**The following are notes that I (Marilyn) made while doing all the experiments of density vs. hubness scores. I am also working on the purity experiments. Note that I still need to make this into an “article” style writing. Also, I was trying to use Latex notation for easy conversion.**

To investigate the relationship between density and hubness, we created 30,60, and 100 intrinsic dimensional data. For each dimension, we formulated two types of data sets: (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. For each type of data set, we made one of the clusters be 1000 points and the other of varying sizes (2000, 3000, 4000, and 5000) to emulate varying densities. Hence, we have a total of 24 data sets.

For our experiments, we will denote a data set from the Gaussian spheres as $G\_{d,N}$ and a data set from the Uniform cubes as $U\_{d,N}$, where $d$ is the dimension and $N$ is the number of points. In general, any $d$-dimensional data of $N$ point will be denoted as $X\_{d,N}$.

Define the $k$-nearest neighbor of some data point $x$, denoted $x\_k$, as the kth point from $x$ not including itself. One possible sample density estimator can be defined as $q\_k(x) = \frac{1}{||x-x\_k||}$. We will use KDE since the other one is too closely related to N\_k and could cause to over evaluate the relationship between density and hubness score.

General observations:

1. G\_{d,N} – If you hold dimension and density constant, and plot $k$ against the number of hubs in the data, you can see a monotonic increase for all of them excluding G\_{30,6000} and G\_{60,3000}. These data sets present a dip. Figures available: G\_{30,6000} (one of the dips), G\_{60,6000} (monotonic increase example), G\_{100,5000} (strictly monotonic example)
2. U\_{d,N} – Using the same formulation as in number (1) but on the uniform data, you can observe 7 of the 12 experiments showed the same monotonic increase as the Gaussians, but four (U\_{30,3000},U\_{30,4000}, U\_{100,4000},U\_{100,6000}) had a dip and one(U\_{30,6000}) had an inverse dip. Figures available: U\_{30,6000} (the reverse dip), U\_{60,3000} (monotonic increase example), U\_{100,4000} (a dip example)

The plots were made using $ka=16,kL=100$ for the density estimation, $\hat{q}$. We calculated the hubness score, or $N\_k$, using $k=10$ because for most of the experiments, not the most lenient results (i.e. it is not a maxima) and because the closest to the $ka$ chosen for $\hat{q}$ calculations. Since the points come from synthetic data, the color corresponds to the true labels. Also, 60-dimensional data was chosen for simplicity.

Observations about the global hubs vs. global sample density

1. Both, the hubness threshold (horizontal black line) and the density threshold (vertical black line) were calculated by adding 2 standard deviations to the mean of values, i.e. $mean(\hat{q} (X\_{d,N}))+2\*std(\hat{q}(X\_{d,N}))$ and $mean(N\_k(X\_{d,N}))+2\*std(N\_k(X\_{d,N}))$.
2. Note the strong positive correlation.
3. Unif Mean proportions (top left, top right, bottom left, bottom right): 1.2304, 3.1762, 94.9089, 0.6845
4. Unif Corresponding standard deviations (top left, top right, bottom left, bottom right): 0.2451, 0.2689, 0.4075, 0.3151
5. Gauss Mean proportions (top left, top right, bottom left, bottom right): 0.7494, 3.4130, 94.9131, 0.9245
6. Gauss Corresponding standard deviations (top left, top right, bottom left, bottom right): 0.2401, 0.3001, 0.5020, 0.2784
7. Gauss - All 36 experiments had a larger proportion of hubs in the high density region: hubs are more likely to be in a low density region are not exactly the same thing.
8. Unif - All 36 experiments had a larger proportion of hubs in the high density region: hubs are more likely to be in a low density region are not exactly the same thing.
9. All 72 experiments show that high density regions are mostly composed of hubs: if a point is in a high density region, it is most likely a hub.
10. Figures available: G\_{60,3000}, G\_{60,5000}, U\_{60,3000}, U\_{60,5000}
11. Questions: can we always find a hub of all clusters in the high density region?

Observations about the local hubs vs. local sample density

NOTE: I just discovered a mistake! Let me see how I solve this

NEVERMIND! It was not a mistake. This is all correct!

1. The same procedure as before was used to plot the threshold lines in these plots. However, for each cluster, instead of using all of the points, we only used the points in that particular cluster. For example, for cluster 1, we only used $q\_{15}(1:1000)$ for the density threshold calculation and $N\_{k}(1:1000)$ for the hubness threshold calculation
2. Gauss – When local hubs were calculated, similar results resulted to the global hubs: In about 64% of the experiments, most of the hubs landed in the high density region
3. Unif – In these data sets local hubs actually made a difference. While the global hubs were mainly in the global non-dense regions, for about 85% of the experiments, most of the local hubs actually lie in the local “dense” region.
4. Figures available: G\_{60,3000}, G\_{60,5000}, U\_{60,3000}, U\_{60,5000}

To answer the questions on purity and reverse nearest neighbors, let us first define these two terms. Let $a$ be the distance of $x \in X\_{d,N}$ to its kth nearest neighbor in data set $Y\_{d,M}$. Then the reverse k-nearest neighbors of $x$ can be defined as $ r\_k(x,Y\_{d,M})= \{ y\in Y\_{d,M}| ||x-y|| \leq a \}$. Note that for some computations, it is possible for $ X\_{d,N} = Y\_{d,N}$. Now, let $\mathcal{C} = \{\hat{C}\_1,\cdots,\hat{C}\_p\}$ be the set of clusters found via some method. Let $\mathbb{C} = \{C\_1,\cdots C\_q\}$ be the set of true classes of the data. Then purity can be defined as:

\[

\mbox{purity } = \frac{1}{N} \sum\_i \max\_j |\hat{C}\_i \cap C\_j|

\]

The first question deals with the relationship between purity and hubness score. Suppose that you chose all the points in the data set such that $N\_k > \lambda$ and that you know, either from a ground truth or clustering method (in our case we know the ground truth), the labels for these. Hence, for each point $x \in X\_{d,N}$ that passes the threshold, we have N\_k and a cluster label c\_i, and we can calculate r\_k(x,X\_{d,N}) and assign these points the cluster label c\_i. Once that is done for all points that passed the threshold, we can compute purity.

NOTE: I’m trying to resolve the question of a point being the reverse knn for two points in different clusters. Never mind! This actually doesn’t matter since we are calculating purity at each $\lambda$

Observations about Purity and Hubness Score:

1. Since we are dealing with reverse nearest neighbors, important factors that make sense will come into play are: dimension (since distance is involved), density (since the more dense, the closer the k nearest neighbors of a point will be), and $k$ (since a bigger $k$ could potentially bring points not in the cluster as part of a knn)
2. To give interesting results, I made a new data set for this particular test, which I called SQmoons-dim60-6000 and will denote as $S\_{60,6000}$. The two clusters for each of these data sets have about the same density.
3. The plots for these experiments show the purity levels given threshold values of $N\_k$ and the black vertical line show 2 standard deviations away from the mean of $N\_k$
4. Purity was 1 for all except one the experiments ($U\_{100,5000}$ with $k=50$ which started being 1 from 0.116 std’s and difference was $\mathcal{O}^{-6}$) of the data sets $U\_{d,N}$, no matter the combination of dimension, density, $k$ or threshold value of $N\_k$
5. Purity was 1 for almost all the experiments of the data sets $G\_{d,N}$, except for $G\_{60,3000}$, $G\_{60,4000}$,$G\_{60,5000}$, $G\_{100,3000}$, $G\_{100,4000}$, $G\_{100,5000}$, $G\_{100,6000}$, all with $k = 50$
6. For the $G\_{d,N}$ experiments that didn’t have purity=1, the difference was in $\mathcal{O}(10^{-5})$ and most of which (except for $G\_{100,5000}$ with $k=50$, which occurred at 2.08 std’s) occurred before 2 standard deviations (typically where we put the hub threshold)
7. However the $S\_{60,6000}$ data set shows how problems can arise if the two clusters are near each other.
   1. First thing to notice is that for $k$ too large, purity=1 is never achieved
   2. If $k$ is too large, it is not clear that a larger $N\_k$ value will give better purity results, as shown in the plot with $k=10$; however, there is this “leveling out” to the maximum value for all three experiments
   3. For $k=10$ and $k=50$, it takes about 16.5 standard deviations above the mean to achieve the level out into the maximum purity level, and for $k=5$, it takes about 11.4 std’s
   4. However, for $k$ small enough, purity achieves the value of 1, and there is a slight upward trend
   5. The question is then, what is “small enough”? do we care about finding an “optimal” $k$, or are we good with a particular one that seems to be working?

**Note: Figures are in the GitHub repository in: doc/icermfinal/fig.**