Empirical estimation of entropy functionals with confidence

Kumar Sricharan, Department of EECS, University of Michigan
Raviv Raich, School of EECS, Oregon State University
Alfred O. Hero III, Department of EECS, University of Michigan

Abstract

Nonparametric estimation of functionals of density from finite number of samples is an important tool in domains such as statistics, signal processing and machine learning. While several estimators have been proposed in literature, the performance of these estimators is not known. We propose a kNN class of plug-in estimators for estimating non-linear functionals of density, such as entropy, mutual information and support set dimension. The plug-in estimators are designed to automatically incorporate boundary corrections for densities with finite support. Based on the statistical properties of kNN density estimators, we derive the bias and variance of the plug-in estimator in terms of the sample size, the dimension of the samples and the underlying probability distribution. We also establish a central limit theorem for the plug-in estimators. Based on these results, we specify the optimal choice of tuning parameters for minimum mean square error. The theory is illustrated by applications to problems such as intrinsic dimension estimation and structure discovery in high dimensional data.

1 Introduction

Functionals of densities have important applications in domains such as statistics, signal processing and machine learning. Divergence between densities is an important example of such functionals. For example, the Jensen difference [21] and mutual information [26] are used as similarity measures in multimodal image registration, information fusion and other pattern recognition problems. Oftentimes, we do not have access to the density functions, but rather to sample realizations of the underlying density. In this context, nonparametric estimators of functionals of densities from finite number of samples becomes important.

These estimation problems can be treated as specific instances of estimation of non-linear functionals of the density $f(x)$ of the form $\int g(f(x), x)f(x)d\mu(x)$. Bickel and Ritov [1] treat the problem for the specific case of $\int f^2(x)d\mu(x)$, which was generalized to $\int g(f, x)f(x)d\mu(x)$ for arbitrary $g(.)$ by Birge and Massart [3]. They show that for sufficiently smooth densities,

the best possible rate that can be achieved is $O(1/n)$ and suggest estimators that can achieve this rate. However, the estimators proposed by these authors are quite intricate and in general difficult to implement. Several other estimators of entropy measures of the form $\int g(f(x))f(x)d\mu(x)$ have been proposed in literature for specific instances of $g(.)$. These include estimators based on entropic graphs [11], gap estimators [25] and nearest neighbor distances [18]. While these estimators have been shown to be consistent, results on rates of convergence of these estimators are in general unavailable. Hero et.al. [11], who provide minimax rates of convergence, are an exception.

We present a simple class of estimators based on kNN graphs to estimate these non-linear functionals $\int g(f(x), x)f(x)d\mu(x)$ for high dimensional data. Our class of estimators exploit a close relation between density estimation and the geometry of proximity neighborhoods in the data sample. For our proposed class of estimators, we will present an asymptotic statistical analysis of the bias and variance. In addition, we will provide results on weak convergence of these class of estimators. These results are useful for choosing estimator tuning parameters and for predicting fundamental performance limits of these estimators.

The results in this report improve existing results on nearest neighbor estimators available in literature. While our results apply to arbitrary smooth functionals $g(.)$, the authors of [23, 18, 16] only deal with the functionals $g(u) = \log(u)$ and $g(u) = u^{\alpha-1}$. Evans et.al. [8] on the other hand analyze only positive moments of the k-NN distances ($g(u) = u^k, k \in \mathbb{N}$). The authors of [23, 18, 8] show that the estimators they propose are asymptotically unbiased. Evans et.al. [9] show that the variance is bounded by the rate $O(k^5/T)$. From our analysis, we are able to establish the exact rates of decay of the bias and the variance. Finally, CLT for k-NN estimators of Rényi entropy was alluded to by Leonenko et.al. [18] by inferring from experimental results. We successfully establish a CLT for k-NN estimators of arbitrary functionals, including Rényi entropy.

We will illustrate the usefulness of our theory by applying it to diverse applications including intrinsic dimension estimation and factor graph structure discovery. Our results on asymptotic theory of the estimators will be used to predict performance of these applications.

2 Plug-in estimators

We are interested in estimating non-linear functionals $G(f)$ of d-dimensional multi-variate densities f with support S, where $G(f)$ has the specific form

$$G(f) = \int g(f(x), x)f(x)d\mu(x) = \mathbb{E}[g(f(x), x)], \tag{1}$$

for some smooth function $g(x, y)$. Here, μ denotes the Lebesgue measure and $\mathbb{E}$ denotes statistical expectation w.r.t density f. We require that the density f be uniformly bounded away from 0 and finite on the support $\mathcal{S}$, i.e., there exist constants ϵ_0, ϵ_∞ such that $0 < \epsilon_0 < \epsilon_\infty < \infty$ such that $\epsilon_0 \leq f(x) \leq \epsilon_\infty \ \forall x \in \mathcal{S}$. We assume that i.i.d realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ are available from the density f.

2.1 Notation

We will use bold face type to indicate random variables and random vectors and regular type face for constants. We denote the expectation operator by the symbol $\mathbb{E}$ and the bias of an estimator by $\mathbb{B}$. We also denote conditional expectation given $\mathbf{Z}$ using the notation $\mathbb{E}_{\mathbf{Z}}$.

Define the variance operator as

$$\mathbb{V}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2],$$

and the covariance operator as

$$Cov[\mathbf{X}, \mathbf{Y}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])].$$

In addition to the parameters N and M, we have a parameter k which characterizes both the uniform and the kNN density estimates. We talk about asymptotic behavior of the plug-in estimates under the following limiting conditions:

- $k/M \to 0$
- $k \to \infty$
- $N \to \infty$

As shorthand, we will collectively denote the above behavior by $\Delta(k, N, M) \to 0$.

Under these asymptotic conditions, we use the following order notation:

- $a = o(b)$: a is dominated by b asymptotically.
- $a = O(b)$: a is bounded above by b (up to a constant factor) asymptotically.
- $a = \Theta(b)$: a is bounded above and below by b (up to constant factors) asymptotically.
- $a =\sim (b)$: a is equal to b (up to a constant factor) asymptotically.

2.2 Plug-in estimators

We assume we have $T = N + M$ i.i.d realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ from the density f. We begin by defining an *oracle estimate*,

$$\tilde{\mathbf{G}}(f) \;=\; \left(\frac{1}{N} \sum_{i=1}^{N} g(f(\mathbf{X}_i), \mathbf{X}_i) \right). \tag{2}$$

The *oracle estimate* $\tilde{G}(f)$ is a unbiased and consistent estimator of $G(f)$. To get a *plug-in estimate*, we estimate density at the N points $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ using the M realizations

$\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ and plug the estimated density values into the *oracle estimate* in Eq.2 to estimate $G(f)$. The *plug-in estimate* is therefore given by

$$\hat{\mathbf{G}}(f) \quad = \quad \left(\frac{1}{N} \sum_{i=1}^{N} g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) \right). \tag{3}$$

The *plug-in estimate* is consistent if the density estimate $\hat{\mathbf{f}}$ is an consistent estimator of f.

2.3 Density estimation

We use two popular density estimation methods for plug-in estimation: (a) Kernel density estimator (with uniform kernel) and (b) k nearest neighbor density estimator. In the main body, we briefly discuss these k-NN density estimator. Full details can be found in Appendix A and Appendix B respectively.

Let $\mathbf{d}_X^{(k)}$ denote the Euclidean distance between a point X and its k-th nearest neighbor amongst $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$. The k-NN region is $\mathbf{S_k}(X) = \{Y : d(X,Y) \leq \mathbf{d}_X^{(k)}\}$ and the volume of the k-NN region is $\mathbf{V_k}(X) = \int_{\mathbf{S_k}(X)} dZ$. The standard k-NN density estimator [17] is defined as $\hat{\mathbf{f}}_\mathbf{k}(X) = \frac{k-1}{M\mathbf{V}_k(X)}$. If a probability density function has bounded support, the kNN balls centered at points close to the boundary are often are truncated at the the boundary as shown in Fig. 2. As a consequence of this truncation, we show that k-NN density estimates near the boundaries of the support suffer from significant bias. Define the set $\mathcal{B}$ to be the set of boundary points where the kNN ball is truncated by the boundary of the support of the density. We will show that the bias of the standard k-NN density estimate is of order $O((k/M)^{2/d})$ in the interior and is of order $O(1)$ at these boundary points. We propose the following method for compensating the bias of k-NN density estimates near the boundaries of the support for general multivariate data without any prior knowledge of the support of the density. This compensation is done in two stages: (i) we identify the set of boundary points $\mathcal{B}$ using a non-parametric algorithm based on k-nearest neighbors and (ii) we estimate corrected densities at these points by estimating densities at interior points which are close to the boundary points. For a boundary point $\mathbf{X}_i \in \mathcal{B}$, $i \in \{1, \ldots, N\}$ the corrected density estimate is given by

$$\hat{\mathbf{f}}_C(\mathbf{X}_i) = \begin{cases} 2\hat{\mathbf{f}}_\mathbf{k}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}_\mathbf{k}(\mathbf{X}_{p(i)}) & 2\hat{\mathbf{f}}_\mathbf{k}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}_\mathbf{k}(\mathbf{X}_{p(i)}) > 0 \\ \hat{\mathbf{f}}_\mathbf{k}(\mathbf{X}_{n(i)}) & 2\hat{\mathbf{f}}_\mathbf{k}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}_\mathbf{k}(\mathbf{X}_{p(i)}) \leq 0 \end{cases}$$

where $\mathbf{X}_{n(i)}$ is the closest interior sample boundary pair to $\mathbf{X}_{(i)}$, $n(i) \in \{1, \ldots, N\}$. We show that the bias for the corrected density estimate $\hat{\mathbf{f}}_C$ at the boundary is of the same order $O((k/M)^{2/d})$ as in the interior. The details can be found in the Appendix C.

We now define the boundary compensated k-NN density estimator at $\mathbf{X}_{(i)}$, $i \in \{1, \ldots, N\}$ to be

$$\hat{\mathbf{f}}_k(\mathbf{X}_i) = \begin{cases} \hat{\mathbf{f}}_\mathbf{k}(\mathbf{X}_i) & \mathbf{X}_i \in \mathcal{B}^c \\ \hat{\mathbf{f}}_C(\mathbf{X}_i) & \mathbf{X}_i \in \mathcal{B} \end{cases}$$

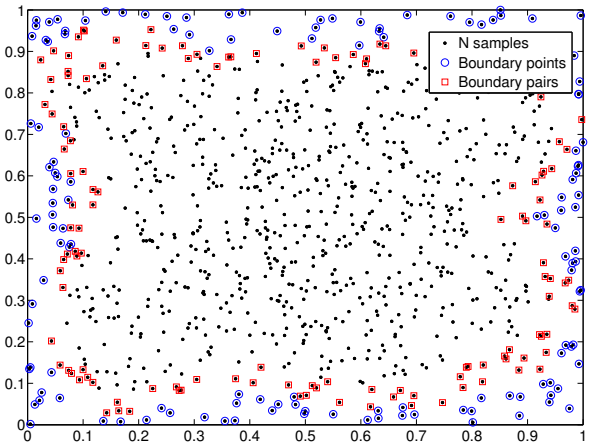


Figure 1: Detection of boundary points for 2D beta distribution.

Henceforth, when we refer to k-NN density estimates, it is to be understood that we are referring to the boundary compensated versions.

3 Main results

In this section, we state the main results concerning plug-in estimators that are established in this article. We assume we have $T = N + M$ i.i.d. realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_{N+M}\}$ from the density f. The plug-in estimate is given by

$$\hat{\mathbf{G}}(f) \quad = \quad \left(\frac{1}{N} \sum_{i=1}^{N} g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) \right). \tag{4}$$

3.1 Regularity conditions

The regularity conditions listed below are discussed in greater detail in Appendix D.

Polynomial growth rate

We assume that k grows polynomially in M, i.e. $k = M^{\alpha}$ for $\alpha \in (0, 1)$.

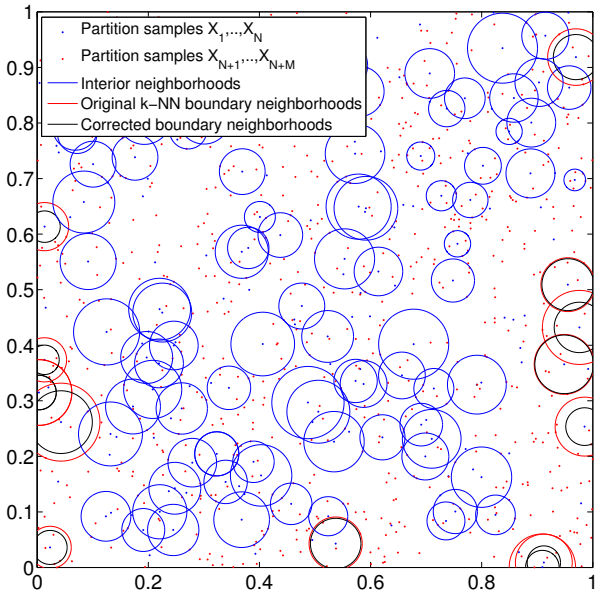


Figure 2: kNN balls centered around a subsample of 2D uniformly distributed points. Note that the original k-NN balls centered at points close to boundary (red) over spill the boundary. The modified k-NN neighborhoods (black) corresponding to the corrected corrected density estimate $\hat{\mathbf{f}}_C$ compensate for the over spill.

Conditions on density

We require that the density f be uniformly bounded away from 0 and finite on the support $\mathcal{S}$, i.e., there exist constants ϵ_0, ϵ_∞ such that $0 < \epsilon_0 < \epsilon_\infty < \infty$ such that $\epsilon_0 \leq f(x) \leq \epsilon_\infty$ $\forall x \in \mathcal{S}$.

We assume that the density f has continuous partial derivatives of order $2r$ in the interior of the support $\mathcal{S}$ where r satisfies the condition $2r(1-\alpha)/d > 1$. We also assume that the functional $g(x, y)$ has λ partial derivatives w.r.t. x, where λ satisfies the condition $\alpha\lambda > 1$.

Conditions on functional

Finally we require that the functional $g(x, y)$ satisfy the following properties. Let $g'(x, y)$, $g''(x, y)$, $g'''(x, y)$ and $g''''(x, y)$ denote the first, second, third and fourth partial derivatives of $g(x, y)$ w.r.t the first argument x. We assume that the absolute value of the functional $g(x, y)$ and its partial derivatives are strictly bounded away from ∞ in the range $\epsilon_0 < x < \epsilon_\infty$ for all y. Let $\mathbf{Y}$ denote a random variable with density f.

3.2 kNN plug-in estimator

We list theorems on the bias, variance and central limit of kNN plug-in estimators. Equivalent results have been shown for uniform kernel plug-in estimators in Appendix F. Let $\hat{\mathbf{G}}_k(f)$ denote the kNN plug-in estimator and $\mathbf{Y}$ be a random variable with density f.

Theorem 3.1. *Suppose that the density f, the functional g and the density estimate $\hat{\mathbf{f}}_k$ satisfy the necessary conditions listed above. The bias of the plug-in estimator $\hat{\mathbf{G}}_k(f)$ is then given by*

$$\mathbb{B}_k(f) \;=\; c_1 \left(\frac{k}{M}\right)^{2/d} + c_2 \left(\frac{1}{k}\right) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right),$$

where the constant $c_1 = \mathbb{E}[g'(f(\mathbf{Y}), \mathbf{Y}) f^{-2/d}(\mathbf{Y})(\Gamma^{(2/d)}((n+2)/2) tr[\nabla^2(f(\mathbf{Y}))])]$, and the constant $c_2 = \mathbb{E}[f^2(\mathbf{Y}) g''(f(\mathbf{Y}), \mathbf{Y})/2]$ are constants which depend on the underlying density f.

Theorem 3.2. *Suppose that the density f, the functional g and the density estimate $\hat{\mathbf{f}}_k$ satisfy the necessary conditions listed above. The variance of the plug-in estimator $\hat{\mathbf{G}}_k(f)$ is given by*

$$\mathbb{V}_k(f) \;=\; c_4 \left(\frac{1}{N}\right) + c_5 \left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right),$$

where the constant $c_4 = \mathbb{V}[g(f(\mathbf{Y}), \mathbf{Y})]$ and the constant $c_5 = \mathbb{V}[f(\mathbf{Y}) g'(f(\mathbf{Y}), \mathbf{Y})]$ depend on the underlying density f.

Proof. We briefly sketch the proof here. The above theorems have been stated more generally and proved in Appendix D.

The principal idea here involves Taylor series expansions of the functional $g(\hat{\mathbf{f}}(X), X)$ about the true value $g(f(X), X)$, and subsequently (a) using the moment properties of density estimates listed earlier to obtain the leading terms, and (b) bounding the remainder term in the Taylor series and showing that it can be ignored in comparison to the leading terms. □

Theorem 3.3. *Suppose that the density f, the functional g and the density estimate $\hat{\mathbf{f}}_k$ satisfy the necessary conditions listed above. The asymptotic distribution of the plug-in estimator $\hat{\mathbf{G}}_k(f)$ is given by*

$$\lim_{\Delta(k,N,M)\to 0} Pr\left(\frac{\hat{\mathbf{G}}_k(f) - \mathbb{E}[\hat{\mathbf{G}}_k(f)]}{\sqrt{\mathbb{V}[f(\mathbf{Y}) g'(f(\mathbf{Y}), \mathbf{Y})]/N}} \leq \alpha\right) = Pr(\mathbf{Z} \leq \alpha),$$

where $\mathbf{Z}$ is a standard normal random variable.

Proof. Define the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ for any fixed M as

$$\mathbf{Y}_{M,i} = g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) - \mathbb{E}[g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)].$$

The key idea here is to recognize that $\mathbf{Y}_{M,i}$ are exchangeable random variables. Blum et.al. [4] showed that for exchangeable 0 mean, unit variance random variables $\mathbf{Z_i}$, the sum $\mathbf{S}_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Z}_i$ converges in distribution to $N(0,1)$ if and only if $Cov(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $Cov(\mathbf{Z}_1^2, \mathbf{Z}_2^2) = 0$. In our case,

$$\begin{aligned} Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) &= O(1/M), \\ Cov(\mathbf{Y}_{M,i}^2, \mathbf{Y}_{M,j}^2) &= O(1/M). \end{aligned}$$

As M gets large, we then have that $Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) \to 0$ and $Cov(\mathbf{Y}_{M,i}^2, \mathbf{Y}_{M,j}^2) \to 0$. We then extend the work by Blum et.al. to show that convergence in distribution to $N(0,1)$ holds in our case as both N and M get large. These ideas are rigorously treated in Appendix E.

□

The CLT for k-NN estimators of Rényi entropy was alluded to by Leonenko et.al. [18] by inferring from experimental results. Theorem 3.3 establishes the CLT for k-NN estimators of arbitrary functionals, including Rényi entropy. This result allows one to define approximate finite sample confidence intervals on the estimated values of the functionals and define p-values .

4 Analysis of M.S.E

The general form of the **bias** for the estimators discussed above is of the form:

$$\mathbb{B}(f) = c_1 \left(\frac{k}{M}\right)^{2/d} + c_2 \left(\frac{1}{k}\right). \tag{5}$$

The general form of the **variance** for the estimators discussed above is of the form:

$$\mathbb{V}(f) = c_4 \left(\frac{1}{N}\right) + c_5 \left(\frac{1}{M}\right). \tag{6}$$

In both the above expressions, we ignore the higher order terms for the sake of analysis presented below. This gives us the general form of the **mean square error** as

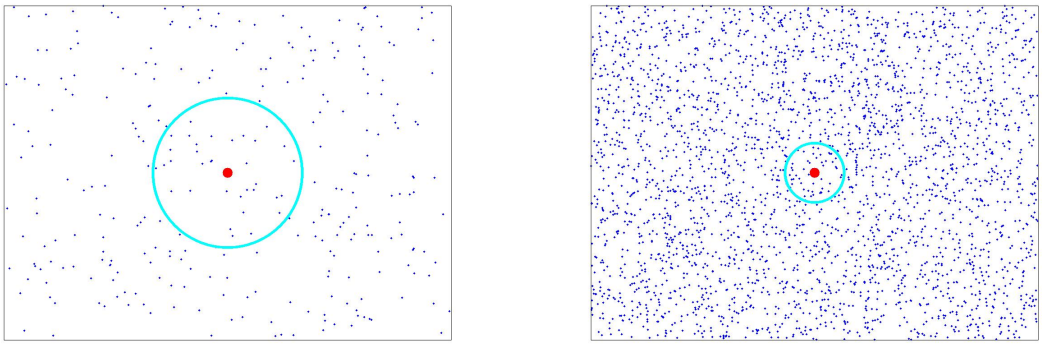


Figure 3: Asymptotics. Variation of density estimate with increasing k and M

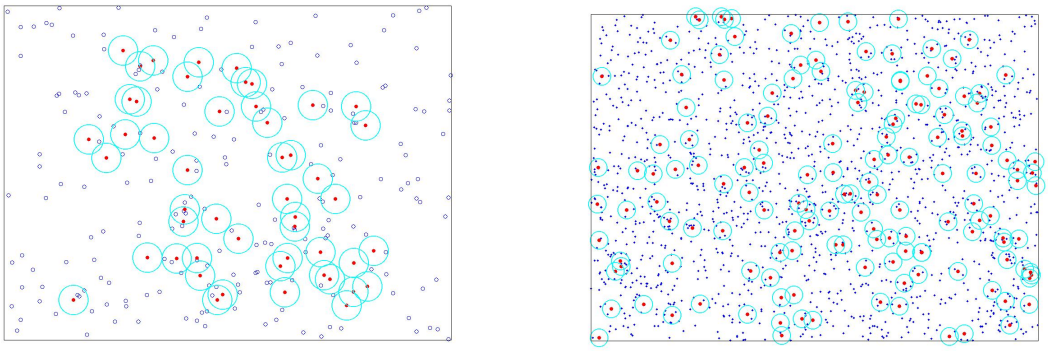


Figure 4: Asymptotics. Variation of plug-in estimate with increasing k, M and N

$$\begin{aligned}\mathbb{M}(f) &= \mathbb{B}^2(f) + \mathbb{V}(f) \\ &= \left(c_1\left(\frac{k}{M}\right)^{2/d} + c_2\left(\frac{1}{k}\right)\right)^2 \\ &+ c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right). \end{aligned} \tag{7}$$

From Eq.5 we see that we need we need $k \to \infty$ and $k/M \to 0$ for the estimator to be unbiased. Likewise from Eq.6 we see that we need we need $N \to \infty$ and $M \to \infty$ for the variance of the estimator to converge to 0. Figures 3 and 4 illustrate the asymptotic behavior of the density estimate and the plug-in estimate with increasing sample size.

4.1 Optimal choice of parameters

In this section, we obtain optimal values for k,M and N for minimum M.S.E.

4.1.1 Optimal choice of k

Minimizing the M.S.E. over k is equivalent to minimizing the square of the bias over k. We observe that the constants c_1 and c_2 can possibly have opposite signs. We consider two separate cases: $c_1c_2 > 0$ and $c_1c_2 < 0$. In either case the optimal choice of k is given by

$$k_{opt} = \arg\min_k |\mathbb{B}(f)| = \lfloor k_0 M^{\frac{2}{2+d}} \rfloor, \tag{8}$$

where $\lfloor x \rfloor$ is the closest integer to x and we have defined the constant $k_0 = (|c_2|d/2|c_1|)^{\frac{d}{d+2}}$ when $c_1c_2 > 0$ and $k_0 = (|c_2|/|c_1|)^{\frac{d}{d+2}}$ when $c_1c_2 < 0$. When $c_1c_2 > 0$, the bias evaluated at k_{opt} is $b_0^+ M^{\frac{-2}{2+d}}(1 + o(1))$ where the constant $b_0^+ = c_1 k_0^{2/d} + c_2/k_0$.

Let $k_{frac} = k_0 M^{\frac{2}{2+d}} - k_{opt}$. When $c_1c_2 < 0$, we see that $c_1((k_{frac}+k_{opt})/M)^{2/d}+c_2/(k_{frac}+k_{opt})$ is equal to zero. When this happens a higher order asymptotic analysis is required, yielding (see Appendix E):

$$\begin{aligned}
\mathbb{B}(\hat{\mathbf{G}}(f)) &= c_1 \left(\frac{k}{M}\right)^{2/d} + c_2 \left(\frac{1}{k}\right) \\
&\quad + h_1 \left(\frac{k}{M}\right)^{4/d} + h_2 \left(\frac{1}{k^2}\right) + h_3 \left(\left(\frac{k}{M}\right)^{2/d} \frac{1}{k}\right) \\
&\quad + o\left(\left(\frac{k}{M}\right)^{4/d} + \frac{1}{k^2} + \left(\frac{k}{M}\right)^{2/d} \frac{1}{k}\right)
\end{aligned}$$

where the constants are given by

$$h_1 = \mathbb{E}[(1/2)g''(f(\mathbf{Y}))h^2(X) + g'(f(\mathbf{Y}))h_o(\mathbf{Y})],$$

$$h_2 = \mathbb{E}[(2/3)g'''(f(\mathbf{Y}))f^3(\mathbf{Y})]$$

and

$$h_3 = (1 - 2/d)\mathbb{E}[g''(f(\mathbf{Y}))f(\mathbf{Y})c(\mathbf{Y})].$$

The bias evaluated at k_{opt} is then given by $b_0^- M^{\frac{-4}{2+d}}(1 + o(1))$ where the constant $b_0^- = h_1 k_0^{4/d} + (h_2 + c_2 k_{frac})/k_0^2 + (h_3 + 2c_1 k_{frac}/d)k_0^{2/d-1}$. In practice, the constants c_1 and c_2 have to be estimated with error of order $o(1/k + (k/M)^{2/d})$ or smaller for the leading terms to cancel using the optimal choice of $k_{opt} = \lfloor k_0 M^{\frac{2}{2+d}} \rfloor$, where k_0 depends on the estimated values of c_1 and c_2.

We note that b_0^2, c_4 and c_5 are all non-negative constants.

4.1.2 Optimal choice of N as a function of M

For $d = 1$, the leading terms in the M.S.E are given by

$$\mathbb{M}(f) \approx c_4 \left(\frac{1}{N}\right) + c_5 \left(\frac{1}{M}\right). \tag{9}$$

For $d = 2$, the leading terms in the M.S.E are given by

$$\mathbb{M}(f) \approx c_4 \left(\frac{1}{N}\right) + (c_5 + b_0^2) \left(\frac{1}{M}\right). \tag{10}$$

For large dimensions ($d > 6$), the optimal choice of N as a function of M for minimum M.S.E. is then given by

$$N_{opt} = \begin{cases} \lfloor N_0^+ M_{opt}^{\frac{6+d}{2(2+d)}} \rfloor & c_1 c_2 > 0 \\ \lfloor N_0^- M_{opt}^{\frac{10+d}{2(2+d)}} \rfloor & c_1 c_2 < 0 \end{cases}$$

where the constant N_0^+ is given by $N_0^+ = \frac{\sqrt{c_4(2+d)}}{2b_0^+}$ and the constant $N_0^- = \frac{\sqrt{c_4(2+d)}}{2\sqrt{2}b_0^-}$.

4.1.3 Optimal M.S.E.

For the optimal choices of k and N, the M.S.E in terms of M is given by

$$\mathbb{M}(f) = b_0^2 M^{\frac{-4}{2+d}} + \frac{(c_4)}{N_0} M^{\frac{-(6+d)}{2(2+d)}} + c_5 M^{-1}. \tag{11}$$

4.1.4 Relation between k_{opt} and N_{opt}

Now consider the ratio of k_{opt} to N_{opt}.

For $d = 1$ we have,

$$\frac{k_{opt}}{N_{opt}} = \frac{k_0}{N_0} \left(\frac{1}{M}\right)^{1/3}, \tag{12}$$

and for $d \geq 2$, we have

$$\frac{k_{opt}}{N_{opt}} = \frac{k_0}{N_0} \sqrt{\frac{1}{M}}, \tag{13}$$

The ratio of k_{opt} to N_{opt} therefore goes to 0 as M goes to ∞.

4.2 Discussion on optimal parameter choices

1. **Choice of** k: The optimal k grows at a smaller rate as compared to the total number of samples M used for density estimation. Furthermore, the rate at which k grows as compared to M decreases as the dimension d increases. This can be explained by observing that the choice of k primarily controls the bias of the entropy estimator. For a fixed choice of k and M ($k < M$), we expect the bias in the density estimates (and correspondingly in the entropy estimates) to increase as the dimension increases. For fixed M, to ensure optimal bias, we would therefore require that the density estimates are based on realizations which lie in smaller neighborhoods as the dimension increases. This in turn corresponds to choosing a smaller k relative to M as the dimension d grows.

2. **Choice of** N: For large dimensions ($d \geq 2$), the optimal choice of N (the number of samples used for estimating entropy) grows at a smaller rate as compared to M (the number of samples used for density estimation). This agrees with our intuition that in higher dimensions, density estimation is the more difficult problem as compared to the problem of entropy estimation when the density is known, and therefore a greater fraction of the total realizations available should be used for estimation of the density.

4.2.1 Comparison of rates

We note that for high dimensions ($d > 6$), $N_{opt} = o(M_{opt})$, which in turn implies that $M_{opt} = \Theta(T)$. This then implies that the optimal bias decays as $b_0^+(T^{\frac{-2}{2+d}})(1 + o(1))$ when $c_1c_2 > 0$ and $b_o^-(T^{\frac{-4}{2+d}})(1 + o(1))$ when $c_1c_2 < 0$. In addition, the optimal variance decays as $c_5(1/T)(1 + o(1))$. To date, rates of convergence of the bias for Shannon and Rényi entropy estimators have been explicitly provided only by Liitiäinen et.al. [16]. Liitiäinen et.al. have a leading term of order $\Theta(T^{-1/d})$ which arises due to boundary effects. This term can be eliminated either by using the weighted estimator in [16] or by using boundary corrected k-NN density estimates. Liitiäinen et.al. show the rate of decay of the subsequent term to be $o(T^{-1/d})$. From our analysis, we know that the optimal bias decays at the exact rate $\Theta(T^{\frac{-2}{2+d}})$ when $c_1c_2 > 0$ and $\Theta(T^{\frac{-4}{2+d}})$ when $c_1c_2 < 0$. Evans et.al. [9] have previously shown that the variance of k-NN based functionals is bounded above by the rate $O(k^5/T)$. Our result is an improvement in that we are able to provide an sharper rate of $\Theta(1/T)$.

5 Application to estimation of Shannon and Rényi entropy and Shannon MI

5.1 Estimation of entropy

In this section, we specifically discuss estimation of Shannon and Rényi entropies. The Shannon entropy of a density f is given by $H = -\int f \log f d\mu$, while the Rényi-α entropy is given by $H_\alpha = (1-\alpha)^{-1} \log \int f^\alpha d\mu$. For Shannon entropy, we choose the functional $g(u) = -\log(u)$ to obtain the plug-in estimator $\hat{H}$. For estimating Rényi entropy, we first estimate the integral $I_\alpha = \int f^\alpha d\mu$ by correspondingly choosing $g(u) = u^{\alpha-1}$ to obtain the plug-in estimator $\hat{\mathbf{I}}_\alpha$. We subsequently obtain the Rényi entropy estimator $\hat{H}_\alpha = (1-\alpha)^{-1}\log(\hat{\mathbf{I}}_\alpha)$. Assuming that the regularity conditions are satisfied, the bias and variance of the plug-in estimators $\hat{H}$ and $\hat{\mathbf{I}}_\alpha$ are given by theorems 3.1 and 3.2 above and the asymptotic distribution is described by theorem 3.3. We can then obtain expressions for the bias and variance of $\hat{\mathbf{H}}_\alpha$ by using a Taylor series expansion of the log function about I_α. We can similarly obtain a central limit theorem for $\hat{\mathbf{H}}_\alpha$ using the Delta method.

The functional form of the bias, variance and central limit theorem for these estimators is given by (5), (5) and (5) respectively. The constants for the Shannon entropy estimate $\hat{H}$ are given by $c_1 = \mathbb{E}[-f^{-(1+2/d)}(\mathbf{Y})(\Gamma^{(2/d)}((d+2)/2)tr[\nabla^2(f(\mathbf{Y}))])]$, $c_2 = 0.5$, $c_4 = \mathbb{V}[\log(f(\mathbf{Y}))]$ and $c_5 = 0$. The corresponding constants for the Rényi entropy estimate $\hat{H}_\alpha$ are given by $c_1 = -(1/I_\alpha)\mathbb{E}[f^{(\alpha-2-2/d)}(\mathbf{Y})(\Gamma^{(2/d)}((d+2)/2)tr[\nabla^2(f(\mathbf{Y}))])]$, $c_2 = (-1/2I_\alpha)(\alpha-2)\mathbb{E}[f^{\alpha-1}(\mathbf{Y})]$, $c_4 = (1/(1-\alpha)I_\alpha)^2\mathbb{V}[f^{\alpha-1}(\mathbf{Y})]$ and $c_5 = (\alpha-1)c_4$.

Nearest neighbor estimators of Shannon entropy and Rényi entropy [18, 16] have been previously proposed in literature. Liitiäinen et.al. [16] provide finite sample bias rates for the estimators proposed by Leonenko et.al. and propose weighted versions of the estimators proposed by Leonenko et.al. to improve rate of convergence of the bias. Denote the *data split* versions of the Shannon entropy estimator ((3.20) in [18]) and the Rényi entropy estimator ((3.13) in [18]) proposed by Leonenko et.al. by $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}}_\alpha$ respectively. We have the following relations: $\tilde{\mathbf{H}} = \hat{\mathbf{H}} + [\log(k-1) - \Psi(k-1)]$ and $\tilde{\mathbf{H}}_\alpha = (1-\alpha)^{-1}\log\tilde{\mathbf{I}}_\alpha$ where $\tilde{\mathbf{I}}_\alpha = (1/[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}])\hat{\mathbf{I}}_\alpha$.

An important distinction between our estimators and the estimators of Leonenko et.al. [18] is that we require the additional condition that the bandwidth k to grow to ∞ for asymptotic unbiasedness. This can be understood as follows: if we do not ignore the $o(1/k)$ terms in the expression for bias, we can show

$$\mathbb{E}[\hat{\mathbf{H}}] = I + [\log(M) - \Psi(M)] - [\log(k-1) - \Psi(k-1)] + c_1(k/M)^{2/d} + o((k/M)^{2/d}) \quad (14)$$

and

$$\mathbb{E}[\hat{\mathbf{I}}_\alpha] = [(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}]I_\alpha + c_1(k/M)^{2/d} + o((k/M)^{2/d}) \quad (15)$$

Note that $[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}] \to 1$ and $\Psi(k-1) = \log(k-1) - 1/(2k-2) + O(1/k^2)$ as $k \to \infty$. From the above equations we see that the scale factor $[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-$

$1)^{\alpha-1}]$ and the additive factor $[\log(k-1) - \Psi(k-1)]$ account for the $O(1/k)$ terms in the expressions for bias, thereby removing the requirement that $k \to \infty$ for asymptotic unbiasedness.

5.2 Estimation of Shannon Mutual information

The joint entropy of random vectors $\mathbf{X}$ and $\mathbf{Y}$ with joint density f_{XY} is given by

$$H(\mathbf{X}, \mathbf{Y}) = -\int f_{XY} \log(f_{XY}) d\mu, \tag{16}$$

where f_{XY} is the joint density of $\mathbf{X}$ and $\mathbf{Y}$. The Shannon MI between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ is then given by

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}). \tag{17}$$

We use a classic plug-in estimator to estimate MI from $N+M$ d-dimensional i.i.d samples $\{(\mathbf{X_i}, \mathbf{Y_i}); i = 1, \ldots, N+M\}$ of the underlying joint density f_{XY}. We estimate the Shannon MI by estimating the individual entropies. We estimate the joint Shannon entropy $H(\mathbf{X}, \mathbf{Y})$ from samples using the *plug-in* estimate

$$\hat{\mathbf{H}}(\mathbf{X}, \mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} -\log(\hat{\mathbf{f}}_{\mathbf{XY}}(\mathbf{X_i}, \mathbf{Y_i})), \tag{18}$$

where $\hat{\mathbf{f}}_{\mathbf{XY}}$ is a k nearest neighbor density estimate (kNN) estimated using the remaining M samples.

The kNN density estimate [17] is given by

$$\hat{\mathbf{f}}_{\mathbf{XY}}(X, Y) = \frac{k-1}{M \mathbf{V_k}(X, Y)}, \tag{19}$$

where $\mathbf{V_k}(X, Y)$ is the volume corresponding to the kth nearest neighbor distance between the point of density estimation (X, Y) and the M i.i.d samples $\{(\mathbf{X_i}, \mathbf{Y_i}); i = N+1, \ldots, N+M\}$.

We estimate the marginal entropies by first obtaining estimates of the marginal density using kNN density estimates

$$\hat{\mathbf{f}}_{\mathbf{X}}(X) = \frac{k-1}{M \mathbf{V_k}(X)}, \tag{20}$$

where $\mathbf{V_k}(X)$ is the volume corresponding to the kth nearest neighbor distance between the point of density estimation X and the M i.i.d samples $\{\mathbf{X_i}; i = N+1, \ldots, N+M\}$, and then plugging the estimated marginals into Eq. 21.

$$\hat{\mathbf{H}}(\mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} -\log(\hat{\mathbf{f}}_{\mathbf{X}}(\mathbf{X_i})). \tag{21}$$

Denote the estimated MI by $\hat{\mathbf{I}}$.

$$\hat{\mathbf{I}} = \hat{\mathbf{H}}(\mathbf{X}) + \hat{\mathbf{H}}(\mathbf{Y}) - \hat{\mathbf{H}}(\mathbf{X}, \mathbf{Y}). \tag{22}$$

We make the assumption that f_{XY} is three times continuously differentiable. Under this assumption, as in the case of entropy, we can show the following results on bias, variance and asymptotic distribution. Note that the results here require cross moments between density estimates of the joint and marginal densities, which while not discussed in this report, can be obtained in exactly the same manner as computing cross moments between the same density.

Bias The bias of the plug-in estimator $\hat{\mathbf{I}}$ is given by

$$\text{Bias}(\hat{\mathbf{I}}) = c_{b1}\left(\frac{k}{M}\right)^{2/d} + c_{b2}\left(\frac{1}{k}\right) + o\left(\left(\frac{k}{M}\right)^{2/d} + \frac{1}{k}\right),$$

where

$$\begin{aligned} c_{b1} &= \mathbb{E}\Big[-c_d f_{XY}^{-(d+2)/d}(\mathbf{X}, \mathbf{Y}) tr[\nabla^2(f_{XY}(\mathbf{X}, \mathbf{Y}))]\Big], \\ c_{b2} &= 0.5, \end{aligned}$$

are constants which depend on the underlying density f_{XY} and the constant $c_d = (\Gamma^{(2/d)}((d+2)/2))/(\pi(d+2))$.

Variance The variance of the plug-in estimator $\hat{\mathbf{I}}$ is given by

$$\text{Var}(\hat{\mathbf{I}}) = c_v\left(\frac{1}{N}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right),$$

where

$$c_v \;=\; Var\left[\log\left(\frac{f_X(\mathbf{X}) f_Y(\mathbf{Y})}{f_{XY}(\mathbf{X}, \mathbf{Y})}\right)\right],$$

is a constant which depends on the underlying density.f_{XY}.

Asymptotic distribution Let $\mathbf{Z}$ be a standard normal random variable. Then,

$$\lim_{N,M\to\infty} Pr\left(\frac{\sqrt{N}(\hat{\mathbf{I}} - \mathbb{E}[\hat{\mathbf{I}}])}{\sqrt{c_v}} \leq \alpha\right) = Pr(\mathbf{Z} \leq \alpha).$$

6 Simulations

We consider the problem of Shannon entropy estimation for a 2 dimensional distribution. We consider two different types of densities:

1. Uniform distribution
2. 2 dimensional mixture density $f_m = pf_\beta + (1-p)f_u$; f_β: Beta density with parameters a=4,b=4; f_u: Uniform density; Mixing ratio $p = 0.8$

The first set of simulation results illustrates that the corrections suggested for density estimates close to boundaries indeed works. The second set verifies the theoretical results on the bias, variance and central limit theorem.

6.1 Boundary correction

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 for these two distributions. This is shown in Fig. 5.

The uniform density clearly suffers from boundary effects. As discussed earlier, the theoretically predicted bias (ignoring higher order terms) for this estimator is $1/2k$ for the uniform distribution. 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, there is significant discrepancy between the bias observed in the uncorrected estimator and the predicted bias, as we should expect. In fact, the bias of the uncorrected estimator increases with increasing k, which is in direct contrast to the theoretically predicted trend of $1/2k$. This can be attributed to the fact that as k increases, the fraction of boundary points increase, which in turn results in the bias contribution from these boundary points to increase.

On the other hand, 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 much smaller as compared to the uniform density (because the probability density has very small mass towards the boundary of the support of the density). As a result, the boundary corrected estimator does not show any significant improvement over the uncorrected estimator in this case.

6.2 Validation of theory

To validate our theory, we once again use the 2D mixture density defined above. In the first experiment, we plot experimentally obtained and theoretically computed bias for fixed N,M. The results are shown in Fig. 6(a). The theoretically predicted optimal choice of k minimizes the experimentally obtained bias curve. Our theory can therefore be used to

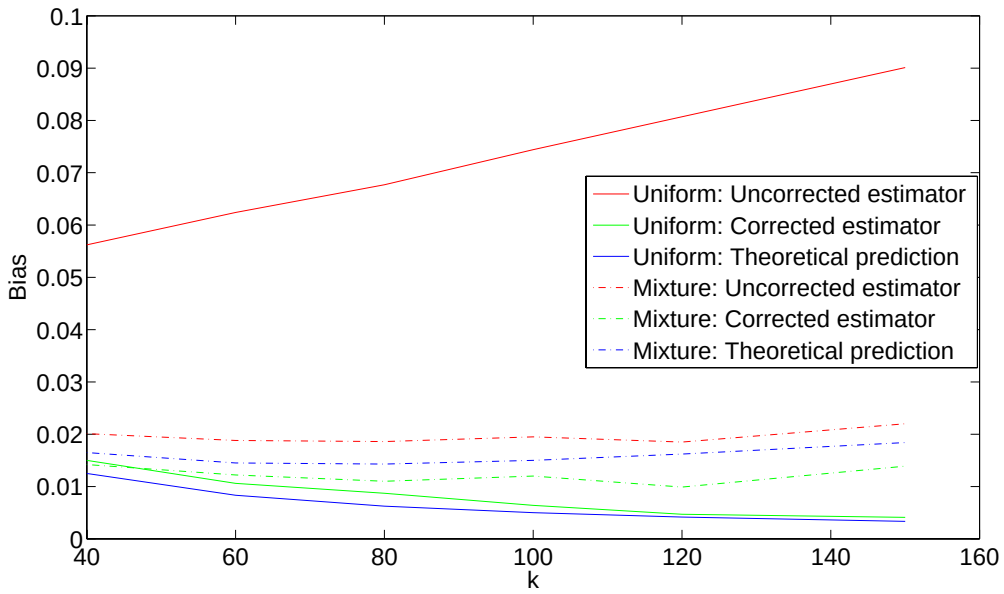
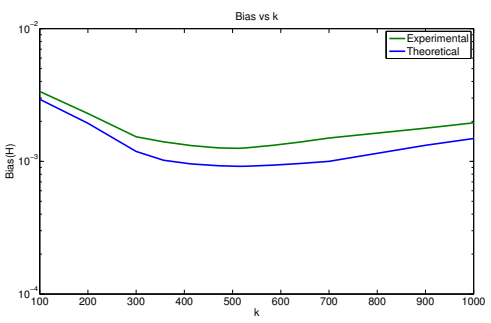
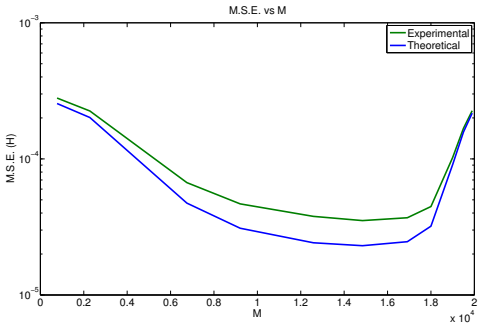


Figure 5: Variation of bias of estimated Shannon entropy vs k for fixed N = 1000,M=9000.



(a) Optimal choice of k. $k_{opt} = 513$.



(b) Optimal choice of M and $N = T - M$. $M_{opt} = 14686$, $N_{opt} = 5314$.

Figure 6: Optimal parameter selection.

specify bandwidth parameters for minimum bias. In the next experiment, we plot experimentally obtained and theoretically computed MSE for fixed T. The results are shown in Fig. 6(b). The theoretically predicted optimal choice of M and N minimizes the experimentally obtained MSE curve. Our theory can therefore be used to specify optimal partitioning of sample space for minimum MSE. Finally, we show the Q-Q plot of the normalized MI estimate and the standard normal distribution in Fig. 6.2. The linear Q-Q plot validates our theorem on asymptotic normality of the plug-in estimator.

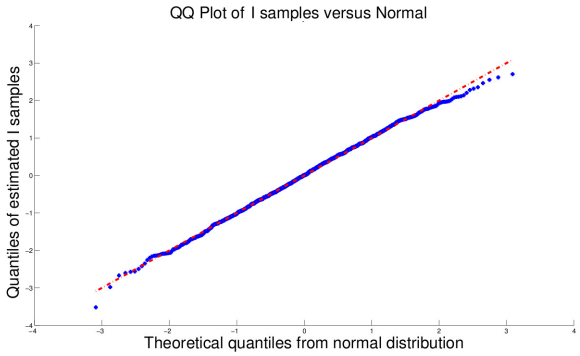
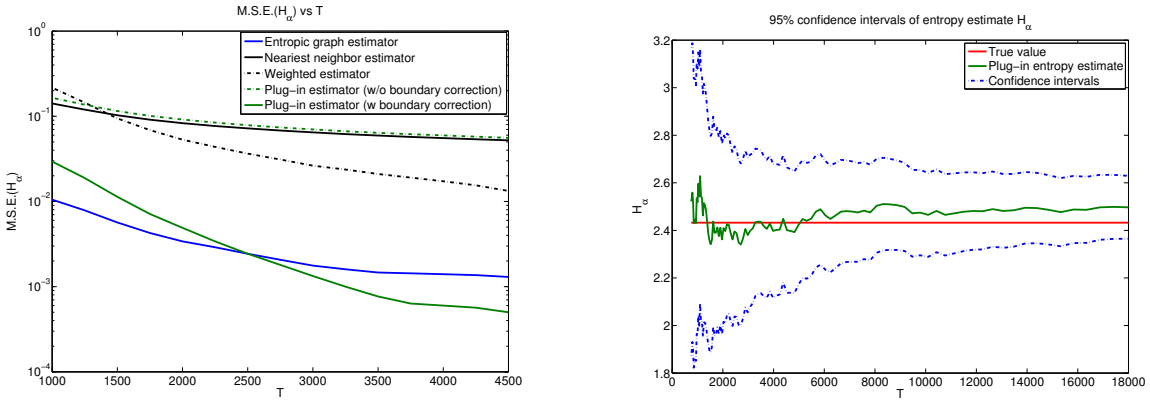


Figure 7: Asymptotic normal distribution of plug-in estimator for Shannon entropy.



(a) Variation of M.S.E.

(b) Confidence intervals.

Figure 8: Variation of M.S.E. with sample size and corresponding confidence intervals for Rényi plug-in estimator.

6.3 Comparison of estimators

For comparison purposes, we estimate Rényi entropy using the estimator $\hat{H}_\alpha$ described in Section 5.1 for the choice $\alpha = 0.5$. We compare the M.S.E. performance of our estimator with the entropic graph estimator of Hero et.al. [11], the k-nearest neighbor estimator of Leonenko et.al. [18] and the weighted k-nearest neighbor estimator of Liitiäinen et.al. [16].

Finally, using the CLT, we plot the 95% confidence intervals for our estimator as a function of sample size in Fig. 8(b).

6.4 Effect of dimension

As a final experiment, we plot the ratio of the variance against the squared bias and the ratio of the optimal choice of k over M and the optimal choice of N over M as a function

of dimension for the Rényi plug-in estimator. The results in Fig. 9 show that the bias becomes dominant as dimension increaces as predicted by theory. In addition, the optimal neighborhood size k/M and the ratio of optimal samples allocated for functional estimation N compared to density estimation M decreace as a function of dimension in accordance with theory.

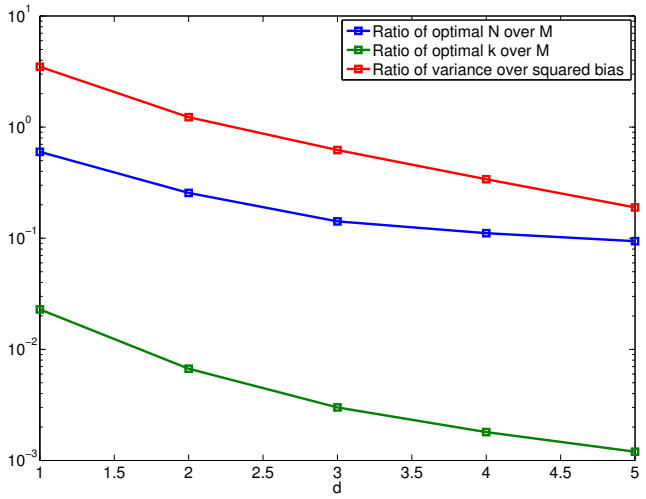


Figure 9: Variation of ratio of variance w.r.t bias, and optimal parameter choices as a function of dimension d for Rényi plug-in estimator. $T = 10000$.

7 Application to anomaly detection in wireless sensor networks

We apply our theory to the problem of anomaly detection in wireless sensor networks. The experiment was set up on a Mica2 platform, which consists of 14 sensor nodes randomly deployed inside and outside a lab room. Wireless sensors communicate with each other by broadcasting and the received signal strength (RSS), defined as the voltage measured by a receiver's received signal strength indicator circuit (RSSI), was recorded for each pair of transmitting and receiving nodes. There were $14 \times 13 = 182$ pairs of RSSI measurements over a 30 minute period, and each sample was acquired every 0.5 sec. During the measuring period, students walked into and out of lab at random times, which caused anomaly patterns in the RSSI measurements. Finally, a web camera was employed to record activity for ground truth.

The mission of this experiment is to use the 182 RSS sequences to detect any intruders (anomalies). We note that the ground truth indicator is only for evaluating the detecting

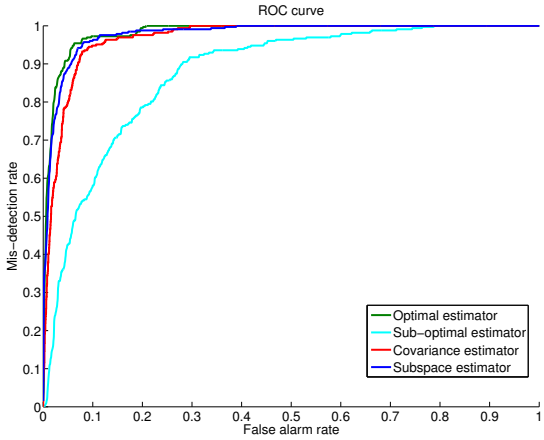


Figure 10: ROC curve for sub-optimal and optimal entropy estimators. The performance of the optimal estimator is clearly superior to the performance of the sub-optimal estimator.

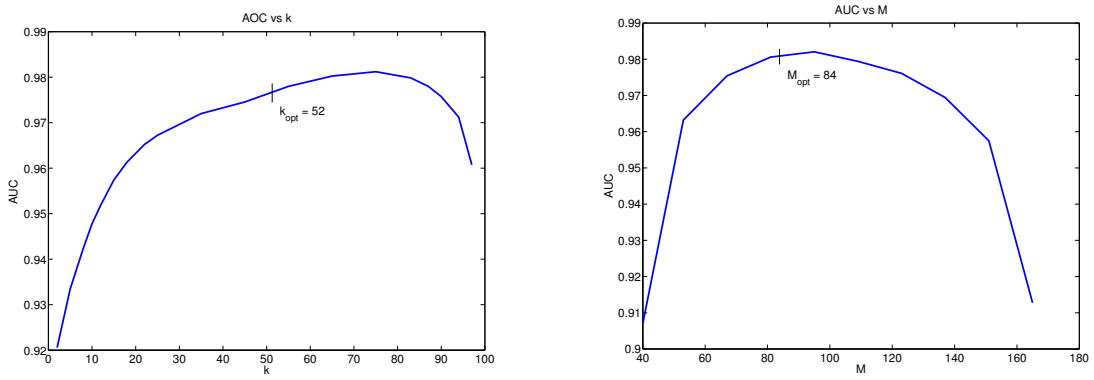
performance and the detection schemes presented here are conduct in a completely unsupervised way. To remove the temperature drifts of receivers we pre-process the data by removing their local mean values. Let $x_i[n]$ be the n-th sample of the i-th signal and denote $x[n] = (x_1[n], \ldots, x_{182}[n])'$. Due to temperature drifts, certain trends exist in $x[n]$. We de-trend the data by $y[n] = x[n] - \bar{x}[n]$ yielding $y[n]$ for anomaly detection, where $\bar{x}[n] = (2m+1)^{-1} \sum_{i=n-m}^{n+m} x[n]$ is the local mean value. We set $m = 50$ in this experiment.

We now estimate the Rényi entropy $H_i[n]$ for the choice $\alpha = 0.5$ for each 1-dimensional sequence $y[n]$ using the estimator $\hat{H}_\alpha$ described in Section 5.1. We perform anomaly detection by thresholding the entropy estimate $H_i[n]$. A time sample is regarded to be anomalous if the entropy estimate $H_i[n]$ exceeds a specified threshold. We estimate the Shannon entropy for different choices of parameters $\{k, N, M\}$ including the optimal choice $\{k_{opt}, N_{opt}, M_{opt}\}$.

In this experiment, the estimated constants c_1 and c_2 are of opposite signs. The optimal choice of k is given by $k_{opt} = \lfloor k_0 M^{\frac{2}{2+d}} \rfloor$ where $k_0 = (|c_2|/|c_1|)^{\frac{d}{d+2}}$. Because the dimension $d = 1$, the optimal choice of N and M is given by $N_{opt} = \sqrt{c_4/c_5} M_{opt}$ which in turn implies $M_{opt} = (1/(1+\sqrt{c_4/c_5})T)$ and $N_{opt} = T - M_{opt} = (\sqrt{c_4/c_5}/(1+\sqrt{c_4/c_5})T)$. We find the optimal partition to be $M_{opt} = 84$, $N_{opt} = 182 - 84 = 98$. The corresponding k_{opt} for this optimal partition is $k_{opt} = 46$.

ROC curves corresponding to a sub-optimal estimator and the optimal plug-in estimator are shown in Fig. 10 in addition to the ROC curves using the subspace method of Lakhina et.al. [13] and the covariance based estimator of Chen et.al. [5]. It is clear that the detection performance using the optimal estimator is superior to the performance using the sub-optimal estimator and is marginally better than the subspace subspace and covariance based estimators of Lakhina et.al. and Chen et.al. respectively.

In Fig. 11(a), the Area Under the ROC curve (AUC) is shown for varying choices of k for fixed partition $\{N = 72, M = 110\}$. In Fig. 11(b), AUC is shown for varying choices of



(a) AUC for different choices of k with fixed partition N, M.

(b) AUC for different choices of M. For each choice of M, k was chosen optimally.

Figure 11: Variation of AUC with parameters $\{k, N, M\}$.

partition $\{N, M\}$ with k being chosen optimally for each partition.

8 Application to structure discovery

Discovering structural dependencies among random variables from a multivariate sample is an important task in signal processing, pattern recognition and machine learning. Based on dependence relationships, the density function of the variables can be modeled using factor graphs. When the sample is highly structured, the corresponding factor graph configuration is sparse. Sparse factor graphs correspond to joint multivariate distributions which separate into a parsimonious product of few lower dimensional distributions. The inherent low-dimensional nature of this product leads to a compact representation of the variables having sparse factor graph configurations.

In practice, these structure dependencies have to be discovered from sample realizations of the multivariate distribution. Discovering dependencies when parametric probability density models are not known a priori is an important restriction of the above problem. For parametric distribution estimates, the errors are of order $O(1/N)$ if the true distribution is included in the parametric model. If not, a non-vanishing bias will dominate the error yielding an even higher error than that of a nonparametric distribution estimate (e.g. kNN estimates). In this restricted setting, recourse is therefore taken to nonparametric methods.

Chow et.al. [6] proposed an elegant solution to structure discovery of Markov tree distributions and provided a nonparametric algorithm to obtain the optimal tree. Ihler et.al. [12] developed the method of nonparametric hypothesis tests for structure discovery.

Nonparametric methods, while asymptotically consistent, can uncover incorrect factor graph structure when estimated from a finite number of samples. This is distinctly true for small

sample sizes. While consistency is an important qualitative property, there is clearly an important motivation for quantitative characterization of performance in structure discovery. In this work, we analyze factor graph structure discovery in the finite sample size setting.

We present a class of k-nearest neighbor (kNN) based nonparametric geometric algorithms to discover factor graph structure among variables. We provide results on mean square error of the nonparametric estimates, which can be optimized over free parameters, thereby guaranteeing improved correct structure discovery. In addition, we provide confidence intervals on these nonparametric estimates to determine the probability of false error in choosing an incorrect structure model. These results are an direct extension of our work on optimized nonparametric estimates of divergence measures introduced earlier.

As a consequence of our statistical analysis, we introduce the notion of dependence-based **dimension** for factor graph models and show that comparing models within the same dimension class is an easier task with lower probability of false error as compared to comparing models across different dimensions.

8.1 Factor graphs

Factor graphs are bipartite graphs used to represent factorizations of probability density functions. Consider a set of variables $\underline{X} = \{X_1, X_2, \ldots, X_T\}$ and let $\{S_j \subseteq \{X_1, X_2, \ldots, X_n\}, j = 1, \ldots, m\}$ be a set of subsets of $\underline{X}$. Let $g(X_1, \ldots, X_T)$ denote a probability density function on the random vector $\underline{X}$. For the factorization $g(X_1, \ldots, X_T) = \prod_{j=1}^{m} f_j(S_j)$ of the density function, the corresponding factor graph $G = (\underline{X}, \underline{F}, E)$ consists of variable vertice's $\underline{X}$, factor vertices's $\underline{F} = \{f_1, f_2, \ldots, f_m\}$, and edges E. The edges in the factor graph depend on the factorization as follows: there is an undirected edge between factor vertex f_j and variable vertex X_k when $X_k \subseteq S_j$.

8.2 Factor graph discovery

Problem statement: Consider a set of factor graphs $\{g_i(X_1, \ldots, X_T), i = 1, \ldots, I\}$. We seek to find the factor graph configuration from this set that best models the data.

The Kullback-Leibler (KL) divergence measure induces a **geometry** on the space of probability distributions. On this induced geometry, we naturally define the best factor graph configuration g_o to be the one closest to the actual distribution $p(X_1, \ldots, X_T)$ in terms of KL divergence (c.f. [6]).

$$g_o = \arg\min_{g_i} KL(p||g_i) = \arg\min_{g_i} H_c(p, g_i), \tag{23}$$

where $H_c(p, g_i) = -\int p \log g_i$ is the cross-entropy between p and g_i. In practice, these cross-entropy terms have to be estimated from the finite data sample. **Errors in estimation of cross-entropy terms can result in incorrect factor graph discovery**.

The problem considered by [6] is a specific instance of discovering factor graph structure. For the class of Markov tree factor graphs considered by [6], the cross entropy reduces to a sum of pairwise Shannon mutual information terms between variables with edges in the Markov tree. In their work, they empirically estimate the mutual information terms from the data using nonparametric estimators which are consistent. However, they do not take into account the error in the mutual information estimates when estimated from finite samples.

8.3 Disjoint factor graph discovery

In order to illustrate the effect of nonparametric estimation from finite sample size on factor graph discovery, we restrict our attention to disjoint factor graphs ([12]). For $i = 1, \ldots, I$, let

$$g_i(X_1, X_2, \ldots, X_T) = \prod_{j=1}^{m} p(S_j^{(i)}), \tag{24}$$

where $S_j^{(i)} \cap S_k^{(i)} = \phi$ whenever $j \neq k$, and $p(.)$ denotes the marginal density function. In this case of disjoint factor graphs, the cross-entropy takes the following simple form:

$$H_c(p, g_i) = \sum_j H(S_j^{(i)}), \tag{25}$$

where $H(S_j^{(i)})$ is the Shannon entropy of the variables $S_j^{(i)}$ under the true distribution p.

For example, consider the disjoint factor graph $g(X_1, \ldots, X_5) = p(X_1, X_2)p(X_3)p(X_4, X_5)$. The cross-entropy for this factor graph is given by $H_c(p, g) = H(X_1, X_2)+H(X_3)+H(X_4, X_5)$.

Consider two disjoint factor graph configurations: (a) $n(X_1, \ldots, X_T) = \prod_{i=1}^{m_1} f(R_i)$ and (b) $l(X_1, \ldots, X_T) = \prod_{j=1}^{m_2} f(S_j)$. Denote the dimension of R_i by d_i^n and S_j by d_j^l. We note that $\sum_{i=1}^{m_1} d_i^{(n)} = \sum_{j=1}^{m_2} d_j^{(l)} = T$. Based on the above formulation, in order to compare the two potential factor graph models n and l, we need to compare the respective cross-entropy terms. The cross entropy test is stated below.

Cross entropy test: The cross entropy test to compare between models n and l is given by

$$H_c(p, n) - H_c(p, l) = \sum_{i=1}^{m_1} H(R_i) - \sum_{j=1}^{m_2} H(S_j) \gtrless 0. \tag{26}$$

We estimate these entropy terms in the test statistic $H_c(p, n) - H_c(p, l)$ from sample realizations using kNN plug-in estimators introduced earlier.

8.4 Errors in factor graph discovery

To illustrate the effect of estimation error in factor graph discovery, again consider the two factor graph models $n(X_1, \ldots, X_T) = \prod_{i=1}^{m_1} f(R_i)$ and $l(X_1, \ldots, X_T) = \prod_{j=1}^{m_2} f(S_j)$.

The cross entropy test (Eq. 25) between models n and l is $H_c(p,n) - H_c(p,l) \gtrless 0$. We replace this optimal cross entropy test with the following **surrogate** cross entropy test:

$$\hat{H}_c(p,n) - \hat{H}_c(p,l) = \sum_{i=1}^{m_1} \hat{H}(R_i) - \sum_{j=1}^{m_2} \hat{H}(S_j) \gtrless 0. \tag{27}$$

where we estimate entropy terms $\hat{H}(R_i)$ or $\hat{H}(S_j)$ using independent realizations of the underlying density p. To elaborate, if we have V samples $\{\underline{X}^{(1)}, \ldots, \underline{X}^{(V)}\}$ from the density p, we partition these V samples into $m_1 + m_2$ disjoint subsets of size $N + M$ each. This implies that $N + M \approx V/(m_1 + m_2)$. We then use each subset to estimate entropy using the partitioning strategy as discussed earlier.

Denote the coefficients corresponding to the entropy estimate $\hat{H}(R_i)$ of the subset of variables R_i in the factor graph model n by $c_{n_i 1}$, $c_{n_i 2}$ and $c_{n_i 4}$. Using the theorems established in this report, we have the following results:

Mean: The mean of this surrogate test statistic is then given by

$$\begin{aligned} \mathbb{E}_p[\hat{H}_c(p,n) - \hat{H}_c(p,l)] &= H_c(p,n) - H_c(p,l) \\ &+ \sum_{i=1}^{m_1} c_{n_i 1} \left(\frac{k}{M}\right)^{2/d_i^{(n)}} - \sum_{j=1}^{m_2} c_{l_j 1} \left(\frac{k}{M}\right)^{2/d_j^{(l)}} \\ &+ \sum_{i=1}^{m_1} c_{n_i 2}/k - \sum_{j=1}^{m_2} c_{l_j 2}/k. \end{aligned} \tag{28}$$

Variance: The variance of the surrogate test statistic is then given by the sum of the variance of the individual entropy estimates (by independence)

$$\mathbb{V}_p[\hat{H}_c(p,n) - \hat{H}_c(p,l)] = \left(\sum_{i=1}^{m_1} c_{n_i 4} + \sum_{j=1}^{m_2} c_{l_j 4}\right)\left(\frac{1}{N}\right). \tag{29}$$

Weak convergence: Again, by independence of the individual entropy estimates, we have the following weak convergence law

$$\lim_{N,M\to\infty} Pr\left(\frac{\sqrt{N}(\hat{H}_c(p,n) - \hat{H}_c(p,l) - \mathbb{E}_p[\hat{H}_c(p,n) - \hat{H}_c(p,l)])}{\sqrt{\mathbb{V}_p[\hat{H}_c(p,n) - \hat{H}_c(p,l)]}} \leq \alpha\right) = Pr\left(Z \leq \alpha\right), \tag{30}$$

where Z is standard normal.

8.5 Discussion

From the above expressions for the mean, variance and weak convergence law of the surrogate test statistic, we make the following observations:

1. The bias term is dependent on the dimension of the factors of the factor graph models $d_i^{(n)}$ and $d_j^{(l)}$. The variance term is independent of dimension. Furthermore, it is clear that the bias term dominates the MSE as the dimension of the factors grows.

2. For better performance in discovering factor graph structure using cross entropy tests, it is clear that we want the MSE of the surrogate test statistic to be small. A significant route to achieving this is to get the bias from each factor graph cross entropy estimate in the estimated test statistic to cancel. This is to say, we want

$$\begin{aligned} \mathbb{E}_p[\hat{H}_c(p,n) - \hat{H}_c(p,l)] &\approx H_c(p,n) - H_c(p,l) \\ \Rightarrow \mathbb{E}_p[\hat{H}_c(p,n)] - \hat{H}_c(p,n) &\approx \mathbb{E}_p[\hat{H}_c(p,l)] - \hat{H}_c(p,l) \\ \Rightarrow \sum_{i=1}^{m_1} c_{n_i 1} \left(\frac{k}{M}\right)^{2/d_i^{(n)}} + \sum_{i=1}^{m_1} c_{n_i 2}/k &\approx \sum_{j=1}^{m_2} c_{l_j 1} \left(\frac{k}{M}\right)^{2/d_j^{(l)}} + \sum_{j=1}^{m_2} c_{l_j 2}/k. \end{aligned} \quad (31)$$

3. This cancellation effect will be maximized when the dimensions of the factor graph subsets R_i and S_j match. That is to say, we want $m_1 = m_2$ and furthermore $d_i^{(n)} = d_j^{(l)}$. In this case, the bias from each cross entropy estimate are of the same order and will nearly cancel.

   On the other hand, when there is a mismatch in dimension, the bias from one cross entropy estimate will dominate the bias from the other cross entropy estimate, resulting in significant bias in the surrogate test statistic.

   In both these cases, the variance of the surrogate test statistic will be of the same order $O(1/N)$.

4. This gives rise to notion of multivariate dimension for factor graphs. Index the factorizations according to the vector $E = [e_1, e_2, ..., e_p]$, where e_i is an integer between 0 and T that counts the number of factors of order i, i.e. involving a marginal density over i variables. The **dimension** E of factor graph configurations partitions the factor graphs into equivalence classes having nearly constant cross entropy estimate bias.

   For two factor graph models n and l with dimensions E_n and E_l, we will refer to n as a higher dimensional model relative to l if the last non-zero entry of $E_n - E_l$ is positive.

5. As discussed earlier, the bias will not be a significant factor when comparing models over an equivalence class having fixed values of E. On the other hand, the bias will be significant when comparing models across different values of E, resulting in higher probability of error in factor graph discovery.

6. Prior knowledge of the equivalence class will therefore translate into much improved performance in factor graph discovery as compared to prior knowledge that mixes between equivalence classes.

7. We note that the number of samples required to maintain a constant level of bias grows **geometrically** with dimension E.

8. Using the expressions for the bias and variance of the surrrogate test statistic, we can optimize over the free parameters: (a) the choice of partition N and M for fixed total sample size $N + M$ and (b) the choice of bandwidth parameter k, for minimum MSE.

9. Using the weak convergence law, we can theoretically predict the probability of choosing model n over model l using the surrogate cross entropy test.

8.6 Experiment

We illustrate the implications of our analysis with a toy example. Let $f_\beta(x, a, b, d)$ denote a beta density of dimension d with parameters a and b. Now let $f_\mu(x, d) = 0.5 f_\beta(x, 5, 2, d) + 0.5 f_\beta(x, 2, 5, d)$ be a mixture of beta densities. When $d > 1$, the mixing of densities ensures there is strong dependence between the variates.

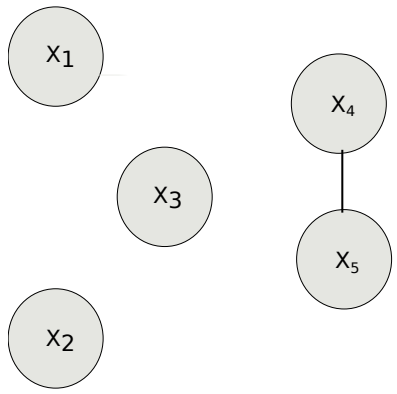


Figure 12: True factor graph representation of the 5-dimensional joint density $p(X_1, \ldots, X_5) = f_\mu(X_1, 1) f_\mu(X_2, 1) f_\mu(X_3, 1) f_\mu(X_4, X_5, 2)$.

We draw $V = 10^5$ independent sample realizations from the joint density $p(X_1, \ldots, X_5) = f_\mu(X_1, 1) f_\mu(X_2, 1) f_\mu(X_3, 1) f_\mu(X_4, X_5, 2)$.

	E	True	False
l	$[1, 0, 0, 1, 0]$	$f(X_1, X_2, X_4, X_5) f(X_3)$	$f(X_1, X_2, X_3, X_4) f(X_5)$
m	$[1, 2, 0, 0, 0]$	$f(X_1, X_2) f(X_4, X_5) f(X_3)$	$f(X_1, X_3) f(X_2, X_4) f(X_5)$
n	$[3, 1, 0, 0, 0]$	$f(X_4, X_5) f(X_1) f(X_2) f(X_3)$	$f(X_2, X_4) f(X_1) f(X_3) f(X_5)$

Experiment The table above shows six different factor graph models. We compare each true model against each false model. Denote the true models by l_T, m_T and n_T and the corresponding false models by l_F, m_F and n_F. We note that the true cross entropy terms $H_c(p, l_T) = H_c(p, m_T) = H_c(p, n_T)$ and $H_c(p, l_L) = H_c(p, m_L) = H_c(p, n_L)$. This guarantees level playing field when comparing each true model against each false model using the surrogate cross entropy test.

For the surrogate cross entropy test, we set $N = .2 * 10^4$, $M = .8 * 10^4$ and $k = 20$. We note that the maximum value of $m_1 + m_2$ for the above set of tests is 8 and that $V/8 > (N + M)$. This choice of N and M therefore ensures that there are enough samples V to guarantee sufficient number of independent samples for estimating individual entropies (see Section 5).

The table below lists the probability (experimental/theoretical prediction[1]) of choosing the false model over the true model for the various tests.

Same true vs Same false	l_T vs l_F	m_T vs m_F	n_T vs n_F
Error (Exp/Theor)	0.071/0.032	0.067/0.066	0.068/0.028
High true vs Low false	l_T vs m_F	l_T vs n_F	m_T vs n_F
Error (Exp/Theor)	0/0	0/0	0/0
Low true vs High false	m_T vs l_F	n_T vs l_F	n_T vs m_F
Error (Exp/Theor)	0.689/0.732	0.995/1.000	0.691/0.665

Explanation For the class of models above, the set of constants $\{c_{n_i 1}, c_{l_j 1}\}$ are always negative. As a result, when comparing a high dimensional model to a low dimensional model, the additional bias will strongly tilt the test statistic towards the higher dimensional model. As a result, there is a greater chance of detecting the higher dimension model in the surrogate cross entropy test, irrespective of whether the higher dimensional model is true or false.

To elaborate, when the high dimensional model is true and the low dimensional model is false, the bias will further tilt the test statistic towards the high dimensional model, resulting in zero false detections. On the other hand, when the low dimensional model is true, the bias in the surrogate test statistic deviates towards the high dimensional model, resulting in a high number of false detections. When we compare factor graph models within the same class of dimension, the bias from the cross entropy estimates for each model nearly cancel, resulting in a surrogate test statistic with much smaller bias as compared to the above two cases. As a result, the number of false detections is correspondingly low when comparing models within the same dimension.

By the same argument, for factor graph models where the set of constants $\{c_{n_i 1}, c_{l_j 1}\}$ are positive, we can conclude that the surrogate test statistic will be biased towards lower dimensional models.

9 Application to intrinsic dimension estimation

In this work we introduce a new dimensionality estimator that is based on fluctuations of the sizes of nearest neighbor balls centered at a subset of the data points. In this respect it is similar to Costa's k-nearest neighbor (kNN) graph dimension estimator [7] and to Farahmand's

[1]The theoretical prediction requires estimation of constants $c_{l_i 1}, c_{l_i 2}$ and $c_{l_i 3}$. These constants were estimated from the data using oracle Monte Carlo methods which utilized the true form of the density p. In practice, when the true form of p is never known, we adopt methods given by [22] to estimate these constants from data.

dimension estimator based on nearest neighbor distances [10]. The estimator can also be related to the Leonenko's Rényi entropy estimator [14]. However, unlike these estimators, our new dimension estimator is derived directly from a mean squared error (M.S.E.) optimality condition for partitioned kNN estimators of multivariate density functionals. This guarantees that our estimator has the best possible M.S.E. convergence rate among estimators in its class. Empirical experiments are presented that show that this asymptotic optimality translates into improved performance in the finite sample regime.

9.1 Problem formulation

Let $\mathcal{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_T\}$ be T independent and identically distributed sample realizations in $\mathbb{R}^D$ distributed according to density f. Assume the random vectors in $\mathcal{Y}$ are constrained to lie on a d-dimensional Riemannian submanifold $\mathcal{S}$ of $\mathbb{R}^D$ ($d < D$). We are interested in estimating the intrinsic dimension d.

9.2 Log-length statistics

Let $\gamma > 0$ be any arbitrary number and $\alpha = \gamma/d$. Partition the T samples in $\mathcal{Y}$ into two disjoint sets $\mathcal{X}$ and $\mathcal{Z}$ of size $\lfloor T/2 \rfloor$ each. Denote the samples of $\mathcal{X}$ as $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_{\lfloor T/2 \rfloor}\}$ and $\mathcal{Z}$ as $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_{\lfloor T/2 \rfloor}\}$.

Partition $\mathcal{X}$ into N 'target' and M 'reference' samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{\lfloor T/2 \rfloor}\}$ respectively with $N + M = \lfloor T/2 \rfloor$. Partition $\mathcal{Z}$ in an identical manner. Now consider the following statistics based on the partitioning of sample space:

$$\mathbf{L}_{\mathbf{k}}(\mathcal{X}) = \frac{\gamma}{N} \sum_{i=1}^{N} \log \left(\mathbf{R}_{\mathbf{k}}(\mathbf{X}_i) \right),$$

where $\mathbf{R}_{\mathbf{k}}(\mathbf{X}_i)$ is the Euclidean k nearest neighbor (kNN) distance from the target sample $\mathbf{X}_i$ to the M reference samples $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{\lfloor T/2 \rfloor}\}$. This partitioning of samples is illustrated in Fig. 13.

9.3 Relation to kNN density estimates

Under the condition that k/M is small, the Euclidean kNN distance $\mathbf{R}_{\mathbf{k}}(\mathbf{X}_i)$ approximates the kNN distance on the submanifold $\mathcal{S}$. The kNN density estimate [19] of f at $\mathbf{X}_i$ based on the M samples $\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}$ is then given by

$$\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{\mathbf{i}}) = \frac{k-1}{M} \frac{1}{c_d \mathbf{R}_{\mathbf{k}}(\mathbf{X}_i)^d} = \frac{k-1}{M} \frac{1}{\mathbf{V}_{\mathbf{k}}(\mathbf{X}_i)},$$

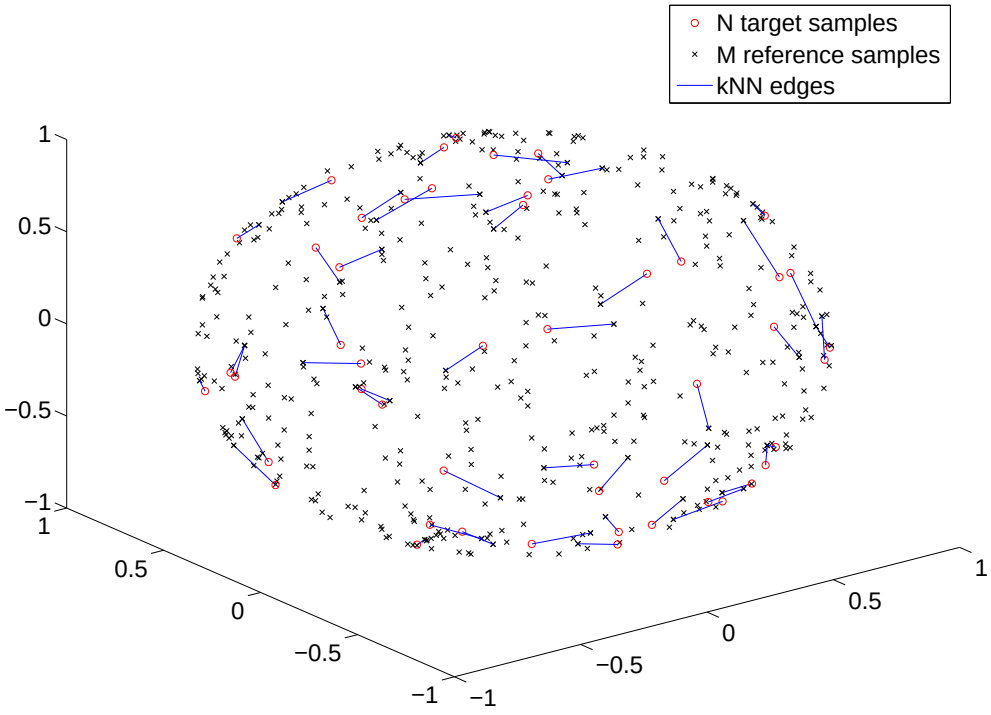


Figure 13: kNN edges on sphere manifold with uniform distribution for $d = 2$, $D = 3$, and $k = 5$.

where c_d is the volume of the unit ball in d dimensions and therefore $\mathbf{V_k}(\mathbf{X_i})$ is the volume of the kNN ball. This implies that $\mathbf{L_k}(\mathcal{X})$ can be rewritten as follows:

$$\begin{aligned}
\mathbf{L_k}(\mathcal{X}) &= \frac{\gamma}{N}\sum_{i=1}^{N}\log\left(\mathbf{R_k}(\mathbf{X}_i)\right) \\
&= \log\left(\frac{k-1}{Mc_d}\right)^{\alpha} + \frac{1}{N}\sum_{i=1}^{N}\log\left(\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_i)\right)^{-\alpha} \\
&= \alpha\log(k-1) - \frac{\alpha}{N}\sum_{i=1}^{N}\log\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_i) \\
&\quad -\alpha\log(c_d M). \qquad (32)
\end{aligned}$$

As eq. (32) indicates, the log-length statistics is linear with respect to $\log(k-1)$ with a slope of α. This prompts the idea of estimating α (and later d) from the slope of $\mathbf{L_k}(\mathcal{X})$ as a function of $\log(k-1)$.

9.4 Intrinsic dimension estimate based on varying bandwidth k

Let k_1 and k_2 be two different choices of bandwidth parameters. Let $\mathbf{L_{k_1}}(\mathcal{X})$ and $\mathbf{L_{k_2}}(\mathcal{Z})$ be the length statistics evaluated at bandwidths k_1 and k_2 using data $\mathcal{X}$ and $\mathcal{Z}$ respectively. A natural choice for the estimate of α would then be

$$\begin{aligned}\hat{\alpha} &= \frac{\mathbf{L_{k_2}}(\mathcal{Z}) - \mathbf{L_{k_1}}(\mathcal{X})}{\log(k_2 - 1) - \log(k_1 - 1)} \\ &= \alpha + \frac{\nu}{N}\sum_{i=1}^{N}\left(\log \hat{\mathbf{f}}_{\mathbf{k_2}}(\mathbf{Z}_i) - \log \hat{\mathbf{f}}_{\mathbf{k_1}}(\mathbf{X}_i)\right) \\ &= \alpha + \nu(\hat{\mathbf{E}}_{\mathbf{k_2}}(\mathcal{Z}) - \hat{\mathbf{E}}_{\mathbf{k_1}}(\mathcal{X})),\end{aligned}$$

where

$$\hat{\mathbf{E}}_{\mathbf{k}}(\mathcal{X}) = \frac{1}{N}\sum_{i=1}^{N}\log(\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X_i})),$$

and $\nu = -\alpha/\log((k_2 - 1)/(k_1 - 1))$. The intrinsic dimension estimate is related to $\hat{\alpha}$ by the simple relation $\hat{\mathbf{d}} = \gamma/\hat{\alpha}$.

9.5 Statistical properties of intrinsic dimension estimate

We can relate the error in estimation of α to the error in dimension estimation as follows:

$$\begin{aligned}\hat{\mathbf{d}} - d &= \gamma\left(\frac{1}{\hat{\alpha}} - \frac{1}{\alpha}\right) \\ &= \gamma\frac{\alpha - \hat{\alpha}}{\hat{\alpha}\alpha} \\ &= -\frac{\gamma}{\alpha^2}(\hat{\alpha} - \alpha) + o(\hat{\alpha} - \alpha).\end{aligned}$$

Define $\kappa = -\gamma\nu/\alpha^2$. We recognize that the density functional estimate $\hat{\mathbf{E}}_{\mathbf{k}}(\mathcal{X})$ is in the form of the plug-in estimators introduced in this report. Using the results on the bias, variance and asymptotic distribution of the density functional estimate $\hat{\mathbf{E}}_{\mathbf{k}}(\mathcal{X})$ established in this report and the above relation between the errors $\hat{\mathbf{d}} - d$ and $\hat{\alpha} - \alpha$, we then have the following statistical properties for the estimate $\hat{\mathbf{d}}$:

Estimator bias

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{d}}] - d &= \kappa c_{b_1}\left(\left(\frac{k_2}{M}\right)^{2/d} - \left(\frac{k_1}{M}\right)^{2/d}\right) \\ &+ \kappa c_{b_2}\left(\left(\frac{1}{k_2}\right) - \left(\frac{1}{k_1}\right)\right) \\ &+ o\left(\frac{1}{k_1} + \frac{1}{k_2} + \left(\frac{k_1}{M}\right)^{2/d} + \left(\frac{k_2}{M}\right)^{2/d}\right).\end{aligned}$$

Estimator variance

$$\mathbb{V}(\hat{\mathbf{d}}) \;=\; 2\kappa^2 c_v \left(\frac{1}{N}\right) + o\left(\frac{1}{M}+\frac{1}{N}\right).$$

Central limit theorem

Let $\mathbf{Z}$ be a standard normal random variable. Then,

$$\lim_{N,M\to\infty} Pr\left(\frac{\hat{\mathbf{d}}-\mathbb{E}[\hat{\mathbf{d}}]}{\sqrt{2\kappa^2 c_v/N}} \le \alpha\right) = Pr(\mathbf{Z}\le\alpha).$$

9.6 Optimal selection of parameters

We have theoretical expressions for the mean square error (M.S.E) of the dimension estimate $\hat{\mathbf{d}}$, which we can optimize over the free parameters k_1, k_2, N and M [24]. We restrict our attention to the case $k_2 = 2k$; $k_1 = k$. The M.S.E. of $\hat{\mathbf{d}}$ (ignoring higher order terms) is given by

$$\begin{aligned}
\text{M.S.E.}(\hat{\mathbf{d}}) &= (\mathbb{E}[\hat{\mathbf{d}}]-d)^2 + \mathbb{V}[\hat{\mathbf{d}}] \\
&= \left(C_{b_1}\left(\frac{k}{M}\right)^{2/d} + C_{b_2}\left(\frac{1}{k}\right)\right)^2 \\
&+ \; C_v\left(\frac{1}{N}\right). \qquad (33)
\end{aligned}$$

where $C_{b_1} = \kappa 2^{(2/d-1)}$, $C_{b_2} = \kappa/4$ and $C_v = 2\kappa^2 c_v$.

Optimal choice of bandwidth

The optimal value of k w.r.t the M.S.E. is given by

$$k_{opt} \;=\; \lfloor k_0 M^{\frac{2}{2+d}} \rfloor. \qquad (34)$$

where the constant $k_0 = (|C_{b_2}|d/2|C_{b_1}|)^{\frac{d}{d+2}}$.

Optimal partitioning of sample space

Under the constraint that $N+M = \lfloor T/2 \rfloor$ is fixed, the optimal choice of N as a function of M is then given by

$$N_{opt} = \lfloor N_0 M^{\frac{6+d}{2(2+d)}} \rfloor, \qquad (35)$$

where the constant $N_0 = \frac{\sqrt{C_v(2+d)}}{2b_0}$.

9.7 Improved estimator based on correlated error

Consider the following alternative estimator for α:

$$\begin{aligned}\tilde{\alpha} &= \frac{\mathbf{L_{k_2}}(\mathcal{X}) - \mathbf{L_{k_1}}(\mathcal{X})}{\log(k_2 - 1) - \log(k_1 - 1)} \\ &= \alpha + \kappa(\hat{\mathbf{E}}_{\mathbf{k_2}}(\mathcal{X}) - \hat{\mathbf{E}}_{\mathbf{k_1}}(\mathcal{X})),\end{aligned}$$

and the corresponding density estimate $\tilde{\mathbf{d}}$ which satisfies

$$\tilde{\mathbf{d}} - d = -\frac{\gamma}{\alpha^2}(\tilde{\alpha} - \alpha) + o(\hat{\alpha} - \alpha),$$

where both the length statistics at bandwidths k_1 and k_2 are evaluated using the same sample X. The density functional estimates $\hat{\mathbf{E}}_{\mathbf{k_1}}(\mathcal{X})$ and $\hat{\mathbf{E}}_{\mathbf{k_2}}(\mathcal{X})$ will be highly correlated (as compared to the independent quantities $\hat{\mathbf{E}}_{\mathbf{k_1}}(\mathcal{X})$ and $\hat{\mathbf{E}}_{\mathbf{k_2}}(\mathcal{Z})$). This implies that the variance of the difference $\hat{\mathbf{E}}_{\mathbf{k_2}}(\mathcal{X}) - \hat{\mathbf{E}}_{\mathbf{k_1}}(\mathcal{X})$ will be smaller when compared to $\hat{\mathbf{E}}_{\mathbf{k_2}}(\mathcal{Z}) - \hat{\mathbf{E}}_{\mathbf{k_1}}(\mathcal{X})$, (while the expectation remains the same).

Since the estimator bias is unaffected by this modification, the variance reduction suggests that $\tilde{d}$ will be an improved estimator as compared to $\hat{\mathbf{d}}$ in terms of M.S.E.. In order to obtain statistical properties for the improved estimator $\tilde{\mathbf{d}}$ (equivalent to the properties developed in Section 9.5 for the original estimator $\hat{\mathbf{d}}$), we need to analyze the joint distribution between $\hat{\mathbf{f}}_{\mathbf{k_1}}(X_i)$ and $\hat{\mathbf{f}}_{\mathbf{k_2}}(X_j)$ for two distinct values k_1 and k_2. Our theory, at present, cannot address the case of distinct bandwidths k_1 and k_2.

Since the estimate $\tilde{\mathbf{d}}$ has smaller M.S.E. compared to $\hat{\mathbf{d}}$, M.S.E. predictions for the estimate $\hat{\mathbf{d}}$ can serve as upper bounds on the M.S.E. performance of the improved estimate $\tilde{\mathbf{d}}$.

9.8 Simulations

We generate $T = 10^5$ samples $\mathcal{B}$ drawn from a $d = 2$ mixture density $f_m = .8f_\beta + .2f_u$, where f_β is the product of two 1 dimensional marginal beta distributions with parameters $\alpha = 2$, $\beta = 2$ and f_u is a uniform density in 2 dimensions. These samples are then projected to a 3-dimensional hyperplane in $\mathbb{R}^3$ by applying the transformation $\mathcal{Y} = U\mathcal{B}$ where U is a 3×2 random matrix whose columns are orthonormal. We apply our intrinsic dimension estimates on the samples $\mathcal{Y}$.

Optimal selection of free parameters

In our first experiment, we theoretically compute the optimal choice of k for a fixed partition with $M = 3.5 \times 10^4$ and $N = 1.5 \times 10^4$. We then show the variation of the theoretical and experimental M.S.E. of the estimate $\hat{\mathbf{d}}$ and the experimental M.S.E. of the improved estimate $\tilde{\mathbf{d}}$ with changing bandwidth k in Fig. 14. In our second experiment, we compute

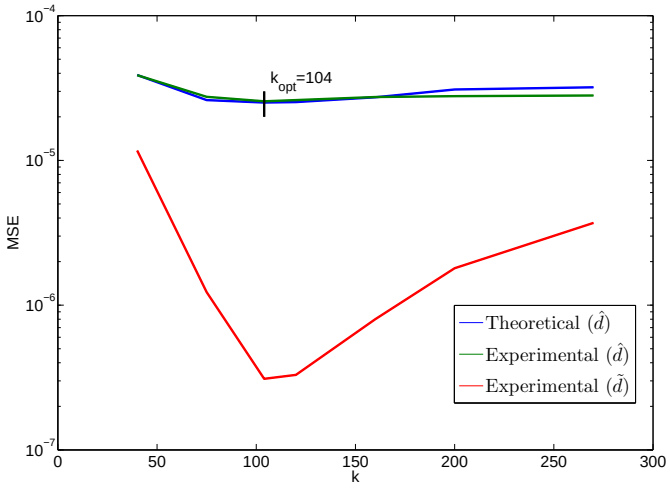


Figure 14: Comparison of theoretically predicted and experimental M.S.E. for varying choices of k. The experimental performance of the estimator $\hat{\mathbf{d}}$ is in excellent agreement with the theoretical expression and, as predicted by our theory, the modified estimator $\tilde{\mathbf{d}}$ significantly outperforms $\hat{\mathbf{d}}$.

the optimal partition according to eq. (35) and show the variation of M.S.E. with varying choices of partition in Fig. 15.

From our experiments, we see that there is good agreement between our theory and simulations. As a consequence, we find the theoretically predicted optimal choices of $k, N$and M to minimize the observed M.S.E.. In addition, as predicted by our theory, the modified estimator $\tilde{\mathbf{d}}$ significantly outperforms $\hat{\mathbf{d}}$. The theoretically predicted M.S.E. for $\hat{\mathbf{d}}$ therefore serves as a strict upper bound for the M.S.E. of the improved estimator $\tilde{\mathbf{d}}$.

Comparison of dimension estimation methods

We compare the performance of our proposed dimension estimators to the estimated proposed by Frahmand et. al. [10] (denote as $\hat{\mathbf{d}}_f$) and Costa et. al. [7] (denote as $\hat{\mathbf{d}}_j$).

Expressions for the optimal bandwidth k (eq. (34)) and partition N, M (eq. (35)) depend on the unknown intrinsic dimension d and constants c_{b_1}, c_{b_2} and c_v which depend on unknown density f. The constants c_{b_1}, c_{b_2} and c_v can be estimated from the data using plug-in methods similar to the ones used by Raykar et. al. [22] for optimal bandwidth selection for kernel density estimation . To establish the potential advantages of our dimension estimators we compare an omniscient optimal form of our estimator, for which the true values of these constants are known, to a suboptimal form of our estimator that does not know the constants.

For the optimal estimator, we theoretically compute the optimal choice for k, N and M

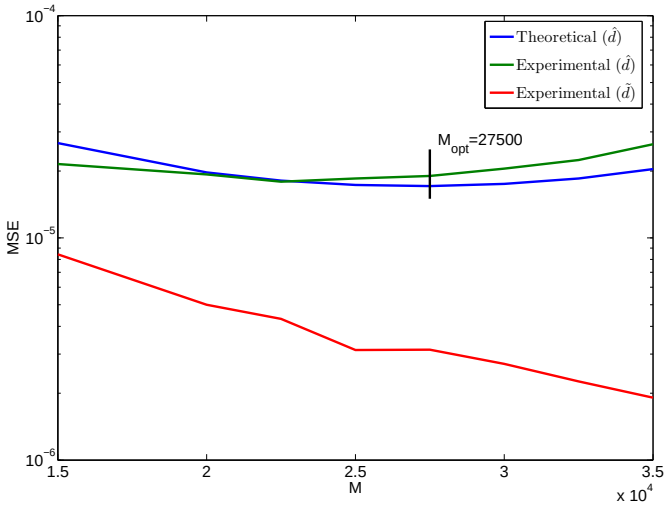


Figure 15: Comparison of theoretically predicted and experimental M.S.E. for varying choices of M. The experimental performance of the estimator $\hat{\mathbf{d}}$ is in excellent agreement with the theoretical expression and, as predicted by our theory, the modified estimator $\tilde{\mathbf{d}}$ significantly outperforms $\hat{\mathbf{d}}$.

for different choices of total sample size T (sub-sampled from the initial 10^5 samples), and use these optimal parameters for the estimators $\hat{\mathbf{d}}$ and $\tilde{\mathbf{d}}$. We use this optimal choice of bandwidth k for the estimators $\hat{\mathbf{d}}_f$ and $\hat{\mathbf{d}}_j$ as well (partitioning not applicable). For the suboptimal estimator, we arbitrarily choose the parameters as follows: fixed $k = 20$, $N = T/50$, $M = \lfloor T/2 \rfloor - N$.

The performance of these estimators as a function of sample size T is shown in Fig. 16. Estimators with optimal choice of parameters are indicated in solid line, and the suboptimal estimators are indicated in dashed lines.

From our experiments we see that the performance of the original estimator $\hat{\mathbf{d}}$ with suboptimal choice of parameters is marginally inferior when compared to the estimator with optimal choice of parameters. This does not hold for the other estimators as can be expected since the parameters are optimized w.r.t. the performance of $\hat{\mathbf{d}}$.

We note that the improved estimator $\tilde{\mathbf{d}}$ outperforms all other estimators while the performance of our original estimator $\hat{\mathbf{d}}$ is sandwiched between $\hat{\mathbf{d}}_f$ and $\hat{\mathbf{d}}_j$. We conjecture that the performance of $\hat{\mathbf{d}}_j$ is superior to $\hat{\mathbf{d}}$ for the same reason that $\tilde{\mathbf{d}}$ outperforms $\hat{\mathbf{d}}$: correlated error between different length statistics.

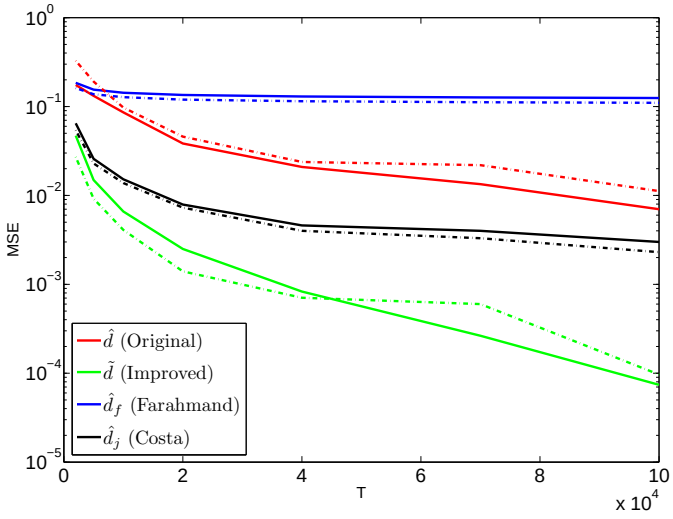


Figure 16: Comparison of performance of dimension estimates (Solid line: Optimal (optimal choice of k,N and M as per eq. (34) and eq. (35)); Dashed line: Suboptimal (fixed $k = 20$, $N = T/50$, $M = \lfloor T/2 \rfloor - N$)): The proposed improved kNN distance estimator outperforms all other estimators considered.

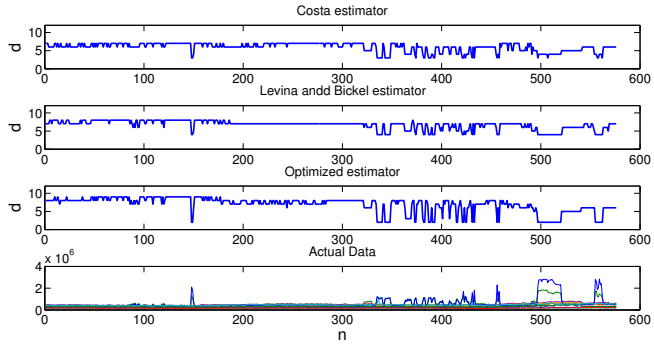


Figure 17: Comparison of performance of dimension estimates for anomaly detection in Abilene network data.

Anomaly detection in Abilene network data

Anomalies can be detected in router netowrks by estimating the local dimension at each time point and monitoring change in dimension. The data used is the number of packets sent by each of the 11 routers on the abiline network between January 1-2, 2005. A sample is taken every 5 minutes, leading to 576 samples with an extrinsic dimension pf 11.

The performance of different dimension estimators is shown in Fig. 17. We know that sim-

ulataneous peaks in router traffic should imply strong correlation between the routers and therefore lower intrinsic dimension. This behaviour is clearly reflected better by the optimized estimator as compared to the estimator of Costa et. al. [7] and Levina and Bickel [15].

10 Conclusion

Development of theoretical performance predictions for estimators of functionals of densities is important. We proposed plug-in estimators for smooth non-linear functionals of densities. We derived the bias, variance and mean square error of the estimator in terms of the sample size, the dimension of the samples and the underlying probability distribution. In addition, we developed a Gaussian central limit result for these estimators. In addition, we established rates of convergence of these plug-in estimators to the Gaussian distribution.

Our theory has two important by-products: (1) We established similarity between the moments of kNN density estimates and kernel density estimates. This in turn implies that plug-in estimators based on kNN density estimators and kernel density estimators have asymptotically equal rates of convergence. (2) We developed an algorithm for detection and correction of density estimates at boundary points for densities with finite support. This correction helps reduce the bias of density estimates at the boundaries of the support of the density, thereby reducing the overall bias of the plug-in estimators.

We verified the validity of our theorems through simulations. We applied the results developed in this report to specify optimal choice of bandwidth parameters and optimal partitioning of data samples, one part of which is used for density estimation while the remaining is used for functional estimation. We applied our theory to obtain statistical convergence results on estimators of entropy and mutual information and illustrated the applicability of our theory for determining confidence intervals of MI estimates.

We applied our theory to the problem of estimating Shannon entropy and Shannon mutual information. Furthermore, we used the Shannon entropy estimator to discover structure in high dimensional data and to determine the intrinsic dimension of data samples.

As a consequence of our work, we can tune the parameters of the plug-in estimator for optimal performance. Furthermore, we can specify the necessary sample size required to obtain requisite accuracy. This in turn can be used to predict and optimize performance in higher level applications like structure discovery and dimension estimation. This is not possible using current estimation methods in literature and underlines the significance of the results established in this work.

Appendices

A Uniform kernel density estimation

Throughout this section, we will derive results on moments of the uniform kernel density estimates for points in the set $\mathcal{S}' = \{X : \mathbf{S_u}(X) \subset \mathcal{S}\}$. This definition implies that the density f has continuous partial derivatives of order $2r$ in the uniform ball neighborhood for each $X \in \mathcal{S}'$ where r satisfies the condition $2r(1-t)/d > 1$. This excludes the set of points close to the boundary of the support, where the continuity assumption of the density is not satisfied. We will deal with these points in Appendix C.

Let $\mathbf{X}_1, .., \mathbf{X}_M$ denote M i.i.d realizations of the density f. We will assume that f is continuously differentiable evrywhere in the interior of the sWe seek to estimate the density at X from the M i.i.d realizations $\mathbf{X}_1, .., \mathbf{X}_M$. Let c_d denote the volume of a unit hyper-sphere in d dimensions. The uniform kernel density estimator is defined as follows:

A.1 Uniform kernel density estimator

The *uniform kernel* density estimator is defined below. The volume of the uniform kernel is given by

$$V_u(X) = \frac{k}{M}, \tag{36}$$

and the kernel region is given by

$$S_u(X) = \{Y : c_d||X - Y||^d \leq V_u\}. \tag{37}$$

$\mathbf{l_u}(X)$ denotes the number of points falling in $S_u(X)$

$$\mathbf{l_u}(X) = \Sigma_{i=1}^{M} 1_{X_i \in S_u(X)}, \tag{38}$$

and the *uniform kernel* density estimator is defined by

$$\hat{\mathbf{f}}_\mathbf{u}(X) = \frac{\mathbf{l_u}(X)}{MV_u(X)}. \tag{39}$$

The *coverage* of the *uniform kernel* is defined as

$$U(X) = \int_{S_u(X)} f(z)dz = \mathbb{E}[1_{\mathbf{Z} \in S_u(X)}]. \tag{40}$$

We observe that $\mathbf{l_u}(X)$ is a binomial random variable with parameters M and $U(X)$. Figure 18 illustrates the *uniform kernel* density estimate.

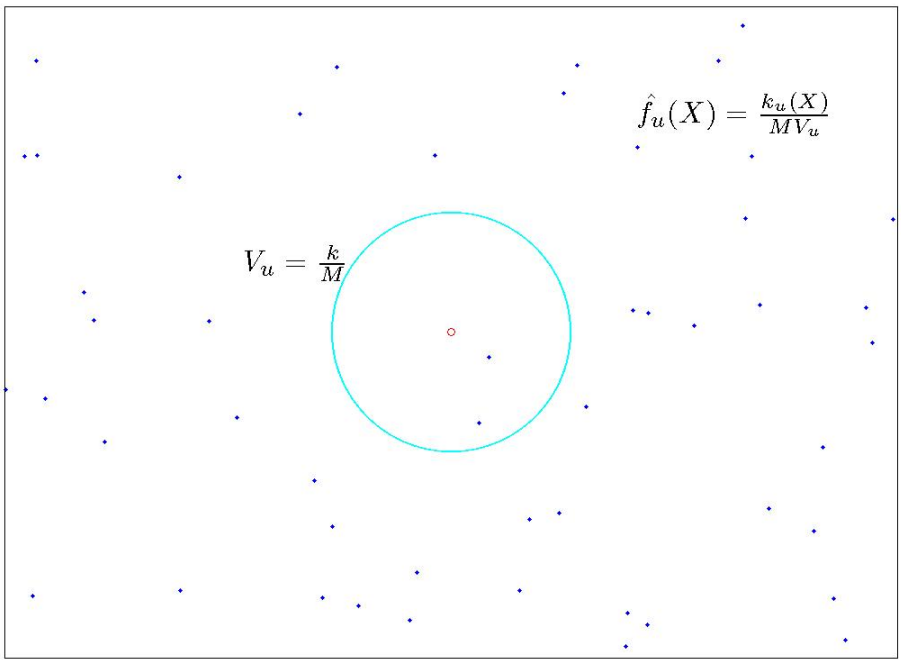


Figure 18: Uniform kernel density estimator.

A.2 Taylor series expansion of coverage

We assume that the density f has continuous partial derivatives of third order in a neighborhood of X. For small volumes $V_u(X)$ (which is equivalent to the condition that k/M is small), we can represent the coverage function $U(X)$ by using a third order Taylor series expansion of f about about X [19].

$$\begin{aligned} U(X) &= \int_{S_u(X)} f(Z)dZ \\ &= f(X)V_u(X) + c(X)V_u^{1+2/d}(X) + o(V_u^{1+2/d}(X)) \\ &= f(X)\frac{k}{M} + c(X)\left(\frac{k}{M}\right)^{1+2/d} + o\left(\left(\frac{k}{M}\right)^{1+2/d}\right), \end{aligned} \tag{41}$$

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

A.3 Concentration inequalities for uniform kernel density

Because $\mathbf{l_u}(X)$ is a binomial random variable, we can apply standard Chernoff inequalities to obtain concentration bounds on the density estimate. $\mathbf{l_u}(X)$ is a binomial random variable with parameters M and $U(X)$.

A.3.1 Concentration around true density

For $0 < p < 1/2$,

$$Pr(\mathbf{l_u}(X) > (1+p)MU(X)) \leq e^{-MU(X)p^2/4}, \tag{42}$$

and

$$Pr(\mathbf{l_u}(X) < (1-p)MU(X)) \leq e^{-MU(X)p^2/4}. \tag{43}$$

Using the Taylor expansion of coverage, we then have

$$Pr(\hat{\mathbf{f}}_\mathbf{u}(X) > (1+p)(f(X) + O((k/M)^{2/d}))) \leq\sim e^{-p^2kf(X)/4}, \tag{44}$$

and

$$Pr(\hat{\mathbf{f}}_\mathbf{u}(X) < (1-p)(f(X) + O((k/M)^{2/d}))) \leq\sim e^{-p^2kf(X)/4}. \tag{45}$$

This then implies that

$$Pr(\hat{\mathbf{f}}_\mathbf{u}(X) > (1+p)f(X)) \leq\sim e^{-p^2kf(X)/4}, \tag{46}$$

and

$$Pr(\hat{\mathbf{f}}_\mathbf{u}(X) < (1-p)f(X)) \leq\sim e^{-p^2kf(X)/4}. \tag{47}$$

Let $\mathbf{X}$ be a random variable with density f independent of the M i.i.d realizations $\mathbf{X}_1, .., \mathbf{X}_M$. Then,

$$\begin{aligned} Pr(\hat{\mathbf{f}}_\mathbf{u}(\mathbf{X}) > (1+p)f(\mathbf{X})) &= \mathbb{E}_\mathbf{X}[Pr(\hat{\mathbf{f}}_\mathbf{u}(\mathbf{X}) > (1+p)f(\mathbf{X}))] \\ &\leq \mathbb{E}[\sim (e^{-p^2kf(\mathbf{X})/4})] \\ &= \sim e^{-p^2k/4}, \end{aligned} \tag{48}$$

and

$$\begin{aligned} Pr(\hat{\mathbf{f}}_\mathbf{u}(\mathbf{X}) < (1-p)f(\mathbf{X})) &= \mathbb{E}_\mathbf{X}[Pr(\hat{\mathbf{f}}_\mathbf{u}(\mathbf{X}) < (1-p)f(\mathbf{X}))] \\ &\leq \mathbb{E}[\sim (e^{-p^2kf(\mathbf{X})/4})] \\ &= \sim e^{-p^2k/4}. \end{aligned} \tag{49}$$

A.3.2 Concentration away from 0

We can also bound the density estimate away from 0 as follows:

$$\begin{aligned} Pr(\hat{\mathbf{f}}_{\mathbf{u}}(\mathbf{X}) = 0) &= \mathbb{E}_{\mathbf{X}}[Pr(\hat{\mathbf{f}}_{\mathbf{u}}(\mathbf{X}) = 0] \\ &= \mathbb{E}[(1 - U(X))^M] \\ &= \mathbb{E}[(1 - (kf(X) + o(k)/M)^M] \\ &= \mathbb{E}[((1 - (kf(X) + o(k)/M)^{M/(kf(X)+o(k))})^{kf(X)+o(k)}] \\ &= \mathbb{E}[\sim (1/e)^{kf(X)+o(k)}] \\ &= \sim e^{-k}. \end{aligned} \tag{50}$$

A.4 Central Moments

Define the error function of the uniform kernel density,

$$\mathbf{e}_{\mathbf{u}}(X) = \hat{\mathbf{f}}_{\mathbf{u}}(X) - \mathbb{E}[\hat{\mathbf{f}}_{\mathbf{u}}(X)]. \tag{51}$$

The probability mass function of the binomial random variable $\mathbf{l}_{\mathbf{u}}(X)$ is given by

$$Pr(\mathbf{l}_{\mathbf{u}}(X) = l_x) = \binom{M}{l_x} (U(X))^{l_x} (1 - U(X))^{M - l_x}.$$

Since $\mathbf{l}_{\mathbf{u}}(X)$ is a binomial random variable, we can easily obtain moments of the uniform kernel density estimate. These are listed below.

First Moment:

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{f}}_{\mathbf{u}}(X)] - f(X) &= \frac{M}{k} U(X) - f(X) \\ &= c(X) \left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{2/d}\right). \end{aligned} \tag{52}$$

Second Moment:

$$\begin{aligned} \mathbb{V}[\hat{\mathbf{f}}_{\mathbf{u}}(X)] &= \mathbb{E}[\mathbf{e}_{\mathbf{u}}^2(X)] \\ &= \frac{M}{k^2} U(X)(1 - U(X)) \\ &= f(X)\frac{1}{k} + o\left(\frac{1}{k}\right). \end{aligned} \tag{53}$$

Higher Moments: For any integer $r \geq 3$,

$$\mathbb{E}[\mathbf{e}_u^r(X)] = O\left(\frac{1}{k^{r/2}}\right). \tag{54}$$

A.5 Covariance

Let X and Y be two distinct points. Clearly the density estimates at X and Y are not independent. We expect the density estimates to have positive covariance if X and Y are close and have negative covariance if X and Y are far. This is illustrated in Figure 19.

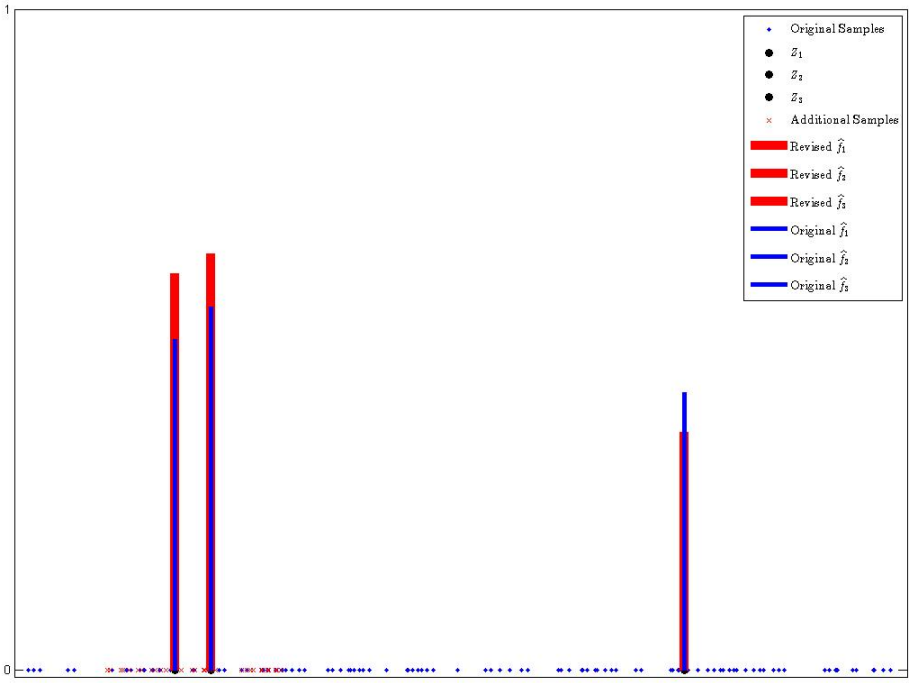


Figure 19: Covariance between uniform kernel density estimates.

Observe that the uniform kernels are disjoint for the set of points given by $\Psi_u := \{X, Y\} : ||X - Y|| \geq 2(k/c_d M)^{1/d}$, and have finite intersection on the complement of Ψ_u. Indeed we will show that when the uniform balls intersect (and therefore X and Y are close), the density estimates have positive covariance and that they have negative covariance when the uniform kernels are disjoint. Intersecting and disjoint balls are illustrated in Figure 20.

Define,

$$U(X, Y) := \mathbb{E}[1_{\mathbf{Z} \in S_u(X)} 1_{\mathbf{Z} \in S_u(Y)}]. \tag{55}$$

Intersecting balls

Lemma A.1. *For a fixed pair of points* $\{X, Y\} \in \Psi_u$,

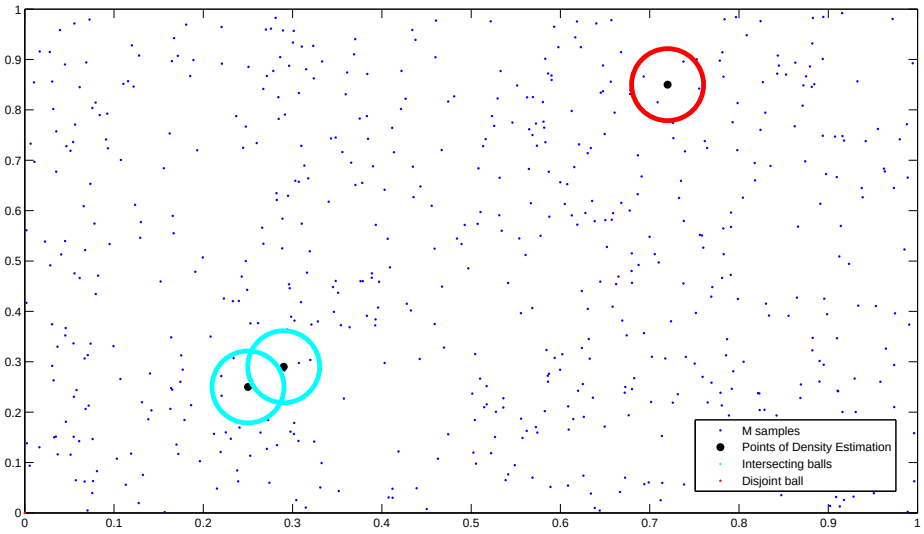


Figure 20: Intersecting and disjoint balls.

$$Cov[\mathbf{e_u}(X), \mathbf{e_u}(Y)] = \frac{-f(X)f(Y)}{M} + o\left(\frac{1}{M}\right).$$

Proof. For $\{X, Y\} \in \Psi_u$, we have that $1_{\mathbf{Z}\in S_u(X)} 1_{\mathbf{Z}\in S_u(Y)} = 0$ and therefore $U(X, Y) = 0$. We then have,

$$\begin{aligned}
Cov[\mathbf{e_u}(X), \mathbf{e_u}(Y)] &= \mathbb{E}[(\hat{\mathbf{f}}_\mathbf{u}(X) - \mathbb{E}[\hat{\mathbf{f}}_\mathbf{u}(X)])(\hat{\mathbf{f}}_\mathbf{u}(Y) - \mathbb{E}[\hat{\mathbf{f}}_\mathbf{u}(Y)])] \\
&= \frac{M}{k^2}\mathbb{E}[(1_{\mathbf{Z}\in S_u(X)} - U(X))(1_{\mathbf{Z}\in S_u(Y)} - U(Y))] \\
&= \frac{M}{k^2}\mathbb{E}[1_{\mathbf{Z}\in S_u(X)} 1_{\mathbf{Z}\in S_u(Y)} - U(X)U(Y)] \\
&= \frac{M}{k^2}(U(X, Y) - U(X)U(Y)) \\
&= -\frac{M}{k^2}[U(X)U(Y)] = \frac{-f(X)f(Y)}{M} + o\left(\frac{1}{M}\right).
\end{aligned}$$

□

Disjoint balls For $\{X, Y\} \in \Psi_u^c$, there is no closed form expression for the covariance. However we have the following lemmas:

Let $R_u(X)$ and $R_u(Y)$ denote the (constant and equal) radii of the uniform balls respectively. Define $\aleph(||X-Y||/R_u(X)) = V(S_u(X) \cap S_u(Y))/V_u(X)$ where $V(S_u(X) \cap S_u(Y))$ is the volume of the intersection of the two balls.

We observe that,

$$\begin{aligned}
\aleph(||X-Y||/R_u(X)) &= V(S_u(X) \cap S_u(Y))/V_u(X) \\
&= \frac{V[1_{\mathbf{Z} \in B(0,R_u(X))} 1_{\mathbf{Z} \in B(||Y-X||,R_u(Y))}]}{V_u(X)} \\
&= \frac{V[1_{\mathbf{Z} \in B(0,1)} 1_{\mathbf{Z} \in B(||Y-X||/R_u(X),1)}]}{V[1_{\mathbf{Z} \in B(0,1)}]} \\
&= O(1). \qquad (56)
\end{aligned}$$

Because f is assumed to be continuous, we have

$$U(X,Y) = \mathbb{E}[1_{\mathbf{Z} \in S_u(X)} 1_{\mathbf{Z} \in S_u(Y)}] = [f(X) + o(1)] V(S_u(X) \cap S_u(Y)). \qquad (57)$$

Lemma A.2. *For a fixed pair of points* $\{X, Y\} \in {\Psi_u}^c$,

$$Cov[\mathbf{e_u}(X), \mathbf{e_u}(Y)] = O(1/k).$$

Proof.

$$\begin{aligned}
\frac{M}{k^2} U(X,Y) &= \frac{M}{k^2} [f(X) + o(1)] V(S_u(X) \cap S_u(Y)) \\
&= \frac{f(X) + o(1)}{k} \frac{V(B_X \cap B_Y)}{V_u(X)} \\
&= \frac{f(X) + o(1)}{k} \aleph(||X-Y||/R_u(X)) \\
&= \frac{f(X)}{k} \aleph(||X-Y||/R_u(X)) + o(1/k) \\
&= O(1/k).
\end{aligned}$$

Therefore,

$$\begin{aligned}
Cov[\mathbf{e_u}(X), \mathbf{e_u}(Y)] &= \mathbb{E}[(\hat{\mathbf{f}}_\mathbf{u}(X) - \mathbb{E}[\hat{\mathbf{f}}_\mathbf{u}(X)])(\hat{\mathbf{f}}_\mathbf{u}(Y) - \mathbb{E}[\hat{\mathbf{f}}_\mathbf{u}(Y)])] \\
&= \frac{M}{k^2} (U(X,Y) - U(X)U(Y)) \\
&= \frac{M}{k^2} U(X,Y) - \frac{M}{k^2} U(X)U(Y) \\
&= O(1/k) - \Theta(1/M) \\
&= O(1/k).
\end{aligned}$$

□

Lemma A.3.

$$\int_y U(X,y)dy = [f(X) + o(1)]V_u(X)^2.$$

Proof. We note that for $U(X,y) \neq 0$, we need $\{X,y\} \in \Psi_u^c$. We therefore have, $f(y) = f(X) + o(1)$.

$$\begin{aligned}
\int_y U(X,y)dy &= \int [f(X) + o(1)]V(S_u(X) \cap S_u(Y))dy \\
&= V_u(X)[f(X) + o(1)] \int \aleph(||X - y||/R_u(X))dy \\
&= V_u(X)[f(X) + o(1)]R_u(X)^d \int \aleph(||y||/R_u(X))d(y/R_u(X)) \\
&= V_u(X)[f(X) + o(1)]\frac{V_u(X)}{c_d} \int \aleph(||y||/R_u(X))d(y/R_u(X)) \\
&= [f(X) + o(1)]\frac{V_u^2(X)}{c_d} \int \aleph(\delta)d(\delta).
\end{aligned}$$

The integral $\int \aleph(\delta)d(\delta)$ can be shown to be equal to c_d for all dimensions d.
We then have,

$$\begin{aligned}
\int_y U(X,y)dy &= [f(X) + o(1)]V_u^2(X) \\
&= [f(X) + o(1)]\left(\frac{k}{M}\right)^2.
\end{aligned}$$

□

Lemma A.4. *Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density f. Then,*

$$Cov[\gamma_1(\mathbf{X})\mathbf{e_u}(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e_u}(\mathbf{Y})] = \frac{Cov[\gamma_1(\mathbf{X})f(\mathbf{X}), \gamma_2(\mathbf{X})f(\mathbf{X})]}{M} + o(1/M).$$

Proof.

$$\begin{aligned} Cov[\gamma_1(\mathbf{X})\mathbf{e_u}(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e_u}(\mathbf{Y})] &= \mathbb{E}\Big[\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})(\hat{\mathbf{f}}_\mathbf{u}(X) - \mathbb{E}[\hat{\mathbf{f}}_\mathbf{u}(X)])(\hat{\mathbf{f}}_\mathbf{u}(Y) - \mathbb{E}[\hat{\mathbf{f}}_\mathbf{u}(Y)])\Big] \\ &= \frac{1}{MV_u(X)V_u(Y)}\mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})(U(\mathbf{X},\mathbf{Y}) - U(\mathbf{X})U(\mathbf{Y}))] \\ &= \frac{1}{MV_u^2(X)}\mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})U(\mathbf{X},\mathbf{Y})] \\ &- \frac{1}{MV_u^2(X)}\mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})U(\mathbf{X})U(\mathbf{Y})] \\ &= I - II. \end{aligned}$$

$$II = \frac{1}{M}\left(\mathbb{E}[\gamma_1(\mathbf{X})f(\mathbf{X})]\mathbb{E}[\gamma_2(\mathbf{Y})f(\mathbf{Y})]\right).$$

$$\begin{aligned} I &= \frac{1}{MV_u^2(X)}\mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})U(\mathbf{X},\mathbf{Y})] \\ &= \frac{1}{MV_u^2(X)}\int\int \gamma_1(x)\gamma_2(y)f(x)f(y)U(x,y)dxdy. \end{aligned}$$

Now for $U(x,y) \neq 0$, we need $\{x,y\} \in \Psi_u^c$. We therefore have, $\gamma_2(y)f(y) = \gamma_2(x)f(x) + o(1)$. We then have,

$$\begin{aligned} I &= \frac{1}{MV_u^2(X)}\int\int [\gamma_1(x)\gamma_2(x)f^2(x) + o(1)]U(x,y)dxdy \\ &= \frac{1}{MV_u^2(X)}\int [\gamma_1(x)\gamma_2(x)f^2(x) + o(1)]\left(\int U(x,y)dy\right)dx \\ &= \frac{1}{MV_u^2(X)}\int [\gamma_1(x)\gamma_2(x)f^2(x) + o(1)]\left((f(x) + o(1))V_u(x)^2\right)dx \\ &= \frac{1}{M}\int [\gamma_1(x)\gamma_2(x)f^2(x) + o(1)](f(x) + o(1))dx \\ &= \frac{1}{M}\left(\mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})] + o(1)\right) \\ &= \frac{1}{M}\mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})] + o(1/M). \end{aligned}$$

□

A.6 Higher cross moments

Disjoint balls We have the following results concerning higher cross moments for disjoint balls:

Lemma A.5. *Let q,r be positive integers satisfying $q + r > 2$. For a fixed pair of points $\{X, Y\} \in {\Psi_u}^c$,*

$$Cov(\mathbf{e}_{\mathbf{u}}^q(X), \mathbf{e}_u^r(Y)) \;=\; o(1/M).$$

Proof. For a fixed pair of points $\{X, Y\} \in {\Psi_u}^c$, the joint probability mass function of the functions $\mathbf{l_u}(X)$,$\mathbf{l_u}(Y)$ is given by

$$Pr(\mathbf{l_u}(X) = l_x, \mathbf{l_u}(Y) = l_y) = 1_{l_x+l_y \le M} \binom{M}{l_x, l_y} (U(X))^{l_x} (U(Y))^{l_y} (1 - U(X) - U(Y))^{M-l_x-l_y}.$$

We also have from chernoff inequalities for binomial random variables that

$$Pr((1-p)k < \mathbf{l_u}(X) < (1+p)k) = 1 - e^{-p^2 k},$$
$$Pr((1-p)k < \mathbf{l_u}(Y) < (1+p)k) = 1 - e^{-p^2 k}.$$

Denote the high probability event χ by $(1-p)k < \mathbf{l_u}(X), \mathbf{l_u}(Y) < (1+p)k$. Define $\hat{\mathbf{l}}_{\mathbf{u}}(X)$, $\hat{\mathbf{l}}_{\mathbf{u}}(Y)$ to be binomial random variables with parameters $\{U(X),M-q\}$ and $\{U(Y),M-r\}$ respectively. The covariance between powers of density estimates is then given by

$$\begin{aligned}
& Cov(\hat{\mathbf{f}}_{\mathbf{u}}^q(X), \hat{\mathbf{f}}_{\mathbf{u}}^r(Y)) = \frac{1}{k^{q+r}} Cov(\mathbf{l}_{\mathbf{u}}^q(X), \mathbf{l}_{\mathbf{u}}^r(Y)) \\
= \;& \frac{1}{k^{q+r}} \sum l_x^q l_y^r Pr(\mathbf{l_u}(X) = l_x, \mathbf{l_u}(Y) = l_y) - \frac{1}{k^{q+r}} \sum l_x^q l_y^r Pr(\mathbf{l_u}(X) = l_x) Pr(\mathbf{l_u}(Y) = l_y) \\
= \;& \sum_{\chi} \frac{l_x^q l_y^r}{k^{q+r}} \left[Pr(\mathbf{l_u}(X) = l_x, \mathbf{l_u}(Y) = l_y) - Pr(\mathbf{l_u}(X) = l_x) Pr(\mathbf{l_u}(Y) = l_y) \right] + O(e^{-p^2 k}) \\
= \;& \sum_{\chi} \frac{f^q(X) f^r(Y) l_x^q l_y^r U^q(X) U^r(Y)}{k^{q+r} (l_x \times \ldots \times l_x - q + 1)(l_y \times \ldots \times l_y - r + 1)} \times \\
& [(M \times \ldots \times M - (q+r-1)) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(X) = l_x, \hat{\mathbf{l}}_{\mathbf{u}}(Y) = l_y) \\
& -(M \times \ldots \times M - q + 1)(M \times \ldots \times M - r + 1) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(X) = l_x) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(Y) = l_y)] \\
+ \;& o(1/M) \\
= \;& \left(\frac{f^q(X) f^r(Y)}{M^{q+r}} + O\left(\frac{1}{k M^{q+r}} \right) \right) \times \\
& \sum_{\chi} [(M \times \ldots \times M - (q+r-1)) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(X) = l_x, \hat{\mathbf{l}}_{\mathbf{u}}(Y) = l_y) \\
& -(M \times \ldots \times M - (q-1))(M \times \ldots \times M - (r-1)) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(X) = l_x) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(Y) = l_y)] \\
+ \;& o(1/M)
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{f^q(X)f^r(Y)}{M^{q+r}} + O\left(\frac{1}{kM^{q+r}}\right)\right) \times \\
&\quad [(M \times \ldots \times M - (q+r-1)) - (M \times \ldots \times M - (q-1))(M \times \ldots \times M - (r-1))] \\
&+ o(1/M) \\
&= \frac{-qrf^q(X)f^r(Y)}{M} + o\left(\frac{1}{M}\right).
\end{aligned}
$$

Then, the covariance between the powers of the error function is given by

$$
\begin{aligned}
Cov(\mathbf{e}_{\mathbf{u}}^q(X), \mathbf{e}_u^r(Y)) &= Cov((\hat{\mathbf{f}}_{\mathbf{u}}(X) - \mathbb{E}[\hat{\mathbf{f}}_{\mathbf{u}}(X)])^q, (\hat{\mathbf{f}}_{\mathbf{u}}(Y) - \mathbb{E}[\hat{\mathbf{f}}_{\mathbf{u}}(Y)])^r) \\
&= \sum_{a=1}^{q}\sum_{b=1}^{r}\binom{q}{a}\binom{r}{b}(-\mathbb{E}[\hat{\mathbf{f}}_{\mathbf{u}}(X)])^a(-\mathbb{E}[\hat{\mathbf{f}}_{\mathbf{u}}(Y)])^b Cov(\hat{\mathbf{f}}_{\mathbf{u}}^a(X), \hat{\mathbf{f}}_{\mathbf{u}}^b(Y)) \\
&= \sum_{a=1}^{q}\sum_{b=1}^{r}\binom{q}{a}\binom{r}{b}[(-f(X))^a(-f(Y))^b + o(1)]Cov(\hat{\mathbf{f}}_{\mathbf{u}}^a(X), \hat{\mathbf{f}}_{\mathbf{u}}^b(Y)) \\
&= -f^q(X)f^r(Y)\sum_{a=1}^{q}\sum_{b=1}^{r}\binom{q}{a}\binom{r}{b}\frac{(-1)^a a(-1)^b b}{M} + o\left(\frac{1}{M}\right) \\
&= 1_{\{q=1,r=1\}}\left(\frac{-f(X)f(Y)}{M}\right) + o(1/M) \\
&= o(1/M).
\end{aligned}
$$

where the last step follows from the condition that $q + r > 2$.

□

Intersecting balls For $\{X, Y\} \in {\Psi_u}^c$, we have the following bounds

Lemma A.6. *Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density f. Also let the indicator function $1_{\Delta_u}(X, Y)$ denote the event $\Delta_u : \{X, Y\} \in {\Psi_u}^c$. For q,r positive integers satisfying $q + r > 1$,*

$$
\mathbb{E}[\mathbf{1}_{\boldsymbol{\Delta}_{\mathbf{u}}}(\mathbf{X}, \mathbf{Y})\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_{\mathbf{u}}^q(\mathbf{X})\mathbf{e}_{\mathbf{u}}^r(\mathbf{Y})] = o\left(\frac{1}{M}\right), \tag{58}
$$

Proof. For $1_{\Delta_u}(X,Y) \neq 0$, we have $\{X,Y\} \in \Psi_u^c$. Then,

$$\begin{aligned}
& \mathbb{E}[\mathbf{1_{\Delta_u}}(\mathbf{X},\mathbf{Y})\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e_u^q}(\mathbf{X})\mathbf{e_u^r}(\mathbf{Y})] \\
= & \mathbb{E}[\mathbf{1_{\Delta_u}}(\mathbf{X},\mathbf{Y})\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\mathbf{X},\mathbf{Y}}[\mathbf{e_u^q}(X)\mathbf{e_u^r}(Y)]] \\
\leq & \mathbb{E}\left[\mathbf{1_{\Delta_u}}(\mathbf{X},\mathbf{Y})\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\sqrt{\mathbb{E}_{\mathbf{X}}[\mathbf{e_u^{2q}}(X)]\mathbb{E}_{\mathbf{Y}}[\mathbf{e_u^{2r}}(Y)]}\right] \\
= & \mathbb{E}\left[\mathbf{1_{\Delta_u}}(\mathbf{X},\mathbf{Y})\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})O\left(\frac{1}{k^{q+r/2}}\right)\right] \\
= & \int\left[O\left(\frac{1}{k^{q+r/2}}\right)(\gamma_1(x)\gamma_2(x)+o(1))\right]\left(\int\Delta_u(x,y)dy\right)dx \\
= & \int\left[O\left(\frac{1}{k^{q+r/2}}\right)(\gamma_1(x)\gamma_2(x)+o(1))\right]\left(2^d\frac{k}{M}\right)dx \\
= & o\left(\frac{1}{M}\right).
\end{aligned}$$

where the bound is obtained using the Cauchy-Schwarz inequality and using Eq.54. □

We can succinctly state the results derived in the last two lemmas in the form of the following lemma:

Lemma A.7. *Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. Let $\mathbf{X}_1,..,\mathbf{X}_M,\mathbf{X},\mathbf{Y}$ denote $M+2$ i.i.d realizations of the density f. If q,r are positive integers satisfying $q+r>2$*

$$Cov[\gamma_1(\mathbf{X})\mathbf{e_u^q}(\mathbf{X}),\gamma_2(\mathbf{Y})\mathbf{e_u^r}(\mathbf{Y})] = o(1/M).$$

Proof. The result for the case $q=1$, $r=1$ was established earlier in Lemma A.4.

$$Cov[\gamma_1(\mathbf{X})\mathbf{e_u^q}(\mathbf{X}),\gamma_2(\mathbf{Y})\mathbf{e_u^r}(\mathbf{Y})] = I + D,$$

where 'I' stands for the contribution form the intersecting balls and 'D' for the contribution from the dis-joint balls. I and D are given by

$$\begin{aligned}
I &= \mathbb{E}[\mathbf{1_{\Delta_u}}(\mathbf{X},\mathbf{Y})Cov\left[\gamma_1(X)\mathbf{e_u^q}(X),\gamma_2(Y)\mathbf{e_u^r}(Y)\right]], \\
D &= \mathbb{E}[(\mathbf{1}-\mathbf{1_{\Delta_u}}(\mathbf{X},\mathbf{Y}))Cov\left[\gamma_1(X)\mathbf{e_u^q}(X),\gamma_2(Y)\mathbf{e_u^r}(Y)\right]].
\end{aligned}$$

We have already established in the previous lemma that

$$I = o\left(\frac{1}{M}\right).$$

Now,

$$\begin{aligned}
D &= \mathbb{E}[(1-\mathbf{1_{\Delta_u}}(\mathbf{X},\mathbf{Y}))\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\mathbf{X},\mathbf{Y}}[Cov(\mathbf{e_u^q}(X),\mathbf{e_u^r}(Y))]] \\
&= \mathbb{E}[(1-\mathbf{1_{\Delta_u}}(\mathbf{X},\mathbf{Y}))\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})o(1/M)] \\
&= o\left(\frac{1}{M}\right).
\end{aligned} \tag{59}$$

This concludes the proof. □

B k-NN density estimation

Throughout this section, we will derive results on moments of k-NN density estimates for points in the set $\mathcal{S}' = \{X : \mathbf{S_k}(X) \subset \mathcal{S}\}$. This definition implies that the density f has continuous partial derivatives of order $2r$ in the k-NN ball neighborhood for each $X \in \mathcal{S}'$ where r satisfies the condition $2r(1-t)/d > 1$. This excludes the set of points close to the boundary of the support, where the continuity assumption of the density is not satisfied. We will deal with these points in Appendix C.

B.1 Concentration inequality for coverage probability

It has been previously established that $\mathbf{P}(X)$ has a beta distribution with parameters k, $M-k+1$. [19]. 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,

$$Be(P|k, M-k+1) = 1 - Bi(k-1|M,P). \tag{60}$$

The following Chernoff bounds for binomial random variables have also been established previously. When $k < MP$, $Bi(k|M,P) \leq exp[-(MP-k)^2/2PM]$, and when $k > MP$, $1 - Bi(k|M,P) \leq exp[-(MP-k)^2/2PM]$. We therefore have that for some $0 < p < 1/2$,

$$Pr((1-p)(k-1)/M < \mathbf{P}(X) < (p+1)(k-1)/M) = O(e^{-p^2k/2}). \tag{61}$$

B.2 Taylor series expansion of coverage probability

Let $X \in \mathcal{S}'$. We can then represent the coverage function $\mathbf{P}(X)$ in terms of the volume of the k-NN ball $\mathbf{V_k}(X)$ by expanding the density f in a Taylor series about X [19].

$$\begin{aligned}\mathbf{P}(X) &= \int_{\mathbf{S_k}(X)} f(z)dz \\ &= f(X)\mathbf{V_k}(X) + c(X){\mathbf{V_k}}^{1+2/d}(X) + \sum_{i=2}^{r-1} c_i(X){\mathbf{V_k}}^{1+2i/d}(X) + c_r(\tilde{X}){\mathbf{V_k}}^{1+2r/d}(X)\end{aligned} \tag{62}$$

where $c(X) = \Gamma^{(2/d)}(\frac{n+2}{2})tr[\nabla^2(f(X))]$ and $c_r(\tilde{X})$ is the coefficient of the reminder term. Also define $h(X) = c(X)f^{-2/d}(X)$. We note that r satisfies the condition $2r(1-\alpha)/d > 1$. We can now rearrange terms to get the following representation of $1/\mathbf{V_k}(X)$ [19].

$$\frac{1}{\mathbf{V_k}(X)} = \frac{f(X)}{\mathbf{P}(X)} + \frac{h(X)}{\mathbf{P}^{1-2/d}(X)} + \sum_{t\in T} h_t(X)\mathbf{P}^{1-t}(X) + \mathbf{h_r}(X) \tag{63}$$

where T is some countable set with $\inf\{T\} = 4/d$ and $\mathbf{h_r}(X) = o(1/\mathbf{P}^{1-2r/d}(X))$. Let $\natural(X)$ denote the event $(1-p_k)(k-1)/M < \mathbf{P}(X) < (p_k+1)(k-1)/M$ where $p_k = 1/(k^{\delta_k/2})$ with $\delta_k = 1 - (\log(\log k)/\log k)$. From the concentration inequality, $1 - Pr(\natural(X)) = O(e^{-p_k^2 k/2})$. From the choice of p_k, $O(e^{-p_k^2 k/2}) = o(1/k^a)$ for arbitrarily large values of a. From the polynomial growth condition on k which specifies $k = M^\alpha$, we have $o(1/k^a) = o(1/M^{\alpha a})$. We can then summarize that $1 - Pr(\natural(X)) = \mathbb{E}[1_{\natural^c(X)}] = o(1/M^a)$ for arbitrarily large values of a.

B.3 Bias of the k-NN density estimates

Let $X \in \mathcal{S}'$. Using the above Taylor series expansion, it has been previously shown [19] that the bias of the k-NN density estimate is given by

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

This gives

$$\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})(\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}) - f(\mathbf{X}))] = \mathbb{E}[\gamma(\mathbf{X})h(\mathbf{X})]\left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{2/d}\right).$$

B.4 Approximation to the k-NN density estimator

Define the *coverage* density estimate to be,

$$\hat{\mathbf{f}}_{\mathbf{c}}(X) = f(X)\frac{k-1}{M}\frac{1}{\mathbf{P}(X)}. \tag{64}$$

We see that the estimate $\hat{\mathbf{f}}_{\mathbf{c}}(X)$ is not tractable. We also note that the two estimates - $\hat{\mathbf{f}}_{\mathbf{c}}(X)$ and $\hat{\mathbf{f}}_{\mathbf{k}}(X)$ - are identical in the case of the uniform density. Define the error functions $\mathbf{e_c}(X) = \hat{\mathbf{f}}_{\mathbf{c}}(X) - \mathbb{E}[\hat{\mathbf{f}}_{\mathbf{c}}(X)]$ and $\mathbf{e_k}(X) = \hat{\mathbf{f}}_{\mathbf{k}}(X) - \mathbb{E}[\hat{\mathbf{f}}_{\mathbf{k}}(X)]$. Note that the coverage density estimate corresponds to the leading term in the Taylor series expansion of the volume. We can therefore write

$$\hat{\mathbf{f}}_{\mathbf{k}}(X) = \hat{\mathbf{f}}_{\mathbf{c}}(X) + \sum_t \left(\frac{k-1}{M}\right)^t h_t(X) \times \left(\frac{k-1}{M}\right)^{1-t} (1/\mathbf{P}^{1-t}(X)) + \frac{k-1}{M}\mathbf{h_r}(X) \tag{65}$$

We can then establish the following lemmas:

Lemma B.1. *Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}$ denote $M+1$ i.i.d realizations of the density f. Let q be any positive integer. Let $\gamma(X)$ be any arbitrary continuous function satisfying that $\mathbb{E}[\gamma(\mathbf{X})]$ is finite. Then,*

$$\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] \;=\; \mathbb{E}[\gamma(\mathbf{X})\mathbf{e}_c^q(\mathbf{X})](1+o(1)) + o(1/M)$$

Lemma B.2. *Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M+2$ i.i.d realizations of the density f.*

$$\begin{aligned} &Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e_k}^q(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e_k}^r(\mathbf{Y})\big] \\ &= Cov[\gamma_1(\mathbf{X})\mathbf{e}_c^q(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}_c^r(\mathbf{Y})] + o(1/M). \end{aligned}$$

As a consequence of these lemma, for $X \in \mathcal{S}'$, we can compute all central and cross moments of the k-NN density $\hat{\mathbf{f}}_\mathbf{k}(X)$ up to $o(1/M)$ by equivalently computing the corresponding moments for the coverage density estimate. We will first prove the above lemmas and subsequently work on obtaining the exact rates for the coverage density estimate.

Define the operator $\mathcal{M}(\mathbf{Z}) = \mathbf{Z} - \mathbb{E}[\mathbf{Z}]$ and the terms $\mathbf{e_t}(X) = \mathcal{M}(\sum_t((k-1)/M)h_t(X)(1/\mathbf{P}^{1-t}(X)))$ and $\mathbf{e_r}(X) = \mathcal{M}(((k-1)/M)\mathbf{h_r}(X))$. Note that for $X \in \mathcal{S}'$, $\mathbf{e_k}(X) = \mathbf{e_c}(X) + \mathbf{e_t}(X) + \mathbf{e_r}(X)$. Also define $\mathbf{e_t}(X) = \mathcal{M}(((k-1)/M)^{1-t}(1/\mathbf{P}^{1-t}(X)))$. We will next establish moment properties of the coverage function.

B.5 Moments of coverage function

Since $\mathbf{P}(X)$ is a beta random variable, the probability density function of $\mathbf{P}(X)$ is given by

$$f(p_X) = \frac{M!}{(k-1)!(M-k)!} p_X^{k-1}(1-p_X)^{M-k}.$$

Under the event $\natural(X)$, we can clearly see that $\mathbb{E}[1_{\natural(X)}\mathbf{P}^{-t}(X)] = \Theta((k/M)^{-t})$. For large enough k, M, we also see that $\mathbb{E}[\mathbf{P}^{-2t}(X)]$ is bounded between 0 and 1, which implies that $\mathbb{E}[1_{\natural^c(X)}\mathbf{P}^{-t}(X)] = o(1/M^{a/2})$ using Cauchy-Schwartz and the concentration inequality. This then gives $\mathbb{E}[\mathbf{P}^{-t}(X)] = \Theta((k/M)^{-t})$. We then get $\mathbb{E}[1_{\natural(X)}\mathbf{e}_t^q(X)] = O\left(k^{-(\delta_k q/2)}\right)$. We can again bound $\mathbb{E}[1_{\natural^c(X)}\mathbf{e}_t^q(X)]$ by $o(1/M^{a/2})$ using Cauchy-Schwartz inequality and the concentration bound. This gives $\mathbb{E}[\mathbf{e}_t^q(X)] = O\left(k^{-(\delta_k q/2)}\right)$. Noting that $\delta_k \to 1$ as $k \to \infty$ gives

$$\mathbb{E}[\mathbf{e}_t^q(X)] = O(k^{-q/2}). \tag{66}$$

Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M+2$ i.i.d realizations of the density f. Before we address this, we seek to answer the following question: For which set of pair of points $\{X, Y\}$ are the k-NN balls disjoint?

Intersecting and disjoint balls Define $\Psi_\epsilon := \{X, Y\} : ||X - Y|| \geq R_\epsilon(X) + R_\epsilon(Y)$ where $R_\epsilon(X)$ and $R_\epsilon(Y)$ are the ball radii corresponding to coverages $Q_\epsilon(X) = Q_\epsilon(Y) = (1+p_k)((k-1)/M)$. We will now show that for $\{X, Y\} \in \Psi_\epsilon$, the k-NN balls will be disjoint with exponentially high probability. Let $\mathbf{d}_\mathbf{X}^{(\mathbf{k})}$ and $\mathbf{d}_\mathbf{Y}^{(\mathbf{k})}$ denote the k-NN distances from X and Y and let $\boldsymbol{\Upsilon}$ denote the event that the k-NN balls intersect. For $\{X, Y\} \in \Psi_\epsilon$, we then

have

$$\begin{aligned} Pr(\boldsymbol{\Upsilon}) &= Pr(\mathbf{d_X^{(k)}} + \mathbf{d_Y^{(k)}} \geq ||X - Y||) \\ &\leq Pr(\mathbf{d_X^{(k)}} + \mathbf{d_Y^{(k)}} \geq R_\epsilon(X) + R_\epsilon(Y)). \\ &\leq Pr(\mathbf{d_X^{(k)}} \geq R_\epsilon(X)) + Pr(\mathbf{d_Y^{(k)}} \geq R_\epsilon(Y)) \\ &= Pr(\mathbf{P}(X) \geq (p_k + 1)((k-1)/M)) + Pr(\mathbf{P}(Y) \geq (p_k + 1)((k-1)/M)) \\ &= O(e^{-p_k^2 k/3}), \end{aligned} \tag{67}$$

where the last inequality follows from the concentration inequality. We conclude that for $\{X, Y\} \in \Psi_\epsilon$, the probability of intersection of k-NN balls centered at X and Y decays exponentially in $p_k^2 k$. Stated in a different way, we have shown that for a given pair of points $\{X, Y\}$, if the ϵ balls around these points are disjoint, then the k-NN balls will be disjoint with exponentially high probability. Let $\Delta_\epsilon(X, Y)$ denote the event $\{X, Y\} \in \Psi_\epsilon^c$.

Let $\{X, Y\} \in \Psi_\epsilon$ and let q, r be non-negative integers satisfying $q + r > 1$. The event that the k-NN balls intersect is given by $\boldsymbol{\Upsilon} := \{\mathbf{d_X^{(k)}} + \mathbf{d_Y^{(k)}} > ||X - Y||\}$. The joint probability distribution of $\mathbf{P}(X)$ and $\mathbf{P}(Y)$ when the k-NN balls do not intersect $=: \boldsymbol{\Upsilon^c}$ is given by

$$f_{\boldsymbol{\Upsilon^c}}(p_X, p_Y) = \frac{M!}{(k-1)!^2(M-2k)!}(p_X p_Y)^{k-1}(1 - p_X - p_Y)^{M-2k}.$$

Figure 21 shows the distribution of the M samples when the k-NN balls are disjoint.

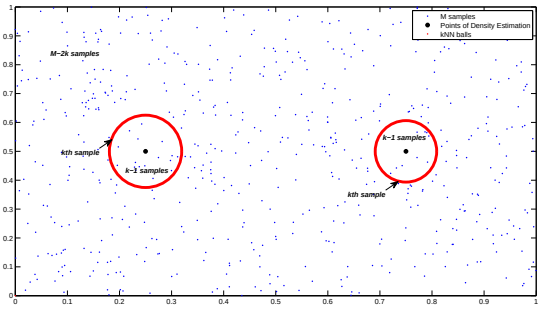


Figure 21: Distribution of samples when k-NN balls are disjoint.

Define

$$i(p_X, p_Y) = \frac{\Gamma(t)\Gamma(\beta)\Gamma(\gamma)}{\Gamma(t + \beta + \gamma)} p_X^{t-1} p_Y^{\beta-1}(1 - p_X - p_Y)^{\gamma-1},$$

and note that

$$\int_{p_X=0}^{1} \int_{p_Y=0}^{1} 1_{\{p_X + p_Y \leq 1\}} i(p_X, p_Y) dp_X dp_Y = 1.$$

Now note that $i(p_X, p_Y)$ corresponds to the density function $f_{\boldsymbol{\Upsilon^c}}(p_X, p_Y)$ for the choices $t = k$, $\beta = k$ and $\gamma = M - 2k + 1$. Furthermore, for $\{X, Y\} \in \Psi_\epsilon$, the set $\mathcal{C} := \{p_X, p_Y\} : (1 - p_k)(k -$

$1)/M \leq p_X, p_Y \leq (1+p_k)(k-1)/M$ is a subset of the region $\mathcal{T} := \{p_X, p_Y\} : 0 \leq p_X, p_Y \leq 1;$ $p_X + p_Y \leq 1$. Note that $\mathbb{E}[1_{\mathcal{C}}] = 1 - o(1/M^a)$. This implies that expectations over the region $\mathcal{R} := \{p_X, p_Y\} : 0 \leq p_X, p_Y \leq 1;$ should be of the same order as the expectations over $\mathcal{T}$ with differences of order $o(1/M^a)$. In particular,

$$\mathbb{E}[1/\mathbf{P}^t(X)\mathbf{P}^\beta(Y)] = \mathbb{E}[1_{\mathcal{T}}/\mathbf{P}^t(X)\mathbf{P}^\beta(Y)] + o(1/M^a).$$

From the joint distribution representation, we also get

$$\frac{\mathbb{E}[1_{\mathcal{T}}/\mathbf{P}^t(X)\mathbf{P}^\beta(Y)]}{\mathbb{E}[1/\mathbf{P}^t(X)]\mathbb{E}[1/\mathbf{P}^\beta(Y)]} = -\frac{t\beta}{M} + o(1/M).$$

Now observe that

$$\begin{aligned}
&\left(\frac{k-1}{M}\right)^{t+\beta} Cov(1/\mathbf{P}^t(X), 1/\mathbf{P}^\beta(Y)) \\
&= \left(\frac{k-1}{M}\right)^{t+\beta} [\mathbb{E}[1/\mathbf{P}^t(X)\mathbf{P}^\beta(Y)] - \mathbb{E}[1/\mathbf{P}^t(X)]\mathbb{E}[1/\mathbf{P}^\beta(Y)]] \\
&= \left(\frac{k-1}{M}\right)^{t+\beta} \mathbb{E}[1/\mathbf{P}^t(X)]\mathbb{E}[1/\mathbf{P}^\beta(Y)] \left[\frac{\mathbb{E}[1/\mathbf{P}^t(X)\mathbf{P}^\beta(Y)]}{\mathbb{E}[1/\mathbf{P}^t(X)]\mathbb{E}[1/\mathbf{P}^\beta(Y)]} - 1\right] \\
&= (1 + o(1/k)) \left[1 - \frac{t\beta}{M} + o(1/M) - 1\right] \\
&= -\left(\frac{t\beta}{M}\right) + o(1/M). \qquad (68)
\end{aligned}$$

Then, the covariance between the powers of the error function $\mathbf{e}_t$ is given by

$$\begin{aligned}
Cov(\mathbf{e}_t^q(X), \mathbf{e}_\beta^r(Y)) &= \left(\frac{k-1}{M}\right)^{tq+\beta r} Cov\left(\left[\frac{1}{\mathbf{P}^t(X)} - \mathbb{E}\left[\frac{1}{\mathbf{P}^t(X)}\right]\right]^q, \left[\frac{1}{\mathbf{P}^\beta(Y)} - \mathbb{E}\left[\frac{1}{\mathbf{P}^\beta(Y)}\right]\right]^r\right) \\
&= \sum_{a=1}^{q}\sum_{b=1}^{r} \binom{q}{a}\binom{r}{b}[(-1)^{a+b} + o(1)] \left(\frac{k-1}{M}\right)^{ta+\beta b} Cov(1/\mathbf{P}^{ta}(X), 1/\mathbf{P}^{\beta b}(Y)) \\
&= -t\beta \sum_{a=1}^{q}\sum_{b=1}^{r} \binom{q}{a}\binom{r}{b}\frac{(-1)^a a(-1)^b b}{M} + o\left(\frac{1}{M}\right) \\
&= 1_{\{q=1,r=1\}}\left(\frac{-t\beta}{M}\right) + o(1/M) \\
&= 1_{\{q=1,r=1\}}\Theta(1/M) + o(1/M). \qquad (69)
\end{aligned}$$

B.6 Analysis of central terms

$$\begin{aligned}
\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] &= \mathbb{E}\left[\gamma(\mathbf{X})\mathbb{E}_{\mathbf{X}}[1_{\{X\in\mathcal{S}'\}}\mathbf{e}_k^q(X)]\right] \\
&= \mathbb{E}\left[\gamma(\mathbf{X})\mathbb{E}_{\mathbf{X}}[1_{\{X\in\mathcal{S}'\}}(\mathbf{e_c}(X) + \mathbf{e_t}(X) + \mathbf{e_r}(X))^q]\right] \\
&= \mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbb{E}_{\mathbf{X}}[(\mathbf{e_c}(X) + \mathbf{e_t}(X) + \mathbf{e_r}(X))^q]\right]
\end{aligned}$$

Let us focus on the inner expectation first. From the analysis in the previous section on $\mathbf{e_t}(X)$, it is easy to see that $\mathbb{E}[\mathbf{e_r}^l(X)] = O((k/M)^{2rl/d}) = o(1/M)$ for any $l > 1$. Similarly, we can see that $\mathbb{E}[\mathbf{e_c}^l(X)] = O(k^{-l/2})$. Now, we can write $\mathbf{e_t}^l(X)$ as a sum of terms of the form $\prod_t((k/M)^t h_t(X)\mathbf{e_t}^{l_t}(X))$ where $\sum_t l_t = l$. The coefficients in the product form $(k/M)^t = o(1)$ while each $\mathbf{e_t}^{l_t}(X)$ term contributes $O(k^{-l_t/2})$ by (66). By repeatedly using Cauchy-Schwartz, we can show that the expectation of each of these terms and therefore $\mathbb{E}[\mathbf{e_t}^l(X)]$ is $o(k^{-l/2})$.

We note that $\mathbf{e_k}^q(X)$ will contain terms of the form $(\mathbf{e_c}(X) + \mathbf{e_t}(X))^l(\mathbf{e_r}(X))^{q-l}$. If $l \neq q$, the expectation of this term can be bounded as follows:

$$\begin{aligned} |\mathbb{E}[(\mathbf{e_t}(X))^l(\mathbf{e_r}(X))^{q-l}]| &\leq \sqrt{\mathbb{E}[(\mathbf{e_c}(X) + \mathbf{e_t}(X))^{2l}]\mathbb{E}[(\mathbf{e_r}(X))^{2(q-l)}]} \\ &= O(1) \times o(1/M) = o(1/M). \end{aligned}$$

Let us concentrate on the case $l = q$. In this case, $\mathbf{e_k}^q(X)$ will contain terms of the form $(\mathbf{e_c}(X))^m(\mathbf{e_t}(X))^{q-m}$. For $q \neq m$, we then have

$$\begin{aligned} |\mathbb{E}[(\mathbf{e_c}(X))^m(\mathbf{e_t}(X))^{q-m}]| &\leq \sqrt{\mathbb{E}[(\mathbf{e_c}(X))^{2l}]\mathbb{E}[(\mathbf{e_t}(X))^{2(q-l)}]} \\ &= O(k^{-m/2}) \times o(k^{-(q-m)/2}) = o(k^{-q/2}). \end{aligned}$$

Noting that $\mathbb{E}[\mathbf{e_c}^q(X)] = O(k^{-q/2})$ gives us

$$\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] = \mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_c^q(\mathbf{X})\big](1 + o(1)) + o(1/M)$$

Finally, we have that $\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}^c}] = O((k/M)^{1/d})$. Using this fact and Cauchy-Schwartz therefore gives us Lemma B.1.

B.7 Analysis of cross terms

Similarly,

$$\begin{aligned} &Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e_k}^q(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e_k}^r(\mathbf{Y})\big] \\ &= Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})(\mathbf{e_c}(\mathbf{X}) + \mathbf{e_t}(\mathbf{X}) + \mathbf{e_r}(\mathbf{X}))^q, 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})(\mathbf{e_c}(\mathbf{Y}) + \mathbf{e_t}(\mathbf{Y}) + \mathbf{e_r}(\mathbf{Y}))^r\big]. \end{aligned}$$

Using the same arguments as in the previous section, we can show that the contribution of terms with $\mathbf{e_r}(\mathbf{X})$ or $\mathbf{e_r}(\mathbf{Y})$ is $o(1/M)$. We can then reduce,

$$\begin{aligned} &Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e_k}^q(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e_k}^r(\mathbf{Y})\big] \\ &= Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})(\mathbf{e_c}(\mathbf{X}) + \mathbf{e_t}(\mathbf{X}))^q, 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})(\mathbf{e_c}(\mathbf{Y}) + \mathbf{e_t}(\mathbf{Y}))^r\big] + o(1/M). \\ &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^c(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})Cov_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e_c}(X) + \mathbf{e_t}(X))^q, (\mathbf{e_c}(Y) + \mathbf{e_t}(Y))^r]] \\ &+\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})Cov_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e_c}(X) + \mathbf{e_t}(X))^q, (\mathbf{e_c}(Y) + \mathbf{e_t}(Y))^r]] + o(1/M) \\ &= I + II + o(1/M). \end{aligned}$$

Now note that $\mathbf{e_k}^q(X)$ will contain terms of the form $(\mathbf{e_c}(X))^m(\mathbf{e_t}(X))^{q-m}$. For $q \neq m$, the term $(\mathbf{e_c}(X))^m(\mathbf{e_t}(X))^{q-m}$ will be a sum of terms of the form $(k/M)^{q-m-\beta}\tilde{h}(X) \times (k/M)^{m+\beta}\mathbf{P}^{-(m+\beta)}(X)$ for arbitrary $\beta < q-m$.

For $\{X, Y\} \in \Psi_\epsilon$, the covariance term $Cov[(\mathbf{e_c}(X))^m(\mathbf{e_t}(X))^{q-m}, (\mathbf{e_c}(Y))^n(\mathbf{e_t}(Y))^{r-m}]$ will be $o(1/M)$ if either $m < q$ or $n < r$ by (68) and recognizing that the coefficients $(k/M)^{q-m-\beta} = o(1)$ for $m < q$. On the other hand, if $m = q$ and $n = r$, $Cov[(\mathbf{e_c}(X))^q, (\mathbf{e_c}(Y))^r] = 1_{\{q=1,r=1\}}O(1/M) + o(1/M)$ by (68) and recognizing that the error $\mathbf{e}_c$ is a special instance of $\mathbf{e}_t$ and subsequently invoking (69).

For $\{X, Y\} \in \Psi_\epsilon^c$, the covariance term $Cov[(\mathbf{e_c}(X))^m(\mathbf{e_t}(X))^{q-m}, (\mathbf{e_c}(Y))^n(\mathbf{e_t}(Y))^{r-m}]$ using (66) and Cauchy-Schwartz can be shown to be $o(k^{-(q+r)/2})$. On the other hand, if $m = q$ and $n = r$, $Cov[(\mathbf{e_c}(X))^q, (\mathbf{e_c}(Y))^r] = O(k^{-(q+r)/2})$.

We therefore have

$$\begin{aligned}
&I = \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^\mathbf{c}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})Cov_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e_c}(X) + \mathbf{e_t}(X))^q, (\mathbf{e_c}(Y) + \mathbf{e_t}(Y))^r]] \\
&= \mathbb{E}\big[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^\mathbf{c}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\left(1_{\{q=m=1,r=n=1\}}O(1/M) + o(1/M)\right)\big] \\
&= 1_{\{q=m=1,r=n=1\}}O(1/M) + o(1/M).
\end{aligned}$$

where the last step follows from the probability of the region Ψ_ϵ^c being $1 - O(k/M) = O(1)$. Similarly,

$$\begin{aligned}
&II = \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})Cov_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e_c}(X) + \mathbf{e_t}(X))^q, (\mathbf{e_c}(Y) + \mathbf{e_t}(Y))^r]] \\
&= \mathbb{E}\left[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\left(O\left(\frac{1_{\{q=m,r=n\}}}{k^{q+r/2}}\right) + o\left(\frac{1}{k^{q+r/2}}\right)\right)\right] \\
&= \int\left[\left(O\left(\frac{1_{\{q=m,r=n\}}}{k^{q+r/2}}\right) + o\left(\frac{1}{k^{q+r/2}}\right)\right)(\gamma_1(x)f(x)\gamma_2(x) + o(1))\right]\left(\int \Delta_\epsilon(x,y)f(y)dy\right)dx \\
&= \int\left[\left(O\left(\frac{1_{\{q=m,r=n\}}}{k^{q+r/2}}\right) + o\left(\frac{1}{k^{q+r/2}}\right)\right)(\gamma_1(x)f(x)\gamma_2(x) + o(1))\right]O\left(\frac{k}{M}\right)dx \\
&= 1_{\{q=m=1,r=n=1\}}O(1/M) + o(1/M).
\end{aligned}$$

where the last but one step follows from the probability of the region Ψ_ϵ being $O(k/M)$. This then gives

$$\begin{aligned}
&Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e_k}{}^q(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e_k}{}^r(\mathbf{Y})\big] \\
&= Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})(\mathbf{e_c}(\mathbf{X}))^q, 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})(\mathbf{e_c}(\mathbf{Y}))^r\big] + o(1/M).
\end{aligned}$$

Using the fact that $\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}^c}] = O((k/M)^{1/d})$ and Cauchy-Schwartz inequality gives us Lemma B.2. In the next section, we will investigate the central and cross moments of the coverage density estimate.

B.8 Central moments

$P(X)$ has a beta distribution with parameters $k, M-k+1$. This implies

$$\mathbb{E}[\gamma(\mathbf{X})\mathbf{e}_c^q(\mathbf{X})] = 1_{\{q=2\}}\mathbb{E}\big[\gamma(\mathbf{X})f^2(\mathbf{X})\big]\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right).$$

B.9 Cross Moments for the Coverage density estimate

In the previous section, we showed

$$Cov[\gamma_1(\mathbf{X})\mathbf{e}_c^q(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}_c^r(\mathbf{Y})] = \left(1_{\{q=1,r=1\}}O(1/M) + o(1/M)\right).$$

We now concentrate on the case $\{q=1, r=1\}$. We separately analyze the case for disjoint balls and intersecting balls:

$$\begin{aligned}
&Cov[\gamma_1(\mathbf{X})\mathbf{e}_c(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}_c(\mathbf{Y})] = \mathbb{E}[[\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_c(\mathbf{X})\mathbf{e}_c(\mathbf{Y})]] \\
&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}}[\mathbf{e}_c(X), \gamma_2(Y)\mathbf{e}_c(Y)]] \\
&+\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}}[\mathbf{e}_c(X), \gamma_2(Y)\mathbf{e}_c(Y)]] \\
&= I + II.
\end{aligned}$$

- The Disjoint balls case: For $\{X, Y\} \in \Psi_\epsilon$, we can explicitly evaluate the cross-correlation between the coverage density estimates using (68)as follows:

$$\begin{aligned}
I &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})Cov_{\{\mathbf{X},\mathbf{Y}\}}[\mathbf{e}_c(X), \mathbf{e}_c(Y)]] \\
&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})]\left(-1/M + o(1/M)\right) \\
&= \mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})]\left(-1/M + o(1/M)\right) \\
&= -\mathbb{E}[\gamma_1(\mathbf{X})]\mathbb{E}[\gamma_2(\mathbf{Y})]\frac{1}{M} + o(1/M).
\end{aligned}$$

  where the last but one step follows by applying the Cauchy-Schwartz inequality and subsequently using the fact that $\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}] = 1 - o(1)$.

- The Intersecting balls case: For $\{X, Y\} \in \Psi_\epsilon^c$, we will directly show that the cross-correlations of the coverage and the ϵ ball density estimates are identical up to leading terms (without explicitly evaluating the cross-correlation between the coverage density estimates) and then make use of the results developed for the ϵ ball density estimate to obtain corresponding rates for the k-NN estimates.

ϵ ball density estimate

In order to estimate cross moments for the coverage (and thereby k-NN density estimates), we first introduce the ϵ *ball* density estimator. The ϵ *ball* density estimator is an oracle

uniform kernel density estimator with varying bandwidth which depends on the unknown density f. Let the volume of the kernel be $V_\epsilon(X)$ and the corresponding kernel region be $S_\epsilon(X) = \{Y : c_d||X - Y||^d \leq V_\epsilon(X)\}$. The volume is chosen such that the coverage $Q_\epsilon(X) = \int_{S_\epsilon(X)} f(z)dz$ is set to $(1 + p_k)((k-1)/M)$ where p_k is a function of k which we choose such that $p_k^2 k \to 0$ polynomially in k as $k \to \infty$. Let $\mathbf{l}_\epsilon(X)$ denote the number of points among $\{\mathbf{X}_1, .., \mathbf{X}_M\}$ falling in $S_\epsilon(X)$: $\mathbf{l}_\epsilon(\mathbf{X}) = \Sigma_{i=1}^M 1_{\mathbf{X}_i \in S_\epsilon(X)}$. The ϵ *ball* density estimator is defined as

$$\hat{\mathbf{f}}_\epsilon(X) = \frac{\mathbf{l}_\epsilon(\mathbf{X})}{MV_\epsilon(X)}. \tag{70}$$

Also define the error $\mathbf{e}_\epsilon(X)$ as $\mathbf{e}_\epsilon(X) = \hat{\mathbf{f}}_\epsilon(X) - \mathbb{E}[\hat{\mathbf{f}}_\epsilon(X)]$. It is then possible to prove the following lemma. This involves computing volumes of intersections of hyper spheres.

Lemma B.3. *Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density f. Then,*

$$\mathbb{E}\left[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\mathbf{e}_\epsilon(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_\epsilon(\mathbf{Y})\right] = \mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})]\left(\frac{1}{M} + o\left(\frac{1}{M}\right)\right).$$

We would like to now establish that the cross-correlations of the coverage density estimator and the ϵ ball density estimator are equal up to the leading terms. We will first show that,

Lemma B.4. $E[\mathbf{e_c}(X)\mathbf{e}_\epsilon(X)] = f^2(X)\left(\frac{1}{k} + o\left(\frac{1}{k}\right)\right)$.

Proof. We begin by establishing the conditional density and expectation of $\hat{\mathbf{f}}_\epsilon(X)$ given $\hat{\mathbf{f}}_\mathbf{c}(X)$:

Conditional density

We have $\mathbf{l}_\epsilon(X) = \Sigma_{i=1}^M 1_{X_i \in S_\epsilon(X)}$. Since we are dealing with a fixed X, we drop the dependence on X and denote the k-NN coverage by $\mathbf{P}$ and the ϵ ball coverage by Q. Let $\mathbf{q} = Q/\mathbf{P}$ and $\mathbf{r} = (Q - \mathbf{P})/(1 - \mathbf{P})$. We have the following expressions for conditional densities and expectations [20]:

$$\mathbf{P}r\{l_\epsilon = l/\mathbf{P}; \mathbf{P} > Q\} = \begin{cases} \binom{k-1}{l}\mathbf{q}^l(1-\mathbf{q})^{k-1-l} & l = 0, 1, \ldots, k-1 \\ 0 & l = k, k+1, \ldots, M \end{cases}$$

$$\mathbf{P}r\{l_\epsilon = l/\mathbf{P}; \mathbf{P} \leq Q\} = \begin{cases} 0 & l = 0, 1, \ldots, k-1 \\ \binom{M-k}{l-k}\mathbf{r}^{l-k}(1-\mathbf{r})^{M-l} & l = k, k+1, \ldots, M \end{cases}$$

Conditional Expectation

$$\mathbb{E}[l_\epsilon = l/\mathbf{P}; \mathbf{P} > Q] = (k-1)q = (k-1)Q/\mathbf{P}$$
$$\mathbb{E}[l_\epsilon = l/\mathbf{P}; \mathbf{P} \leq Q] = k + r(M-k) = \left(\frac{1-Q}{1-\mathbf{P}}\right)(k-M) + M$$

Joint Expectation

Denote the density of the coverage $\mathbf{P}$ by $f_{k,M}(p)$. Also let $\hat{\mathbf{P}}$ be the coverage corresponding to the $k-2$ nearest neighbor in a total field of $M-3$ points. We can show that

$$\begin{aligned}
\mathbb{E}[\mathbf{e_c}(X)\mathbf{e}_\epsilon(X)] &= \mathbb{E}[\hat{\mathbf{f}}_\epsilon(X)\hat{\mathbf{f}}_\mathbf{c}(X)] - \mathbb{E}[\hat{\mathbf{f}}_\mathbf{c}(X)]\mathbb{E}[\hat{\mathbf{f}}_\epsilon(X)] \\
&= \mathbb{E}\left[\left(\left(\frac{1-Q}{\mathbf{P}(1-\mathbf{P})}\right)(k-M) + M/\mathbf{P}\right)1_{\mathbf{P}\leq Q}\right] \\
&\quad + \frac{f^2(X)(k-1)}{kM}\mathbb{E}\left[\left((k-1)Q/\mathbf{P}^2\right)1_{\mathbf{P}>Q}\right] - \frac{f^2(X)}{k}MQ. \\
&= \frac{f^2(X)}{k}\frac{(M-1)(M-2)}{(k-2)(M-k)} \times \\
&\quad \mathbb{E}[(1-Q\hat{\mathbf{P}})(k-M) + M\hat{\mathbf{P}}(1-\hat{\mathbf{P}})] - \frac{f^2(X)}{k}MQ \\
&\quad + \mathbb{E}[((k-1)Q(1-\hat{\mathbf{P}}) - (1-Q\hat{\mathbf{P}})(k-M) + M\hat{\mathbf{P}}(1-\hat{\mathbf{P}}))(1_{\hat{\mathbf{P}}>Q})] \\
&= C \times (I - II + III).
\end{aligned}$$

We can show that $C \times (I - II) = \frac{f^2(X)}{k}(1-Q)$ using the fact that $\hat{\mathbf{P}}$ has a beta distribution. Note that from the definition of $Q = ((1+p_k)(k-1)/M)$, from the concentration inequality we have that $\mathbb{E}[1_{\hat{\mathbf{P}}>Q}] = O(e^{-p_k^2 k/6})$. The remainder $(C \times III)$ can be simplified and bounded using the Cauchy-Schwartz inequality and the concentration inequality to show $C \times III = o(1/M)$.

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{e_c}(X)\mathbf{e}_\epsilon(X)] &= \frac{f^2(X)}{k}(1-Q) + O(e^{-p_k^2 k/6}). \\
&= \frac{f^2(X)}{k} - \frac{f^2(X)}{M} + o\left(\frac{1}{M}\right). \qquad (71)
\end{aligned}$$

□

Now denote $(\mathbf{e_c}(X) - \mathbf{e}_\epsilon(X))^2 =: \mathbf{E}(X)$. Note that $\mathbb{E}[E(X)] = \mathbb{E}[\mathbf{e_c}(X)^2] - 2E[\mathbf{e_c}(X)\mathbf{e}_\epsilon(X)] + \mathbb{E}[\mathbf{e}_\epsilon(X)^2]$. We know that $E[\mathbf{e_c}(X)^2] = f^2(X)\frac{1}{k} + o(1/k)$ and $E[\mathbf{e}_\epsilon(X)^2] = f^2(X)(1/k + o(1/k))$. Then from the above Lemma B.4, we have that $\mathbb{E}[E(X)] = o(1/k)$. This result should mean that $\mathbf{e_c}(X)$ and $\mathbf{e}_\epsilon(X)$ are almost perfectly correlated. Intuitively this would makes sense in that if more than k points fall in ϵ-ball to give a density estimate that is higher than the mean, then the k-NN distance will be correspondingly smaller and the coverage density estimate will also be higher than the mean and vice versa.

We can now write the covariance between the coverage density estimates in terms of the

covariance between the ϵ ball estimates as follows:

$$\begin{aligned}
&\mathbb{E}[\mathbf{e_c}(X)\mathbf{e_c}(Y)] = \mathbb{E}[(\mathbf{e}_\epsilon(X) + [\mathbf{e_c}(X) - \mathbf{e}_\epsilon(X)])(\mathbf{e}_\epsilon(Y) + [\mathbf{e_c}(Y) - \mathbf{e}_\epsilon(Y)])] \\
&= \mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)] + \mathbb{E}[\mathbf{e}_\epsilon(X)(\mathbf{e_c}(Y) - \mathbf{e}_\epsilon(Y))] + \\
&+\mathbb{E}[\mathbf{e}_\epsilon(Y)(\mathbf{e_c}(X) - \mathbf{e}_\epsilon(X))] + \mathbb{E}[(\mathbf{e_c}(X) - \mathbf{e}_\epsilon(X))(\mathbf{e_c}(Y) - \mathbf{e}_\epsilon(Y))] \\
&= I + II + III + IV. \qquad (72)
\end{aligned}$$

To establish the similarity of the cross-correlation's, we would like to find out that terms II,III and IV are negligible compared to term I. Using Cauchy-Schwartz, we can bound each of the terms II, III and IV in terms of $\mathbb{E}[\mathbf{E}(X)]$ as $|II| \leq \sqrt{\mathbb{E}[\mathbf{E}(Y)]\mathbb{E}[\mathbf{e}_\epsilon{}^2(X)]}$, $|III| \leq \sqrt{\mathbb{E}[\mathbf{E}(X)]\mathbb{E}[\mathbf{e}_\epsilon{}^2(Y)]}$ and $|IV| \leq \sqrt{\mathbb{E}[\mathbf{E}(X)]\mathbb{E}[\mathbf{E}(Y)]}$. Note that the above application of Cauchy-Schwartz helps *decouple* the problem of joint expectation of density estimates located at two *different* points $X$and Y to a problem of estimating the error $\mathbf{E}$ between two different density estimates at the *same* point(s). We then have that all the three terms II, III and IV are $f(X)f(Y)o(1/k)$ and therefore,

$$\mathbb{E}[\mathbf{e_c}(X)\mathbf{e_c}(Y)] = \mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)] + o(1/k),$$

for any X and Y. For this result to be useful, we would want $\mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)]$ to be orders of magnitude larger than the error $o(1/k)$, which is indeed the case for $\{X, Y\} \in \Psi_\epsilon{}^c$ since $\mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)] = O(1/k)$ for such X and Y. We can then use this lemma and the previously established results on co-variance of ϵ-ball density estimates to obtain the corresponding result for coverage density estimates:

Lemma B.5. *Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density f. Then,*

$$\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e_c}(\mathbf{X})\mathbf{e_c}(\mathbf{Y})] \;=\; \mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})]\left(\frac{1}{M} + o\left(\frac{1}{M}\right)\right).$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e_c}(\mathbf{X})\mathbf{e_c}(\mathbf{Y})] \\
&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\mathbf{X},\mathbf{Y}}[\mathbf{e_c}(X)\mathbf{e_c}(Y)]] \\
&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})(\mathbb{E}_{\mathbf{X},\mathbf{Y}}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)] + f(X)f(Y)o(1/k)) \\
&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_\epsilon(\mathbf{X})\mathbf{e}_\epsilon(\mathbf{Y})] \\
&+\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})f(\mathbf{X})f(\mathbf{Y})]o(1/M) \\
&= \mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})]\left(\frac{1}{M} + o\left(\frac{1}{M}\right)\right).
\end{aligned}$$

In the last but one step, we obtain o(1/M) for the second term by recognizing that the volume $V(\Psi_\epsilon{}^c) = O(k/M)$ and $O(k/M) \times o(1/k) = o(1/M)$. □

This implies that

$$\begin{aligned}
&Cov[\gamma_1(\mathbf{X})\mathbf{e}_c(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}_c(\mathbf{Y})] \\
&= Cov[\gamma_1(\mathbf{X})f(\mathbf{X}), \gamma_2(\mathbf{Y})f(\mathbf{Y})] \left(\frac{1}{M} + o\left(\frac{1}{M}\right)\right).
\end{aligned}$$

C Boundary correction for density estimates

In the previous section, we established results for points in the level set $\mathcal{S}' = \{X : \mathbf{S_k}(X) \subset \mathcal{S}\}$ for the k-NN density estimator. In this section we extend results to the entire set $\mathcal{S}$.

C.1 Bias of k-NN density estimator in the interior

We showed that the bias at a point X in the interior of the density $\mathcal{S}'$ is given by

$$\mathbb{E}[\hat{\mathbf{f}}_\mathbf{k}(X)] - f(X) \;=\; h(X)\left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{2/d}\right). \tag{73}$$

C.2 Bias in the k-NN density estimator near boundary

If a probability density function has bounded support, the kNN balls centered at points close to the boundary are often are truncated at the the boundary as shown in Fig. 2. Let

$$\alpha_k(X) = \frac{\int_{\mathbf{S_k}(X)\cap\mathbb{S}} dZ}{\int_{\mathbf{S_k}(X)} dZ}$$

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

$$\mathbb{E}[\hat{\mathbf{f}}_\mathbf{k}(X)] - f(X) \;=\; (1 - \alpha_k(X))f(X) + o(1).$$

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 $k/M \to 0$.

C.3 Boundary corrected k-NN density estimates

Denote the set of N i.i.d. realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ from the density f by $\mathcal{X}_N$. We formally define *boundary points* to be the set of points among $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ whose kNN ball are

truncated by the boundary of the support of the density. Denote the set of boundary points by $\mathcal{B}_{\mathcal{N}}$ and the complementary set of interior points by $\mathcal{I}_{\mathcal{N}}$. 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.

C.4 Boundary point detection

From the concentration inequality and Taylor series expansion of the coverage function, we have

$$1 - Pr\left((1 - p(k,M))\frac{k-1}{M\alpha_k(X)f(X)} \leq \mathbf{V_k}(X) \leq (1 + p(k,M))\frac{k-1}{M\alpha_k(X)f(X)}\right) \leq o(1/M),$$

where $p(k,M)$ is any positive function that satisfies $p(k,M) = \Theta((k/M)^{2/d})$. To detect the boundary points $\mathcal{I}_{\mathcal{N}}$, we pool all the samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_{N+M}\}$ together and construct a K-NN graph where $K = \lfloor k \times (T/M) \rfloor$. This choice of K guarantees that the size of the k-NN ball in the partitioned sample is the same as the the size of the K-NN ball in the pooled sample. Denote the set of interior points in the pooled sample by $\mathcal{I}_{\mathcal{T}}$ and the boundary points by $\mathcal{B}_{\mathcal{T}}$. We can then obtain the set of interior points in $\mathcal{I}_{\mathcal{N}} = \mathcal{X}_{\mathcal{N}} \cap \mathcal{I}_{\mathcal{T}}$.

C.4.1 Identification of interior points $\mathcal{I}_{\mathcal{T}}$

Using the K-NN tree, for each sample $\mathbf{X} \in \mathcal{X}_{\mathcal{T}}$, we compute the number of points in $\mathcal{X}_{\mathcal{T}}$ that have $\mathbf{X}$ as one their l-NN, $l = \{1, \ldots, K\}$. Denote this count as $count(\mathbf{X})$. For any $X \in \mathcal{X}_{\mathcal{T}}$, from the concentration inequality and Taylor series expansion of the coverage function, we have

$$1 - Pr\left((1 - p(K,T))\frac{K-1}{T\alpha_K(X)f(X)} \leq \mathbf{V_K}(X) \leq (1 + p(K,T))\frac{K-1}{T\alpha_K(X)f(X)}\right) \leq o(1/T).$$

This implies that with high probability, the radius of the K-NN ball at X concentrates around $(K - 1/c_d T\alpha_K(X)f(X))^{1/d}$. Let Y be the l-nearest neighbor of X, $l = \{1, \ldots, K\}$. Then we can represent Y as $Y = X + R_K(X)u$ where u is any arbitrary vector with $||u|| \leq 1$.

For X to be one of the K-NN of Y, we need $R_K(Y) \geq ||Y - X||$ or equivalently, $R_K(Y)/R_K(X) \geq ||u||$. Using the concentration inequality for $R_K(X)$ and $R_K(Y)$, a sufficient condition for this is

$$\frac{\alpha_K(X)f(X)}{\alpha_K(Y)f(Y)}(1 - 2p(K,T)) \geq ||u||. \tag{74}$$

Assume that f is Lipschitz continuous with Lipschitz constant $\mathbb{L}$. Then, we have $|f(Y) - f(X)| \leq q(K,T)$ where $q(K,T) = \mathbb{L}(K - 1/c_d T\epsilon_0)^{1/d}$. A further sufficient restriction for X to be one of the K-NN of Y is

$$\frac{\alpha_K(X)}{\alpha_K(Y)}(1 - q(K,T)) \geq ||u||, \tag{75}$$

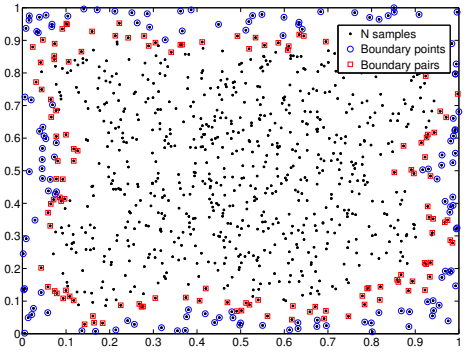


Figure 22: Detection of boundary points for 2D beta distribution.

For interior points, $\alpha_K(X) = 1$. This implies that X will one of the K-NN of Y provided $||u|| \leq 1 - q(K,T)$. From (7) in [18], this implies that $count(\mathbf{X}) \geq K(1-q(K,T))$ whenever $X \in \mathcal{I}_{\mathcal{T}}$. On the other hand, for $X \in \mathcal{B}_{\mathcal{T}}$, $\alpha_K(X) < 1$. It is also clear that for small values of k/M, $\alpha_K(X) < \alpha_K(Y)$ for at least $k/2$ l-NN Y of X. This then implies that $count(\mathbf{X}) < K(1-q(K,T))$ for $X \in \mathcal{B}_{\mathcal{T}}$. We therefore use the threshold $K(1-q(K,T))$ to detect interior points $\mathcal{I}_{\mathcal{T}}$ in the pooled sample. As a final step, we obtain the set of interior points in $\mathcal{I}_{\mathcal{N}} = \mathcal{X}_{\mathcal{N}} \cap \mathcal{I}_{\mathcal{T}}$.

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

1. Construct K-NN tree on $\mathcal{X}$
2. Compute $count(\mathbf{X})$ for each $\mathbf{X} \in \mathcal{X}_{\mathcal{T}}$
3. Detect boundary points $\mathcal{B}_{\mathcal{T}}$:

for each $\mathbf{X} \in \mathcal{X}_{\mathcal{T}}$ **do**
  if $count(\mathbf{X}) < (1 - q(K,T)) * K$ **then**
    $\mathcal{B}_{\mathcal{T}} \leftarrow \mathbf{X}$
  else
    $\mathcal{I}_{\mathcal{T}} \leftarrow \mathbf{X}$
  end if
end for

C.5 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 $f(\mathbf{Y})$.

Then, for each $\mathbf{X}_i \in \mathcal{B}_{\mathcal{N}}$, we identify its nearest neighbor $\mathbf{X}_{n(i)} \in \mathcal{I}_{\mathcal{N}}$, for some $n(i) = \{1, .., N\}$. 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 $O((k/M)^{1/d})$. Fig. C.4.1 depicts the detection of boundary points and pairing of boundary points with interior points. Clearly, the algorithm identifies the boundary points in this example.

Let $\mathbf{X}_i$ be a boundary point. From Eq. 74, 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

$$\hat{\mathbf{f}}_C(\mathbf{X}_i) = \begin{cases} 2\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{p(i)}) & 2\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{p(i)}) > 0 \\ \hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{n(i)}) & 2\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{p(i)}) \leq 0 \end{cases}$$

From the concentration inequality, the event $2\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{p(i)}) \leq 0$ will occur with probability $o(1/M)$.

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 $f'(\mathbf{X}_{n(i)})$

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

Eqs. 9 and 10 imply

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

Because $\mathbf{X}_{n(i)}$ and $\mathbf{X}_{p(i)}$ are located in the interior of the density, by Eq. 73,

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

and therefore

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{f}}_C(\mathbf{X}_i)] &= \mathbb{E}[2\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{n(i)}) - \hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{p(i)})] + o(1/M) \\ &= 2f(\mathbf{X}_{n(i)}) - f(\mathbf{X}_{p(i)}) + O(||h||^2) + o(1/M) \\ &= f(\mathbf{X}_i) + O((k/M)^{2/d}). \end{aligned}$$

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 Eq. 73 and Eq. 74). Thus the compensation has reduced the bias of the estimator from $O(1)$ to $O((k/M)^{2/d})$. In the definition of the corrected estimator in Eq.76, $\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{n(i)})$ and $\hat{\mathbf{f}}_{\mathbf{k}}(\mathbf{X}_{p(i)})$ are standard k-NN density estimates. It therefore follows that the variance and other central and cross moments of the corrected density estimate will continue to decay at the same rate as the standard k-NN density estimates in the interior.

Note that prior to the correction, the contribution to the bias from the interior was $O((k/M)^{2/d})$ while the contribution from the boundary was $O(1) \times O((k/M)^{1/d}) = O((k/M)^{1/d})$. Given these identical rates and that the probability of a point being in a boundary region is $O((k/M)^{1/d}) = o(1)$, the contribution of the corrected estimator to the overall bias, variance and other cross moments are negligible compared to the contribution from the interior. As a result we can now generalize the results from Appendix B to include the boundary regions.

Bias

$$\mathbb{E}[\gamma(\mathbf{X})(\hat{\mathbf{f}}_k(\mathbf{X}) - f(\mathbf{X}))] = \mathbb{E}[\gamma(\mathbf{X})h(\mathbf{X})] \left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{2/d}\right).$$

Central moments

$$\mathbb{E}[\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})] = 1_{\{q=2\}}\mathbb{E}\big[\gamma(\mathbf{X})f^2(\mathbf{X})\big] \left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right).$$

Cross moments

$$\begin{aligned} &Cov[\gamma_1(\mathbf{X})\mathbf{e}_k^q(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}_k^r(\mathbf{Y})] \\ &= 1_{\{q=1,r=1\}} Cov[\gamma_1(\mathbf{X})f(\mathbf{X}), \gamma_2(\mathbf{Y})f(\mathbf{Y})] \left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right). \end{aligned}$$

Equivalent corrections exist for the uniform kernel density estimator and will be left to the reader.

D Proof of theorems on bias and variance

In this section we provide results on the bias and variance of the k-NN plug-in estimators, and therefore the mean square error (M.S.E). We assume we have $T = N + M$ i.i.d. realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_{N+M}\}$ from the density f. The plug-in estimate is given by

$$\hat{\mathbf{G}}(f) \;=\; \left(\frac{1}{N}\sum_{i=1}^{N} g(\hat{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)\right). \tag{76}$$

D.1 Bias

Theorem D.1. *The bias of the plug-in estimator $\hat{\mathbf{G}}_k(f)$ is then given by*

$$\mathbb{B}(\hat{\mathbf{G}}(f)) \;=\; c_1 \left(\frac{k}{M}\right)^{2/d} + c_2 \left(\frac{1}{k}\right) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right),$$

where the constants $c_1 = \mathbb{E}[g'(f(\mathbf{Y}), \mathbf{Y}) f^{-2/d}(\mathbf{Y})(\Gamma^{(2/d)}((n+2)/2) tr[\nabla^2(f(\mathbf{Y}))])]$, *and* $c_2(\delta) = \mathbb{E}[f^2(\mathbf{Y}) g''(f(\mathbf{Y}), \mathbf{Y})/2]$ *depend on the underlying density* f.

Proof. Let $\mathbf{Z}$ be a random variable with density f. By the continuity of $g'''(x, y)$, we can construct the following third order Taylor series of $g(\hat{\mathbf{f}}_k(\mathbf{Z}), \mathbf{Z})$ around the conditional expected value $\mathbb{E}_Z[\hat{\mathbf{f}}_k(\mathbf{Z})]$.

$$\begin{aligned} g(\hat{\mathbf{f}}_k(\mathbf{Z}), \mathbf{X}) &= g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) + g'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - f(\mathbf{Z})) + \frac{1}{2} g''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - f(\mathbf{Z}))^2 \\ &+ \frac{1}{6} g^{(3)}(\zeta_{\mathbf{Z}}, \mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - f(\mathbf{Z}))^3, \end{aligned}$$

where we are guaranteed the existence of $\zeta_{\mathbf{Z}} \in (g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}), g(\hat{\mathbf{f}}_k(\mathbf{Z}), \mathbf{Z})$ by the mean value theorem. This gives,

$$\begin{aligned} &\mathbb{E}[(g(\hat{\mathbf{f}}_k(\mathbf{Z}), \mathbf{Z}) - g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}))] = \mathbb{E}\Big[g'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})])\Big] \\ &+\mathbb{E}\left[\frac{1}{2} g''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})])^2\right] + \mathbb{E}\left[\frac{1}{6} g^{(3)}(\zeta_{\mathbf{Z}}, \mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})])^3\right] \\ &= 0 + \mathbb{E}[f^2(\mathbf{Y}) g''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Y})/2] \left(\left(\frac{1}{k}\right) + o\left(\left(\frac{1}{k}\right)\right)\right) + \mathbb{E}\left[\frac{1}{6} g^{(3)}(\zeta_{\mathbf{Z}}, \mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - f(\mathbf{Z}))^3\right] \\ &= c_2\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right) + \mathbb{E}\left[\frac{1}{6} g^{(3)}(\zeta_{\mathbf{Z}}, \mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - f(\mathbf{Z}))^3\right]. \end{aligned}$$

where the last but one step follows from the joint continuity of $g^{(3)}(x, y)$ (in the interval $x \in (\delta, f_\infty)$) and the moment properties of density estimates. Let $\Delta(\mathbf{Z}) = \frac{1}{6} g^{(3)}(\zeta_{\mathbf{Z}}, \mathbf{Z})$. From the consistency of k-NN density estimates, it follows that $\Delta(\mathbf{Z})$ converges in probability to $\frac{1}{6} g^{(3)}(f(\mathbf{Z}), \mathbf{Z})$. This combined with the fact that $\Delta(\mathbf{Z})$ is uniformly bounded implies that $\mathbb{E}[\Delta^2(\mathbf{Z})] = O(1)$. By Cauchy-schwartz,

$$\begin{aligned} \left|\mathbb{E}\left[\frac{1}{6}\Delta(\mathbf{Z})(\hat{\mathbf{f}}(\mathbf{Z}) - f(\mathbf{Z}))^3\right]\right| &\leq \sqrt{\mathbb{E}\left[\frac{1}{36}\Delta^2(\mathbf{Z})\right] \mathbb{E}\left[(\hat{\mathbf{f}}(\mathbf{Z}) - f(\mathbf{Z}))^6\right]} \\ &= o\left(\frac{1}{k}\right). \end{aligned}$$

From the analysis of bias of k-NN density estimates and using Taylor series, the second term can similarly be reduced to

$$\mathbb{E}[g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z})] = c_1\left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{2/d}\right)$$

□

D.2 Variance

Theorem D.2. *The variance of the plug-in estimator $\hat{\mathbf{G}}_k(f)$ is given by*

$$\mathbb{V}(\hat{\mathbf{G}}(f)) \;=\; c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M}+\frac{1}{N}\right),$$

where the constants $c_4 = \mathbb{V}[g(f(\mathbf{Y}),\mathbf{Y})]$ and $c_5 = \mathbb{V}[f(\mathbf{Y})g'(f(\mathbf{Y}),\mathbf{Y})]$ depend on the underlying density f.

Proof. Let $\mathbf{Z}$ be a random variable with density f. By the continuity of $g^{(\lambda)}(x,y)$, we can construct the following Taylor series of $g(\hat{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z})$ around the conditional expected value $\mathbb{E}_Z[\hat{\mathbf{f}}_k(\mathbf{Z})]$.

$$\begin{aligned} g(\hat{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z}) &= g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}) + g'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})(\hat{\mathbf{f}}_k(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})]) \\ &+ \left(\sum_{i=2}^{\lambda-1} \frac{g^{(i)}(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})}{i!}(\hat{\mathbf{f}}_k(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})])^i\right) \\ &+ \frac{g^{(\lambda)}(\xi_{\mathbf{Z}},\mathbf{Z})}{\lambda!}(\hat{\mathbf{f}}_k(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})])^\lambda, \end{aligned}$$

where we are guaranteed the existence of $\xi_{\mathbf{Z}} \in (g(f(\mathbf{Z})), g(\hat{\mathbf{f}}_k(\mathbf{Z})))$ by the mean value theorem. Denote $(g'(\xi_{\mathbf{Z}},\mathbf{Z}))/\lambda!$ by $\Psi(\mathbf{Z})$. Define the operator $\mathcal{M}(\mathbf{Z}) = \mathbf{Z} - \mathbb{E}[\mathbf{Z}]$ and

$$\begin{aligned} p_i &= M(g(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}_k(\mathbf{X}_i)],\mathbf{X_i})), \\ q_i &= M(g'(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}_k(\mathbf{X}_i)],\mathbf{X_i})(\hat{\mathbf{f}}(\mathbf{X_i}) - \mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}_k(\mathbf{X}_i)])), \\ r_i &= M\left(\sum_{i=2}^{\lambda} \frac{g^{(i)}(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}_k(\mathbf{X}_i)],\mathbf{X_i})}{i!}(\hat{\mathbf{f}}(\mathbf{X_i}) - \mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}_k(\mathbf{X}_i)])^i\right), \\ s_i &= M\left(\Psi(\mathbf{X_i})(\hat{\mathbf{f}}_k(\mathbf{X_i}) - \mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}_k(\mathbf{X}_i)])^\lambda\right). \end{aligned}$$

The variance of the estimator $\hat{\mathbf{G}}(f)$ is given by

$$\begin{aligned} \mathbb{V}(\hat{\mathbf{G}}(f)) &= \mathbb{E}[(\hat{\mathbf{G}}(f) - \mathbb{E}[\hat{\mathbf{G}}(f)])^2] \\ &= \frac{1}{N}\mathbb{E}\big[(p_1+q_1+r_1+s_1)^2\big] \\ &+ \frac{N-1}{N}\mathbb{E}[(p_1+q_1+r_1+s_1)(p_2+q_2+r_2+s_2)]. \end{aligned}$$

Because $\mathbf{X}_1, \mathbf{X}_2$ are independent, we have $\mathbb{E}[(p_1)(p_2+q_2+r_2+s_2)] = 0$. We also have,

$$\begin{aligned} \mathbb{E}\big[(p_1+q_1+r_1+s_1)^2\big] &= \mathbb{E}[{p_1}^2] + o(1) \\ &= \mathbb{V}[g(\mathbb{E}_{\mathbf{Y}}[\hat{\mathbf{f}}(\mathbf{Y})],\mathbf{Y})] + o(1) \\ &= c_4 + o(1). \end{aligned}$$

From the results on cross moments for density estimates we then have the following results.

$$\mathbb{E}[q_1 q_2] \;=\; c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right),$$

and additionally $\mathbb{E}[q_1 r_2] = o\left(\frac{1}{M}\right)$ and $\mathbb{E}[r_1 r_2] = o\left(\frac{1}{M}\right)$. Also note that q_1 and s_2 are 0 mean random variables. This implies that

$$\begin{aligned}
\mathbb{E}[q_1 s_2] &= \mathbb{E}\left[q_1 \Psi(\mathbf{X_2})(\hat{\mathbf{f}}(\mathbf{X_2}) - \mathbb{E}_{\mathbf{X}_2}[\hat{\mathbf{f}}_k(\mathbf{X}_2)])^{\lambda}\right] \\
&= \mathbb{E}\left[q_1 \Psi(\mathbf{X_2})(\hat{\mathbf{f}}(\mathbf{X_2}) - \mathbb{E}_{\mathbf{X}_2}[\hat{\mathbf{f}}_k(\mathbf{X}_2)])^{\lambda}\right] \\
&\leq \sqrt{\mathbb{E}\left[\Psi^2(\mathbf{X_2})\right]\mathbb{E}\left[q_1^2(\hat{\mathbf{f}}(\mathbf{X_2}) - \mathbb{E}_{\mathbf{X}_2}[\hat{\mathbf{f}}_k(\mathbf{X}_2)])^{2\lambda}\right]} \\
&= \sqrt{\mathbb{E}\left[\Psi^2(\mathbf{Z})\right]} o\left(\frac{1}{k^{\lambda}}\right)
\end{aligned}$$

We can show that $\mathbb{E}\left[\Psi^2(\mathbf{Z})\right] = O(1)$ (in an identical manner to showing $\mathbb{E}\left[\Delta^2(\mathbf{Z})\right] = O(1)$ in the proof for the result on bias). Note that from the polynomial growth condition on k, $o\left(\frac{1}{k^{\lambda}}\right) = o(1/M)$. In a similar manner, we can show $\mathbb{E}[r_1 s_2] = o\left(\frac{1}{M}\right)$ and $\mathbb{E}[s_1 s_2] = o\left(\frac{1}{M}\right)$. This implies that

$$\begin{aligned}
\mathbb{V}(\hat{\mathbf{G}}(f)) &= \frac{1}{N}\mathbb{E}\left[{p_1}^2\right] + \frac{(N-1)}{N}\mathbb{E}[q_1 q_2] + o\left(\frac{1}{M} + \frac{1}{N}\right). \\
&= c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right).
\end{aligned}$$

□

E Asymptotic normality

In this section we provide results on the asymptotic distribution of the plug-in estimators. We assume we have $T = N + M$ i.i.d. realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_{N+M}\}$ from the density f. The plug-in estimate is given by

$$\hat{\mathbf{G}}(f) \;=\; \left(\frac{1}{N}\sum_{i=1}^{N} g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)\right). \tag{77}$$

E.1 Interchangeable random variables

Define the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ for any fixed M as

$$\mathbf{Y}_{M,i} = g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) - \mathbb{E}[g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)], \tag{78}$$

and define the sum $\mathbf{S_{N,M}}$ as

$$\mathbf{S}_{N,M} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Y}_{M,i}, \tag{79}$$

where the indices N and M explicitly stress the dependence of the sum $\mathbf{S}_{N,M}$ on the number of random variables $N+M$. We let $\hat{\mathbf{f}}$ be either the uniform kernel or the kNN density estimates. We observe that the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ belong to an interchangeable process for all values of M.

E.2 Covariance properties

From the regularity conditions listed in Section 3.1, we obtain the following lemma on the covariance properties of this interchangeable process.

Lemma E.1.

$$\begin{aligned}
\mathbb{V}(\mathbf{Y}_{M,i}) &= \mathbb{V}(g(f(\mathbf{X}), \mathbf{X})) + o(1), \\
Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) &= \frac{\mathbb{V}(g'(f(\mathbf{X}), \mathbf{X}) f(\mathbf{X}))}{M} + o\left(\frac{1}{M}\right), \\
Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}) &= \frac{\mathbb{V}(g'(f(\mathbf{X}), \mathbf{X})(g(f(\mathbf{X}), \mathbf{X}) - \mathbb{E}[g(f(\mathbf{X}), \mathbf{X})]) f(\mathbf{X}))}{M} + o\left(\frac{1}{M}\right).
\end{aligned}$$

Proof. From the properties of the density estimates listed in Section 3.1,

$$\mathbb{V}(\mathbf{Y}_{M,i}) = \mathbb{V}g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) = \mathbb{V}(g(f(\mathbf{X}), \mathbf{X})) + o(1),$$

and

$$Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) = Cov(g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i), g(\hat{\mathbf{f}}(\mathbf{X}_j), \mathbf{X}_j)) = \frac{\mathbb{V}(g'(f(\mathbf{X}), \mathbf{X}) f(\mathbf{X}))}{M} + o\left(\frac{1}{M}\right).$$

Define $d(x, y) = g(x, y)(g(x, y) - c)$, where the constant $c = \mathbb{E}[g(\hat{\mathbf{f}}(\mathbf{X}_1), \mathbf{X}_1)]$. Then, using the properties of the density estimates listed in Section 3.1,

$$\begin{aligned}
Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}) &= Cov(d(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i), d(\hat{\mathbf{f}}(\mathbf{X}_j), \mathbf{X}_j)) \\
&= \frac{\mathbb{V}(d'(f(\mathbf{X}), \mathbf{X}) f(\mathbf{X}))}{M} + o\left(\frac{1}{M}\right) \\
&= \frac{\mathbb{V}(g'(f(\mathbf{X}), \mathbf{X})(g(f(\mathbf{X}), \mathbf{X}) - \mathbb{E}[g(f(\mathbf{X}), \mathbf{X})]) f(\mathbf{X}))}{M} + o\left(\frac{1}{M}\right).
\end{aligned}$$

□

E.3 CLT for Interchangeable Processes

Let $\{\mathbf{Z}_i; i = 1, 2, \ldots\}$ be an interchangeable stochastic process with 0 mean and variance ρ. Blum et.al.[4] showed that the random variable $\mathbf{S}_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Z}_i$ converges in distribution to $N(0, \rho)$ if and only if $Cov(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $Cov(\mathbf{Z}_1^2, \mathbf{Z}_2^2) = 0$.

E.3.1 De Finetti's Theorem

De Finetti showed that the probability measure of any interchangeable process is a mixture of probability measures each consisting of independent and identically distributed random variables. Specifically, let $\mathcal{F}$ be the class of one dimensional distribution functions and for each pair of real numbers x and y define

$$\mathcal{F}(x, y) = \{F \in \mathcal{F} | F(x) \leq y\}. \tag{80}$$

Let $\mathbb{B}$ be the Borel field of subsets of $\mathcal{F}$ generated by the class of sets $\mathcal{F}(x, y)$. Then De Finetti's theorem asserts that for any interchangeable process $\{\mathbf{Z}_i\}$ there exists a probability measure μ defined on $\mathbb{B}$ such that

$$Pr\{\mathbf{B}\} = \int_{\mathcal{F}} Pr_F\{\mathbf{B}\} d\mu(F), \tag{81}$$

for any Borel measurable set defined on the sample space of the sequence $\{\mathbf{Z}_i\}$. Here $Pr\{\mathbf{B}\}$ is the probability of the event $\mathbf{B}$ and $Pr_F\{\mathbf{B}\}$ is the probability of the event B under the assumption that component random variables $\mathbf{X}_i$ of the interchangeable process are independent and identically distributed with distribution F.

E.3.2 Necessary and Sufficient conditions for CLT

The Central Limit theorem [2] is said to hold if for any real number α if we have

$$\lim_{N \to \infty} Pr\{\mathbf{S}_N \leq \alpha\} = \phi(\alpha), \tag{82}$$

where $\phi(\alpha)$ is the distribution function of a Gaussian random variable with 0 mean and variance ρ.

From Eq.81 we have

$$Pr\{\mathbf{S}_N \leq \alpha\} = \int_{\mathcal{F}} Pr_F\{\mathbf{S}_N \leq \alpha\} d\mu(F). \tag{83}$$

For each $F \in \mathcal{F}$ define $m(F)$ and $\sigma^2(F)$ as

$$m(F) = \int_{-\infty}^{\infty} x dF(x), \tag{84}$$

$$\sigma(F) = \int_{-\infty}^{\infty} x^2 dF(x) - \rho. \tag{85}$$

Also define

$$\kappa(F) = \int_{-\infty}^{\infty} |x - m(F)|^3 dF(x), \tag{86}$$

and for all real numbers m and non-negative real numbers σ^2 let $\digamma_{m,\sigma^2}$ be the set of $F \in \digamma$ for which $m(F) = m$ and $\sigma^2(F) = \sigma^2$.

Blum et.al show that the process $\{\mathbf{Z}_i\}$ will satisfy the CLT if and only if $\mu(\digamma_{0,0}) = 1$. Furthermore, they show that the condition $\mu(\digamma_{0,0}) = 1$ is equivalent to the condition that $Cov(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $Cov(\mathbf{Z}_1^2, \mathbf{Z}_2^2) = 0$.

We state the classical Berry-Esseen Theorem for i.i.d sequences in the next section.

E.3.3 Classical Berry-Esseen Theorem

Let $\mathbf{X}_1, \mathbf{X}_2, ...,$ be i.i.d. random variables with $E(\mathbf{X}_1) = 0$, $E(\mathbf{X}_1^2) = \sigma^2 > 0$, and $E(|\mathbf{X}_1|^3) = \kappa < \infty$. Also, let $\mathbf{S}_n = (\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_n)/\sqrt{n}$ be the sample mean, with F_n the C.D.F of S_n, and ψ the C.D.F of Gaussian distribution with 0 mean and variance σ^2. Then there exists a positive constant C such that for all x and n,

$$|F_n(x) - \psi(x)| \leq \frac{C\kappa}{\sigma^3\sqrt{n}}. \tag{87}$$

E.4 CLT for Asymptotically Uncorrelated Interchangeable processes

Consider the array of random variables $\{\mathbf{Y}_{M,i}\}$ defined earlier. We make the observation that $Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j})$ and $Cov(\mathbf{Y}_{M,i}^2, \mathbf{Y}_{M,j}^2)$ approach 0 as $M \to \infty$. We refer to this array as an interchangeable process that is asymptotically uncorrelated in M.

We will now establish that the Central Limit Theorem will hold for the sum $\mathbf{S}_{N,M} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Y}_{M,i}$ as M approaches ∞.

E.4.1 Details

Denote $\mathbb{V}(\mathbf{Y}_{M,i})$ by ρ. Let $\kappa = \mathbb{E}[|\mathbf{Y}_{M,i}|^3]$, the absolute third order central moment, be finite. From the definition of $\kappa(F)$, it is clear that $\kappa(F) \leq \kappa$ for any $F \in \digamma$, and therefore $\kappa(F)$ is finite for any $F \in \digamma$.

Let $\delta_\mu(M)$ and $\delta_\sigma(M)$ be a strictly positive functions parameterized by M such that

$$\delta_\mu(M) = o(1); \frac{1}{M\delta_\mu(M)} = o(1), \tag{88}$$

$$\delta_\sigma(M) = o(1); \frac{1}{M\delta_\sigma(M)} = o(1). \tag{89}$$

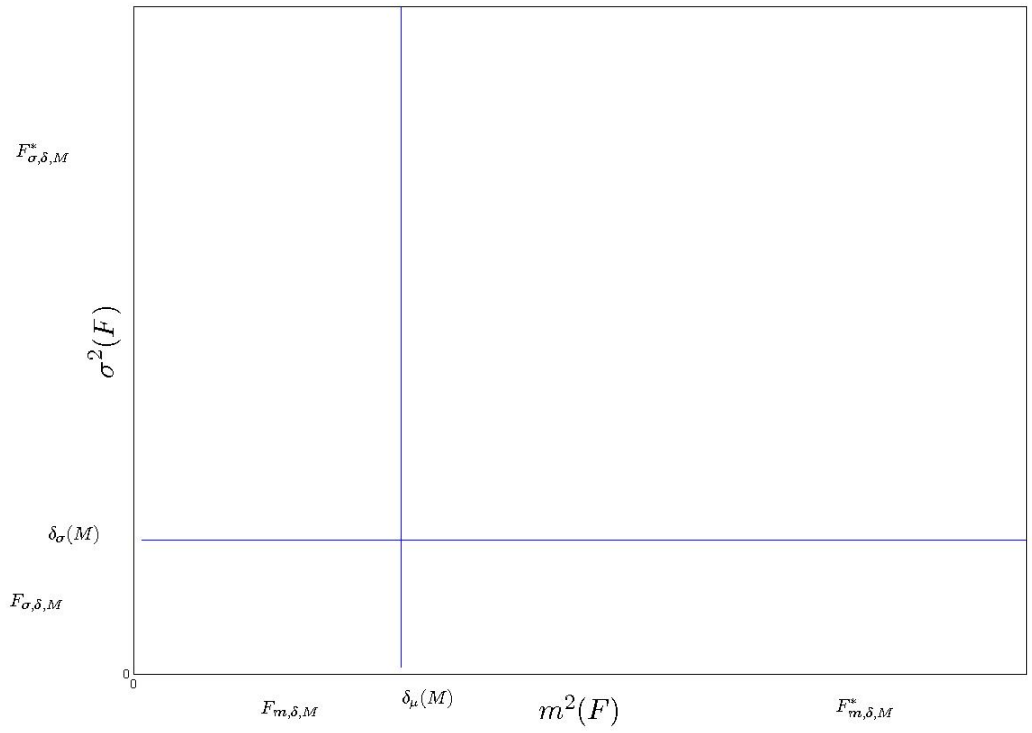


Figure 23: Partition.

Denote the set of $F \in \digamma$ with

1. $\{m^2(F) \geq \delta_\mu(M)\}$ by $\digamma_{m,\delta,M}$; $\mu_{m,\delta,M} := \mu(\digamma_{m,\delta,M})$.
2. $\{\sigma^2(F) \geq \delta_\sigma(M)\}$ by $\digamma_{\sigma,\delta,M}$; $\mu_{\sigma,\delta,M} := \mu(\digamma_{\sigma,\delta,M})$.

3. $\{m^2(F) \in (0, \delta_\mu(M))\}$ by $\digamma^*_{m,\delta,M}$; $\mu^*_{m,\delta,M} := \mu(\digamma^*_{m,\delta,M})$.

4. $\{\sigma^2(F) \in (0, \delta_\sigma(M))\}$ by $\digamma^*_{\sigma,\delta,M}$; $\mu^*_{\sigma,\delta,M} := \mu(\digamma^*_{\sigma,\delta,M})$.

We have from Eq.81 that

$$\int_{\digamma} m^2(F) d\mu(F) = Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) \tag{90}$$

$$\int_{\digamma} [\mathbb{E}_F[\mathbf{Z}^2 - \rho]]^2 d\mu(F) = Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}) \tag{91}$$

Applying the Chebyshev inequality, we get

$$\delta_\mu(M)\mu_{m,\delta,M} \leq Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}), \tag{92}$$

$$\delta_\sigma(M)\mu_{\sigma,\delta,M} \leq Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}). \tag{93}$$

From Eq.92 and Eq.88 we therefore have that $\mu_{m,\delta,M}$ and $\mu_{\sigma,\delta,M} \to 0$ as $M \to \infty$. From the definition of $\digamma^*_{m,\delta,M}$ and $\digamma^*_{\sigma,\delta,M}$, we also have that $\mu^*_{m,\delta,M}$ and $\mu^*_{\sigma,\delta,M} \to 0$ as $M \to \infty$.

We first establish the following lemma,

Lemma E.2.

$$\lim_{N,M\to\infty} \int_{\digamma_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) = \phi(\alpha),$$

where $\phi(.)$ is the distribution function of a Gaussian random variable with mean 0 and variance ρ.

Proof. For $F \in \digamma_{0,0}$, the mean and variance are respectively given by 0 and ρ respectively. We then have the following upper and lower bounds -

Lower bound

$$\begin{aligned}
&\int_{\digamma_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) \\
&\int_{\digamma_{0,0}} \left[\phi(\alpha) - \frac{C\kappa(F)}{\rho^3\sqrt{N}}\right] d\mu(F) \\
&\geq \phi(\alpha)\, \mu(\digamma_{0,0}) - \frac{C\kappa}{\rho^3\sqrt{N}}.
\end{aligned}$$

Upper bound

$$\begin{aligned}
&\int_{F_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) \\
&\leq \int_{F_{0,0}} \left[\phi(\alpha) + \frac{C\kappa(F)}{\rho^3\sqrt{N}}\right] d\mu(F) \\
&\leq \phi(\alpha)\,\mu(F_{0,0}) + \frac{C\kappa}{\rho^3\sqrt{N}}.
\end{aligned}$$

We also have

$$1 - (\mu_{m,\delta,M} + \mu_{\sigma,\delta,M} + \mu^*_{m,\delta,M} + \mu^*_{\sigma,\delta,M}) \leq \mu(F_{0,0}) \leq 1,$$

and therefore

$$\lim_{M\to\infty} \mu(F_{0,0}) = 1.$$

Combining the above equations, we get

$$\lim_{N,M\to\infty} \int_{F_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) = \phi(\alpha).$$

□

In order to apply the Berry-Esseen bounds, we require $\kappa = \mathbb{E}[|\mathbf{Y}_{M,i}|^3]$ to be finite. We show that a sufficient condition to guarantee this is to require $\mathbb{E}[|g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)|^3]$ to be finite. To see this,

$$\begin{aligned}
\kappa &= \mathbb{E}[|\mathbf{Y}_{M,i}|^3] \\
&= \mathbb{E}[|g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) - \mathbb{E}[g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)]|^3] \\
&\leq \mathbb{E}[(|g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)| + |\mathbb{E}[g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)]|)^3] \\
&\leq 8\mathbb{E}[|g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)|^3]. \qquad (94)
\end{aligned}$$

We now state and prove the Central Limit theorem for asymptotically uncorrelated Interchangeable processes.

Theorem E.3. *Suppose that the density f, the functional g and the density estimate $\hat{\mathbf{f}}$ satisfy the conditions listed in Section 3.1. Further suppose $\mathbb{E}[|g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)|^3]$ is finite. The asymptotic distribution of the plug-in estimator $\hat{\mathbf{G}}(f)$ is given by*

$$\lim_{\Delta(k,N,M)\to 0} Pr\left(\frac{\hat{\mathbf{G}}(f) - \mathbb{E}[\hat{\mathbf{G}}(f)]}{\sqrt{\mathbb{V}[g(f,x)]/N}} \leq \alpha\right) = Pr(\mathbf{Z} \leq \alpha),$$

where $\mathbf{Z}$ *is a standard normal random variable.*

Proof. Since all probability measures are bounded between 0 and 1, we have the following upper and lower bounds

Lower bound

$$\begin{aligned} Pr\{\mathbf{S}_{N,M} \leq \alpha\} &= \int_{\mathcal{F}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) \\ &\geq \int_{\mathcal{F}_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F). \end{aligned}$$

Upper bound

$$\begin{aligned} Pr\{\mathbf{S}_{N,M} \leq \alpha\} &= \int_{\mathcal{F}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) \\ &\leq \int_{\mathcal{F}_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) + \\ &\mu_{m,\delta,M} + \mu_{\sigma,\delta,M} + \mu^*_{m,\delta,M} + \mu^*_{\sigma,\delta,M}. \end{aligned}$$

The above bounds along with Lemma E.2 gives the required result. □

E.5 Berry-Esseen bounds

We now establish Berry-Esseen bounds for the case where $\frac{N}{M} \to 0$. In particular, we assume that there exists a $\delta : 0 < \delta < 1$, such that $N = O(M^\delta)$. We also assume that the interchangeable process has finite absolute third order moment $E(|\mathbf{Z}_{M,i}|^3) = \rho_M < \infty \vee M$.

E.5.1 Details

Define the subset $\tilde{\mathcal{F}}$ of $\mathcal{F}$ as follows: $\tilde{\mathcal{F}} = \mathcal{F} - \{\mathcal{F}_{m,\delta,M} \bigcup \mathcal{F}_{\sigma,\delta,M}\}$.

We recognize that for $F \in \tilde{\mathcal{F}}$, we have

$$\begin{aligned} -\sqrt{\delta_\mu(M)} \leq m(F) \leq \sqrt{\delta_\mu(M)}, \\ -\sqrt{\delta_\sigma(M)} \leq \sigma(F) \leq \sqrt{\delta_\sigma(M)}. \end{aligned}$$

The mean and variance of $Y_{M,i}$ under the distribution F are given by $m(F)$ and $\sigma(F) + \rho - m^2(F)$ respectively.

As in the previous section, let ϕ be the distribution function of a Gaussian random variable with 0 mean and ρ variance.

Lower bound

$$
\begin{aligned}
&Pr\{\mathbf{S}_{N,M} \leq \alpha\} = \int_{\mathcal{F}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) \\
&\geq \int_{\tilde{F}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) \\
&\geq \int_{\tilde{F}} \left[\phi\left(\frac{\alpha - \sqrt{N} m(F)}{1 + (\sigma(F) - m^2(F))/\rho} \right) - \frac{C\kappa(F)}{(\sigma(F) + \rho - m^2(F))^3 \sqrt{N}} \right] d\mu(F) \\
&\geq \phi\left(\frac{\alpha - \sqrt{N\delta_\mu(M)}}{1 + (\sqrt{\delta_\sigma(M)})/\rho} \right) \mu(\tilde{F}) - \int_{\tilde{F}} \frac{C\kappa(F)}{(\rho - \sqrt{\delta_\sigma(M)} - \delta_\mu(M))^3 \sqrt{N}} d\mu(F) \\
&\geq \phi\left(\frac{\alpha - \sqrt{N\delta_\mu(M)}}{1 + (\sqrt{\delta_\sigma(M)})/\rho} \right) \mu(\tilde{F}) - \frac{C\kappa}{(\rho - \sqrt{\delta_\sigma(M)} - \delta_\mu(M))^3 \sqrt{N}}.
\end{aligned}
$$

Upper bound

Denote $\mu(\tilde{F}^c) := \tilde{\mu}$. We note that $\tilde{\mu} \leq \mu_{m,\delta,M} + \mu_{\sigma,\delta,M}$.

$$
\begin{aligned}
&Pr\{\mathbf{S}_{N,M} \leq \alpha\} = \int_{\mathcal{F}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) \\
&\leq \int_{\tilde{F}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) + \tilde{\mu} \\
&\leq \int_{\tilde{F}} \left[\phi\left(\frac{\alpha - \sqrt{N} m(F)}{1 + (\sigma(F) - m^2(F))/\rho} \right) + \frac{C\kappa(F)}{(\sigma(F) + \rho - m^2(F))^3 \sqrt{N}} \right] d\mu(F) + \tilde{\mu} \\
&\leq \phi\left(\frac{\alpha + \sqrt{N\delta_\mu(M)}}{1 - (\sqrt{\delta_\sigma(M)} + \delta_\mu(M))/\rho} \right) \mu(\tilde{F}) + \int_{\tilde{F}} \frac{C\kappa(F)}{(\rho + \sqrt{\delta_\sigma(M)})^3 \sqrt{N}} d\mu(F) + \tilde{\mu} \\
&\leq \phi\left(\frac{\alpha - \sqrt{N\delta_\mu(M)}}{1 - (\sqrt{\delta_\sigma(M)} + \delta_\mu(M))/\rho} \right) \mu(\tilde{F}) + \frac{C\kappa}{(\rho + \sqrt{\delta_\sigma(M)})^3 \sqrt{N}} + \mu_{m,\delta,M} + \mu_{\sigma,\delta,M} \\
&\leq \phi\left(\frac{\alpha - \sqrt{N\delta_\mu(M)}}{1 - (\sqrt{\delta_\sigma(M)} + \delta_\mu(M))/\rho} \right) \mu(\tilde{F}) + \frac{C\kappa}{(\rho + \sqrt{\delta_\sigma(M)})^3 \sqrt{N}} + \frac{1}{M\delta_\mu(M)} + \frac{1}{M\delta_\sigma(M)}.
\end{aligned}
$$

The result in Theorem E.3 therefore gives us that the appropriately normalized plug-in estimator $S_{N,M}$ converges in distribution to a normal random variable. Also for the case where N grows slower than M. we have established Berry-Esseen type bounds on the error.

F Uniform kernel based plug-in estimator

In this section, we will state the main results concerning uniform kernel plug-in estimators. The proofs for these results rely on the propoerties of the uniform kernel density estimates established in Appendix A and proofs for equivalent results for the k-NN plug-in estimators. Let $\hat{\mathbf{f}}_u$ denote the boundary corrected uniform kernel density estimate. Denote the uniform kernel plug-in estimator by

$$\hat{\mathbf{G}}_{\mathbf{u}}(f) \;=\; \left(\frac{1}{N}\sum_{i=1}^{N} g(\hat{\mathbf{f}}_u(\mathbf{X}_i), \mathbf{X}_i)\right). \tag{95}$$

Let $\mathbf{Y}$ denote a random variable with density function f.

F.1 Results

Corollary F.1. *Suppose that the density f, the functional g and the density estimate $\hat{\mathbf{f}}_u$ satisfy the necessary conditions listed above. The bias of the plug-in estimator $\hat{\mathbf{G}}_u(f)$ is then given by*

$$B_u(f) \;=\; c_1\left(\frac{k}{M}\right)^{2/d} + c_2\left(\frac{1}{k}\right) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right),$$

where $c_1 = \mathbb{E}[g'(f(\mathbf{Y}), \mathbf{Y})c(\mathbf{Y})]$, $c_2 = \mathbb{E}[g''(f(\mathbf{Y}), \mathbf{Y})f(\mathbf{Y})/2]$ are constants which depend on the underlying density f.

Corollary F.2. *Suppose that the density f, the functional g and the density estimate $\hat{\mathbf{f}}_u$ satisfy the necessary conditions listed above. The variance of the plug-in estimator $\hat{\mathbf{G}}_u(f)$ is given by*

$$\mathbb{V}_u(f) \;=\; c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right),$$

where $c_4 = \mathbb{V}[g(f(\mathbf{Y}), \mathbf{Y})]$ and $c_5 = \mathbb{V}[f(\mathbf{Y})g'(f(\mathbf{Y}), \mathbf{Y})]$ are constants which depend on the underlying density f.

Corollary F.3. *Suppose that the density f, the functional g and the density estimate $\hat{\mathbf{f}}_u$ satisfy the necessary conditions listed above. Further suppose $\mathbb{E}[|g(f)|^3]$ is finite. The asymptotic distribution of the plug-in estimator $\hat{\mathbf{G}}_u(f)$ is given by*

$$\lim_{\Delta(k,N,M)\to 0} Pr\left(\frac{\hat{\mathbf{G}}_u(f) - \mathbb{E}[\hat{\mathbf{G}}_u(f)]}{\sqrt{\mathbb{V}[f(\mathbf{Y})g'(f(\mathbf{Y}), \mathbf{Y})]/N}} \leq \alpha\right) = Pr(\mathbf{Z} \leq \alpha),$$

where $\mathbf{Z}$ is a standard normal random variable.

References

[1] P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhya: The Indian Journal of Statistics*, 50:381–393, October 1988.

[2] P. Billingsley. *Probability and Measure (Third ed.)*. John Wiley and sons, 1995.

[3] L. Birge and P. Massart. Estimation of integral functions of a density. *The Annals of Statistics*, 23(1):11–29, 1995.

[4] J.R. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher. Central limit theorems for interchangeable processes. *Canadian Journal of Mathematics*, June 1957.

[5] Y. Chen, A. Wiesel, and A. O. Hero. Robust shrinkage estimation of high-dimensional covariance matrices. submitted to IEEE Trans. on Signal Process., preprint available in arXiv:1009.5331.

[6] C. I. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.

[7] 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*, pages 417–422, 2005.

[8] A. Jones D. Evans and W. M. Schmidt. Asymptotic moments of nearest neighbor distance distributions. *Proceedings of the Royal Society A*, 458:2839–2849, 2008.

[9] D. Evans. A law of large numbers for nearest neighbor statistics. *Proceedings of the Royal Society A*, 464:3175–3192, 2008.

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

[11] 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.

[12] A. T. Ihler, J. W. Fisher III, and A. S. Willsky. Nonparametric hypothesis tests for statistical dependency. *IEEE Transactions on Signal Processing*, 52(8):2234–2249, August 2004.

[13] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *In ACM SIGCOMM*, pages 217–228, 2005.

[14] N. Leonenko, L. Prozanto, and V. Savani. A class of rényi information estimators for multidimensional densities. *Annals of Statistics*, 36:2153–2182, 2008.

[15] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, Cambridge, MA, 2005.

[16] E. Liitiäinen, A. Lendasse, and F. Corona. A boundary corrected expansion of the moments of nearest neighbor distributions. *Random Structures and Algorithms*, 37:223–247, September 2010.

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

[18] 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.

[19] Y. P. Mack and M. Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1 – 15, 1979.

[20] D. S. Moore and J. W. Yackel. Consistency properties of nearest neighbor density function estimators. *The Annals of Statistics*, 1977.

[21] H. Neemuchwala and A. O. Hero. Image registration in high dimensional feature space. *Proc. of SPIE Conference on Electronic Imaging, San Jose*, January 2005.

[22] V. C. Raykar and R. Duraiswami. Fast optimal bandwidth selection for kernel density estimation. In J. Ghosh, D. Lambert, D. Skillicorn, and J. Srivastava, editors, *Proceedings of the sixth SIAM International Conference on Data Mining*, pages 524–528, 2006.

[23] H. Singh, N. Misra, and V. Hnizdo. Nearest neighbor estimators of entropy. *The Annals of Statistics*, 2005.

[24] K. Sricharan, R. Raich, and A. O. Hero. Global performance prediction for divergence-based image registration criteria. In *Proc. IEEE Workshop on Statistical Signal Processing*, 2009.

[25] B. van Es. Estimating functionals related to a density by class of statistics based on spacing. *Scandinavian Journal of Statistics*, 1992.

[26] P. Viola and W.M. Wells. Alignment by maximization of mutual information. *Proc. of 5th Int. Conf. on Computer Vision, MIT*, 1:16–23, 1995.