# The Hubness Phenomenon in High Dimensional Spaces

Name of First Author and Name of Second Author

**Abstract** Each chapter should be preceded by an abstract (10–15 lines long) that summarizes the content. The abstract will appear *online* at www.SpringerLink.com and be available with unrestricted access. This allows unregistered users to read the abstract as a teaser for the complete chapter. As a general rule the abstracts will not appear in the printed version of your book unless it is the style of your particular book or that of the series to which your book belongs.

Please use the 'starred' version of the new Springer `abstract` command for typesetting the text of the online abstracts (cf. source file of this chapter template `abstract`) and include them with the source files of your manuscript. Use the plain `abstract` command if the abstract is also to appear in the printed version of the book.

## 1 Introduction

[Describe the phenomenon, recent observations, and open questions we are investigating here]

Name of First Author
Name, Address of Institute, e-mail: name@email.address

Name of Second Author
Name, Address of Institute e-mail: name@email.address

## 2 Background and Related Work

## 3 Intrinsic Dimensionality via Hubness

### *3.1 Skewness vs. Feature Ranking - What is the Right Way to Rank Features?*

[Local Skewness plots, etc.]

### *3.2 Supervised vs. Unsupervised Methods*

[Compare methods for feature relevance, cluster properties in original and subspace, etc.]

## 4 Hubs, Density, and Clustering

To investigate the relationship between density, hubness scores, and clustering, data sets of 30, 60, and 100 relevant dimensions were created. For each of these dimensions, two types of data sets were generated: (1) two Gaussian spheres separated by 10 units in the first coordinate and (2) two uniform cubes separated by one unit along its first dimension. In order to emulate varying densities, one cluster was designed to contain 1000 points and the second varied in size (2000, 3000, 4000, and 5000). There was a resulting total of 24 data sets given the different dimensions, distribution of points, and varying densities. An additional non-convex data cloud in 60 dimensions composed of two close by clusters with similar density was created to measure clustering potential. For the rest of the chapter, the Gaussian data sets will be denoted as $G_{d,N}$, the uniform ones as $U_{d,N}$, where $d$ is the dimension and $N$ is the number of points, and the non-convex data cloud as $S_{60,6000}$. In general, any $d$-dimensional data of $N$ points will be denoted as $X_{d,N}$.

### *4.1 Hubness and Data Density*

The first question to be addressed is whether hubness scores and density are the same measure. Suppose that there is a low density area, high density area, a low hubness score area and a high hubness score area, as shown in Fig. 1. Intuition dictates that if density and hubness score were correlated, all the points will lie within the first (high density and high hubness scores) and third (low density and low hubness scores) quadrants. However, if the points are all over the four quadrants, then it might imply a more complicated relationship between density and hubness scores.

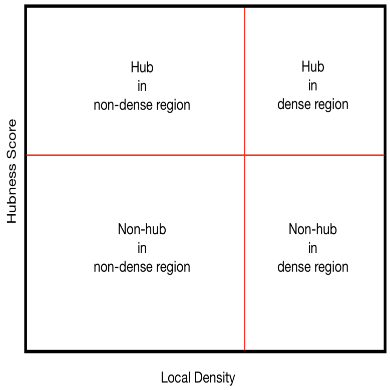
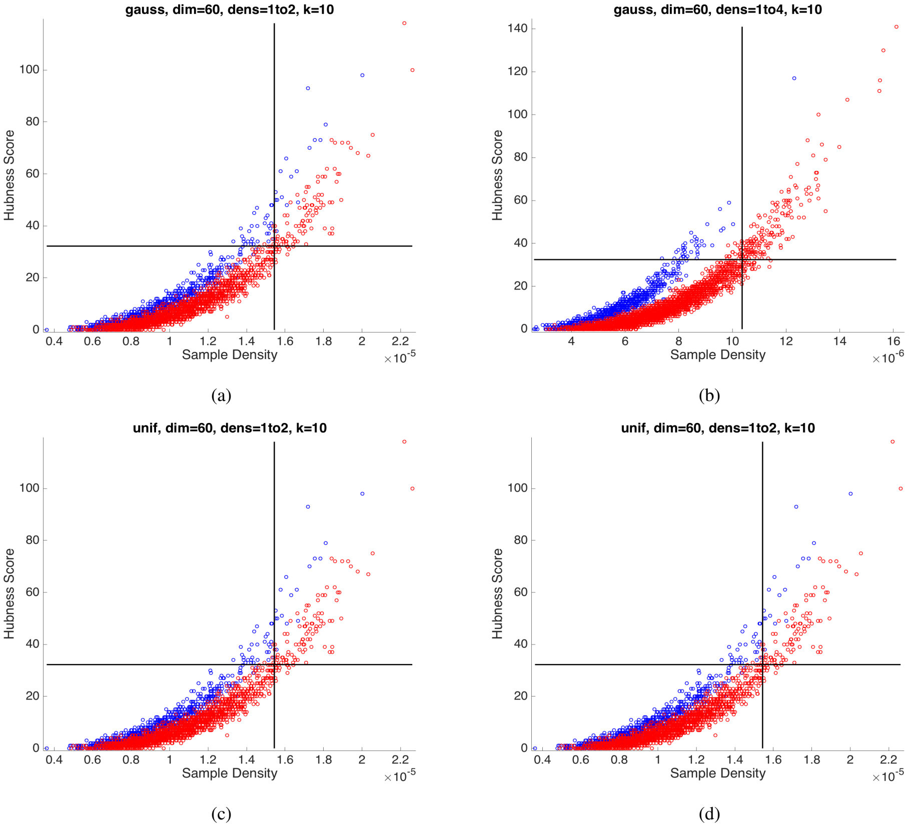


**Fig. 1** An illustration of the four quadrants, which are formed by density and hubness scores thresholds, that will be seen in the following experiments where density will be plotted against hubness scores.

To estimate density, we used the Kernel Density Estimator described in [**?**] with number of nearest neighbors being 100 and the bandwidth nearest neighbors being 16. The KDE is known to be an accurate consistent estimator. To determine what is “high density” a threshold value of two standard deviations above the mean was calculated. Also, the hubness scores $N_k$ were calculated for all the data sets with $k = 5, 10$, and 50. Taking into account the different types of data, densities, dimensions and $k$’s, there were 72 total plots.

The first set of experiments used a global threshold for the density, i.e. the mean and standard deviation were calculated for the density of all the points regardless of class membership. Some representative results are shown in Figure **??**. In these figures, global thresholds for density and hubness scores were plotted in black. The blue and red represent the class membership of each point, which is known from how the data was created.

While in these results points actually appeared in all four quadrants, there is a strong positive correlation, shown in Figure **??**. In all 72 experiments, there is an average of 94.91% of the data that lies in the third quadrant (non-hub and low density), which is due to how we defined the thresholds. The interesting observations come from the proportions of data in the other quadrants. For the uniform data sets, 72.26% of all the hubs were in the dense region, while for the gaussian data sets, it is 82.34%. Similarly, 78.82% and 82.64% of the “dense” points in the gaussian and uniform data sets, respectively, are also hubs. Therefore hubs are most likely in dense regions and dense regions are largely composed of hubs. An interesting fact to note is that another density estimator is one over the distance to the $k$-nearest neighbor, which is related to the hubness score. This fact and the evidence from our



(a)

(b)

(c)

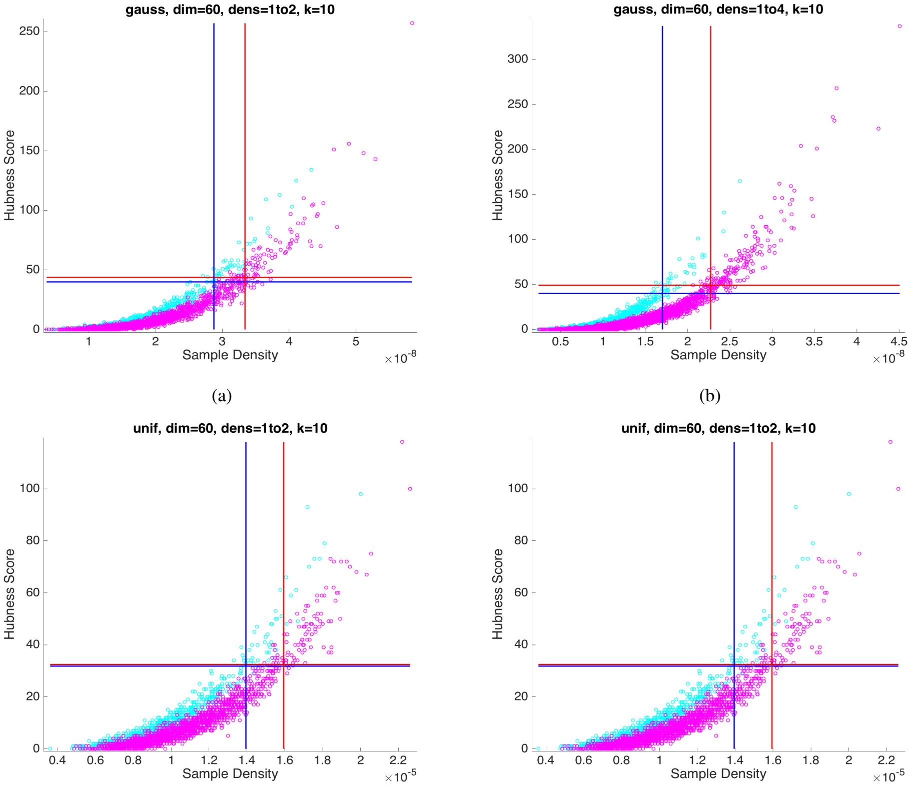
(d)

**Fig. 2** KDE, $\hat{q}$, plotted against hubness score, $N_{10}$, for data sets (a) $G_{60,3000}$, (b) $G_{60,5000}$, (c) $U_{60,3000}$, and (d) $U_{60,5000}$. The true class is depicted in the two colors of the points and the global thresholds are plotted in black.

experiments, we hypothesize that hubness will be closely related to density and in the future will be trying to mathematically express and prove this relationship.

The same experiments were performed but this time with “local thresholds”, meaning that the same $N_k$ and $\hat{q}$ were used, but this time the mean and standard deviation were calculated for each class. The thresholds for the light blue cluster were plotted in dark blue and the ones for the pink one were plotted in red. Some of the resulting plots can be seen in Fig 3. For these experiments, the same but stronger pattern was observed. Taking into account all the experiments, an average of 85% of the hubs were in the high density region, and 84.74% of the points in the high density region were hubs. The standard deviation for both of these averages was about 0.07. This again supports that hubness scores and density are closely related.

While the local thresholds are informative since it removes the effects of the difference in densities, in the real world one can only see the global thresholds. Since hubs are being used to do clustering procedures, it is reassuring to see how



(a)

(b)

(c)

(d)

**Fig. 3** KDE, $\hat{q}$, plotted against hubness score, $N_{10}$, for data sets (a) $G_{60,3000}$, (b) $G_{60,5000}$, (c) $U_{60,3000}$, and (d) $U_{60,5000}$. The true class is depicted in the two colors: pink and blue. Class thresholds are plotted in red (for the pink class) and blue (for the blue class).

even though high density regions can lack points from the smaller cluster, hubs appear from both clusters. Also, there is not a big difference between the "local" and "global" thresholds, so this gives a sense of "robustness" to density that needs to be further explored and exploited.

Another interesting observation from all the experiments, both the local and global thresholds, are the "wings" formed by the two clusters. The experiments show that the bigger the difference in densities between the two clusters, the more of a gap there is, and hence forming the "wings." Due to the curse of dimensionality, these wings are more common in the 30 and 60 dimensional data than they are in the 100 dimensional data. However, it would be interesting to further explore what gives rise to the "wings" and under what other conditions these are seen.

[Distance plots, Correlation of hubs to density, etc.]

### 4.2 Hubness and Purity

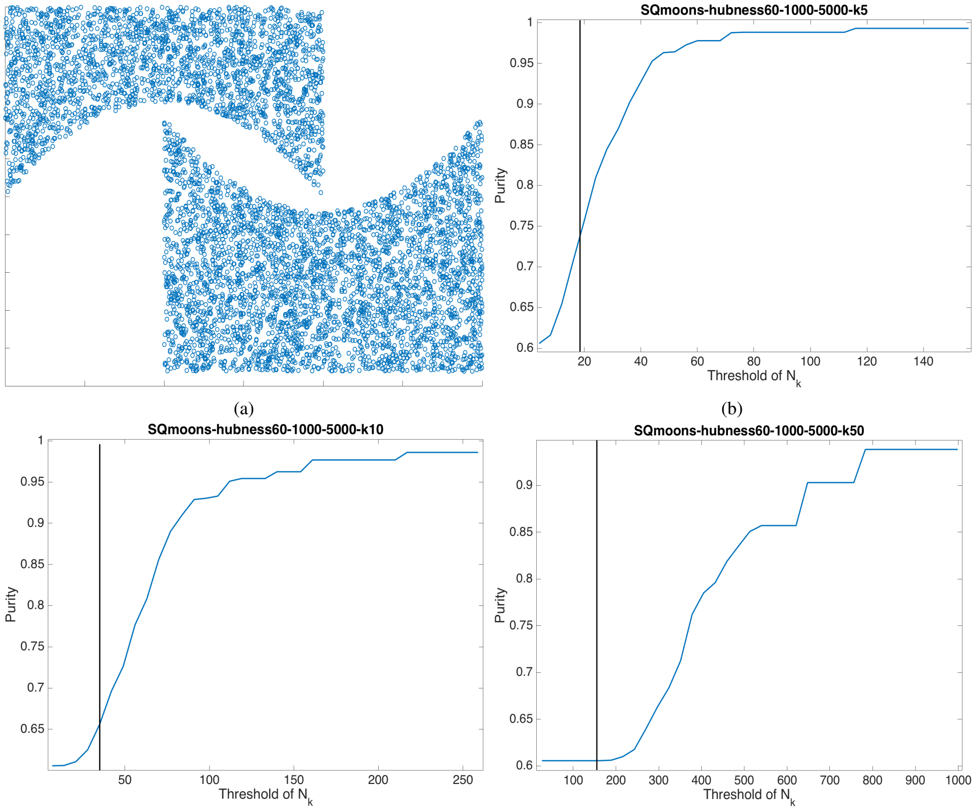
Hubs have the potential to be used for clustering. For example, it is possible to cluster only the hubs and then assign their corresponding reverse $k$-nearest neighbors (RkNN), or those points that have hubs as their $k$-nearest neighbors, the same labeling as the hubs. Hence, we explored the relationship between hubness score and purity, where the latter is defined as:

$$\text{purity } = \frac{1}{N}\sum_{i}\max_{j}|\hat{C}_i \cap C_j|$$

where $\mathscr{C} = \{\hat{C}_1, \cdots, \hat{C}_p\}$ is the set of calculated clusters and $\mathbb{C} = \{C_1, \cdots C_q\}$ is the set of true classes in the data.

To explore the relationship, all the points in the dataset which have $N_k > \lambda_i$ were chosen and their corresponding true labels were recorded. For our experiments, $\lambda_i$ is iteratively increased by half a standard deviation of $N_k$. Also, the true labels $\mathscr{C}$ are used in order to avoid calculation errors, as this is an exploratory experiment. For each point that passed the hubness score threshold, their RkNNs were assigned the same cluster label, and their purity was calculated.

The plots for these experiments in Figure 4 show the purity levels given threshold values of $N_k$, and the black vertical line show 2 standard deviations away from the mean of $N_k$. For the Gaussian data sets, purity was 1 for 25 experiments, $1-\mathscr{O}(10^{-4})$ for 8 experiments, and $1-\mathscr{O}(10^{-3})$ for the remaining 3, which all corresponded to 100 dimensions and $k = 50$. For the Uniform data sets, purity was 1 for all except 2 experiments, both of which had purity values of $1-\mathscr{O}(10^{-4})$. The results of $G_{d,N}$ and $U_{d,N}$ are not very surprising since in the convex and well separated cases, the k-nearest neighbors are going to be other points in the same cluster. Therefore, good training labels on these types of data will lead to high purity using something as simple as kNN. There was a more interesting behavior for the non-convex data, shown in Figure 4(a). While the Gaussian and Uniform plots are not all monotonic increasing, it turns out that the non-convex data set does show a monotonic increasing relation between the $N_k$ threshold and the purity level. This shows that the $\frac{1}{N}$ from the purity measure is influencing more the non-covex data set than the convex ones. As expected, the non-convex data sets showed lower purity than the convex ones, which can be seen in the plots that never reached purity level of 1. Also, the non-convex data shows the advantage of taking $k$ to be a small number since there is a better purity for $k = 5$ compared to $k = 50$ and even for $k = 10$. It is interesting to note the steep increase that seems to be delayed as $k$ increases, which is something to be further studied. Looking at Figure 4(b)-(d), there are two flat regions that look like two steps towards the end of the domain. The first one corresponds to when all the hubs selected only come from one cluster and the flat region at the end corresponds to when only one point is higher than the threshold selected. Since the goal is to cluster, the interesting region is the one before these plateaus. In this region, purity levels higher than 85% appear well above the threshold for hubs, which correspond to the black lines in Figure 4. This means that a better threshold value needs to be chosen than the one that currently defines the “hubs.”



(a) (b)

(c) (d)

**Fig. 4** (a) The first 2 coordinates of the non-convex data set. The hubness score with $k$ = (a)5, (b)10, (c)50 plotted against purity. The black vertical line shows 2 standard deviations above the mean of the hubness score so that all the values to the right will be considered hubs.

## 5 Discussion

## 6 Conclusion and Future Work