{\Gamma(-1/2)} \cdot \frac{{}_1F_1(3/2;\alpha+2;z)}{\Gamma(\alpha+2)}\,.$$

For $\alpha > 0$ and $z < 0$, from Equation 5 it follows that ${}_1F_1(3/2;\alpha+2;z) > 0$. Since $\Gamma(-1/2) < 0$ and all remaining terms are positive, it follows that $-L_{-3/2}^{(\alpha+1)}(z) > 0$. Thus, $L_{-1/2}^{(\alpha)}(z)$ is monotonically increasing in z.

(b) Proofs that $L_{1/2}^{(\alpha)}(z)$ is positive and monotonically decreasing are very similar to the proofs in part (a), and will be omitted. To address concavity, we observe the second derivative of $L_{1/2}^{(\alpha)}(z)$:

$$\frac{\partial^2}{\partial z} L_{1/2}^{(\alpha)}(z) = L_{-3/2}^{(\alpha+2)}(z) = \frac{\Gamma(\alpha+3/2)}{\Gamma(-1/2)} \cdot \frac{{}_1F_1(3/2;\alpha+3;z)}{\Gamma(\alpha+3)}\,.$$

Similarly to part (a), from Equation 5, Definition 7, and basic properties of the gamma function it follows that $L_{-3/2}^{(\alpha+2)}(z) < 0$ for $\alpha > 0$ and $z < 0$. Thus, $L_{1/2}^{(\alpha)}(z)$ is concave in z. ■

Lemma 10 *For* $\alpha > 0$ *and* $z < 0$, $L_{1/2}^{(\alpha+1)}(z) \approx L_{1/2}^{(\alpha)}(z)$.

Proof From the recurrence relation in Equation 4 we obtain

$$L_{1/2}^{(\alpha+1)}(z) = L_{-1/2}^{(\alpha+1)}(z) + L_{1/2}^{(\alpha)}(z).$$

Therefore, to prove the lemma it needs to be shown that for $z < 0$,

$$\lim_{\alpha\to\infty} L_{-1/2}^{(\alpha)}(z) = 0. \tag{6}$$

From Definition 8 and Equation 5 we have

$$L_{-1/2}^{(\alpha)}(z) = \frac{e^{-z}}{\Gamma(1/2)} \cdot \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha+1)} \cdot {}_1F_1(\alpha+1/2;\alpha+1;-z).$$

From the asymptotic expansion by Fujikoshi (2007, p. 16, adapted),

$${}_1F_1\left(\frac{1}{2}n;\frac{1}{2}(n+b);x\right) = e^x\left(1+O(n^{-1})\right), \tag{7}$$

where n is large and $x \geq 0$, it follows that $\lim_{\alpha\to\infty} {}_1F_1(\alpha+1/2;\alpha+1;-z) < \infty$. Thus, to prove Equation 6 and the lemma it remains to be shown that

$$\lim_{\alpha\to\infty} \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha+1)} = 0. \tag{8}$$

As in the proof of Lemma 6, from the inequalities derived by Haagerup (1982) we have

$$\sqrt{\beta-1} \leq \sqrt{2}\,\frac{\Gamma\left(\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta}{2}\right)} \leq \sqrt{\beta},$$

where $\beta > 1$. Applying inversion and substituting β with $2\alpha+1$ yields the desired limit. ■

Lemma 11 *For* $\alpha > 1/2$ *and* $z < 0$:

(a) $\lim_{z\to-\infty} L_{-3/2}^{(\alpha)}(z) = 0$, *and*

(b) $\lim_{\alpha\to\infty} L_{-3/2}^{(\alpha)}(z) = 0$.

Proof (a) From Definition 8 we have

$$L_{-3/2}^{(\alpha)}(z) = \frac{\Gamma(\alpha-1/2)}{\Gamma(-1/2)} \cdot \frac{{}_1F_1(3/2;\alpha+1;z)}{\Gamma(\alpha+1)}. \tag{9}$$

The following property (Abramowitz and Stegun, 1964, p. 504, Equation 13.1.5),

$${}_1F_1(a;b;z) = \frac{\Gamma(b)}{\Gamma(b-a)}(-z)^a\left(1+O(|z|^{-1})\right) \quad (z<0),$$

when substituted into Equation 9, taking $a = 3/2$ and $b = \alpha + 1$, yields

$$L_{-3/2}^{(\alpha)}(z) = \frac{1}{\Gamma(-1/2)}(-z)^{-3/2}\left(1 + O(|z|^{-1})\right). \tag{10}$$

From Equation 10 the desired limit directly follows.

(b) The proof of part (b) is analogous to the proof of Lemma 10, that is, Equation 6. From Definition 8 and Equation 5, after applying the expansion by Fujikoshi (2007) given in Equation 7, it remains to be shown that

$$\lim_{\alpha\to\infty} \frac{\Gamma(\alpha - 1/2)}{\Gamma(\alpha + 1)} = 0. \tag{11}$$

Since $\Gamma(\alpha - 1/2) < \Gamma(\alpha + 1/2)$ for every $\alpha \geq 2$, the desired limit in Equation 11 follows directly from Equation 8. ■

5.1.6 THE MAIN RESULT

This section restates and proves our main theoretical result.

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 $d \in \mathbb{N}^+$, $c_1, c_2 \leq 0$, $c_1 < c_2$, 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*

$$\Delta\mu_d(\lambda_{d,1}, \lambda_{d,2}) = \mu_{\chi(d,\lambda_{d,2})} - \mu_{\chi(d,\lambda_{d,1})},$$

where $\mu_{\chi(d,\lambda_{d,i})}$ is the mean of the noncentral Chi distribution with d degrees of freedom and noncentrality parameter $\lambda_{d,i}$ ($i \in \{1,2\}$).

There exists $d_0 \in \mathbb{N}$ such that for every $d > d_0$,

$$\Delta\mu_d(\lambda_{d,1}, \lambda_{d,2}) > 0, \tag{12}$$

and

$$\Delta\mu_{d+2}(\lambda_{d+2,1}, \lambda_{d+2,2}) > \Delta\mu_d(\lambda_{d,1}, \lambda_{d,2}). \tag{13}$$

Proof To prove Equation 12, we observe that for $d > 2$,

$$\begin{aligned}
\Delta\mu_d(\lambda_{d,1}, \lambda_{d,2}) &= \mu_{\chi(d,\lambda_{d,2})} - \mu_{\chi(d,\lambda_{d,1})} \\
&= \sqrt{\frac{\pi}{2}}\, L_{1/2}^{(\frac{d}{2}-1)}\left(-\frac{\lambda_{d,2}^2}{2}\right) - \sqrt{\frac{\pi}{2}}\, L_{1/2}^{(\frac{d}{2}-1)}\left(-\frac{\lambda_{d,1}^2}{2}\right) \\
&> 0,
\end{aligned}$$

where the last inequality holds for $d > 2$ because $\lambda_{d,1} < \lambda_{d,2}$, and $L_{1/2}^{(d/2-1)}(z)$ is a monotonically decreasing function in $z < 0$ for $d/2 - 1 > 0$ (Lemma 9).

In order to prove Equation 13, we will use approximate values of noncentrality parameters $\lambda_{d,1}$ and $\lambda_{d,2}$. Let $\widehat{\lambda}_{d,1} = \sqrt{d} + c_1/\sqrt{2}$, and $\widehat{\lambda}_{d,2} = \sqrt{d} + c_2/\sqrt{2}$. From Lemma 6 it follows that $\widehat{\lambda}_{d,1} \approx \lambda_{d,1}$ and $\widehat{\lambda}_{d,2} \approx \lambda_{d,2}$. Thus, by proving that there exists $d_2 \in \mathbb{N}$ such that for every $d > d_2$,

$$\Delta\mu_{d+2}(\widehat{\lambda}_{d+2,1}, \widehat{\lambda}_{d+2,2}) > \Delta\mu_d(\widehat{\lambda}_{d,1}, \widehat{\lambda}_{d,2}), \tag{14}$$

we prove that there exists $d_1 \in \mathbb{N}$ such that for every $d > d_1$ Equation 13 holds. The existence of such d_1, when approximations are used as function arguments, is ensured by the fact that $L_{1/2}^{(\alpha)}(z)$ is a monotonically decreasing *concave* function in z (Lemma 9), and by the transition from α to $\alpha+1$ having an insignificant impact on the value of the Laguerre function for large enough α (Lemma 10). Once Equation 14 is proven, Equations 12 and 13 will hold for every $d > d_0$, where $d_0 = \max(2, d_1)$.

To prove Equation 14, from Equation 2 it follows we need to show that

$$\begin{aligned} & L_{1/2}^{(d/2)}\left(-\frac{1}{2}\left(\sqrt{d+2}+c_2/\sqrt{2}\right)^2\right) - L_{1/2}^{(d/2)}\left(-\frac{1}{2}\left(\sqrt{d+2}+c_1/\sqrt{2}\right)^2\right) \\ & \quad > \; L_{1/2}^{(d/2-1)}\left(-\frac{1}{2}\left(\sqrt{d}+c_2/\sqrt{2}\right)^2\right) - L_{1/2}^{(d/2-1)}\left(-\frac{1}{2}\left(\sqrt{d}+c_1/\sqrt{2}\right)^2\right). \end{aligned} \tag{15}$$

We observe the second derivative of $L_{1/2}^{(\alpha)}(z)$:

$$\frac{\partial^2}{\partial z} L_{1/2}^{(\alpha)}(z) = L_{-3/2}^{(\alpha+2)}(z).$$

Since $L_{-3/2}^{(\alpha+2)}(z)$ tends to 0 as $z \to -\infty$, and tends to 0 also as $\alpha \to \infty$ (Lemma 11), it follows that the two Laguerre functions on the left side of Equation 15 can be approximated by a linear function with an arbitrary degree of accuracy for large enough d. More precisely, since $L_{1/2}^{(\alpha)}(z)$ is monotonically decreasing in z (Lemma 9) there exist $a, b \in \mathbb{R}$, $a > 0$, such that the left side of Equation 15, for large enough d, can be replaced by

$$\begin{aligned} & -a\left(-\frac{1}{2}\left(\sqrt{d+2}+c_2/\sqrt{2}\right)^2\right)+b - \left(-a\left(-\frac{1}{2}\left(\sqrt{d+2}+c_1/\sqrt{2}\right)^2\right)+b\right) \\ & \quad = \; \frac{a}{2}\left(\sqrt{d+2}+c_2/\sqrt{2}\right)^2 - \frac{a}{2}\left(\sqrt{d+2}+c_1/\sqrt{2}\right)^2. \end{aligned} \tag{16}$$

From Lemma 10 it follows that the same linear approximation can be used for the right side of Equation 15, replacing it by

$$\frac{a}{2}\left(\sqrt{d}+c_2/\sqrt{2}\right)^2 - \frac{a}{2}\left(\sqrt{d}+c_1/\sqrt{2}\right)^2. \tag{17}$$

After substituting the left and right side of Equation 15 with Equations 16 and 17, respectively, it remains to be shown that

$$\begin{aligned} & \frac{a}{2}\left(\sqrt{d+2}+c_2/\sqrt{2}\right)^2 - \frac{a}{2}\left(\sqrt{d+2}+c_1/\sqrt{2}\right)^2 \\ & \quad > \frac{a}{2}\left(\sqrt{d}+c_2/\sqrt{2}\right)^2 - \frac{a}{2}\left(\sqrt{d}+c_1/\sqrt{2}\right)^2. \end{aligned} \tag{18}$$

Multiplying both sides by $\sqrt{2}/a$, moving the right side to the left, and applying algebraic simplification reduces Equation 18 to

$$\left(c_2 - c_1\right)\left(\sqrt{d+2} - \sqrt{d}\right) > 0,$$

which holds for $c_1 < c_2$, thus concluding the proof. ■

5.2 Discussion

This section will discuss several additional considerations and related work regarding the geometry of high-dimensional spaces and the behavior of data distributions within them. First, let us consider the geometric upper limit to the number of points that point $\mathbf{x}$ can be a nearest neighbor of, in Euclidean space. In one dimension, this number is 2, in two dimensions it is 5, while in 3 dimensions it equals 11 (Tversky and Hutchinson, 1986). Generally, for Euclidean space of dimensionality d this number is equal to the *kissing number*, which is the maximal number of hyperspheres that can be placed to touch a given hypersphere without overlapping, with all hyperspheres being of the same size.[14] Exact kissing numbers for arbitrary d are generally not known, however there exist bounds which imply that they progress exponentially with d (Odlyzko and Sloane, 1979; Zeger and Gersho, 1994). Furthermore, when considering k nearest neighbors for $k > 1$, the bounds become even larger. Therefore, only for very low values of d geometrical constraints of vector space prevent hubness. On the other hand, for higher values of d hubness may or may not occur, and the geometric bounds, besides providing "room" for hubness (even for values of k as low as 1) do not contribute much in fully characterizing the hubness phenomenon. Therefore, in high dimensions the behavior of *data distributions* needed to be studied.

We focus the rest of the discussion around the following important result,[15] drawing parallels with our results and analysis, and extending existing interpretations.

Theorem 12 (Newman and Rinott, 1985, p. 803, Theorem 3, adapted)

Let $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$, $i = 0, \ldots, n$ be a sample of $n+1$ i.i.d. points from distribution $\mathbf{F}(\mathbf{X})$, $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$. Assume that $\mathbf{F}$ is of the form $\mathbf{F}(\mathbf{X}) = \prod_{k=1}^{d} \mathcal{F}(X_k)$, that is, the coordinates $X_1, \ldots, X_d$ are i.i.d. Let the distance measure be of the form $D(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sum_{k=1}^{d} g(x_k^{(i)}, x_k^{(j)})$. Let $N_1^{n,d}$ denote the number of points among $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ whose nearest neighbor is $\mathbf{x}^{(0)}$.

Suppose $0 < \mathrm{Var}(g(X,Y)) < \infty$ and set

$$\beta = \mathrm{Correlation}(g(X,Y), g(X,Z)), \tag{19}$$

where X, Y, Z are i.i.d. with common distribution $\mathcal{F}$ (the marginal distribution of X_k).
(a) *If $\beta = 0$ then*

$$\lim_{n\to\infty} \lim_{d\to\infty} N_1^{n,d} = \mathrm{Poisson}(\lambda = 1) \ \textit{in distribution} \tag{20}$$

and

$$\lim_{n\to\infty} \lim_{d\to\infty} \mathrm{Var}(N_1^{n,d}) = 1. \tag{21}$$

(b) *If $\beta > 0$ then*

$$\lim_{n\to\infty} \lim_{d\to\infty} N_1^{n,d} = 0 \ \textit{in distribution} \tag{22}$$

while

$$\lim_{n\to\infty} \lim_{d\to\infty} \mathrm{Var}(N_1^{n,d}) = \infty. \tag{23}$$

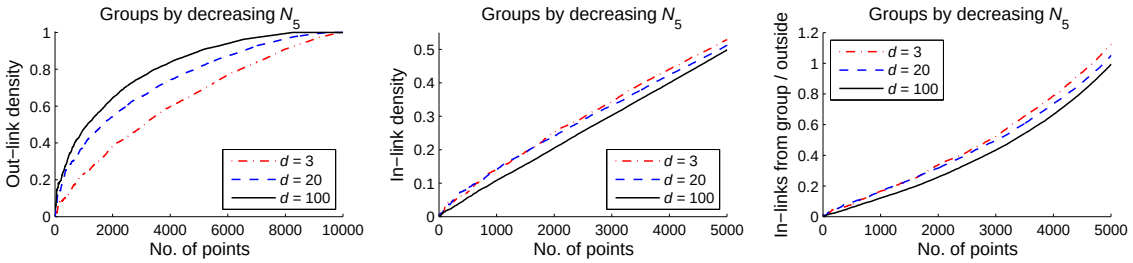
14. If ties are disallowed, it may be necessary to subtract 1 from the kissing number to obtain the maximum of N_1.
15. A theorem that is effectively a special case of this result was proven previously (Newman et al., 1983, Theorem 7) for continuous distributions with finite kurtosis and Euclidean distance.

What is exceptional in this theorem are Equations 22 and 23. According to the interpretation by Tversky et al. (1983), they suggest 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 N_1 equaling 0, and a small proportion of points with high N_1 values, that is, hubs.[16] Trivially, Equation 23 also holds for N_k with $k > 1$, since for any point $\mathbf{x}$, $N_k(\mathbf{x}) \geq N_1(\mathbf{x})$.

The setting involving i.i.d. normal random data and Euclidean distance, used in our Theorem 1 (and, generally, any i.i.d. random data distribution with Euclidean distance), fulfills the conditions of Theorem 12 for Equations 22 and 23 to be applied, since the correlation parameter $\beta > 0$. This correlation exists because, for example, if we view vector component variable X_j ($j \in \{1, 2, \ldots, d\}$) and the distribution of data points within it, if a random point drawn from X_j is closer to the mean of X_j it is more likely to be close to other random points drawn from X_j, and vice versa, producing the case $\beta > 0$.[17] Therefore, Equations 22 and 23 from Theorem 12 directly apply to the setting studied in Theorem 1, providing asymptotic evidence for hubness. However, since the proof of Equations 22 and 23 in Theorem 12 relies on applying the central limit theorem to the (normalized) distributions of pairwise distances between vectors $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ ($0 \leq i \neq j \leq n$) as $d \to \infty$ (obtaining limit distance distributions which are Gaussian), the results of Theorem 12 are inherently asymptotic in nature. Theorem 1, on the other hand, describes the behavior of distances in finite dimensionalities,[18] providing the means to characterize the behavior of N_k in high, but finite-dimensional space. What remains to be done is to formally connect Theorem 1 with the skewness of N_k in finite dimensionalities, for example by observing point $\mathbf{x}$ with a fixed position relative to the data distribution mean (the origin) across dimensionalities, in terms of being at distance $\mu_{\chi(d)} + c\sigma_{\chi(d)}$ from the origin, and expressing how the probability of $\mathbf{x}$ to be the nearest neighbor (or among the k nearest neighbors) of a randomly drawn point changes with increasing dimensionality.[19] We address this investigation as a point of future work.

Returning to Theorem 12 and the value of β from Equation 19, as previously discussed, $\beta > 0$ signifies that the position of a vector component value makes a difference when computing distances between vectors, causing some component values to be more "special" than others. Another contribution of Theorem 1 is that it illustrates how the individual differences in component values combine to make positions of whole vectors more special (by being closer to the data center). On the other hand, if $\beta = 0$ no point can have a special position with respect to all others. In this case, Equations 20 and 21 hold, which imply there is no hubness. This setting is relevant, for example, to points being generated by a Poisson process which spreads the vectors uniformly over $\mathbb{R}^d$, where all positions within the space (both at component-level and globally) become basically equivalent. Although it does not directly fit into the framework of Theorem 12, the same principle can be used to explain the absence of hubness for normally distributed data and cosine distance from Section 3.1: in this setting no vector is more spatially central than any other. Equations 20 and 21, which imply no hubness, hold for many more "centerless" settings, including random graphs, settings with exchangeable distances, and d-dimensional toruses (Newman and Rinott, 1985).

16. Reversing the order of limits, which corresponds to having a large number of points relative to the number of dimensions, produces the same asymptotic behavior as in Equations 20 and 21, that is, no hubness, in all studied settings.
17. Similarly to Section 4.2, this argument holds for unimodal distributions of component variables; for multimodal distributions the driving force behind nonzero β is the proximity to a peak in the probability density function.
18. Although the statement of Theorem 1 relies on dimensionality being greater than some d_0 which is finite, but can be arbitrarily high, empirical evidence suggests that actual d_0 values are low, often equaling 0.
19. For $c < 0$ we expect this probability to increase since $\mathbf{x}$ is closer to the data distribution mean, and becomes closer to other points as dimensionality increases.



(a) (b) (c)

Figure 8: (a) Out-link, and (b) in-link densities of groups of hubs with increasing size; (c) ratio of the number of in-links originating from points within the group and in-links originating from outside points, for i.i.d. uniform random data with dimensionality $d = 3,\ 20,\ 100$.

The following two subsections will address several additional issues concerning the interpretation of Theorem 12.

5.2.1 NEAREST-NEIGHBOR GRAPH STRUCTURE

The interpretation of Theorem 12 by Tversky et al. (1983) may be understood in the sense that, with increasing dimensionality, *very few* exceptional points become hubs, while all others are relegated to antihubs. In this section we will empirically examine the structural change of the k-NN graph as the number of dimensions increases. We will also discuss and consolidate different notions node centrality in the k-NN graph, and their dependence on the (intrinsic) dimensionality of data.

First, as in Section 3.1, we consider $n = 10000$ i.i.d. uniform random data points of different dimensionality. Let us observe hubs, that is, points with highest N_5, collected in groups of progressively increasing size: $5, 10, 15, \ldots, 10000$. In analogy with the notion of network density from social network analysis (Scott, 2000), we define group *out-link density* as the proportion of the number of arcs that originate and end in nodes from the group, and the total number of arcs that originate from nodes in the group. Conversely, we define group *in-link density* as the proportion of the number of arcs that originate and end in nodes from the group, and the total number of arcs that *end* in nodes from the group. Figure 8(a, b) shows the out-link and in-link densities of groups of strongest hubs in i.i.d. uniform random data (similar tendencies can be observed with other synthetic data distributions). It can be seen in Figure 8(a) that hubs are more cohesive in high dimensions, with more of their out-links leading to other hubs. On the other hand, Figure 8(b) suggests that hubs also receive more in-links from non-hub points in high dimensions than in low dimensions. Moreover, Figure 8(c), which plots the ratio of the number of in-links that originate within the group, and the number of in-links which originate outside, shows that hubs receive a larger proportion of in-links from non-hub points in high dimensions than in low dimensions. We have reported our findings for $k = 5$, however similar results are obtained with other values of k.

Overall, it can be said that in high dimensions hubs receive more in-links than in low dimensions from both hubs and non-hubs, and that the range of influence of hubs gradually widens as dimensionality increases. We can therefore conclude that the transition of hubness from low to high dimensionalities is "smooth," both in the sense of the change in the overall distribution of N_k, and the change in the degree of influence of data points, as expressed by the above analysis of links.

So far, we have viewed hubs primarily through their exhibited high values of N_k, that is, high degree centrality in the k-NN directed graph. However, (scale-free) network analysis literature often attributes other properties to hubs (Albert and Barabási, 2002), viewing them as nodes that are important for preserving network structure due to their central positions within the *graph*, indicated, for example, by their *betweenness centrality* (Scott, 2000). On the other hand, as discussed by Li et al. (2005), in both synthetic and real-world networks high-degree nodes do not necessarily need to correspond to nodes that are central in the graph, that is, high-degree nodes can be concentrated at the *periphery* of the network and bear little structural significance. For this reason, we computed the betweenness centrality of nodes in k-NN graphs of synthetic and real data sets studied in this paper, and calculated its Spearman correlation with node degree, denoting the measure by $C_{\mathrm{BC}}^{N_k}$. For i.i.d. uniform data ($k=5$), when $d=3$ the measured correlation is $C_{\mathrm{BC}}^{N_5}=0.311$, when $d=20$ the correlation is $C_{\mathrm{BC}}^{N_5}=0.539$, and finally when $d=100$ the correlation rises to $C_{\mathrm{BC}}^{N_5}=0.647$.[20] This suggests that with increasing dimensionality the centrality of nodes increases not only in the sense of higher node degree or spatial centrality of vectors (as discussed in Section 4.2), but also in the structural graph-based sense. We support this observation further by computing, over the 50 real data sets listed in Table 1, the correlation between $C_{\mathrm{BC}}^{N_{10}}$ and $S_{N_{10}}$, finding it to be significant: 0.548.[21] This indicates that real data sets which exhibit strong skewness in the distribution of N_{10} also tend to have strong correlation between N_{10} and betweenness centrality of nodes, giving hubs a broader significance for the structure of the k-NN graphs.

5.2.2 RATE OF CONVERGENCE AND THE ROLE OF BOUNDARIES

On several occasions, the authors of Theorem 12 have somewhat downplayed the significance of equations 22 and 23 (Tversky et al., 1983; Newman and Rinott, 1985), citing empirically observed slow convergence (Maloney, 1983), even to the extent of not observing significant differences between hubness in the Poisson process and i.i.d. uniform cube settings. However, results in the preceding sections of this paper suggest that this convergence is fast enough to produce notable hubness in high-dimensional data. In order to directly illustrate the difference between a setting which provides no possibility for spatial centrality of points, and one that does, we will observe the Poisson process vs. the i.i.d. unit cube setting. We will be focusing on the location of the nearest neighbor of a point from the cube, that is, on determining whether it stays within the boundaries of the cube as dimensionality increases.

Lemma 13 *Let points be spread in $\mathbb{R}^d$ according to a Poisson process with constant intensity $\lambda > 1$. Observe a unit hypercube $C \subset \mathbb{R}^d$, and an arbitrary point $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in C$, generated by the Poisson process. Let $p_{\lambda,d}$ denote the probability that the nearest neighbor of $\mathbf{x}$, with respect to Euclidean distance, is not situated in C. Then,*

$$\lim_{d\to\infty} p_{\lambda,d} = 1\,.$$

Proof Out of the 3^d-1 unit hypercubes that surround C, let us observe only the $2d$ hypercubes that differ from C only in one coordinate. We will restrict the set of considered points to these $2d$ cubes, and prove that the probability that the nearest neighbor of $\mathbf{x}$ comes from one of the $2d$ cubes, $\widehat{p}_{\lambda,d}$, converges to 1 as $d\to\infty$. From this, the convergence of $p_{\lambda,d}$ directly follows, since $p_{\lambda,d} \geq \widehat{p}_{\lambda,d}$.

20. Betweenness centrality is computed on directed k-NN graphs. We obtained similar correlations when undirected graphs were used.

21. When betweenness centrality is computed on undirected graphs, the correlation is even stronger: 0.585.

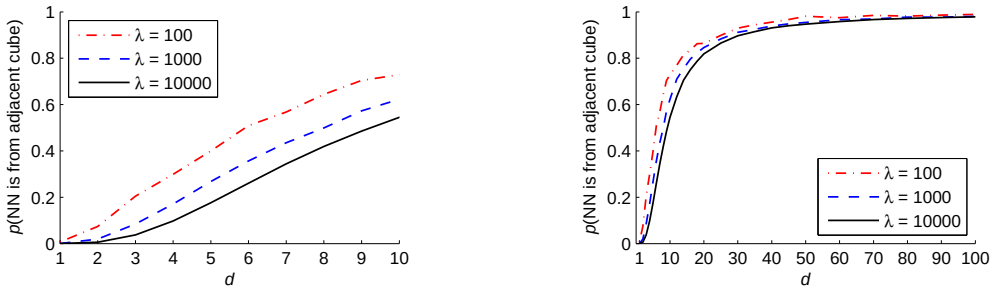


Figure 9: Probability that the nearest neighbor of a point from the unit hypercube originates from one of the adjacent hypercubes, for Poisson processes with $\lambda = 100, 1000,$ and 10000 expected points per hypercube, obtained through simulation and averaged over 10 runs.

Let $\widehat{p}_{\lambda,d}(i)$ denote the probability that the nearest neighbor of $\mathbf{x}$ comes from one of the two hypercubes which differ from C only in the ith coordinate, $i \in \{1,2,\ldots,d\}$. For a given coordinate i and point $\mathbf{x}$, let the *1-dimensional nearest neighbor* of x_i denote the closest y_i value of all other points $\mathbf{y}$ from the Poisson process, observed when all coordinates except i are disregarded. Conversely, for a given coordinate i and point $\mathbf{x}$, let the $(d-1)$*-dimensional nearest neighbor* of $(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d)$ denote the closest point $(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_d)$, obtained when coordinate i is disregarded.

Observing the ith coordinate only, the probability for the 1-dimensional nearest neighbor of x_i to come from one of the surrounding unit intervals is $\widehat{p}_{\lambda,1}$. Although small, $\widehat{p}_{\lambda,1} > 0$. Assuming this event has occurred, let $\mathbf{y} \in \mathbb{R}^d$ be the point whose component y_i is the 1-dimensional nearest neighbor of x_i that is not within the unit interval containing x_i. Let r_λ denote the probability that the remaining coordinates of $\mathbf{y}$, $(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_d)$, constitute a $(d-1)$-dimensional nearest neighbor of $(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d)$, within the confines of C. It can be observed that r_λ is strictly greater than 0, inversely proportional to λ (roughly equaling $1/(\lambda-1)$), and independent of d. Thus, $\widehat{p}_{\lambda,d}(i) \geq \widehat{p}_{\lambda,1} \cdot r_\lambda > 0$.

Let $\widehat{q}_{\lambda,d} = 1 - \widehat{p}_{\lambda,d}$, the probability that the nearest neighbor of $\mathbf{x}$ comes from C (recall the restriction of the location of the nearest neighbor to C and its immediately surrounding $2d$ hypercubes). In light of the above, $\widehat{q}_{\lambda,d} = \prod_{i=1}^{d} \widehat{q}_{\lambda,d}(i) = \prod_{i=1}^{d} (1 - \widehat{p}_{\lambda,d}(i))$. Since each $\widehat{p}_{\lambda,d}(i)$ is bounded from below by a constant strictly greater than 0 (which depends only on λ), each $\widehat{q}_{\lambda,d}(i)$ is bounded from above by a constant strictly smaller than 1. It follows that $\lim_{d\to\infty} \widehat{q}_{\lambda,d} = 0$, and therefore $\lim_{d\to\infty} \widehat{p}_{\lambda,d} = 1$. ■

To illustrate the rate of convergence in Lemma 13, Figure 9 plots the empirically observed probabilities that the nearest neighbor of a point from a unit hypercube originates in one of the $2d$ immediately adjacent unit hypercubes ($\widehat{p}_{\lambda,d}$). It can be seen that the probability that the nearest neighbor comes from outside of the cube quickly becomes close to 1 as dimensionality increases. Please note that due to feasibility of simulation the plots in Figure 9 represent the empirical lower bounds (that is, the empirical estimates of $\widehat{p}_{\lambda,d}$) of the true probabilities from Lemma 13 ($p_{\lambda,d}$) by considering only the $2d$ immediately adjacent hypercubes.

The above indicates that when boundaries are introduced in high dimensions, the setting completely changes, in the sense that new nearest neighbors need to be located inside the boundaries. Under such circumstances, points which are closer to the center have a better chance of becoming nearest neighbors, with the mechanism described in previous sections. Another implication of the above observations, stemming from Figure 9, is that the number of dimensions does not need to be very large compared to the number of points for the setting to change. As for boundaries, they can be viewed as a dual notion to spatial centrality discussed earlier. With Poisson processes and cubes this duality is rather straightforward, however for continuous distributions in general there exist no boundaries in a strict mathematical sense. Nevertheless, since data sets contain a finite number of points, it can be said that "practical" boundaries exist in this case as well.

6. Hubness and Dimensionality Reduction

In this section we elaborate further on the interplay of skewness of N_k and intrinsic dimensionality by considering dimensionality-reduction (DR) techniques. The main question motivating the discussion in this section is whether dimensionality reduction can alleviate the issue of the skewness of k-occurrences altogether. We leave a more detailed and general investigation of the interaction between hubness and dimensionality reduction as a point of future work.

We examined the following methods: principal component analysis—PCA (Jolliffe, 2002), independent component analysis—ICA (Hyvärinen and Oja, 2000), stochastic neighbor embedding—SNE (Hinton and Roweis, 2003), isomap (Tenenbaum et al., 2000), and diffusion maps (Lafon and Lee, 2006; Nadler et al., 2006). Figure 10 depicts the relationship between the percentage of the original number of features maintained by the DR methods and S_{N_k}, for several high-dimensional real data sets (musk1, mfeat-factors, and spectrometer; see Table 1) and i.i.d. uniform random data (with Euclidean distance, $k = 10$, and the same number of neighbors used for isomap and diffusion maps). For PCA, ICA, SNE, and the real data sets, looking from right to left, S_{N_k} stays relatively constant until a small percentage of features is left, after which it suddenly drops (Figure 10(a–c)). It can be said that this is the point where the intrinsic dimensionality of data sets is reached, and further reduction of dimensionality may incur loss of valuable information. Such behavior is in contrast with the case of i.i.d. uniform random data (full black line in Figure 10(a–c)), where S_{N_k} steadily and steeply reduces with the decreasing number of randomly selected features (dimensionality reduction is not meaningful in this case), because intrinsic and embedded dimensionalities are equal. Since PCA is equivalent to metric multidimensional scaling (MDS) when Euclidean distances are used (Tenenbaum et al., 2000), and SNE is a variant of MDS which favors the preservation of distances between nearby points, we can roughly regard the notion of intrinsic dimensionality used in this paper as the minimal number of features needed to account for all *pairwise distances* within a data set. Although ICA does not attempt to explicitly preserve pairwise distances, combinations of independent components produce skewness of N_k which behaves in a way that is similar to the skewness observed with PCA and SNE.

On the other hand, isomap and diffusion maps replace the original distances with distances derived from a neighborhood graph. It can be observed in Figure 10(d, e) that such replacement generally reduces S_{N_k}, but in most cases does not alleviate it completely. With the decreasing number of features, however, S_{N_k} of real data sets in Figure 10(d, e) still behaves in a manner more similar to S_{N_k} of real data sets for PCA, ICA, and SNE (Figure 10(a–c)) than i.i.d. random data (dash-dot black line in Figure 10(d, e)).

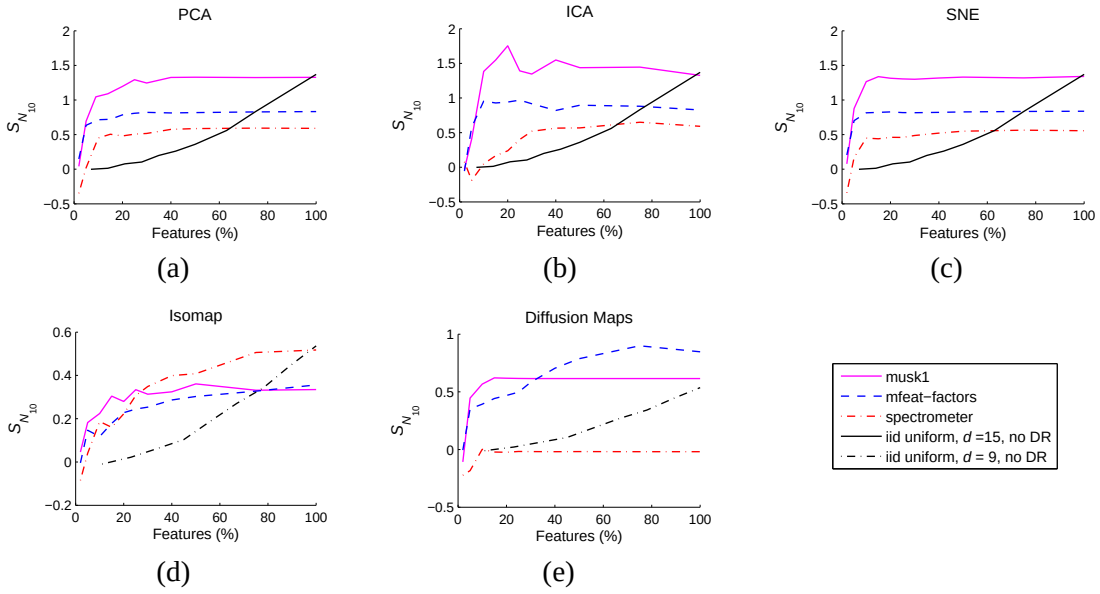


Figure 10: Skewness of N_{10} in relation to the percentage of the original number of features maintained by dimensionality reduction, for real and i.i.d. uniform random data and (a) principal component analysis—PCA, (b) independent component analysis—ICA, (c) stochastic neighbor embedding—SNE, (d) isomap, and (e) diffusion maps.

The above observations signify that, if distances are not explicitly altered (as with isomap and diffusion maps DR methods), that is, if one cares about preserving the original distances, dimensionality reduction may not have a significant effect on hubness when the number of features is above the intrinsic dimensionality. This observation is useful in most practical cases because if dimensionality is reduced below intrinsic dimensionality, loss of information can occur. If one still chooses to apply aggressive dimensionality reduction and let the resulting number of features fall below intrinsic dimensionality, it can be expected of pairwise distances and nearest-neighbor relations between points in the data set to be altered, and hubness to be reduced or even lost. Whether these effects should be actively avoided or sought really depends on the application domain and task at hand, that is, whether and to what degree the original pairwise distances represent valuable information, and how useful are the new distances and neighborhoods after dimensionality reduction.

7. The Impact of Hubness on Machine Learning

The impact of hubness on machine-learning applications has not been thoroughly investigated so far. In this section we examine a wide range of commonly used machine-learning methods for supervised (Section 7.1), semi-supervised (Section 7.2), and unsupervised learning (Section 7.3), that either directly or indirectly use distances in the process of building a model. Our main objective is to demonstrate that hubs (as well as their opposites, antihubs) can have a significant effect on these methods. The presented results highlight the need to take hubs into account in a way equivalent to other factors, such as the existence of outliers, the role of which has been well studied.

7.1 Supervised Learning

To investigate possible implications of hubness on supervised learning, we first study the interaction of k-occurrences with information provided by labels (Section 7.1.1). We then move on to examine the effects of hubness on several well-known classification algorithms in Section 7.1.2.

7.1.1 "GOOD" AND "BAD" k-OCCURRENCES

When labels are present, k-occurrences can be distinguished based on whether labels of neighbors match. We define the number of *"bad" k-occurrences* of $\mathbf{x}$, $BN_k(\mathbf{x})$, as the number of points from data set D for which $\mathbf{x}$ is among the first k NNs, and the labels of $\mathbf{x}$ and the points in question *do not match*. Conversely, $GN_k(\mathbf{x})$, the number of *"good" k-occurrences* of $\mathbf{x}$, is the number of such points where labels do match. Naturally, for every $\mathbf{x} \in D$, $N_k(\mathbf{x}) = BN_k(\mathbf{x}) + GN_k(\mathbf{x})$.

To account for labels, Table 1 includes $\widetilde{BN}_{10}$ (14th column), the sum of all observed "bad" k-occurrences of a data set normalized by $\sum_{\mathbf{x}} N_{10}(\mathbf{x}) = 10n$. This measure is intended to express the total amount of "bad" k-occurrences within a data set. Also, to express the amount of information that "regular" k-occurrences contain about "bad" k-occurrences in a particular data set, $C_{BN_{10}}^{N_{10}}$ (15th column) denotes the Spearman correlation between BN_{10} and N_{10} vectors. The motivation behind this measure is to express the degree to which BN_k and N_k follow a similar distribution.

"Bad" hubs, that is, points with high BN_k, are of particular interest to supervised learning because they carry more information about the location of the decision boundaries than other points, and affect classification algorithms in different ways (as will be described in Section 7.1.2). To understand the origins of "bad" hubs in real data, we rely on the notion of the *cluster assumption* from semi-supervised learning (Chapelle et al., 2006), which roughly states that most pairs of points in a high density region (cluster) should be of the same class. To measure the degree to which the cluster assumption is violated in a particular data set, we simply define the *cluster assumption violation* (CAV) coefficient as follows. Let a be the number of pairs of points which are in different classes but in the same cluster, and b the number of pairs of points which are in the same class and cluster. Then, we define

$$\mathrm{CAV} = \frac{a}{a+b},$$

which gives a number in the $[0,1]$ range, higher if there is more violation. To reduce the sensitivity of CAV to the number of clusters (too low and it will be overly pessimistic, too high and it will be overly optimistic), we choose the number of clusters to be 3 times the number of classes of a particular data set. Clustering is performed with K-means.

For all examined real data sets, we computed the Spearman correlation between the total amount of "bad" k-occurrences, $\widetilde{BN}_{10}$, and CAV (16th column of Table 1) and found it strong (0.85, see Table 2). Another significant correlation (0.39) is observed between $C_{BN_{10}}^{N_{10}}$ and intrinsic dimensionality. In contrast, $\widetilde{BN}_{10}$ and CAV are not correlated with intrinsic dimensionality nor with the skewness of N_{10}. The latter fact indicates that high dimensionality and skewness of N_k are not sufficient to induce "bad" hubs. Instead, based on the former fact, we can argue that there are two, mostly independent, forces at work: violation of the cluster assumption on one hand, and high intrinsic dimensionality on the other. "Bad" hubs originate from putting the two together; that is, the consequences of violating the cluster assumption can be more severe in high dimensions than in low dimensions, not in terms of the total amount of "bad" k-occurrences, but in terms of their distribution, since strong regular hubs are now more prone to "pick up" bad k-occurrences than non-hub points. This is supported by

the positive correlation between $C_{BN_{10}}^{N_{10}}$ and intrinsic dimensionality, meaning that in high dimensions BN_k tends to follow a more similar distribution to N_k than in low dimensions.

7.1.2 INFLUENCE ON CLASSIFICATION ALGORITHMS

We now examine how skewness of N_k and the existence of ("bad") hubs affects well-known classification techniques, focusing on the k-nearest neighbor classifier (k-NN), support vector machines (SVM), and AdaBoost. We demonstrate our findings on a selection of data sets from Table 1 which have relatively high (intrinsic) dimensionality, and a non-negligible amount of "badness" ($\widetilde{BN}_k$) and cluster assumption violation (CAV). Generally, the examined classification algorithms (including semi-supervised learning from Section 7.2) exhibit similar behavior on other data sets from Table 1 with the aforementioned properties, and also with various different values of k (in the general range 1–50, as we focused on values of k which are significantly smaller than the number of points in a data set).

k-nearest neighbor classifier. The k-nearest neighbor classifier is negatively affected by the presence of "bad" hubs, because they provide erroneous class information to many other points. To validate this assumption, we devised a simple weighting scheme. For each point $\mathbf{x}$, we compute its standardized "bad" hubness score:

$$h_B(\mathbf{x},k) = \frac{BN_k(\mathbf{x}) - \mu_{BN_k}}{\sigma_{BN_k}},$$

where μ_{BN_k} and σ_{BN_k} are the mean and standard deviation of BN_k, respectively. During majority voting in the k-NN classification phase, when point $\mathbf{x}$ participates in the k-NN list of the point being classified, the vote of $\mathbf{x}$ is weighted by

$$w_k(\mathbf{x}) = \exp(-h_B(\mathbf{x},k)),$$

thus lowering the influence of "bad" hubs on the classification decision. Figure 11 compares the resulting accuracy of k-NN classifier with and without this weighting scheme for six data sets from Table 1. Leave-one-out cross-validation is performed, with Euclidean distance being used for determining nearest neighbors. The k value for N_k is naturally set to the k value used by the k-NN classifier, and $h_B(\mathbf{x},k)$ is recomputed for the training set of each fold. The reduced accuracy of the unweighted scheme signifies the negative influence of "bad" hubs.

Although "bad" hubs tend to carry more information about the location of class boundaries than other points, the "model" created by the k-NN classifier places the emphasis on describing non-borderline regions of the space occupied by each class. For this reason, it can be said that "bad" hubs are truly bad for k-NN classification, creating the need to penalize their influence on the classification decision. On the other hand, for classifiers that explicitly model the borders between classes, such as support vector machines, "bad" hubs can represent points which contribute information to the model in a positive way, as will be discussed next.

Support vector machines. We consider SVMs with the RBF (Gaussian) kernel of the form:

$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma\|\mathbf{x}-\mathbf{y}\|^2\right),$$

where γ is a data-dependent constant. $K(\mathbf{x},\mathbf{y})$ is a smooth monotone function of Euclidean distance between points. Therefore, N_k values in the kernel space are exactly the same as in the original

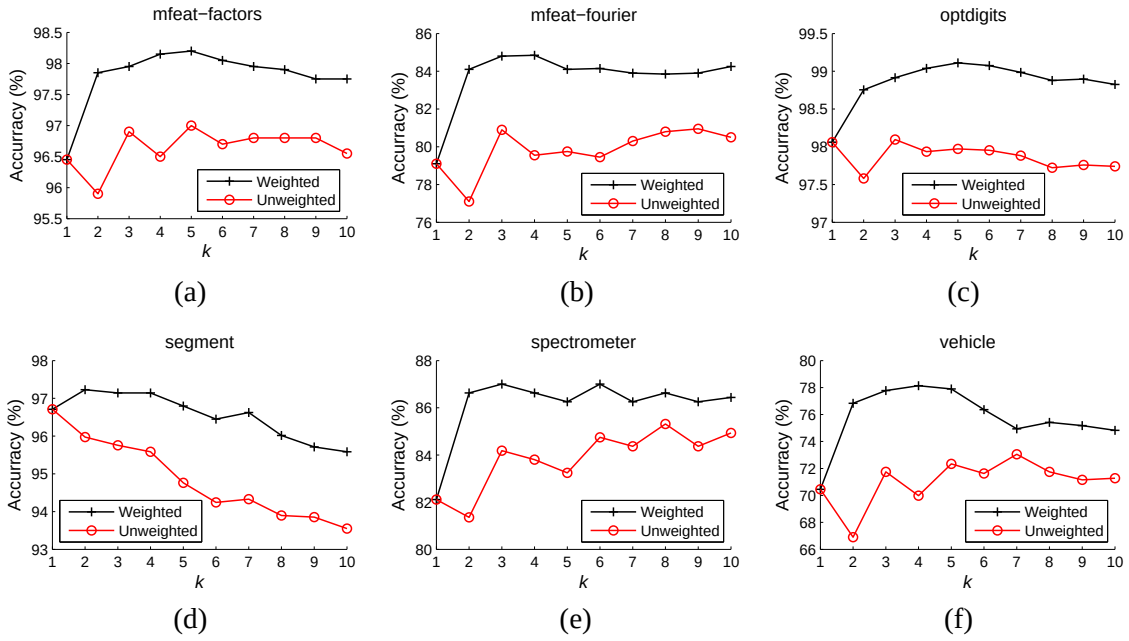


Figure 11: Accuracy of k-NN classifier with and without the weighting scheme.

space.[22] To examine the influence of "bad" hubs on SVMs, Figure 12 illustrates 10-fold cross-validation accuracy results of SVM trained using sequential minimal optimization (Platt, 1999; Keerthi et al., 2001), when points are progressively removed from the training sets: (i) by decreasing BN_k ($k = 5$), and (ii) at random. Accuracy drops with removal by BN_k, indicating that bad hubs are important for support vector machines. The difference in SVM accuracy between random removal and removal by BN_k becomes consistently significant at some stage of progressive removal, as denoted by the dashed vertical lines in the plots, according to the paired t-test at 0.05 significance level.[23]

The reason behind the above observation is that for high-dimensional data, points with high BN_k can comprise good support vectors. Table 3 exemplifies this point by listing the normalized average ranks of support vectors in the 10-fold cross-validation models with regards to decreasing BN_k. The ranks are in the range $[0,1]$, with the value 0.5 expected from a random set of points. Lower values of the ranks indicate that the support vectors, on average, tend to have high BN_k. The table also lists the values of the γ parameter of the RBF kernel, as determined by independent experiments involving 9-fold cross-validation.

AdaBoost. Boosting algorithms take into account the "importance" of points in the training set for classification by weak learners, usually by assigning and updating weights of individual points—the higher the weight, the more attention is to be paid to the point by subsequently trained weak learners. We consider the AdaBoost.MH algorithm (Schapire and Singer, 1999) in conjunction with

22. Centering the kernel matrix changes the N_k of points in the kernel space, but we observed that the overall distribution (that is, its skewness) does not become radically different. Therefore, the following arguments still hold for centered kernels, providing N_k is computed in the kernel space.
23. Since random removal was performed in 20 runs, fold-wise accuracies for statistical testing were obtained in this case by averaging over the runs.

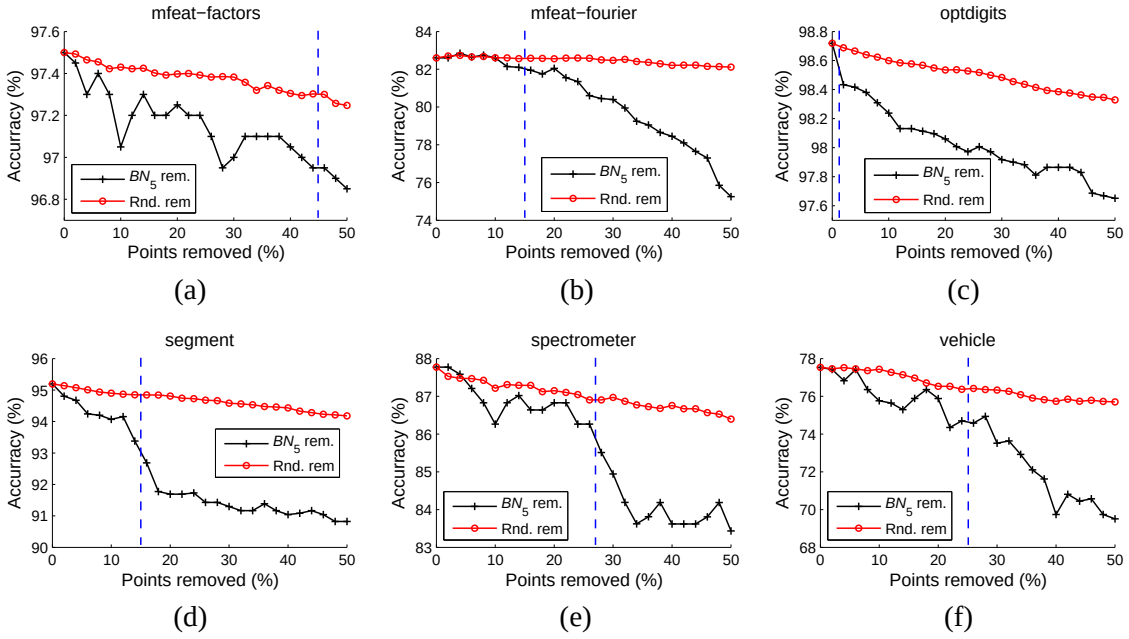


Figure 12: Accuracy of SVM with RBF kernel and points being removed from the training sets by decreasing BN_5, and at random (averaged over 20 runs).

Data set	γ	SV rank	Data set	γ	SV rank
mfeat-factors	0.005	0.218	segment	0.3	0.272
mfeat-fourier	0.02	0.381	spectrometer	0.005	0.383
optdigits	0.02	0.189	vehicle	0.07	0.464

Table 3: Normalized average support vector ranks with regards to decreasing BN_5.

CART trees (Breiman et al., 1984) with the maximal depth of three.[24] We define for each point $\mathbf{x}$ its *standardized hubness* score:

$$h(\mathbf{x},k) = \frac{N_k(\mathbf{x}) - \mu_{N_k}}{\sigma_{N_k}}, \tag{24}$$

where μ_{N_k}, σ_{N_k} are the mean and standard deviation of N_k, respectively. We set the initial weight of each point $\mathbf{x}$ in the training set to

$$w_k(\mathbf{x}) = \frac{1}{1+|h(\mathbf{x},k)|},$$

normalized by the sum over all points, for an empirically determined value of k. The motivation behind the weighting scheme is to assign less importance to both hubs and outliers than other points (this is why we take the absolute value of $h(\mathbf{x},k)$).

Figure 13 illustrates on six classification problems from Table 1 how the weighting scheme helps AdaBoost achieve better generalization in fewer iterations. The data sets were split into training,

24. More precisely, we use the binary "Real AdaBoost" algorithm and the one-vs-all scheme to handle multi-class problems, which is equivalent to the original AdaBoost.MH (Friedman et al., 2000).

validation, and test sets with size ratio 2:1:1, parameter k was chosen based on classification accuracy on the validation sets, and accuracies on the test sets are reported. While it is known that AdaBoost is sensitive to outliers (Rätsch et al., 2001), improved accuracy suggests that hubs should be regarded in an analogous manner, that is, both hubs and antihubs are intrinsically more difficult to classify correctly, and the attention of the weak learners should initially be focused on "regular" points. The discussion from Section 4.4, about hubs corresponding to probabilistic outliers in high-dimensional data, offers an explanation for the observed good performance of the weighting scheme, as both hubs and antihubs can be regarded as (probabilistic) outliers.

To provide further support, Figure 14 depicts binned accuracies of unweighted AdaBoost trained in one fifth of the iterations shown in Figure 13, for points sorted by decreasing N_k. It illustrates how in earlier phases of ensemble training the generalization power with hubs and/or antihubs is worse than with regular points. Moreover, for the considered data sets it is actually the hubs that appear to cause more problems for AdaBoost than antihubs (that is, distance-based outliers).

7.2 Semi-Supervised Learning

Semi-supervised learning algorithms make use of data distribution information provided by unlabeled examples during the process of building a classifier model. An important family of approaches are graph-based methods, which represent data as nodes of a graph, the edges of which are weighted by pairwise distances of incident nodes (Chapelle et al., 2006).

We consider the well-known algorithm by Zhu et al. (2003), whose strategy involves computing a real-valued function f on graph nodes, and assign labels to nodes based on its values. Function f, which exhibits harmonic properties, is obtained by optimizing a quadratic energy function that involves graph edge weights, with the probability distribution on the space of functions f formed using Gaussian fields. For data points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we consider edge weights assigned by the radial basis function (RBF) of the following form:

$$W(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right),$$

where σ is a data-dependent constant. Therefore, large edge weights are assigned between nodes that are close to one another with respect to Euclidean distance.

Taking into account the properties of hubs and antihubs discussed in previous sections, for high-dimensional data sets it can be expected of hubs to be closer to many other points than "regular" points are, and thus carry larger edge weights and be more influential in the process of determining the optimal function f. Conversely, antihubs are positioned farther away from other points, and are expected to bear less influence on the computation of f. Therefore, following an approach that resembles active learning, selecting the initial labeled point set from hubs could be more beneficial in terms of classification accuracy than arbitrarily selecting the initial points to be labeled. On the other extreme, picking the initial labeled point set from the ranks of antihubs could be expected to have a detrimental effect on accuracy.

To validate the above hypothesis, we evaluated the accuracy of the harmonic function algorithm by Zhu et al. (2003) on multiple high-dimensional data sets from Table 1, for labeled set sizes ranging from 1% to 10% of the original data set size, with the test set consisting of all remaining unlabeled points. Because the setting is semi-supervised, we compute the N_k scores of points based on complete data sets, instead of only training sets which was the case in Section 7.1.2. Based on

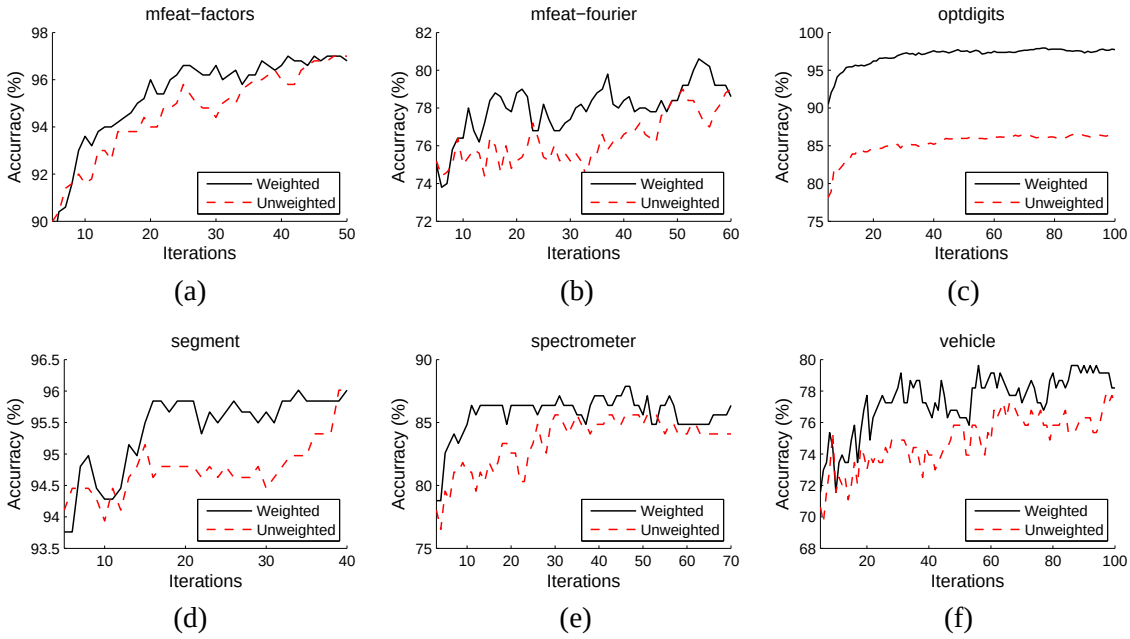


Figure 13: Accuracy of AdaBoost with and without the weighting scheme: (a) $k = 20$, (b) $k = 15$, (c) $k = 10$, (d) $k = 20$, (e) $k = 20$, (f) $k = 40$.

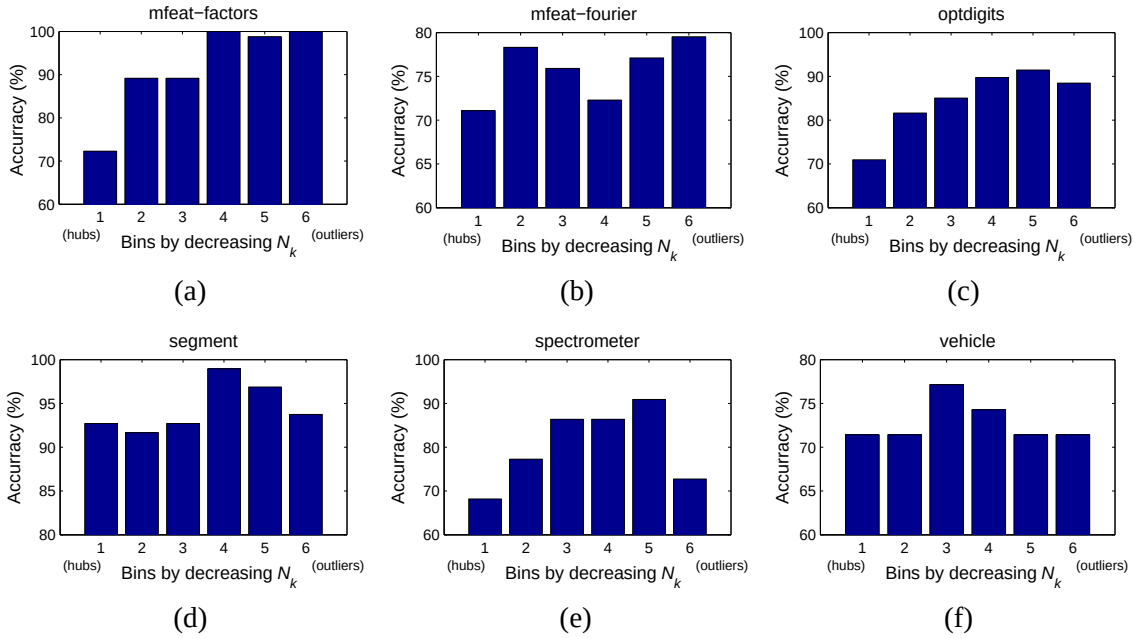


Figure 14: Binned accuracy of AdaBoost by decreasing N_k, at one fifth of the iterations shown in Figure 13.

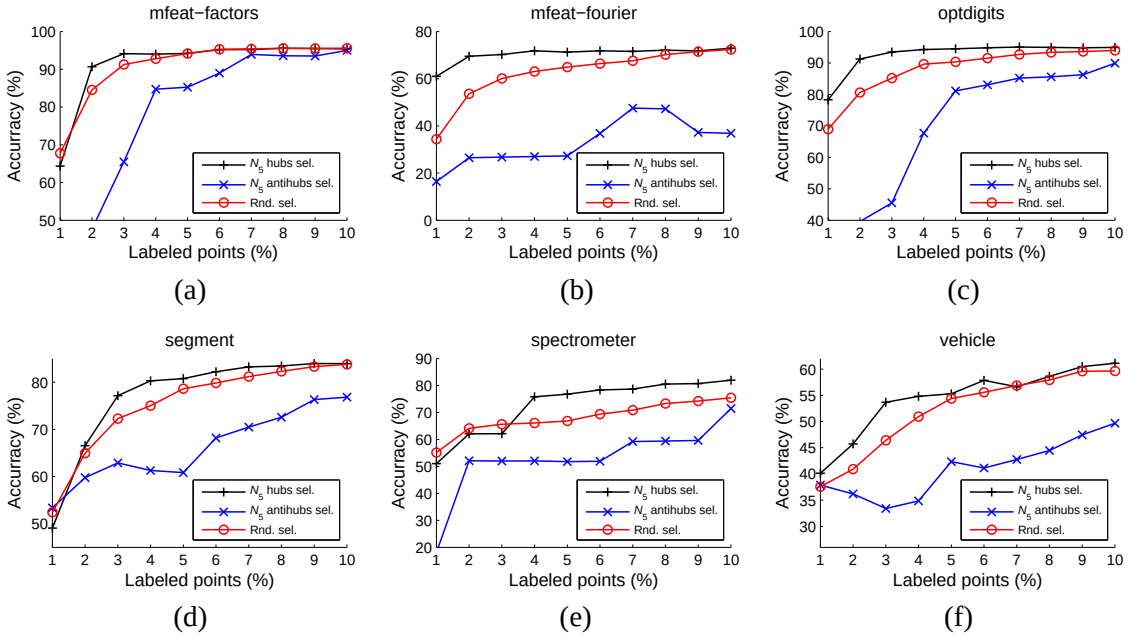


Figure 15: Accuracy of the semi-supervised algorithm by Zhu et al. (2003) with respect to the initial labeled set size as a percentage of the original data set size. Labeled points are selected in the order of decreasing N_5 (hubs first), increasing N_5 (antihubs first), and in random order. Sigma values of the RBF are: (a) $\sigma = 2.9$, (b) $\sigma = 2$, (c) $\sigma = 2.1$, (d) $\sigma = 0.9$, (e) $\sigma = 1.8$, and (f) $\sigma = 0.7$.

the N_k scores ($k = 5$), Figure 15 plots classification accuracies for labeled points selected in the order of decreasing N_5 (we choose to take hub labels first), in the order of increasing N_5 (antihub labels are taken first), and in random order (where we report accuracies averaged over 10 runs).[25] It can be seen that when the number of labeled points is low in comparison to the size of the data sets, taking hub labels first generally produces better classification accuracy. On the other hand, when assigning initial labels to antihubs, accuracy becomes significantly worse, with a much larger labeled set size required for the accuracy to reach that of randomly selected labeled points.

7.3 Unsupervised Learning

This section will discuss the interaction of the hubness phenomenon with unsupervised learning, specifically the tasks of clustering (Section 7.3.1) and outlier detection (Section 7.3.2).

7.3.1 CLUSTERING

The main objectives of (distance-based) clustering algorithms are to minimize intra-cluster distance and maximize inter-cluster distance. The skewness of k-occurrences in high-dimensional data influences both objectives.

25. We determined the σ values of the RBF function in a separate experiment involving 10 runs of random selection of points for the labeled set, the size of which is 10% of the original data set.

Intra-cluster distance may be increased due to points with low k-occurrences. As discussed in Section 4.4, such points are far from all the rest, acting as distance-based outliers. Distance-based outliers and their influence on clustering are well-studied subjects (Tan et al., 2005): outliers do not cluster well because they have high intra-cluster distance, thus they are often discovered and eliminated beforehand. The existence of outliers is attributed to various reasons (for example, erroneous measurements). Nevertheless, the skewness of N_k suggests that in high-dimensional data outliers are also expected due to inherent properties of vector space. Section 7.3.2 will provide further discussion on this point.

Inter-cluster distance, on the other hand, may be reduced due to points with high k-occurrences, that is, hubs. Like outliers, hubs do not cluster well, but for a different reason: they have low inter-cluster distance, because they are close to many points, thus also to points from other clusters. In contrast to outliers, the influence of hubs on clustering has not attracted significant attention.

To examine the influence of both outliers and hubs, we used the popular silhouette coefficients (SC) (Tan et al., 2005). For the ith point, let a_i be the average distance to all points in its cluster (a_i corresponds to intra-cluster distance), and b_i the minimum average distance to points from other clusters (b_i corresponds to inter-cluster distance). The SC of the ith point is $(b_i - a_i)/\max(a_i, b_i)$, ranging between -1 and 1 (higher values are preferred). The SC of a set of points is obtained by averaging the silhouette coefficients of the individual points.

We examined several clustering algorithms, and report results for the spectral algorithm of Meilă and Shi (2001) and Euclidean distance, with similar results obtained for classical K-means, as well as the spectral clustering algorithm by Ng et al. (2002) in conjunction with K-means and the algorithm by Meilă and Shi (2001). For a given data set, we set the number of clusters, K, to the number of classes (specified in Table 1). We select as hubs those points $\mathbf{x}$ with $h(\mathbf{x},k) > 2$, that is, $N_k(\mathbf{x})$ more than two standard deviations higher than the mean (note that $h(\mathbf{x},k)$, defined by Equation 24, ignores labels). Let n_h be the number of hubs selected. Next, we select as outliers the n_h points with the lowest k-occurrences. Finally, we randomly select n_h points from the remaining points (we report averages for 100 different selections). To compare hubs and antihubs against random points, we measure the *relative SC* of hubs (antihubs): the mean SC of hubs (antihubs) divided by the mean SC of random points. For several data sets from Table 1, Figure 16 depicts with bars the relative silhouette coefficients.[26] As expected, outliers have relative SC lower than one, meaning that they cluster worse than random points. Notably, the same holds for hubs, too.[27]

To gain further insight, Figure 16 plots with lines (referring to the right vertical axes) the relative mean values of a_i and b_i for hubs and outliers (dividing with those of randomly selected points). Outliers have high relative a_i values, indicating higher intra-cluster distance. Hubs, in contrast, have low relative b_i values, indicating reduced inter-cluster distance. In conclusion, when clustering high-dimensional data, hubs should receive analogous attention as outliers.

7.3.2 OUTLIER DETECTION

This section will briefly discuss possible implications of high dimensionality on distance-based outlier detection, in light of the findings concerning the hubness phenomenon presented in previous sections. Section 4.4 already discussed the correspondence of antihubs with distance-based outliers

26. Data sets were selected due to their high (intrinsic) dimensionality—similar observations can be made with other high-dimensional data sets from Table 1.
27. The statistical significance of differences between the SC of hubs and randomly selected points has been verified with the paired t-test at 0.05 significance level.

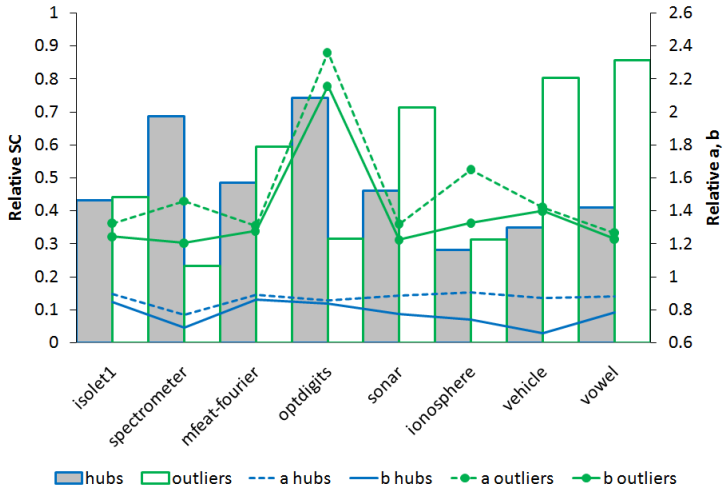


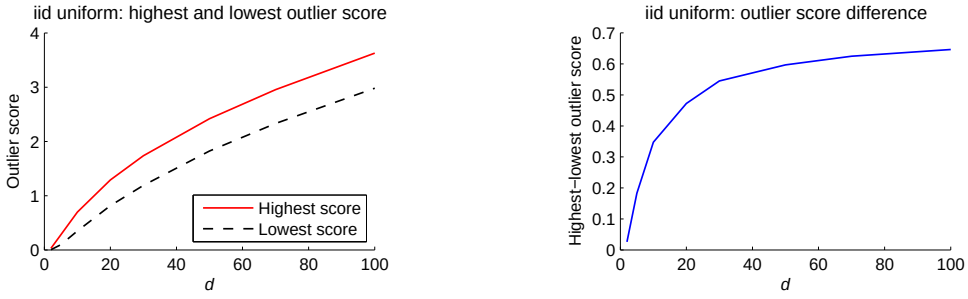
Figure 16: Relative silhouette coefficients for hubs (gray filled bars) and outliers (empty bars). Relative values for a and b coefficients are also plotted (referring to the right vertical axes).

in high dimensions, and demonstrated on real data the negative correlation between N_k and a commonly used outlier score—distance to the kth nearest neighbor. On the other hand, a prevailing view of the effect of high dimensionality on distance-based outlier detection is that, due to distance concentration, every point seems to be an equally good outlier, thus hindering outlier-detection algorithms (Aggarwal and Yu, 2001). Based on the observations regarding hubness and the behavior of distances discussed earlier, we believe that the true problem actually lies in the opposite extreme: high dimensionality induces antihubs that can represent "artificial" outliers. This is because, from the point of view of common distance-based outlier scoring schemes, antihubs may appear to be stronger outliers in high dimensions than in low dimensions, only due to the effects of increasing dimensionality of data.

To illustrate the above discussion, Figure 17(a) plots for i.i.d. uniform random data the highest and lowest outlier score (distance to the kth NN, $k = 5$) with respect to increasing d ($n = 10000$ points, averages over 10 runs are reported). In accordance with the asymptotic behavior of all pairwise distances discussed in previous sections, both scores increase with d. However, as Figure 17(b) shows, the *difference* between the two scores also increases. This implies that a point could be considered a distance-based outlier only because of high dimensionality, since outliers are not expected for any other reason in i.i.d. uniform data (we already demonstrated that such a point would most likely be an antihub). As a consequence, outlier-detection methods based on measuring distances between points may need to be adjusted to account for the (intrinsic) dimensionality of data, in order to prevent dimensionality-induced false positives.

8. Conclusion

In this paper we explored an aspect of the curse of dimensionality that is manifested through the phenomenon of hubness—the tendency of high-dimensional data sets to contain hubs, in the sense of popular nearest neighbors of other points. To the best of our knowledge, hubs and the effects they



(a) (b)

Figure 17: (a) Highest and lowest distance to the 5th nearest neighbor in i.i.d. uniform random data, with respect to increasing d. (b) The difference between the two distances.

have on machine-learning techniques have not been thoroughly studied subjects. Through theoretical and empirical analysis involving synthetic and real data sets we demonstrated the emergence of the phenomenon and explained its origins, showing that it is an inherent property of data distributions in high-dimensional vector space that depends on the intrinsic, rather than embedding dimensionality of data. We also discussed the interaction of hubness with dimensionality reduction. Moreover, we explored the impact of hubness on a wide range of machine-learning tasks that directly or indirectly make use of distances between points, belonging to supervised, semi-supervised, and unsupervised learning families, demonstrating the need to take hubness into account in an equivalent degree to other factors, like the existence of outliers.

Besides application areas that involve audio and image data (Aucouturier and Pachet, 2007; Doddington et al., 1998; Berenzweig, 2007; Hicklin et al., 2005), identifying hubness within data and methods from other fields can be considered an important aspect of future work, as well as designing application-specific methods to mitigate or take advantage of the phenomenon. We already established the existence of hubness and its dependence on data dimensionality on collaborative filtering data with commonly used variants of cosine distance (Nanopoulos et al., 2009), time-series data sets in the context of k-NN classification involving dynamic time warping (DTW) distance (Radovanović et al., 2010b), text data within several variations of the classical vector space model for information retrieval (Radovanović et al., 2010a), and audio data for music information retrieval using spectral similarity measures (Karydis et al., 2010). In the immediate future we plan to perform a more detailed investigation of hubness in the fields of outlier detection and image mining. Another application area that could directly benefit from an investigation into hubness are reverse k-NN queries (which retrieve data points that have the query point $\mathbf{q}$ as one of their k nearest neighbors, Tao et al., 2007).

One concern we elected not to include into the scope of this paper is the efficiency of computing N_k. It would be interesting to explore the interaction between approximate k-NN graphs (Chen et al., 2009) and hubness, in both directions: to what degree do approximate k-NN graphs preserve hubness information, and can hubness information be used to enhance the computation of approximate k-NN graphs for high-dimensional data (in terms of both speed and accuracy).

Possible directions for future work within different aspects of machine learning include a more formal and theoretical study of the interplay between hubness and various distance-based machine-

learning models, possibly leading to approaches that account for the phenomenon at a deeper level. Supervised learning methods may deserve special attention, as it was also observed in another study (Caruana et al., 2008) that the k-NN classifier and boosted decision trees can experience problems in high dimensions. Further directions of research may involve determining whether the phenomenon is applicable to probabilistic models, (unboosted) decision trees, and other techniques not explicitly based on distances between points; and also to algorithms that operate within general metric spaces. Since we determined that for K-means clustering of high-dimensional data hubs tend to be close to cluster centers, it would be interesting to explore whether this can be used to improve iterative clustering algorithms, like K-means or self-organizing maps (Kohonen, 2001). Nearest-neighbor clustering (Bubeck and von Luxburg, 2009) of high-dimensional data may also directly benefit from hubness information. Topics that could also be worth further study are the interplay of hubness with learned metrics (Weinberger and Saul, 2009) and dimensionality reduction, including supervised (Vlassis et al., 2002; Geng et al., 2005), semi-supervised (Zhang et al., 2007), and unsupervised approaches (van der Maaten et al., 2009; Kumar, 2009). Finally, as we determined high correlation between intrinsic dimensionality and the skewness of N_k, it would be interesting to see whether some measure of skewness of the distribution of N_k can be used for estimation of the intrinsic dimensionality of a data set.

Acknowledgments

We would like to thank the authors of all software packages, libraries, and resources used in this research: Matlab with the Statistics Toolbox, Dimensionality Reduction Toolbox (van der Maaten, 2007), the FastICA package (Gävert et al., 2005), GML AdaBoost Toolbox (Vezhnevets, 2006), code for semi-supervised learning with the basic harmonic function (Zhu et al., 2003), SpectraLIB package for symmetric spectral clustering (Verma, 2003; Meilă and Shi, 2001), MatlabBGL graph library (Gleich, 2008); the Weka machine-learning toolkit (Witten and Frank, 2005), and Wolfram Mathematica with invaluable online resources MathWorld (mathworld.wolfram.com/) and the Wolfram Functions Site (functions.wolfram.com/).

We would like to express our gratitude to the action editor, Ulrike von Luxburg, and the anonymous reviewers, for helping to significantly improve the article with their detailed and thoughtful comments. We are also indebted to Zagorka Lozanov-Crvenković for helpful discussions and suggestions regarding the mathematical aspects of the paper.

Miloš Radovanović and Mirjana Ivanović thank the Serbian Ministry of Science for support through project *Abstract Methods and Applications in Computer Science*, no. 144017A. Alexandros Nanopoulos gratefully acknowledges the partial co-funding of his work through the European Commission FP7 project *MyMedia* (www.mymediaproject.org/) under the grant agreement no. 215006.

References

Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables*. National Bureau of Standards, USA, 1964.

Charu C. Aggarwal and Philip S. Yu. Outlier detection for high dimensional data. In *Proceedings of the 27th ACM SIGMOD International Conference on Management of Data*, pages 37–46, 2001.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In *Proceedings of the 8th International Conference on Database Theory (ICDT)*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. Springer, 2001.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

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.

Richard E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

Adam Berenzweig. *Anchors and Hubs in Audio-based Music Similarity*. PhD thesis, Columbia University, New York, USA, 2007.

Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Proceedings of the 7th International Conference on Database Theory (ICDT)*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer, 1999.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.

Sébastien Bubeck and Ulrike von Luxburg. Nearest neighbor clustering: A baseline method for consistent clustering with arbitrary objective functions. *Journal of Machine Learning Research*, 10:657–698, 2009.

Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 96–103, 2008.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006.

Jie Chen, Haw ren Fang, and Yousef Saad. Fast approximate *k*NN graph construction for high dimensional data via recursive Lanczos bisection. *Journal of Machine Learning Research*, 10: 1989–2012, 2009.

Pierre Demartines. *Analyse de Données par Réseaux de Neurones Auto-Organisés*. PhD thesis, Institut national polytechnique de Grenoble, 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 5th International Conference on Spoken Language Processing (ICSLP)*, 1998. Paper 0608.

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 Alfréd Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

Paul F. Evangelista, Mark J. Embrechts, and Boleslaw K. Szymanski. Taming the curse of dimensionality in kernels and novelty detection. In A. Abraham, B. Baets, M. Koppen, and B. Nickolay, editors, *Applied Soft Computing Technologies: The Challenge of Complexity*, volume 34 of *Advances in Soft Computing*, pages 425–438. Springer, 2006.

Damien François. *High-dimensional Data Analysis: Optimal Metrics and Feature Selection*. PhD thesis, Université catholique de Louvain, Louvain, Belgium, 2007.

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–374, 2000.

Yasunori Fujikoshi. Computable error bounds for asymptotic expansions of the hypergeometric function ${}_1F_1$ of matrix argument and their applications. *Hiroshima Mathematical Journal*, 37(1): 13–23, 2007.

Hugo Gävert, Jarmo Hurri, Jaakko Särelä, and Aapo Hyvärinen. The FastICA package for Matlab. http://www.cis.hut.fi/projects/ica/fastica/, 2005.

Xin Geng, De-Chuan Zhan, and Zhi-Hua Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 35(6):1098–1107, 2005.

David Gleich. MatlabBGL: A Matlab graph library. http://www.stanford.edu/∼dgleich/programs/matlab_bgl/, 2008.

Uffe Haagerup. The best constants in the Khintchine inequality. *Studia Mathematica*, 70:231–283, 1982.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

Austin Hicklin, Craig Watson, and Brad Ulery. The myth of goats: How many people have fingerprints that are hard to match? Internal Report 7271, National Institute of Standards and Technology (NIST), USA, 2005.

Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. What is the nearest neighbor in high dimensional spaces? In *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB)*, pages 506–515, 2000.

Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 15*, pages 833–840, 2003.

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.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

Khandoker Tarik-Ul Islam, Kamrul Hasan, Young-Koo Lee, and Sungyoung Lee. Enhanced 1-NN time series classification using badness of records. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, pages 108–113, 2008.

Kiyosi Itô, editor. *Encyclopedic Dictionary of Mathematics*. The MIT Press, 2nd edition, 1993.

Tony Jebara, Jun Wang, and Shih-Fu Chang. Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 441–448, 2009.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, 2nd edition, 1994.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

Ioannis Karydis, Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Looking through the "glass ceiling": A conceptual framework for the problems of spectral similarity. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and K. R. Krishna Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13(3): 637–649, 2001.

Teuvo Kohonen. *Self-Organizing Maps*. Springer, 3rd edition, 2001.

Filip Korn, Bernd-Uwe Pagel, and Christos Faloutsos. On the "dimensionality curse" and the "self-similariy blessing". *IEEE Transactions on Knowledge and Data Engineering*, 13(1):96–111, 2001.

Ch. Aswani Kumar. Analysis of unsupervised dimensionality reduction techniques. *Computer Science and Information Systems*, 6(2):217–227, 2009.

Stéphane Lafon and Ann B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–1403, 2006.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17*, pages 777–784, 2005.

Lun Li, David Alderson, John C. Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.

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

Marina Meilă and Jianbo Shi. Learning segmentation by random walks. In *Advances in Neural Information Processing Systems 13*, pages 873–879, 2001.

Boaz Nadler, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

Alexandros Nanopoulos, Miloš Radovanović, and Mirjana Ivanović. How does high dimensionality affect collaborative filtering? In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys)*, pages 293–296, 2009.

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.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856, 2002.

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

Andrew M. Odlyzko and Neil J. A. Sloane. New bounds on the number of unit spheres that can touch a unit sphere in *n* dimensions. *Journal of Combinatorial Theory, Series A*, 26(2):210–214, 1979.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 271–278, 2004.

Mathew Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.

John C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods—Support Vector Learning*, pages 185–208. MIT Press, 1999.

Miloš Radovanović and Mirjana Ivanović. Document representations for classification of short Web-page descriptions. In *Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, volume 4081 of *Lecture Notes in Computer Science*, pages 544–553. Springer, 2006.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 865–872, 2009.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 186–193, 2010a.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Time-series classification in many intrinsic dimensions. In *Proceedings of the 10th SIAM International Conference on Data Mining (SDM)*, pages 677–688, 2010b.

Gunnar Rätsch, Takashi Onoda, and Klaus-Robert Müller. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, 2001.

Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

John P. Scott. *Social Network Analysis: A Handbook*. Sage Publications, 2nd edition, 2000.

Amit Singh, Hakan Ferhatosmanoğlu, and Ali Şaman Tosun. High dimensional reverse nearest neighbor queries. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM)*, pages 91–98, 2003.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2005.

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

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

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. An introduction to dimensionality reduction using Matlab. Technical Report MICC 07-07, Maastricht University, Maastricht, The Netherlands, 2007.

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

Deepak Verma. SpectraLIB – package for symmetric spectral clustering. http://www.stat.washington.edu/spectral/, 2003.

Alexander Vezhnevets. GML AdaBoost Matlab Toolbox. MSU Graphics & Media Lab, Computer Vision Group, http://graphics.cs.msu.ru/, 2006.

Nikos Vlassis, Yoichi Motomura, and Ben Kröse. Supervised dimension reduction of intrinsically low-dimensional data. *Neural Computation*, 14(1):191–215, 2002.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 2nd edition, 2005.

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.

Kenneth Zeger and Allen Gersho. Number of nearest neighbors in a Euclidean code. *IEEE Transactions on Information Theory*, 40(5):1647–1649, 1994.

Daoqiang Zhang, Zhi-Hua Zhou, and Songcan Chen. Semi-supervised dimensionality reduction. In *Proceedings of the 7th SIAM International Conference on Data Mining (SDM)*, pages 629–634, 2007.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 912–919, 2003.