BOUNDARY COMPENSATED k-NN GRAPHS

Kumar Sricharan[1*], *Raviv Raich*[2], *Alfred O. Hero III*[1]

[1] Department of EECS, University of Michigan, Ann Arbor, MI 48109
[2] School of EECS, Oregon State University, Corvallis, OR 97331
{kksredddy,hero}@umich.edu, raich@eecs.oregonstate.edu

ABSTRACT

The k-nearest neighbor (k-NN) graph conveys local geometry of points in a sample. This attribute has resulted in a wide variety of machine learning applications for k-NN graphs, for e.g., density estimation, manifold learning and non-parametric classification. For samples with finite support, our analysis shows that k-NN density estimators behave differently in the interior of the support as opposed to near the boundary of the support. Motivated by our analysis, we propose improving the behavior of k-NN graphs by thinning its edges near the boundary. We illustrate the advantages of such boundary corrected k-NN graphs for entropy estimation and classification.

1. INTRODUCTION

The k-nearest neighbor (k-NN) graph conveys local geometry of points in a random sample. This attribute has resulted in a variety of applications of k-NN graphs including k-NN based clustering and classification, entropy estimation [1, 2] and intrinsic dimension estimation for manifold learning [3, 4, 5]. This paper introduces a new method to compensate for the bias that occurs when the support of the underlying multivariate density function has a finite boundary.

Consider a large random sample from a continuous multivariate density that is zero outside a bounded region, which is the support of the density. When one constructs the k-NN graph on such a sample the local neighborhoods of the graph behave differently near the boundary of the support. For points well inside the boundary, the k-NN neighbors will be spread almost uniformly around the point. On the other hand, for points close to the boundary of the support, the k-NN neighbors are disproportionately distributed away from the boundary. This phenomenon becomes more striking as the dimension of the multivariate density increases. As a result, the radius of the k-NN neighborhoods tend to be disproportionately larger near the boundary as compared to neighborhoods in the interior. These ideas will be formalized in Section 2 using analysis of the bias of k-NN density estimates.

The bias of finite supported density estimator performance has been previously studied in [6, 7] for kernel density estimates. Corrections have been suggested, primarily for the univariate case. These corrections also assume that the support is known apriori. We perform an similar analysis and propose compensating for the bias of k-NN density estimates for general multivariate data without any prior knowledge of the support of the density.

Motivated by our analysis of k-NN density estimators, we suggest a corrected version of general k-NN graphs which compensates for k-NN graph behavior near the boundary of the support. This general k-NN graph compensation method is applied to several machine learning applications including entropy estimation and k-NN classification.

*Acknowledgement: This work is partially funded by the Air Force Office of Scientific Research, grant number FA9550-09-1-0471.

2. RELATION BETWEEN k-NN DENSITY ESTIMATE AND k-NN GRAPHS

Let $\mathbf{X}_1, .., \mathbf{X}_M$ denote M i.i.d realizations of the density f. Consider a k-NN graph constructed on these M samples. Let $\mathbf{d_k}(\mathbf{X_i})$ denote the Euclidean distance between $\mathbf{X}_i$ and its k-th nearest neighbor amongst $\mathbf{X}_1, .., \mathbf{X}_M$.

The *k-NN* density estimator [8] is defined below. The k-NN region is given by $\mathbf{S_k}(X) = Y : d(X,Y) \leq \mathbf{d_k}(X)$ where $d(X,Y)$ is the Euclidean distance metric between X and Y. The volume of the k-NN region is then given by $\mathbf{V_k}(X) = \int_{\mathbf{S_k}(X)} dZ$. The *$k$-NN* density estimator is then defined as

$$\hat{\mathbf{f}}_{\mathbf{k}}(X) = \frac{k-1}{M\mathbf{V_k}(X)}. \quad (1)$$

We therefore have an *equivalence* relation between a k-NN graph, and the k-NN density estimates constructed using the graph. To correct for boundary effects in the graph, we first analyze boundary effects in the k-NN density estimates and then use this equivalence to specify corrections to the graph.

Throughout the rest of this paper, we focus on the regime where the radius of the k-NN ball (which is $O((k/M)^{1/d})$ [9]) is small. This regime is equivalent to having a large number of samples relative to the dimension d. We note that k-NN methods will work poorly in high-dimensional spaces under small sample sizes and the above operating regime is necessary for k-NN methods to be effective. The consistency of k-NN estimation depends on the assumption that the size of the k-NN neighborhood becomes small relative to the modulus of continuity of the underlying probability density that generates the points. Thus one generally requires a large number of samples before the small estimation error behavior of a consistent estimator kicks-in. Specifically, as compared to low dimension sample space, for high dimensional samples one needs an exponentially greater number of samples to achieve equivalent

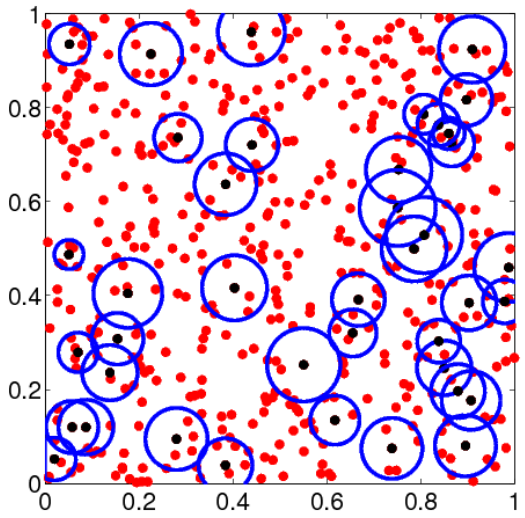

Fig. 1. k-NN balls centered around a subsample of 2D uniformly distributed points. Note that the k-NN balls centered at points close to boundary are truncated by the boundary.

bias. This follows from the fact that k-NN methods [9] require that $(k/M)^{1/d} \to 0$ and $k \to \infty$ for consistency and that the optimal rate is obtained by equalizing $(k/M)^{2/d}$ and $1/k$.

3. BIAS OF k-NN DENSITY ESTIMATES

In this section, we analyze the bias of the k-NN density estimates. We show that the bias decays to 0 as $k/M \to 0$ in the interior of the density. On the other hand, we show that the bias at the support of the density does not converge to zero.

3.1. Taylor series expansion of coverage probability

The coverage probability is defined as $\mathbf{P}(X) = \int_{S_k(X)} f(z)dz$. Assume that the density f has continuous partial derivatives of third order in the k-NN neighborhood $\mathbf{S_k}(X)$ of X. Note that this assumption only needs to hold for points in the interior of the support region. For small volumes $\mathbf{V_k}(X)$ we can then represent the corresponding coverage function $\mathbf{P}(X)$ by expanding the density f in a Taylor series about X [9].

$$\begin{aligned} \mathbf{P}(X) &= \int_{\mathbf{S_k}(X)} f(z)dz \\ &\approx f(X)\mathbf{V_k}(X) + c(X){\mathbf{V_k}}^{1+2/d}(X) \\ \implies \frac{1}{\mathbf{V_k}(X)} &\approx \frac{f(X)}{\mathbf{P}(X)} + \frac{c(X)f^{-2/d}(X)}{\mathbf{P}^{1-2/d}(X)}. \end{aligned} \quad (2)$$

where $c(X) = \Gamma^{(2/d)}(\frac{n+2}{2})tr[\nabla 2(f(X))]$.

3.2. Bias of k-NN density estimator in the interior

It is easy to obtain the bias of k-NN density estimates using the fact that $\mathbf{P}(X)$ has a beta distribution with parameters k, $M-k+1$ [9].

$$\mathbb{E}[\hat{\mathbf{f}}_\mathbf{k}(X)] - f(X) \approx h(X)\left(\frac{k}{M}\right)^{2/d}. \quad (3)$$

3.3. Bias of k-NN density estimator near boundary

If a probability density function has bounded support, the k-NN balls centered at points close to the boundary are often are truncated at the the boundary as shown in Fig. 1. Let

$$\alpha(X) = \frac{\int_{\mathbf{S_k}(X)\cap\mathcal{S}} dZ}{\int_{\mathbf{S_k}(X)} dZ} \quad (4)$$

be the fraction of the volume of the k-NN ball inside the boundary of the support. For interior points, $\alpha(X) = 1$, while for boundary points $\alpha(X)$ can range between 0 and 1, with $\alpha(X)$ closer to 0 when the points are closer to the boundary. For boundary points we then have

$$\mathbb{E}[\hat{\mathbf{f}}_\mathbf{k}(X)] - f(X) \approx (1-\alpha(X))f(X). \quad (5)$$

We therefore see that the bias is much higher at the boundary of the support ($O(1)$) as compared to the interior ($O((k/M)^{2/d})$). Furthermore, the bias at the support does not decay to 0 as $k/M \to 0$.

3.4. Variance of k-NN density estimator

The variance in the interior of the density was shown to be $O(1/k)$ [9]. Identical analysis will reveal that, unlike the bias, the variance will continue to decay at the same rate $O(1/k)$ at the boundary as well. This continues to hold for all higher central moments. This implies that the correction has to account only for the discrepancy in bias and does not have to account for the higher central moments. We suggest a method to correct for the high bias at the boundaries in the next section.

4. BOUNDARY CORRECTED k-NN DENSITY ESTIMATES

We formally define *boundary points* to be the set of points where the k-NN ball is truncated by the boundary of the support of the density. In this section, we suggest a simple way to compensate for this problem. A correction is performed in two stages: (i) Identification of boundary points and (ii) Correction of density estimates at these boundary points. We first establish concentration inequalities on the size of the k-NN ball.

4.1. Concentration inequality for k-NN volume

Consider a binomial random variable with parameters M and P with distribution function $Bi(.|M,P)$ and a beta random variable with parameters k and $M-k+1$ with distribution function $Be(.|k,M-k+1)$. We have the following identity,

$$Be(P|k,M-k+1) = 1 - Bi(k-1|M,P). \quad (6)$$

Using standard Chernoff bounds for binomial r.v.'s and the above relation, the fact that $\mathbf{P}(X)$ has a beta distribution and the relation between $\mathbf{P}(X)$ and $\mathbf{V_k}(X)$, we have the following concentration inequalities on the volume of the k-NN ball: for some $0 < p < 1/2$,

$$Pr\left(\left|\mathbf{V_k}(X) - \frac{k-1}{Mf(X)}\right| \leq p\frac{k-1}{Mf(X)}\right) \leq e^{-\frac{p^2 k}{4}}. \quad (7)$$

4.2. Boundary point detection

Denote the set of M i.i.d. realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$ from the density f by $\mathcal{X}$. Denote the set of boundary points by $\mathcal{B}$ and the set of interior points by $\mathcal{I}$. We categorize the boundary points into two types: (i) **Pure** boundary points: which define the boundary and (ii) **Implied** boundary points: which are defined to be close to the 'pure' boundary points.

4.2.1. Identification of pure boundary points

We construct the k-NN graph on the set of points $\mathcal{X}$. Using the k-NN graph, for each sample $\mathbf{X} \in \mathcal{X}$, we compute the number of points in $\mathcal{X}$ that have $\mathbf{X}$ as one their lNN, $l = \{1, \ldots, k\}$. Denote this count as $count(\mathbf{X})$.

Let $\mathbf{X} \in \mathcal{I}$. For small values of k/M, the size of the k-NN balls are small (7). This implies that the density $f(\mathbf{X})$ over the k-NN ball of $\mathbf{X}$ is nearly constant. We also know that with very high probability, the volume of the k-NN ball of $\mathbf{X}$ is $(k/M)(f(\mathbf{X}) + o(1))$. Denote the lNN, $l = \{1, \ldots, k\}$ of $\mathbf{X}$ by $\mathcal{N}(\mathbf{X})$. We therefore have that the density for any $\mathbf{Y} \in \mathcal{N}(\mathbf{X})$ is $f(X) + o(1)$. This implies that the volume of the k-NN ball of $\mathbf{Y}$ is also $(k/M)(f(\mathbf{X}) + o(1))$. This further implies that for two points $\mathbf{X}, \mathbf{Y} \in \mathcal{I}$, if $\mathbf{X} \in \mathcal{N}(\mathbf{Y})$ then, with high probability, $\mathbf{Y} \in \mathcal{N}(\mathbf{X})$ and vice versa. We therefore have that for points $\mathbf{X} \in \mathcal{I}$, $count(\mathbf{X}) = k + o(k)$ with high probability. On the other hand, for points $\mathbf{Z} \in \mathcal{B}_{\mathcal{P}}$, $count(\mathbf{Z}) \leq k/2 + o(k)$. This then gives us a simple test to detect pure boundary points. In theory, we can use any threshold between $k/2$ and k to detect boundary points. In practice, we set the threshold to be $th = 0.65 * k$.

4.2.2. Identification of implied boundary points

"Implied" boundary points are points close to the pure boundary points whose k-NN balls are truncated by the boundary. Consider the case of k-NN density estimates. Because the pure boundary points define the boundary, this implies that the implied boundary points should have one or more pure boundary points among their k-NN. From the analysis described in the previous section, this in turn is equivalent to the implied boundary points belonging to the $k/2$-NN set of the pure boundary points. We then have the following method for detecting implied boundary points. For the k-NN density estimates, for each $\mathbf{X} \in \mathcal{B}_{\mathcal{P}}$, we obtain its $k/2$-NN $\mathcal{N}(\mathbf{X})$ and add them to the set of implied boundary points $\mathcal{B}_{\mathcal{I}}$.

4.3. Correction of density estimate

The idea for density correction at points close to the boundary is based on the following idea: To estimate the density at a boundary point $\mathbf{X} \in \mathcal{B}$, we find a point $\mathbf{Y} \in \mathcal{I}$ that is close to $\mathbf{X}$. Because of the proximity of $\mathbf{X}$ and $\mathbf{Y}$, $f(\mathbf{X}) \approx f(\mathbf{Y})$. We can then estimate the density at Y instead and use this as an estimate of $f(\mathbf{Y})$.

Algorithm 1 Detect boundary points $\mathcal{B}$

1. Construct k-NN tree on $\mathcal{X}$
2. Compute $count(\mathbf{X})/k$ for each $\mathbf{X} \in \mathcal{X}$
3. Detect pure boundary points $\mathcal{B}_{\mathcal{P}}$:
for each $\mathbf{X} \in \mathcal{X}$ **do**
  if $count(\mathbf{X}) < th$ **then**
    $\mathcal{B}_{\mathcal{P}} \leftarrow \mathcal{B}_{\mathcal{P}} \cup \mathbf{X}$
  else
    $\mathcal{I} \leftarrow \mathcal{I} \cup \mathbf{X}$
  end if
end for
4. Detect implied boundary points $\mathcal{B}_{\mathcal{I}}$:
for each $\mathbf{X} \in \mathcal{B}_{\mathcal{P}}$ **do**
  $\mathcal{B}_{\mathcal{I}} \leftarrow \mathcal{B}_{\mathcal{I}} \cup \mathcal{N}(\mathbf{X})$
end for
5. Output boundary points: $\mathcal{B} \leftarrow \mathcal{B}_{\mathcal{P}} \cup \mathcal{B}_{\mathcal{I}}$

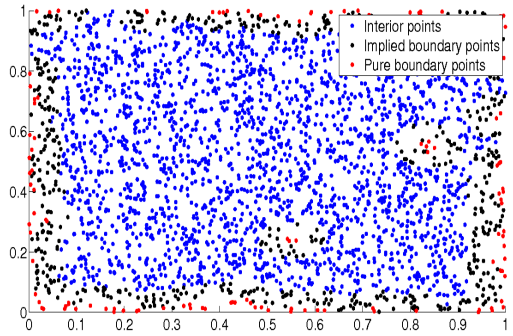

Fig. 2. Detection of boundary points for 2D uniform distribution.

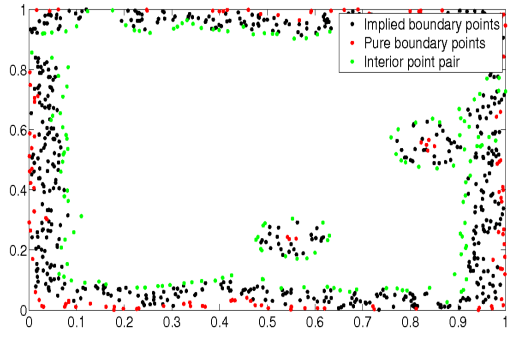

Fig. 3. Pairing of boundary points with interior points for 2D uniform distribution.

Consider a general multivariate setting with T i.i.d. realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ from the density f. First, we detect all the boundary points $\mathcal{B}$ among the total sample set. Then, for each $\mathbf{X}_i \in \mathcal{B}$, we identify its nearest neighbor $\mathbf{X}_{n(i)} \in \mathcal{I}$, for some $n(i) = \{1, .., T\}$. The volume of the region containing the boundary samples is of order $O(k/M)$. This guarantees that the maximum distance between any $\mathbf{X}_i \in \mathcal{B}$, $i = \{1, .., N\}$ and its closest neighbor $\mathbf{X}_{n(i)} \in \mathcal{I}$, for some $n(i) = \{1, .., N\}$, is of order $O((k/M)^{1/d})$.

Let $\mathbf{X}_i$ be a boundary point. From (5), we see that the bias is significant for the density estimate $\hat{\mathbf{f}}(\mathbf{X}_i)$. We suggest an alternative estimator to correct for the bias. Let $\mathbf{X}_{p(i)} = 2\mathbf{X}_{n(i)} - \mathbf{X}_i$. Defining $h = \mathbf{X}_{n(i)} - \mathbf{X}_i$, it is easy to see that $||h|| = O((k/M)^{1/d})$. Define the corrected estimator as

$$\hat{\mathbf{f}}_C(\mathbf{X}_i) = 2\hat{\mathbf{f}}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}(\mathbf{X}_{p(i)}). \tag{8}$$

We claim that this estimator has bias of order $O(||h||^2) = O((k/M)^{2/d})$. This can be shown as follows. Define $v =< h, f'(\mathbf{X}_{n(i)}) >$ as the inner product between h and the first order partial derivatives $f'(\mathbf{X}_{n(i)})$

$$\begin{aligned} f(\mathbf{X}_i) &= f(\mathbf{X}_{n(i)}) - v + O(||h||^2), & (9)\\ f(\mathbf{X}_{p(i)}) &= f(\mathbf{X}_{n(i)}) + v + O(||h||^2). & (10) \end{aligned}$$

(9) and (10) imply

$$f(\mathbf{X}_i) = 2f(\mathbf{X}_{n(i)}) - f(\mathbf{X}_{p(i)}) + O(||h||^2). \tag{11}$$

Because $\mathbf{X}_{n(i)}$ and $\mathbf{X}_{p(i)}$ are located in the interior of the density, by (3),

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{f}}(\mathbf{X}_{n(i)})] &= f(\mathbf{X}_{n(i)}) + O(||h||^2), & (12)\\ \mathbb{E}[\hat{\mathbf{f}}(\mathbf{X}_{p(i)})] &= f(\mathbf{X}_{p(i)}) + O(||h||^2), & (13) \end{aligned}$$

and therefore

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{f}}_C(\mathbf{X}_i)] &= \mathbb{E}[2\hat{\mathbf{f}}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}(\mathbf{X}_{p(i)})]\\ &= 2f(\mathbf{X}_{n(i)}) - f(\mathbf{X}_{p(i)}) + O(||h||^2)\\ &= f(\mathbf{X}_i) + O((k/M)^{2/d}). \end{aligned} \tag{14}$$

The corrected density estimate at the boundary therefore has bias which is of the same order as the bias of the uncorrected density estimate at any interior point (compare to (3) and (5)). Thus the compensation has reduced the bias of the estimator from $O(1)$ to $O((k/M)^{2/d})$. From our earlier observation that all central moments have the same rate behavior at both the interior and the boundary of the support, it trivially follows that the central moments of the boundary corrected estimator have the same rate behavior as the original density estimator.

4.4. Boundary point detection example

Fig. 2 depicts the detection of boundary points in the case of 2 dimensional uniform distribution. Fig. 3 depicts the pairing of boundary points with interior points. Clearly, the algorithm identifies the boundary points in this example.

We note that a small fraction points in the interior of the support have also been detected as boundary points. From the figure, it is clear that this is the result of the particular instantiation of data samples, wherein certain locations in the interior of the density are sparsely populated by samples, thereby creating the illusion of a boundary within the support. The method of compensation of density estimates in (8) ensures that the corrected density estimate at these incorrectly categorized points will closely resemble the original density estimate. Thus, the only drawback of such incorrect classification of interior points as boundary points is the additional computation needed to determine the corrected density estimate for these points.

5. THINNING k-NN GRAPHS

Using the corrected k-NN density estimates and the equivalence relation between density estimates and graphs, we propose corrected k-NN graphs as follows.

For the interior points, we retain the original neighborhood and the corresponding distances. For the boundary points, the corrected k-NN ball radius $\hat{\mathbf{d_k}}(\mathbf{X_i})$ is determined as

$$\begin{aligned} \hat{\mathbf{f}}_C(\mathbf{X}_i) &= 2\hat{\mathbf{f}}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}(2\mathbf{X}_{n(i)} - \mathbf{X}_i)\\ \Rightarrow \frac{1}{(\hat{\mathbf{d_k}}(\mathbf{X_i}))^{\mathbf{d}}} &= \frac{2}{(\mathbf{d_k}(\mathbf{X_{n(i)}}))^{\mathbf{d}}} - \frac{1}{(\mathbf{d_k}(\mathbf{X_{p(i)}})^{\mathbf{d}}} \end{aligned} \tag{15}$$

For each boundary point $\mathbf{X}_i$ in the graph, we now remove the edges from the graph whose length exceeds the corrected k-NN ball radii $\hat{\mathbf{d_k}}(\mathbf{X_i})$. We call this process thinning the k-NN graph. After thinning the number of nearest neighbors in the thinned graph will be less than k. For instance, the pure boundary points should have around $k/2$-NN in the corrected graph.

6. SIMULATIONS

We apply our thinned k-NN graphs to two problems: (i) entropy estimation and (ii) classification. Our simulations show that while the method does indeed breakdown for small sample sizes, we obtain significant performance gains for moderate to large samples sizes that arise in practical problems.

6.1. Entropy estimation

We consider the problem of Shannon entropy estimation for a 2 dimensional distribution. The Shannon entropy of density f is defined to be $H(f) = -\int f \log f d\mu$. We analyze simple partition based plug-in estimators of entropy in [4]. These estimators partition the available T i.i.d realizations from density f into N and M disjoint samples $N + M = T$. The estimator is then defined as

$$\hat{\mathbf{H}}(f) = \left(\frac{1}{N}\sum_{i=1}^{N} -\log(\hat{\mathbf{f}}(\mathbf{X}_i))\right). \tag{16}$$

where $\hat{\mathbf{f}}(\mathbf{X}_i)$ is the k-NN density estimate. When the support of the density has no boundaries, the bias of this estimator is

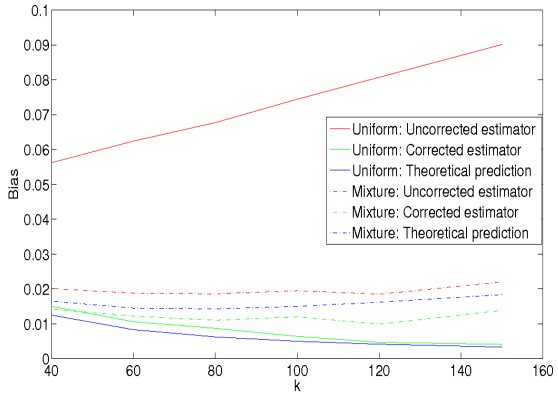

Fig. 4. Variation of bias of estimated entropy with varying k.

given by [4]

$$\mathbb{E}[\hat{\mathbf{H}}(f)] - H(f) = \rho\left(\frac{k}{M}\right)^{2/d} + \left(\frac{1}{2k}\right) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right), \quad (17)$$

where $\rho = \mathbb{E}[f^{-(d+2)/d}(\mathbf{Y})c(\mathbf{Y})]$ where the random variable $\mathbf{Y}$ has density f.

We consider two different types of densities: (a) 2 dimensional uniform distribution and (b) 2 dimensional beta density with parameters a=4,b=4. For a fixed partition of $N = 1000$ and $M = 9000$, we vary the bandwidth parameter k and plot the variation of bias of the entropy estimator using (a) uncorrected and (b) boundary corrected k-NN density estimates. This is shown in Fig. 4.

From the figure, it is clear that the bias corrected entropy estimator agrees well with the theoretical prediction for the uniform distribution. On the other hand, the observed bias for the uncorrected estimator is significantly higher than the predicted bias, as should be expected because of our prediction of boundary effects. For the mixture density, both the uncorrected and corrected estimators agree well with the theoretical prediction. This can be attributed to the fact that for the mixture density, the fraction of boundary points is very small, thereby minimizing the influence of the boundary regions on the entropy estimate.

6.2. k-NN classification

We describe the basic k-NN classification algorithm. An unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. To account for boundary effects, we determine the modified k-NN neighborhood using (15), remove the neighbors which exceed the modified neighborhood size, and assign the label most frequent among the surviving training samples. We will call this the boundary compensated classifier.

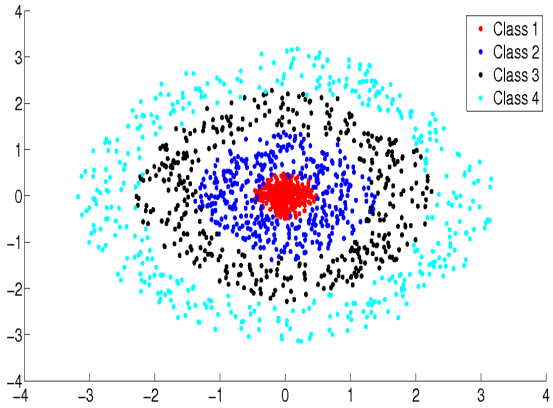

Fig. 5. Concentric circle data.

	1	2	3	4
1	400/400	0/0	0/0	0/0
2	88/88	312/312	0/0	0/0
3	0/0	55/55	341/341	4/4
4	0/0	0/0	148/82	252/318

Table 2. Confusion matrix for concentric circle data (Black: Standard k-NN graph; Blue: Boundary compensated k-NN graph).

6.2.1. A simple example

We consider a simple example where 4 concentric 2D rings constitute 4 different classes of data. Each class consists of 400 samples. The confusion matrix (using the leave-one-out criteria) for the uncompensated and the compensated classifier ($k = 100$) is shown in Table 2.

We note that for the original classifier, while the inner rings (classes 1, 2 and 3) were well classified, the classification performance for the outermost ring (class 4) was relatively worse. This can be attributed to the fact that the boundary points in this data set belong to the outermost ring. From the confusion matrix, we can see that the boundary compensated classifier performs significantly better w.r.t. class 4.

6.2.2. Optical digit recognition

The 'Optical Recognition of Handwritten Digits Data Set' [10] consists of normalized bitmaps of handwritten digits from a preprinted form. This data set has 562 instances of each digit from $0 - 9$. Each instance is characterized by 64 dimensional pixel intensity values. As a first step, we use standard PCA embedding to reduce the dimension to 10. We then normalize these 10 dimensional vectors to unit length. We treat the first 9 dimensions of each normalized vector as our feature vectors f_i. We note that the feature vectors f_i live in a unit hypercube in $\mathbf{R}^9$. A significant fraction of the feature vectors f_i will

	0	1	2	3	4	5	6	7	8	9
0	551/551	0/0	0/0	0/0	2/2	0/0	0/0	0/0	0/0	1/1
1	0/0	558/563	5/4	0/0	1/0	0/0	2/1	1/1	1/0	3/2
2	0/0	3/1	537/549	0/0	0/0	0/0	0/0	4/1	12/5	1/1
3	0/0	3/3	9/6	537/546	0/0	3/3	1/1	4/3	7/4	8/6
4	1/1	1/0	1/1	0/0	555/558	0/0	2/2	1/1	0/0	7/5
5	13/8	2/2	0/0	7/7	0/0	508/519	7/7	0/0	0/0	21/15
6	2/2	2/2	0/0	0/0	1/1	0/0	552/552	0/0	1/1	0/0
7	0/0	0/0	1/0	1/2	1/1	0/0	0/0	549/555	7/3	7/5
8	5/2	18/15	17/18	2/1	3/1	1/1	5/6	1/1	497/505	5/4
9	2/2	6/5	5/5	7/5	1/1	5/7	1/1	11/9	6/7	518/520

Table 1. Confusion matrix for 'Handwritten Digits' dataset (Black: Standard k-NN graph; Blue: Boundary compensated k-NN graph).

lie close to the surface of the hypercube, thereby behaving as boundary points. We apply the standard and boundary compensated k-NN classifiers ($k = 25$) to this data. The confusion matrix for the uncompensated and the compensated classifier is shown in Table 1. The leave-one-out classification error for the uncompensated classifier was found to be 4.59% and improved to 3.59% for the compensated classifier. Using a paired t-test, the p-value for this result was found to be well within a siginificance level of 1%, implying that the improvement in performance is indeed statistically significant.

7. CONCLUSION

We showed that for samples on a finite support, the behavior of the k-NN neighborhoods is different in the interior of the support and the boundary. To resolve this issue, we analyzed and compensated the bias of k-NN density estimates close to the boundary. This in turn helped us define a modified k-NN graph with smaller k-NN neighborhoods for points close to the boundary.

Given the large body of work on boundary compensated kernel density estimates, a particularly important outcome of our work is bias compensated k-NN density estimates. The basic idea for boundary correction introduced in this paper can be extended to kernel density estimates.

Our boundary corrected k-NN graphs can be used in place of standard k-NN graphs whenever the data is suspected to lie on a bounded region. We compared the performance of standard and our modified k-NN graphs in the context of entropy estimation and classification and showed that the modified k-NN graph can significantly outperform the standard k-NN graph.

8. REFERENCES

[1] A. O. Hero, J. Costa, and B. Ma, "Asymptotic relations between minimal graphs and alpha-entropy," *Technical Report CSPL-334 Communications and Signal Processing Laboratory, The University of Michigan*, March 2003.

[2] V. V. Mergel M. N. Goria, N. N. Leonenko and P. L. Novi Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Nonparametric Statistics*, 2004.

[3] J.A. Costa, A. Girotra, and A.O. Hero, "Estimating local intrinsic dimension with k-nearest neighbor graphs," in *2005 IEEE/SP 13th Workshop on Statistical Signal Processing*, 2005, pp. 417–422.

[4] K. Sricharan, R. Raich, and A.O. Hero, "Optimized intrinsic dimension estimation," in *IEEE Intl. Conf on Acoustics, Speech and Signal Processing (ICASSP)*, April 2010.

[5] A.M. Farahmand, C. Sepesvari, and J-Y Audibert, "Manifold-adaptive dimension estimation," *Proc of 24th Intl Conf on Machine Learning*, pp. 265–272, 2007.

[6] M. C. Jones, "Simple boundary correction for kernel density estimation," *Statistics and Computing*, vol. 3, pp. 135–146, 1993.

[7] R.J. Karunamuni and T. Alberts, "On boundary correction in kernel density estimation," *Statistical Methodology*, vol. 2, no. 3, pp. 191 – 212, 2005.

[8] D.O. Loftsgaarden and C.P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, 1965.

[9] K. Fukunaga and L.D. Hostetler, "Optimization of k-nearest-neighbor density estimates," *IEEE Transactions on Information Theory*, 1973.

[10] A. Asuncion and D.J. Newman, "UCI machine learning repository," 2007.