Neighborhood graphs for estimation of density functionals

by

Sricharan Kallur Palli Kumar

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2012

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Professor Jeffrey A. Fessler
Assistant Professor Raj Rao Nadakuditi
Assistant Professor Long Nguyen
Assistant Professor Clayton D. Scott
Assistant Professor Raviv Raich, Oregon State University

To Baba.

ACKNOWLEDGEMENTS

The successful completion of this thesis is in large part due to my advisor, Professor Alfred Hero, whom I would like to thank on several counts. First and foremost, this thesis would not have been possible but for his offer of assistantship. I am forever indebted to him for giving me this opportunity. As a research advisor, his vast knowledge, expertise in several areas, and his vision, were critical to the research work presented in this thesis. I am also grateful to him for his valuable time over the years, and for his reassuring support and encouragement. I thoroughly enjoyed working under him because of his hands-off and accommodating advising style, which gave me the time and freedom to work on research problems I found interesting.

I would also like to thank Professor Raviv Raich for his contribution towards the work in this thesis. He worked closely and patiently with me during my first two years as a research assistant. His involvement and input were invaluable to the research work done in this period.

I am grateful to Professor Jefferey Fessler for his keen interest in my research work over the years, in addition to serving on my qualifying and dissertation committees. I am also indebted to him for his prompt and kind support in my times of need. I also would like to thank the rest of my committee members, Professor Nguyen Long, Professor Clayton Scott and Professor Raj Rao Nadakuditi, for their encouragement, suggestions and thorough feedback.

I would like to extend my thanks to other faculty members and my peers who in some form have shaped the outcome of this thesis and/or contributed to my devel-

opment as a researcher. In particular, I would like to thank Sung Jin Hwang for his significant help in extending my work to data on manifolds, and to Dennis Wei for his input on convex optimization problems for ensemble estimation. I would also like to thank the EE:Systems staff - Becky Turanski, Beth Lawson, Michelle Feldkamp, Ann Pace and Karen Liska - for their help in navigating through red tape and otherwise.

My experience while working towards my PhD. has been significantly enriched by the many interactions I have had with my friends and peers in EE:Systems, including Arnau Puig, Kevin Xu, Joyce Liu, Gyemin Lee, and the rest of Prof. Hero's group members. I would like to especially thank Greg Newstadt for his friendship, and for the time we spent watching and playing sports.

To my friends in Ann Arbor, and in particular - Gowtham, Pradeep, Kaka, Tushar, Hari, Vijai, Raghu, Suhant, Avani, Gandharv and Tejas - thank you for the shared memories, food, entertainment and laughter. A special thanks to the people I have played cricket, racquetball, volleyball and tennis with and against.

I am deeply, deeply thankful to my mother and father, my brother, my grandparents, and the rest of my wonderful family, for their unconditional love and support. I derive a great deal of joy in being able to share this accomplishment with them.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

LIST OF NOTATIONS

k-NN k-nearest neighbor

$f(X)$ Probability density function

S Support of $f(X)$

$\mathcal{B}$ Boundary of support $\mathcal{S}$ of $f(X)$

d Dimension of support $\mathcal{S}$ of $f(X)$

$\mathcal{M}$ Manifold on which support $\mathcal{S}$ is embedded

$r_k(X)$ k-nearest neighbor radius from point centered at X

$S_k(X)$ k-nearest neighbor ball centered at X

$V_k(X)$ Volume of k-nearest neighbor ball centered at X

$\hat{\mathbf{f}}_k(X)$ Standard k-NN density estimator

$\tilde{\mathbf{f}}_k(X)$ Boundary corrected k-NN density estimator

$\hat{\mathbf{f}}_{k,\theta}(X)$ Angular k-NN density estimator

BP-kNNG Bipartite k-nearest neighbor graph

BP-kNN estimator Bipartite k-nearest neighbor estimator

$\hat{\mathbf{G}}_k(f)$ Standard bipartite k-NN functional estimator

$\hat{\mathbf{G}}_{k,BC}(f)$ Standard bipartite k-NN functional estimator with bias correction

$\tilde{\mathbf{G}}_k(f)$ Boundary corrected bipartite k-NN functional estimator

$\hat{\mathbf{G}}_{k,\theta}(f)$ Angular bipartite k-NN functional estimator

$\hat{\mathbf{d}}_s$ Dimension estimator

$\hat{\mathbf{G}}_w$ Weighted ensemble estimator

ABSTRACT

Functionals of densities play a fundamental role in statistics, signal processing, machine learning, information theory and related fields. This class of functionals includes entropy, divergence and mutual information measures of densities, intrinsic dimension of data embedded in manifolds, and minimum volume sets of densities. k-nearest neighbor (k-NN) graph based estimators are widely used for the estimation of these functionals. While several consistent k-NN estimators have been previously proposed for estimating these functionals, general results on rates of convergence of these estimators and confidence intervals on the estimated functional are not available. Since the rate of convergence relates the number of samples to the performance of the estimator, convergence rates have great practical utility.

In this thesis, a new class of estimators based on bipartite k-nearest neighbor graphs is proposed for estimating functionals of probability density functions. This class includes entropy and divergence estimators, intrinsic dimension estimators and estimates of p-values for testing membership of data in minimum volume sets. For this class of estimators, large sample theory is used to characterize performance of the estimators. Specifically, large sample expressions for estimator bias and variance is derived and a central limit theorem for the distribution of the estimators is established. This theory is applied to accurately estimate functionals of interest by optimizing the mean squared error over free parameters, e.g. the number of neighbors k, and obtaining confidence intervals on the estimated functional by invoking the central limit theorem. Furthermore, this theory provides significant insight into the statistical

behavior of these bipartite k-NN estimators, leading to the development of modified k-NN estimators with faster rates of convergence. In particular, a weighted ensemble of bipartite k-NN estimators for functional estimation is proposed, and it is shown using this theory, that the weighted ensemble estimator outperforms the state-of-the-art. This theory can therefore be used to accurately estimate functionals of densities with confidence in a wide variety of applications.

Using this theory, the thesis develops performance-driven algorithms in several applications. First, the theory is applied to determine entropy with confidence to facilitate anomaly detection at desired false alarm rates in wireless sensor networks. Second, the theory is applied to determine complexity of high-dimensional data lying on a manifold, and subsequently applied to fusion and segmentation applications. Finally, the thesis introduces an efficient anomaly detection algorithm based on estimation of p-values of membership in training-sample minimum volume sets using bipartite k-NN graphs.

CHAPTER I

Introduction

1.1 Background

Functionals of probability distributions play a fundamental role in statistics, signal processing, machine learning, information theory and related fields. This class of functionals includes entropy, divergence and mutual information measures of densities [2, 22], intrinsic dimension of data embedded in manifolds [29], and minimum volume (MV) sets of densities [76, 74].

This thesis introduces a new class of estimators based on bipartite k-nearest neighbor (k-NN) graphs for estimating general functionals of probability densities. These tools come with asymptotic expressions for estimator bias and variance that can be used to predict and improve estimator performance. The tools also come with a central limit theorem that can be used to develop confidence intervals and p-values. This thesis illustrates the use of these tools for several applications including: entropy estimation; dimension estimation; and minimum volume set estimation.

1.1.1 Divergence estimation

Information divergence is the distance between probability distributions of different random variables. There are many applications of divergence functional estimation including the following. Divergence based methods for image matching,

image registration and texture classification are developed in [37, 58]. Entropy has been used in Internet anomaly detection [45], data and image compression applications [40], communications [19] and quantization theory [33]. Information flow measures are used in bio-informatics [68] and brain-machine interfaces [65]. Several divergence based nonparametric statistical tests have been developed for testing statistical models including uniformity and normality [85, 24]. Parameter estimation methods based on divergence have been developed in [67].

1.1.1.1 Error exponents

Divergence functionals arise naturally as they specify detection and classification error exponents in asymptotic large sample hypothesis testing problems. For instance, the Kullback-Leibler divergence specifies the exponential rate of decay of error probability for the optimal Neyman-Pearson likelihood test of simple binary hypotheses. Similarly, the Chernoff distance appears as the corresponding error exponent for the optimal Bayesian likelihood ratio test. The relation between divergence functionals and error exponents has motivated divergence maximization approaches to circumvent the intractable problem of direct minimization of the probability of error [21, 43, 5].

1.1.2 Dimension estimation

The intrinsic dimension of sample data roughly characterizes the number of variables needed to explain the phenomenon originating the data and therefore characterizes the complexity of the true underlying probability distribution generating the observed sample data. Intrinsic dimension estimation has been utilized significantly for the purpose of inferring an appropriate projection or embedding dimension in dimension reduction algorithms [29]. Dimension estimation has also been used in tasks where dimension reduction is not the final goal. For instance, intrinsic dimension has been used to analyze the local complexity of signals to detect anomalies in

router networks [14]. Dimension estimation has also been used for image and texture segmentation [16, 63].

1.1.3 MV set estimation

A minimum volume (MV) set of a probability density is a region of minimum size among the regions covering a given probability mass of the density. MV sets provide useful summaries of multi-dimensional functions for many applications including clustering [35, 82], anomaly detection [74, 81, 86], functional neuroimaging [64], bioinformatics [88] and digital elevation mapping [78].

1.2 Previous work

An inherent problem in the aforementioned applications is that we do not have access to the true underlying probability distributions, but rather, have access only to realizations of random variables drawn from the distributions. It is therefore crucial to estimate these functional measures to a high degree of accuracy from the realizations and to quantify estimation error.

1.2.1 Entropy and divergence estimation

The problem of entropy and divergence estimation of densities f_1, f_2 of the form $\int g(f_1(x)/f_2(x), x)f_2(x)dx$ from sample realizations has received significant attention in the mathematical statistics community. These include consistent estimators based on entropic graphs [38, 58, 60], gap estimators [84], nearest neighbor distances [32, 49, 51, 87], Edgeworth approximations [39], convex risk minimization [59] and kernel density estimates [25].

Bickel and Ritov [9] treat the problem for the specific case of $\int f^2(x)dx$ and show convergence at the parametric rate $n^{-1/2}$. This result was generalized to $\int g(f(x), x)dx$

for arbitrary g for univariate densities by Eggermont *et al.* [25] and for certain restricted classes of densities by Birge and Massart [10] and Laurent [47].

Several entropy and divergence estimators based on sums of functionals of k-NN distances have been proposed in the literature. The authors of [77, 32, 51] only deal with estimators corresponding to Shannon entropy ($g(u) = \log(u)$) and Rényi entropy($g(u) = u^{\alpha-1}$). Evans *et al.* [27] on the other hand analyze only positive moments of the k-NN distances ($g(u) = u^k, k \in \mathbb{N}$). Recently, Baryshnikov *et al.* [6] and Wang *et al.* [87] developed k-NN based estimators of divergence $\int g(f_1(x)/f_2(x), x)f_2(x)dx$ when $f_1(.)$ is known and when $f_1(.)$ has to be estimated from samples respectively .

The authors of [77, 32, 27, 87] show that the estimators they propose are asymptotically unbiased and consistent. Finally, CLT for k-NN estimators of Rényi entropy was alluded to by Leonenko *et al.* [32] by inferring from experimental results. The authors of [51] analyze the bias for cases of Shannon and Rényi entropy. For arbitrary smooth functionals $g(.)$, Evans *et al.* [26] show that the variance of the sums of these functionals of k-NN distances is bounded by the rate $O(k^5/T)$. Recently, Baryshnikov *et al.* [6] improved on the results of Evans *et al.* by determining the exact variance up to the leading term for entropy estimation, and divergence estimation when $f_1(.)$ is assumed to be known. Furthermore, they show that the entropy and divergence estimators they propose converge weakly to a normal distribution. Chatterjee [15] proved CLT results for general functionals of k-NN graphs for fixed values of k.

In this thesis, in addition to proving consistency of bipartite k-NN estimators for entropy, divergence and mutual information, we have successfully developed a large sample theory for the MSE and central limit results on the asymptotic distribution of these estimators.

1.2.2 Dimension estimation

A common approach to the problem of intrinsic dimensionality estimation is based on projecting data on subspaces of different dimensions and choosing the intrinsic dimension to be the dimension of the subspace that provides the best fit [41]. An example of this approach includes principal component analysis (PCA) which is applied to estimate dimension of linear subspaces. As these methods do not account for non-linearities, linear methods tend to overestimate intrinsic dimension.

More sophisticated methods are based on the idea of observing the rate of growth of the number of points falling in a fixed ball as a function of the ball radius or the rate of growth of the size of k-nearest neighbor balls for varying values of k. This rate of growth has the property that it grows exponentially in the number of data samples with exponent inversely proportional to the intrinsic dimension. Examples of such estimators include Costa and Hero's k-nearest neighbor (k-NN) graph dimension estimator [20], Levina and Bickel's [50] maximum likelihood estimator and Farahmand *et al.*'s dimension estimator based on nearest neighbor distances [28]. In this thesis we formulate the problem of intrinsic dimension as a divergence functional estimation problem and derive provably better estimators.

1.2.3 MV set estimation

Estimation of minimum volume sets is a difficult problem, especially for high dimensional data. There are two types of approaches to this problem: (1) transform the MV estimation problem to an equivalent density level set estimation problem, which requires estimation of the nominal density; and (2) directly identify the minimal set using function approximation and non-parametric estimation [75, 61, 73]. Both types of approaches involve explicit approximation of high dimensional quantities - the multivariate density function in the first case and the boundary of the minimum volume set in the second - and are therefore not easily applied to high dimensional

problems.

The GEM principle developed by Hero [36], and also used by Zhao *et al.* [89], for determining MV sets circumvents the above difficulties by using the asymptotic theory of random Euclidean graphs instead of function approximation. However, the GEM based K-kNNG minimum volume detection scheme proposed in [36] and the K-LPE scheme proposed in [89] become computationally difficult as the number of data points becomes large. To address this issue, a surrogate L1O-kNNG anomaly detection scheme was proposed in [36]. L1O-kNNG is computationally simpler than K-kNNG, but loses some desirable properties of the K-kNNG, including asymptotic consistency, as shown in this thesis. We introduce a new estimator for level sets that is consistent yet has low computational complexity.

1.2.4 k-NN estimators

Several of the estimators discussed above fall under the class of k-NN graph based estimators. These include entropy and divergence estimators such as entropic graph estimators [38, 58], spacing estimators [84] and k-nearest neighbor based estimators [60, 32, 49, 6], dimension estimators including Costa's [20] and Levina's [50] estimators, and level set estimators proposed by Hero [36] and Zhao *et al.* [89].

The general idea behind these k-NN estimators is that (k-NN) graphs are non-parametric structures that convey local geometry of points in a sample. This attribute has resulted in a wide variety of machine learning applications for k-NN graphs, for e.g., density estimation, manifold learning and non-parametric classification and entropy estimation. The local nature of k-NN graph estimators enables development of k-NN graph estimators for data lying on embedded manifolds.

However, general results on rates of convergence of k-NN graph based estimators are unavailable. This is due to the highly dependent nature of k-NN edges in a data sample, which makes analysis of MSE and asymptotic distribution of k-NN estimators

difficult. This thesis provides results on MSE and asymptotic distributions of k-NN graph estimators by studying the statistical behavior of k-NN neighborhoods in a data sample and exploiting the interchangeable, albeit dependent, nature of k-NN edge lengths.

1.3 Contribution of thesis

In this thesis, we propose a wide class of non-parametric estimators based on bipartite k-nearest neighbor (k-NN) graphs for estimating functionals of densities. The basic construction of the proposed bipartite plug-in estimator is as follows. Given a total of T data samples we split the data into two parts of size N and size M, $N + M = T$. On the part of size M a k-NN density estimate is constructed. The density functional is then estimated by plugging the k-NN density estimate into the functional and approximating the integral by an empirical average over the remaining N samples. This can be thought of as computing the estimator over a bipartite graph with the M density estimation nodes connected to the N integral approximating nodes. This is illustrated in Fig. 1.1.

Our proposed class of estimators include estimators of entropy, divergence and mutual information measures of density functions on Euclidean space (Chapter 2) and on manifolds (Chapter 4), intrinsic dimension estimator of data embedded in manifolds (Chapter 4) and estimator of p-values for testing memberships in level sets (Chapter 5).

For this class of estimators, we derive a large sample theory to characterize performance of the estimators. Specifically, we derive large sample expressions for the bias and variance for this class of estimators and develop a central limit theorem for the distribution of the estimators. Our analysis of MSE of these estimators studies the statistical behavior of k-NN neighborhoods in a data sample. This analysis is not available previously in literature and helped solve a long standing problem of

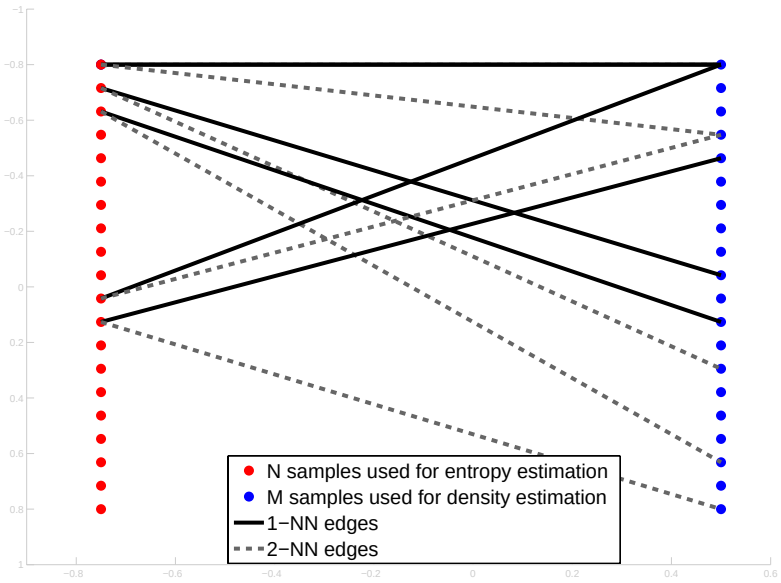


Figure 1.1: Illustration of a 2-NN bipartite graph.

characterizing correlation between k-NN neighborhoods centered at different points. Since the rate of convergence relates the number of samples to the performance of the estimator, convergence rates have great practical utility. To develop the central limit theorem, we prove a general result showing that the distribution of suitably normalized sums of interchangeable processes converges weakly to the standard normal distribution. We subsequently establish that our class of estimators falls under the framework of sums of interchangeable processes, thereby establishing the central limit theorem.

Our analysis led to the development of modified bipartite k-NN estimators with faster rates of convergence. In particular, boundary compensated k-NN estimators, that compensate for bias due to truncation of k-NN neighborhoods near the boundary of the density support, are proposed in Chapter 3 and in Section 6.5. Furthermore, under higher order smoothness assumptions on the density, we have proposed weighted ensembles of bipartite k-NN estimators in Chapter 6, and show that these estimators

have parametric $O(1/n)$ MSE rates of convergence and outperform state-of-the art k-NN estimators.

1.3.1 Entropy, divergence and mutual information estimation

General results on rates of convergence and asymptotic distribution of k-NN estimators of entropy, divergence and mutual information were previously unavailable, except for Shannon and Rényi entropy estimators [32, 49], in which case rates for the bias were provided in [51] and variance and asymptotic distribution were provided in [6].

In Chapter 2, we have successfully developed a large sample theory for the MSE of bipartite k-NN estimators for entropy, divergence and mutual information. The results developed above can be used to improve the rate of convergence of these estimators. For instance, we tune the parameters of the estimator (for example, the number of neighbors k) to achieve optimal mean square error performance in Chapter 2. In the specific case of Shannon and Rényi entropy estimation, in comparison to the estimators proposed by [32, 49, 6], we are able to exploit the bipartite nature of our estimators to improve the rate of convergence (see Section 2.7 and Section 3.5.4 for further details).

We have also developed a general central limit theorem showing the asymptotic distribution of bipartite k-NN estimators to be standard normal. The central limit result can be used to help characterize the error in the estimate via statistical confidence intervals. Furthermore, we have applied this result to do anomaly detection in router networks at desired false alarm rates. Our CLT, in contrast to the result by Chatterjee [15], applies to the case where k grows to ∞ with T, which is a necessary condition for consistency of the proposed bipartite k-NN estimators.

In Chapter 3, we show that for densities with finite support, k-NN density estimators behave differently in the interior of the support as opposed to near the

boundary of the support. Our analysis in Chapter 2 lead to insight about k-NN behavior near the boundary of support of density. Specifically, our analysis showed that k-NN neighborhoods are bloated close to the boundary because of the discontinuity of the density function at the boundary of the support. This results in increased bias in k-NN density estimates near the boundary and in turn increased bias in entropy, divergence and mutual information estimators proposed in Chapter 2. To compensate for this bias, we propose boundary compensated compensated k-NN graphs in Chapter 3 and in Section 6.5, by suitably reducing the size of k-NN neighborhoods near the boundary of the support. We show that the use of boundary compensated graphs reduces the bias from $O(T^{-1/(1+d)})$ to $O(T^{-2/(2+d)})$ without correction factors and from $O(T^{-1/d})$ to $O(T^{-2/d})$ with correction factors.

In Chapter 4, we extend results on entropy and divergence estimation to the case where the support of the density is on a smooth Riemannian manifold embedded in $\mathbb{R}^D$ with intrinsic dimension d.

1.3.2 Dimension estimation

In Chapter 4, 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. The rate of growth function of the k-NN ball size can be directly related to the entropy functional of the underlying data and therefore directly fits in our framework directly.

In this respect it is similar to Costa and Hero's k-nearest neighbor (k-NN) graph dimension estimator [20], Levina and Bickel's [50] maximum likelihood estimator and to Farahmand *et al.*'s dimension estimator based on nearest neighbor distances [28]. The estimator can also be related to the Leonenko *et al.*'s Rényi entropy estimator [49].

However, unlike these estimators, our new dimension estimator is derived directly from a mean squared error (MSE) optimality condition for partitioned k-NN esti-

mators of entropy of multivariate densities on manifolds. This guarantees that our estimator has the best possible MSE 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.

Intrinsic dimension can be used as an indication of the complexity associated with data and can be a useful statistic for discriminating different types of data. We have used this characterization of dimension to detect anomalies in internet network data. We have also used this statistic to analyze geographic formations from hyper-spectral images by computing the spectral complexity of different regions in the formations. This characterization was used to subsequently segment the image (into different regions - for e.g., vegetation, water body, urban area).

1.3.3 MV set testing

In Chapter 5, we propose a novel bipartite k-nearest neighbor graph (BP-kNNG) estimator to estimate p-values for testing membership in minimum volume sets. Our bipartite estimator retains all the desirable theoretical properties of the K-kNNG, while being computationally simpler than the K-kNNG and the surrogate L1O-kNNG detectors. We show that BP-kNNG is asymptotically consistent in recovering the p-value of each test point.

We use the proposed p-value estimator to detect anomalies in data sets. Given a set of normal events, the anomaly detection problem aims to identify unknown, anomalous events that deviate from the normal set. Novelty detection is crucial to various applications where failure to detect anomalous activity could lead to catastrophic outcomes. These include, for instance, detection of faults in mission-critical systems, quality control in manufacturing and medical diagnosis. Learning minimum volume sets of an underlying nominal distribution is a very effective approach to anomaly detection [75, 81, 86]. Experimental results are given that illustrate the

superior performance of BP-kNNG as compared to the L1O-kNNG and other state of the art anomaly detection schemes.

1.3.4 Ensemble Estimation

For d-dimensional data, it is shown that the variance of the k-NN graph estimators decay as $O(T^{-1})$, where T is the sample size, while the bias, because of the curse of dimensionality, decays as $O(T^{-1/(1+d)})$. The squared bias $O(T^{-2/(1+d)})$ therefore dominates the mean square error (MSE) in high dimensions.

To address this large bias in high dimensions, we propose a weighted k-NN estimator in Chapter 6, where the weights serve to lower the bias to $O(T^{-1/2})$, which then ensures convergence of the weighted estimator at the parametric rate of $O(T^{-1/2})$. These weights are determined by solving a convex optimization problem. The proposed weighted ensemble estimators are applied to various problems including density estimation, Shannon and Rényi entropy estimation, intrinsic dimension estimation, and minimum volume set estimation. It is shown by simulation that the weighted ensemble estimators uniformly outperform other, including state-of-the-art, k-NN estimators.

1.4 List of relevant publications

The following publications were produced based on research presented in this thesis:

1. K. Sricharan and A. O. Hero, "*Angular k-NN ensemble estimators*," submitted to the Transactions on Information Theory, March 2012.

2. K. Sricharan and A. O. Hero, "*Ensemble estimators for multivariate entropy estimation*," submitted to the Transactions on Information Theory, March 2012.

3. K. Sricharan, R. Raich and A. O. Hero, "*Statistical estimation of entropy with confidence*," IEEE Transactions on Information Theory, April 2012.

4. K. Sricharan and A. O. Hero, "*Efficient anomaly detection using bipartite k-NN graphs*," In Proc. Advances in Neural Information Processing Systems (NIPS), November 2011.

5. K. Sricharan, R. Raich and A. O. Hero, "*Performance-driven entropic information fusion*," Workshop on Defense Applications of Signal Processing (DASP), July 2011.

6. K. Sricharan, R. Raich and A. O. Hero, "*k-nearest neighbor estimation of entropies with confidence*," IEEE International Symposium on Information Theory (ISIT), April 2011.

7. K. Sricharan and A. O. Hero, "*Weighted k-NN graphs for Rényi entropy estimation in high dimensions*," IEEE Workshop on Statistical Signal Processing (SSP), March 2011.

8. K. Sricharan, R. Raich and A. O. Hero, "*Estimation of non-linear functionals of densities with confidence*," Technical Report, Communications and Signal Processing Laboratory (CSPL), The University of Michigan, December 2010.

9. K. Sricharan, R. Raich and A. O. Hero, "*Boundary compensated k-NN graphs*," IEEE Workshop on Machine Learning in Signal Processing (MLSP), August 2010.

10. K. Sricharan, R. Raich and A. O. Hero, "*Optimized intrinsic dimension estimation using nearest neighbor graphs*," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). September 2009.

11. K. Sricharan, R. Raich and A. O. Hero, "*Global performance prediction for divergence-based image registration*," IEEE Workshop on Statistical Signal Processing (SSP). September 2009.

CHAPTER II

k-NN plug-in estimators of entropy and divergence

2.1 Introduction

Non-linear functionals of the densities f_1, f_2 of the form $\int g(f_1(x)/f_2(x), x)f_2(x)dx$ arise in applications of machine learning, signal processing and statistical estimation. Important examples of such functionals include Shannon and Rényi entropy, Shannon mutual information and other forms of f-divergences. Entropy and divergence based applications for image matching, image registration and texture classification are developed in [37, 58]. Entropy functional estimation is fundamental to independent component analysis in signal processing [56]. Entropy has also been used in Internet anomaly detection [45] and data and image compression applications [40]. Several entropy based nonparametric statistical tests have been developed for testing statistical models including uniformity and normality [85, 24]. Parameter estimation methods based on entropy have been developed in [18, 67].

In these applications, the functional of interest must be estimated empirically from sample realizations of the underlying densities. This problem has received significant attention in the mathematical statistics community. Several estimators of divergence measures have been proposed for general multivariate densities f. These include consistent estimators based on entropic graphs [38, 60], gap estimators [84], nearest neighbor distances [32, 49, 51, 87], kernel density plug-in estimators [1, 25, 9, 34, 10],

Edgeworth approximations [39], orthogonal projections [47] and convex risk minimization [59].

Several of the estimators discussed above fall under the class of k-NN graph based estimators. These include entropy and divergence estimators such as entropic graph estimators [38, 58], spacing estimators [84] and k-nearest neighbor based estimators [60, 32, 49, 6]. However, general results on rates of convergence of these k-NN estimators are unavailable. Since the rate of convergence relates the number of samples to the performance of the estimator, convergence rates have great practical utility. In this paper we derive convergence rates for a class of bipartite k-NN estimators of non-linear functionals.

The authors of [77, 32, 51] only deal with estimators corresponding to Shannon entropy ($g(u) = \log(u)$) and Rényi entropy($g(u) = u^{\alpha-1}$). Evans *et al.* [27] on the other hand analyze only positive moments of the k-NN distances ($g(u) = u^k, k \in \mathbb{N}$). Wang *et al.* and Baryshnikov *et al.* [6] propose estimators based on k-NN distances for estimating the f-divergence between densities. The authors of [77, 32, 27, 87] show that the estimators they propose are asymptotically unbiased and consistent. Finally, CLT for k-NN estimators of Rényi entropy was alluded to by Leonenko *et al.* [32] by inferring from experimental results. The authors of [51] analyze the bias for cases of Shannon and Rényi entropy. For arbitrary smooth functionals $g(.)$, Evans *et al.* [26] show that the variance of the sums of these functionals of k-NN distances is bounded by the rate $O(k^5/T)$. Recently, Baryshnikov *et al.* [6] improved on the results of Evans *et al.* by determining the exact variance up to the leading term (c_k/T for some constant c_k which is a function of k) for entropy estimation. Furthermore, they show that the entropy estimator they propose converges weakly to a normal distribution. However, Baryshnikov *etal* do not analyze the bias of the estimators, nor do they show that the estimators they propose are consistent. Using the results obtained in this paper, we provide an expression for this bias in Section 2.7 and show that the

optimal MSE for Baryshnikov's estimators is $O(T^{-2/(1+d)})$. Chatterjee [15] proved CLT results for general functionals of k-NN graphs for fixed k.

In contrast, the main contribution of this chapter is the analysis of a general class of bipartite k-NN estimators of smooth density functionals. We exploit a close relation between density estimation and the geometry of proximity neighborhoods in the data sample to establish asymptotic statistical analysis of the bias and variance. We then show that the bipartite k-NN estimator is MSE consistent and that the MSE is guaranteed to converge to zero as $T \to \infty$ and $k \to \infty$ with a rate that is minimized for a specific choice of k, M and N as a function of T. Therefore, the thus optimized bipartite k-NN estimator can be implemented without any tuning parameters. In addition a CLT is established that can be used to construct confidence intervals to empirically assess the quality of the bipartite k-NN estimator. Finally, our method of proof is very general and it is likely that it can be extended to kernel density plug-in estimators, f-divergence estimation and mutual information estimation.

An important distinction between the bipartite k-NN estimator and the k-NN estimators of Shannon and Rényi entropy proposed by the authors of [77, 32, 49] is that these latter estimators are consistent for finite k, while the proposed bipartite k-NN estimator require the condition that $k \to \infty$ for MSE convergence. We show in Chapter 3 and Chapter 6 that by exploiting the condition $k \to \infty$, bipartite k-NN estimators with superior rates of convergence can be derived. Also note that our CLT, in contrast to the result by Chatterjee [15], applies to the case where k grows to ∞ with T, which is a necessary condition for consistency of the proposed bipartite k-NN estimators.

Organization

The reminder of the chapter is organized as follows. Section 2.3 formulates the entropy estimation problem and introduces the data-split plug-in estimator. The main

results concerning the bias, variance and asymptotic distribution of these estimators are stated in Section 2.3.1 and the consequences of these results are discussed. The proofs for these results are given in the Appendix. We extend our entropy estimator to estimate divergence in Section 2.4 and mutual information in Section 2.5. We validate our theory using simulation studies in Section 3.5. We discuss our results in Section 2.9.

In Chapter 3, we will illustrate the advantage of using data-split estimators over unsplit estimators for estimating entropy by deriving boundary-corrected k-NN density estimates which account for high bias in density estimates near the boundary of the support of the density. We will show that our data-split boundary corrected k-NN plug-in estimators have lower bias and MSE and are also computationally faster to implement as compared to standard sums of sums of functionals of k-NN distances. We take further advantage of this in Chapter 6 by deriving ensemble estimators based on data-split plug-in estimators which have a parametric rate of convergence.

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 conditional expectation given $\mathbf{Z}$ using the notation $\mathbb{E}_{\mathbf{Z}}$. We also define the variance operator as $\mathbb{V}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2]$ and the covariance operator as $Cov[\mathbf{X}, \mathbf{Y}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])]$. We denote the bias of an estimator by $\mathbb{B}$.

2.2 k-NN density estimate

Let $d(X, Y)$ denote the Euclidean distance between points X and Y and $\mathbf{d}_k(X)$ denote the Euclidean distance between a point X and its k-th nearest neighbor amongst M i.i.d realizations from a density d-dimensional f, $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$. The k-NN region is $\mathbf{S}_k(X) = \{Y : d(X, Y) \leq \mathbf{d}_k(X)\}$ 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 [53] is defined as

$$\hat{\mathbf{f}}_k(X) = \frac{k-1}{M\mathbf{V}_k(X)}.$$

Define the coverage function as $\mathbf{P}(X) = \int_{\mathbf{S}_k(X)} f(Z)dZ$.

Observe that the k-NN density estimate at X will be singular iff the k-nearest neighbors of X in the set $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$ are identically equal to X. This event will occur with probability 0 for any continuous density f. Equivalently stated, when the density f is continuous, the density estimate $\hat{\mathbf{f}}_k(X)$ is almost surely non-singular for any X. While the k-NN density estimate does not integrate to 1 over the support $\mathcal{S}$ of the density, the integral $\mathbb{E}[\int_{X\in\mathcal{S}} \hat{\mathbf{f}}_k(X)dX]$ asymptotically *does* evaluate to 1 as $k \to 0$ and $k/M \to 0$ [30]. The k-NN density estimate evaluated in the interval $[-3, 3]$ using 1000 sample realizations drawn from a 1-dimensional density f uniform in the interval $[-0.5, 0.5]$ is shown in Fig. 2.1.

Analysis of moments of k-NN density estimates is a crucial ingredient in our development of large sample theory for estimators of functionals of densities. This analysis can be found in Appendix B.

2.3 Data-split plug-in estimators of entropy

We are interested in estimating non-linear functionals $G(f)$ of d-dimensional multivariate densities f with support $\mathcal{S}$, where $G(f)$ has the form

$$G(f) = \int 1_{\{x\in\mathcal{S}'\}} g(f(x), x) f(x) d\mu(x) = \mathbb{E}[1_{\{x\in\mathcal{S}'\}} g(f(x), x)],$$

for some smooth function $g(f(x), x)$ and some subset $\mathcal{S}' \subset \mathcal{S}$ of the support $\mathcal{S}$. Let $\mathcal{B}$ denote the boundary of $\mathcal{S}$. Here, μ denotes the Lebesgue measure and $\mathbb{E}$ denotes statistical expectation w.r.t density f. We assume that i.i.d realizations

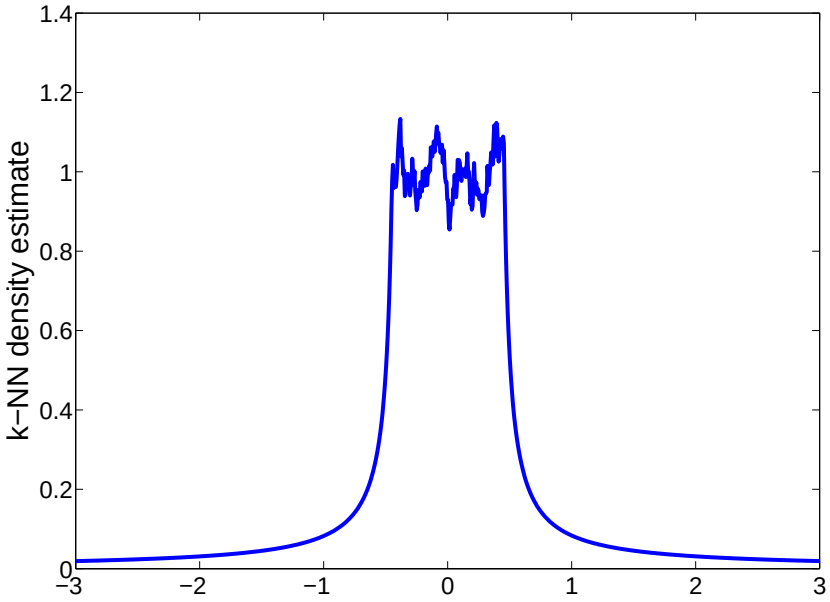


Figure 2.1: k-NN density estimate evaluated in the interval $[-3, 3]$ using 1000 sample realizations drawn from a 1-dimensional density f uniform in the interval $[-0.5, 0.5]$.

$\{\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ are available from the density f.

The plug-in estimator is constructed using a data splitting approach as follows. The data is randomly subdivided into two disjoint parts $\mathcal{X}_N = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\mathcal{X}_M = \{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ of N and M points respectively. In the first stage, we estimate the k-NN density estimator $\hat{\mathbf{f}}_k$ at the N points $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ using the M realizations $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$. Subsequently, we use the N samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ to approximate the functional $G(f)$ to obtain the plug-in estimator:

$$\hat{\mathbf{G}}_k(f) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i). \tag{2.1}$$

We note that the k-NN density estimates at $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ can be determined by constructing a bipartite k-nearest neighbor graph from the set of N samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ to the set of M samples $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ with the k-the nearest

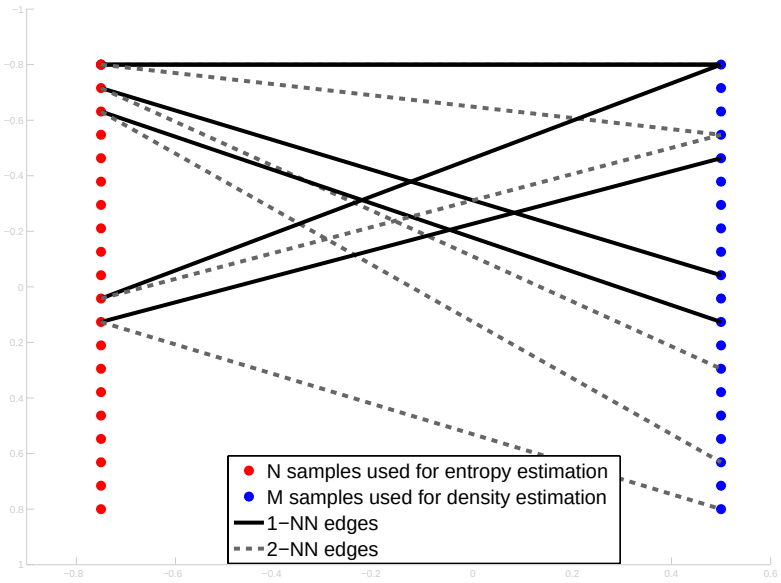


Figure 2.2: Illustration of a 2-NN bipartite graph.

neighbor edge from each $\mathbf{X}_i \in \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ linking the two sets. This is illustrated in Fig. 2.2 and Fig. 2.3.

2.3.1 Assumptions

Let δ be a fixed number in $(2/3, 1)$. For some fixed $0 < \epsilon < 1$, define $p_l = ((k-1)/M)(1-\epsilon)\epsilon_0$ and $p_u = ((k-1)/M)(1+\epsilon)\epsilon_\infty$. Also define $\epsilon_1 = 1/(c_d \mathcal{D}^d)$, where $\mathcal{D}$ is the diameter of the bounded set $\mathcal{S}$ and define $q_l = ((k-1)/M)\epsilon_1$ and $q_u = (1+\epsilon)\epsilon_\infty$. Let $\mathbf{p}$ be a beta random variable with parameters $k, M-k+1$. Let $\mathbf{Y}$, $\mathbf{Z}$ denote i.i.d. random variables with density f and define $c(X) = \Gamma^{(2/d)}((d+2)/2) f^{-2/d}(X) tr[\nabla^2(f(X))]$.

We assume that k grows polynomially in M, i.e. $k = M^\alpha$ for $\alpha \in (0, 1)$. We require that the density f be uniformly bounded away from 0 and finite on the set $\mathcal{S}'$, i.e., there exist constants $\epsilon_0, \epsilon_\infty$ 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

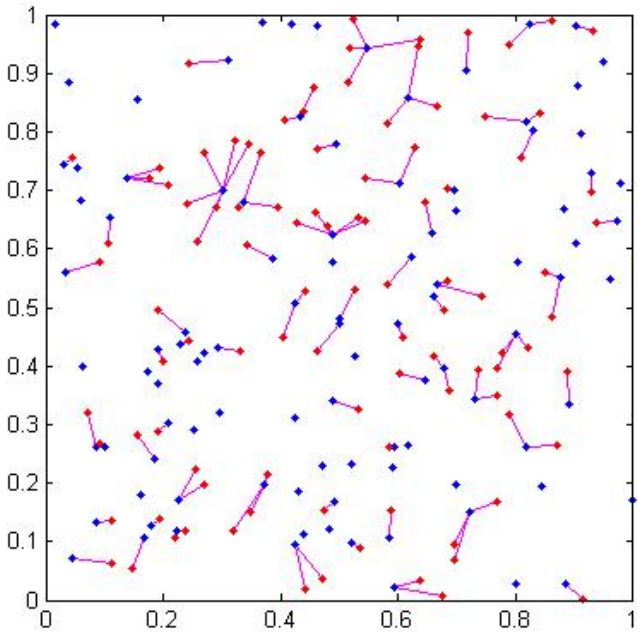


Figure 2.3: 1-NN bipartite graph constructed on 2-d data (red = N samples used for entropy estimation, blue = M samples used for entropy estimation, purple = 1-NN edges).

$2r$ in the interior of the set $\mathcal{S}'$ 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 $\alpha\lambda > 1$. Finally 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. Assume that $\sup_{x\in(q_l,q_u)} |(g^{(r)}/r!)^2(x,y)|e^{-3k^{(1-\delta)}} < \infty$, $\mathbb{E}[\sup_{x\in(p_l,p_u)} |(g^{(r)}/r!)^2(x/\mathbf{p},y)|] < \infty$, for $r = 3, \lambda$.

Define the set $\mathcal{S}_\mathcal{I}$ as follows. For any $X \in \mathcal{S}$, let $r(X)$ be the shortest distance from X to $\mathcal{B}$. For any set $\mathcal{R}$, define the statistic $m(\mathcal{R}) = \int_{x\in\mathcal{R}} r(X)dx$. Let $\mathcal{S}_\mathcal{I} = \mathcal{S}_\mathcal{I}(f)$ then be the set with minimum measure $m(.)$, with probability mass at least $\epsilon_\infty^2 k/M$ wrt the density f. We show in Appendix B that for any $X \in \mathcal{S}_\mathcal{I}$, the probability that the k-NN ball centered at X intersects with the boundary $\mathcal{B}$ is exponentially small. We note that if $\mathcal{S}' \cap \mathcal{B} = \phi$, then for sufficiently small values of k/M, $\mathcal{S}' \cap \mathcal{S}_\mathcal{I} = \mathcal{S}'$.

2.3.2 Bias and Variance

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

$$\begin{aligned}\mathbb{B}(\hat{\mathbf{G}}_k(f)) &= c_0\left(\frac{k}{M}\right)^{1/d} + c_1\left(\frac{k}{M}\right)^{2/d} + c_2\left(\frac{1}{k}\right) \\ &\quad + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{1/d}\right),\end{aligned}$$

where c_0, c_1 *and* c_2 *are constants which depend on the density* f *and the set* $\mathcal{S}'$. *The constant* c_2 *is given by* $c_2 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f^2(\mathbf{Y}) g''(f(\mathbf{Y}), \mathbf{Y})/2]$. *Furthermore,* $c_0 = 0$ *and* $c_1 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} g'(f(\mathbf{Y}), \mathbf{Y}) c(\mathbf{Y})]$ *if and only if* $\mathcal{S}' \subset \mathcal{S}_{\mathcal{I}}$.

Proof. 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 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. We show in appendix B that the k-NN density estimate satisfies assumptions $\mathscr{A}.1$ and $\mathscr{A}.2$ listed in Appendix D, which in turn implies that lemma D.1 holds. This gives:

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{G}}_k(f)] - G(f) &= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\ &\quad + c_2\left(\frac{1}{k}\right) + o(1/k).\end{aligned}$$

Using the properties of standard k-NN density estimates (C.2), we can then show

$$\begin{aligned}&\mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\ &= c_0\left(\frac{k}{M}\right)^{1/d} + c_1\left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{1/d}\right).\end{aligned}$$

This concludes the proof.

□

The source of the leading term $c_0(k/M)^{1/d}$ is due to the fact that if a probability density function has bounded support, the k-NN balls centered at points close to the boundary are often truncated at the the boundary, resulting in estimator bias. For details, please refer Appendix C.

We note that if $c_0 \neq 0$, the bias is minimized by choosing $k = O(M^{1/(1+d)})$, giving the optimal rate of the bias to be $O(M^{-1/(1+d)})$. In Chapter 3, we discuss boundary compensation methods to force $c_0 = 0$ and discuss the optimal choice of k for the case $c_0 = 0$ in detail in section 3.4.1.

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

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

where $c_4 = \mathbb{V}[1_{\{\mathbf{Y}\in\mathcal{S}'\}}g(f(\mathbf{Y}), \mathbf{Y})]$ *and* $c_5 = \mathbb{V}[1_{\{\mathbf{Y}\in\mathcal{S}'\}}f(\mathbf{Y})g'(f(\mathbf{Y}), \mathbf{Y})]$.

Proof. We once again use Taylor series expansion and bound higher order terms. We shown in Appendix B that the k-NN density estimate satisfies assumptions $\mathscr{A}.1$, $\mathscr{A}.2$ and $\mathscr{A}.3$. This in turn implies that lemma D.2 holds. This concludes the proof. □

While Theorem II.1 and Theorem II.2 hold for any choice of positive integers k, M, N, for asymptotic consistency we require that $k \to \infty$, $k/M \to 0$ and $N \to \infty$. This holds for all expressions for bias and variance that are derived in the rest of the thesis. For the optimal choice of $k = O(M^{1/(1+d)})$ to minimize bias, the MSE is given by $O(M^{-2/(1+d)} + 1/M + 1/N)$. The optimal choice of partition N, M to minimize MSE is given by $N_{opt} = O(T^{(3+d)/(2(1+d))})$.

In the higher order terms in Theorem II.1 and Theorem II.2 denoted by $o(.)$, the corresponding constants in front are functionals of derivatives of the density f and the functional g, which by the assumptions in Section 2.3.1 are finite. This observation is true for all expressions for bias and variance that are derived in the rest of the thesis.

2.3.3 Central limit theorem

In addition to the results on bias and variance shown in the previous section, we show that our plug-in estimator, appropriately normalized, weakly converges to the normal distribution. We study the asymptotic behavior of the plug-in estimates under the following limiting conditions: (a) $k/M \to 0$, (b) $k \to \infty$, and (c) $N \to \infty$. As shorthand, we will collectively denote the above limiting assumptions by $\Delta \to 0$.

Theorem II.3. *The asymptotic distribution of the plug-in estimator* $\hat{\mathbf{G}}_k(f)$ *is given by*

$$\lim_{\Delta \to 0} Pr\left(\frac{\hat{\mathbf{G}}_k(f) - \mathbb{E}[\hat{\mathbf{G}}_k(f)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_k(f)]}} \leq \alpha \right) = Pr(\mathbf{S} \leq \alpha),$$

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

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

$$\mathbf{Y}_{M,i} = \frac{1_{\{\mathbf{X}_i \in \mathcal{S}'\}}(g(\hat{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i) - \mathbb{E}[1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)])}{\sqrt{\mathbb{V}[1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)]}},$$

and define the sum $\mathbf{S}_{\mathbf{N},\mathbf{M}}$

$$\mathbf{S}_{N,M} = \frac{\sum_{i=1}^{N} \mathbf{Y}_{M,i}}{\sqrt{\mathbb{V}[\sum_{i=1}^{N} \mathbf{Y}_{M,i}]}},$$

where the indices N and M explicitly stress the dependence of the sum $\mathbf{S}_{N,M}$ on the number of random variables $N + M$. Observe that the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ belong to an 0 mean, unit variance, interchangeable process [11] for all values of M.

We will first show that $Cov(\mathbf{Y}_{M,1}, \mathbf{Y}_{M,2})$ and $Cov(\mathbf{Y}^2_{M,1}, \mathbf{Y}^2_{M,2})$ are $O(1/M)$. Subsequently we will show that the random variable $\mathbf{S}_{N,M}$ converges in distribution to

$N(0,1)$. Let $\mathbf{X}$ be a random variable with density f. Define the function $d(x,y) = g(x,y)(g(x,y)-c)$, where the constant $c = \mathbb{E}[g(\hat{\mathbf{f}}(\mathbf{X}),\mathbf{X})]$.

From (C.4), we have

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

Define $d(x,y) = g(x,y)(g(x,y)-c)$, where the constant $c = \mathbb{E}[g(\hat{\mathbf{f}}_k(\mathbf{X}_1),\mathbf{X}_1)]$. Then,

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

As M gets large, we then have that $Cov(\mathbf{Y}_{M,i},\mathbf{Y}_{M,j}) \to 0$ and $Cov(\mathbf{Y}^2_{M,i},\mathbf{Y}^2_{M,j}) \to 0$. By lemma E.1, we have that $\mathbf{S}_{N,M}$ converges in distribution to $N(0,1)$ as both N and M get large. Finally, note that establishing CLT for

$$\frac{\hat{\mathbf{G}}_k(f) - \mathbb{E}[\hat{\mathbf{G}}_k(f)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_k(f)]}}$$

is equivalent to establishing CLT for the function $\mathbf{S}_{N,M}$. This concludes the proof.

□

The central limit theorem for k-NN estimators of Rényi entropy was alluded to by Leonenko *et al.* [32] by inferring from experimental results. Theorem II.3 establishes the CLT for k-NN estimators of arbitrary functionals, including Rényi entropy.

2.4 Estimation of f-divergences

In this section, we are concerned with the estimation of f-divergences of the form $G(f_1, f_2) = \int 1_{\{x \in \mathcal{S}'\}} g_2(f_1(x)/f_2(x), x) f_2(x) dx$ for smooth functionals g_2 and densities $f_1(x)$, $f_2(x)$ which are bounded away from 0 and ∞ on the set $\mathcal{S}'$. The choice of $g(x) = -\log(x)$ defies Kullback-Leibler information, $g(x) = 2(1 - \sqrt{x})^2$ yields the square of the Hellinger distance, $g(x) = x \log(x)$ yields the log-likelihood ratio statistic or I-divergence of Kullback-Leibler, $g(x) = (x-1)^2/2$ yields the chi-squared divergence,and $g(x) = x^r$ yields Rényi information gain of order $r, r > 0$.

We consider two cases. In the first case, we assume that f_1 is a known density. Then using our estimator $\hat{\mathbf{G}}(f_2)$, it is possible to compute the f-divergence $G(f_1, f_2)$ by defining the corresponding functional g to be $g(x, y) = g_2(f_1(x)/x)$. In this case, theorems II.1, II.2 and II.3 on the bias, variance and CLT will hold provided the necessary conditions on the density f_2 and the functional $g(x, y) = g_2(f_1(x)/x)$ listed in Section 3 are satisfied.

In the second case, we assume that f_1 is also unknown and that $\mathbf{Y}_1, \ldots \mathbf{Y}_{M_1}$ i.i.d realizations are available from the density f_1 and $\mathbf{X}_1, \ldots \mathbf{X}_{N+M_2}$ i.i.d realizations are available from the density f_2. We construct the following estimator:

$$\hat{\mathbf{G}}_k(f_1, f_2) \;=\; \left(\frac{1}{N} \sum_{i=1}^{N} g_2(\hat{\mathbf{f}}_{1k}(\mathbf{X}_i)/\hat{\mathbf{f}}_{2k}(\mathbf{X}_i), \mathbf{X}_i) \right), \tag{2.2}$$

where $\hat{\mathbf{f}}_{1k}$ and $\hat{\mathbf{f}}_{2k}$ are k-NN density estimates constructed from the M_1 samples $\mathbf{Y}_1, \ldots \mathbf{Y}_{M_1}$ and the M_2 samples $\mathbf{X}_{N+1}, \ldots \mathbf{X}_{N+M_2}$ respectively.

2.4.1 Assumptions

Denote the ratio $f_1(x)/f_2(x)$ by $f(x)$. We again assume that k grows polynomially in M, i.e. $k = M^\alpha$ for $\alpha \in (0, 1)$. We assume that the densities f_1, f_2 satisfy (i) that the ratio f be uniformly bounded away from 0 and finite on the set $\mathcal{S}'$

of f_2, i.e., there exist constants ϵ_0, ϵ_∞ such that $0 < \epsilon_0 < \epsilon_\infty < \infty$ such that $\epsilon_0 \leq f(x) \leq \epsilon_\infty \ \forall x \in \mathcal{S}$. (ii) have 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 $g_2(x,y)$ has λ partial derivatives w.r.t. x, where λ satisfies the condition $\alpha\lambda > 1$. Finally we assume that the absolute value of the functional $g_2(x,y)$ 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_2 and define $c_i(X) = \Gamma^{(2/d)}((d+2)/2) f_i^{-2/d}(X) tr[\nabla^2(f_i(X))]$, $i = 1, 2$. Finally, assume that $\sup_{x\in(q_l,q_u)} |(g_2^{(r)}/r!)^2(x,y)| e^{-3k^{(1-\delta)}} < \infty$, $\mathbb{E}[\sup_{x\in(p_l,p_u)} |(g_2^{(r)}/r!)^2(x/\mathbf{p},y)|] < \infty$, for $r = 3, \lambda$.

Define $\mathcal{S}_\mathcal{I} = \mathcal{S}_\mathcal{I}(f_1) \cap \mathcal{S}_\mathcal{I}(f_2)$. We note that if $\mathcal{S}' \cap \mathcal{B} = \phi$, then for sufficiently small values of k/M, $\mathcal{S}' \subset \mathcal{S}_\mathcal{I}$.

2.4.2 Bias and Variance

Theorem II.4. *The bias of the plug-in estimator $\hat{\mathbf{G}}_k(f_1, f_2)$ is given by*

$$\begin{aligned}\mathbb{B}(\hat{\mathbf{G}}_k(f_1,f_2)) &= c_1\left(\frac{k}{M}\right)^{2/d} + c_2\left(\frac{1}{k}\right) \\ &\quad + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right),\end{aligned}$$

where c_0, c_1 and c_2 are constants which depend on the densities f_1, f_2, the functional g and the set $\mathcal{S}'$. The constant c_2 is given by $c_2 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f^2(\mathbf{Y}) g_2''(f(\mathbf{Y}),\mathbf{Y})/2]$. Furthermore, $c_0 = 0$ and $c_1 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f(\mathbf{Y}) g_2'(f(\mathbf{Y}),\mathbf{Y})(c_1(\mathbf{Y})/f_1(\mathbf{Y}) - c_2(\mathbf{Y})/f_2(\mathbf{Y}))]$ if and only if $\mathcal{S}' \subset \mathcal{S}_\mathcal{I}$.

Proof. Define $\hat{\mathbf{f}}(\mathbf{X}_i) = \hat{\mathbf{f}}_{1k}(\mathbf{X}_i)/\hat{\mathbf{f}}_{2k}(\mathbf{X}_i)$. The k-NN density estimators $\hat{f}_1(.)$ and $\hat{f}_2(.)$ satisfy assumptions $\mathscr{A}.1$ and $\mathscr{A}.2$ (see Appendix C.1). This implies that lemma D.3 holds. This gives,

$$\mathbb{E}[\hat{\mathbf{G}}(f_1,f_2)] - G(f_1,f_2) = \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g_2(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}) - g_2(f(\mathbf{Z}),\mathbf{Z}))] + o(1/k)$$
$$+\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\left(g_2'(f(\mathbf{Z}),\mathbf{Z})f(\mathbf{Z}) + g_2''(f(\mathbf{Z}),\mathbf{Z})f(\mathbf{Z})^2\right)\right]\left(\frac{1}{k}\right).$$

We note that

$$g_2(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}) - g_2(f(\mathbf{Z}),\mathbf{Z}) = g_2'(f(\mathbf{Z}),\mathbf{Z})(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})] - f(\mathbf{Z}))(1+o(1))$$
$$= g_2'(f(\mathbf{Z}),\mathbf{Z})(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{1k}(\mathbf{Z})]/\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{2k}(\mathbf{Z})] - f_1(\mathbf{Z})/f_2(\mathbf{Z}))(1+o(1))$$
$$= g_2'(f(\mathbf{Z}),\mathbf{Z})(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{1k}(\mathbf{Z})]/\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{2k}(\mathbf{Z})] - f_1(\mathbf{Z})/f_2(\mathbf{Z}))(1+o(1))$$
$$= g_2'(f(\mathbf{Z}),\mathbf{Z})f(\mathbf{Z})(h_1(\mathbf{Z})/f_1\mathbf{Z} - h_2(\mathbf{Z})/f_2\mathbf{Z})\left(\frac{k}{M}\right)^{2/d}(1+o(1))$$

where the last but one step follows from (C.2). It follows that

$$\mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}) - g(f(\mathbf{Z}),\mathbf{Z}))]$$
$$= c_0\left(\frac{k}{M}\right)^{1/d} + c_1\left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{1/d}\right).$$

This concludes the proof. □

Theorem II.5. *The variance of the plug-in estimator* $\hat{\mathbf{G}}(f_1,f_2)$ *is given by*

$$\mathbb{V}(\hat{\mathbf{G}}_k(f_1,f_2)) \;=\; c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M}+\frac{1}{N}\right),$$

where $c_4 = \mathbb{V}[1_{\{\mathbf{Y}\in\mathcal{S}'\}}g_2(f(\mathbf{Y}),\mathbf{Y})]$ *and* $c_5 = \mathbb{V}[1_{\{\mathbf{Y}\in\mathcal{S}'\}}f(\mathbf{Y})g_2'(f(\mathbf{Y}),\mathbf{Y})]$.

Proof. The k-NN density estimators $\hat{f}_1(.)$ and $\hat{f}_2(.)$ satisfy assumptions $\mathscr{A}.1$, $\mathscr{A}.2$ and $\mathscr{A}.3$ (see Appendix C.1). This implies that lemma D.4 holds. This concludes the proof. □

2.4.3 Central limit theorem

Theorem II.6. *The asymptotic distribution of the plug-in estimator* $\hat{\mathbf{G}}(f_1, f_2)$ *is given by*

$$\lim_{\Delta \to 0} Pr\left(\frac{\hat{\mathbf{G}}_k(f_1, f_2) - \mathbb{E}[\hat{\mathbf{G}}_k(f_1, f_2)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_k(f_1, f_2)]}} \leq \alpha \right) = Pr(\mathbf{S} \leq \alpha),$$

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

Proof. Define $\hat{\mathbf{f}}(\mathbf{X}_i) = \hat{\mathbf{f}}_{1k}(\mathbf{X}_i)/\hat{\mathbf{f}}_{2k}(\mathbf{X}_i)$. Define the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ for any fixed M

$$\mathbf{Y}_{M,i} = \frac{1_{\{\mathbf{X}_i \in \mathcal{S}'\}}(g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) - \mathbb{E}[1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)])}{\sqrt{\mathbb{V}[1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)]}},$$

Using the fact that $\hat{\mathbf{f}}_{1k}(x)$ and $\hat{\mathbf{f}}_{2k}(\mathbf{X}_i)$ are conditionally independent given x, in conjunction with (C.4), it is straightforward to show that $Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) \to 0$ and $Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}) \to 0$. Applying Lemma E.1 concludes the proof. □

2.5 Estimation of f-MI

In this section, we are concerned with the estimation of f-Mutual Information measures of the form $G(f_{12}) = \int 1_{\{x \in \mathcal{S}'\}} g(f_1(x) f_2(y)/f_{12}(x, y), (x, y)) f_{12}(x, y) dx dy$ for smooth functionals g and joint density $f_{12}(x)$ bounded away from 0 and ∞ on the set $\mathcal{S}'$, where $f_1(.)$ and $f_2(.)$ are the marginal densities of f_{12} along x and y. (Note: x and y can be of arbitrary dimension, but have to be disjoint partitions). Denote the $N + M$ i.i.d. realizations from f_{12} by $\{\mathbf{Z}_1, .., \mathbf{Z}_{N+M}\}$ and their marginal components corresponding to f_{12} by $\{\mathbf{Z}_1, .., \mathbf{Z}_{N+M}\}$ and f_{12} by $\{\mathbf{Z}_1, .., \mathbf{Z}_{N+M}\}$ respectively.

We consider the following estimator:

$$\hat{\mathbf{G}}_k(f_{12}) \;=\; \left(\frac{1}{N}\sum_{i=1}^{N} g(\hat{\mathbf{f}}_{1k}(\mathbf{X}_i)\hat{\mathbf{f}}_{2k}(\mathbf{Y}_i)/\hat{\mathbf{f}_{12k}}(\mathbf{Z}_i), \mathbf{Z}_i)\right), \tag{2.3}$$

where $\hat{\mathbf{f}_{12k}}$, (respectively $\hat{\mathbf{f}}_{1k}$ and $\hat{\mathbf{f}}_{2k}$) is the k-NN density estimates constructed from the M samples $\mathbf{Z}_1, \ldots \mathbf{Z}_M$ ($\mathbf{X}_1, \ldots \mathbf{X}_M$ and $\mathbf{Y}_1, \ldots \mathbf{Y}_M$).

For convenience, we will henceforth adopt the following convention where for a point $z = (x, y)$, we will use $f_1(z)$ to denote $f_1(x)$ (respectively $f_2(z)$ to denote $f_2(y)$). Likewise, for a random variable $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, let $\hat{\mathbf{f}}_1(\mathbf{Z})$ denote $\hat{\mathbf{f}}_1(\mathbf{X})$ (respectively $\hat{\mathbf{f}}_2(\mathbf{Z})$ denote $\hat{\mathbf{f}}_2(\mathbf{Y})$).

2.5.1 Assumptions

Denote the ratio $f_1(x)f_2(y)/f_{12}(x, y) = f_1(z)f_2(z)/f_{12}(z)$ by $f(z)$. We again assume that k grows polynomially in M, i.e. $k = M^\alpha$ for $\alpha \in (0, 1)$. We assume that the densities f_1, f_2 satisfy (i) that the ratio f be uniformly bounded away from 0 and finite on the set $\mathcal{S}'$ of f_2, i.e., there exist constants ϵ_0, ϵ_∞ such that $0 < \epsilon_0 < \epsilon_\infty < \infty$ such that $\epsilon_0 \leq f(x) \leq \epsilon_\infty \; \forall x \in \mathcal{S}$. (ii) have 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 $g_2(x, y)$ has λ partial derivatives w.r.t. x, where λ satisfies the condition $\alpha\lambda > 1$. Finally we assume that the absolute value of the functional $g_2(x, y)$ 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_2 and define $c_i(X) = \Gamma^{(2/d)}((d+2)/2) f_i^{-2/d}(X) tr[\nabla^2(f_i(X))]$, $i = 1, 2$. Assume that $\sup_{x\in(q_l,q_u)} |(g^{(r)}/r!)^2(x, y)| e^{-3k^{(1-\delta)}} < \infty$, $\mathbb{E}[\sup_{x\in(p_l,p_u)} |(g^{(r)}/r!)^2(x/\mathbf{p}, y)|] < \infty$, for $r = 3, \lambda$.

Define $\mathcal{S}_\mathcal{I} = \mathcal{S}_\mathcal{I}(f_{12}) \cap \mathcal{S}_\mathcal{I}(f_1) \cap \mathcal{S}_\mathcal{I}(f_2)$. We note that if $\mathcal{S}' \cap \mathcal{B} = \phi$, then for sufficiently small values of k/M, $\mathcal{S}' \subset \mathcal{S}_\mathcal{I}$.

2.5.2 Bias and Variance

Theorem II.7. *The bias of the plug-in estimator* $\hat{\mathbf{G}}_k(f_{12})$ *is given by*

$$\begin{aligned}\mathbb{B}(\hat{\mathbf{G}}_k(f_{12})) &= c_1\left(\frac{k}{M}\right)^{2/d} + c_2\left(\frac{1}{k}\right) \\ &\quad + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right),\end{aligned}$$

where c_0, c_1 *and* c_2 *are constants which depend on the densities* f_1, f_2, *the functional* g *and the set* $\mathcal{S}'$. *The constant* c_2 *is given by* $c_2 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f^2(\mathbf{Y}) g_2''(f(\mathbf{Y}), \mathbf{Y})/2]$. *Furthermore,* $c_0 = 0$ *and* $c_1 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f(\mathbf{Y}) g_2'(f(\mathbf{Y}), \mathbf{Y})(c_1(\mathbf{Y})/f_1(\mathbf{Y}) - c_2(\mathbf{Y})/f_2(\mathbf{Y}))]$ *if and only if* $\mathcal{S}' \subset \mathcal{S}_{\mathcal{I}}$.

Proof. The k-NN density estimators $\hat{f}_{1k}(.)$, $\hat{f}_{2k}(.)$ and $\hat{f}_{12k}(.)$ satisfy assumptions $\mathscr{A}.1$, $\mathscr{A}.2$, $\mathscr{A}.3$ and $\mathscr{A}.4(a)$ (refer section C.1 and section C.2). This implies that lemma D.5 holds. This concludes the proof. □

Theorem II.8. *The variance of the plug-in estimator* $\hat{\mathbf{G}}(f_{12})$ *is given by*

$$\mathbb{V}(\hat{\mathbf{G}}_k(f_{12})) = c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right),$$

where $c_4 = \mathbb{V}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} g(f(\mathbf{Y}), \mathbf{Y})]$ *and* $c_5 = \mathbb{V}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f(\mathbf{Y}) g'(f(\mathbf{Y}), \mathbf{Y})]$.

Proof. The k-NN density estimators $\hat{f}_{1k}(.)$, $\hat{f}_{2k}(.)$ and $\hat{f}_{12k}(.)$ satisfy assumptions $\mathscr{A}.1$, $\mathscr{A}.2$, $\mathscr{A}.3$ and $\mathscr{A}.4(b)$ (refer section C.1 and section C.2). This implies that lemma D.6 holds. This concludes the proof. □

2.5.3 Central limit theorem

Theorem II.9. *The asymptotic distribution of the plug-in estimator* $\hat{\mathbf{G}}(f_1, f_2)$ *is given by*

$$\lim_{\Delta \to 0} Pr\left(\frac{\hat{\mathbf{G}}_k(f_1, f_2) - \mathbb{E}[\hat{\mathbf{G}}_k(f_1, f_2)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_k(f_{12})]}} \leq \alpha \right) = Pr(\mathbf{Z} \leq \alpha),$$

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

Proof. Define $\hat{\mathbf{f}}(\mathbf{X}_i) = \hat{\mathbf{f}}_{1k}(\mathbf{X}_i)\hat{\mathbf{f}}_{2k}(\mathbf{X}_i)/\hat{\mathbf{f}}_{12k}(\mathbf{X}_i)$. Next, define the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ for any fixed M as

$$\mathbf{Y}_{M,i} = \frac{1_{\{\mathbf{X}_i \in \mathcal{S}'\}}(g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) - \mathbb{E}[1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)])}{\sqrt{\mathbb{V}[1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i)]}},$$

From (B.47), it follows that the covariance terms $Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) \to 0$ and $Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}) \to 0$. Applying Lemma E.1 concludes the proof. □

2.6 Bias correction factors

In this section, we restrict our attention to estimation of Shannon and Rényi - α entropy, divergence and Mutual Information (MI). Previously, Goria *et al.* [49], Leonenko *et al.* [32], Wang *et al.* [87] and Pocsoz *et al.* [12] have developed consistent estimators of these quantities. In this chapter, we provide rates for these estimators and establish weak convergence of the same. An important distinction in the case of Shannon and Rényi entropy and divergence estimation is that it suffices for k to grow polynomially in $\log(M)$ (necessary condition for the result on variance to go through) rather than grow polynomially in M. These results are formally stated below.

2.6.1 Main results

For a general function $g(x, y)$, if there exist functions $g_1(k, M)$ and $g_2(k, M)$, such that

$$
\begin{aligned}
&(i) \quad \mathbb{E}[g((k-1)x/M\mathbf{p}, y)] = g(x, y)g_1(k, M) + g_2(k, M) + o(1/M), \\
&(ii) \quad ((k-1)/M)\mathbb{E}[g'((k-1)x/M\mathbf{p}, y)\mathbf{p}^{2/d-1}] = g'(x, y)(k/M)^{2/d} + o((k/M)^{2/d}), \\
&(iii) \quad \lim_{k\to\infty} g_1(k, M) = 1, \\
&(iv) \quad \lim_{k\to\infty} g_2(k, M) = 0,
\end{aligned} \tag{2.4}
$$

then define the BP-kNN plug-in estimator with bias correction as

$$
\hat{\mathbf{G}}_{k,BC}(f) \quad = \quad \frac{\hat{\mathbf{G}}_k(f) - g_2(k, M)}{g_1(k, M)}. \tag{2.5}
$$

2.6.1.1 Bias and Variance

In addition to the assumptions listed in section 2.3.1, assume the growth condition that $k = \Theta((\log(M))^{2/(1-\delta)})$ instead of the condition that $k = \Theta(M^\beta)$. Below the asymptotic bias and variance of the plug-in estimator with bias correction are specified.

Theorem II.10. *The bias of the BPI estimator $\hat{\mathbf{G}}_{k,BC}(f)$ is given by*

$$
\mathbb{B}[\hat{\mathbf{G}}_{k,BC}(f)] \quad = \quad c_0 \left(\frac{k}{M}\right)^{1/d} + o\left(\left(\frac{k}{M}\right)^{1/d}\right). \tag{2.6}
$$

Proof.

$$
\begin{aligned}
&\mathbb{B}(\hat{\mathbf{G}}_{k,BC}(f) = \mathbb{E}[\hat{\mathbf{G}}_{k,BC}(f)] - \int g(f(x),x)f(x)dx \\
&= (\mathbb{E}[g(\hat{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z})] - g_2(k,M))/g_1(k,M) - \int g(f(x),x)f(x)dx \\
&= \mathbb{E}[\mathbb{E}[(g(\hat{\mathbf{f}}_k(\mathbf{Z}),\mathbf{X}) - g_2(k,M))/g_1(k,M) \mid \mathcal{X}_N]] - \int g(f(x),x)f(x)dx \\
&= \mathbb{E}[\mathbb{E}[(g(\hat{\mathbf{f}}_k(\mathbf{X}),\mathbf{X}) - g_2(k,M))/g_1(k,M) \mid \mathcal{X}_N], X \in \mathcal{S}'] \\
&+\mathbb{E}[\mathbb{E}[(g(\hat{\mathbf{f}}_k(\mathbf{X}),\mathbf{X}) - g_2(k,M))/g_1(k,M) \mid \mathcal{X}_N], X \notin \mathcal{S}'] \\
&- \int g(f(x),x)f(x)dx \\
&= I + II. \qquad (2.7)
\end{aligned}
$$

Now, by (B.34),

$$
\begin{aligned}
I &= \mathbb{E}[g(f(\mathbf{X}),\mathbf{X}) + \frac{g'(f(\mathbf{X}),\mathbf{X})h(\mathbf{X})}{g_1(k,M)}(k/M)^{2/d} \\
&= \frac{c_1}{g_1(k,M)}\left(\frac{k}{M}\right)^{2/d} + \frac{c_3}{g_1(k,M)} + o\left(\left(\frac{k}{M}\right)^{2/d}\right).
\end{aligned}
$$

Also, by (2.4), $g_1(k,M) = 1 + o(1)$. This implies that

$$
I = c_1\left(\frac{k}{M}\right)^{2/d} + c_3 + o\left(\left(\frac{k}{M}\right)^{2/d}\right). \qquad (2.8)
$$

On the other hand, for $Z \in \mathcal{S} - \mathcal{S}'$, we have $\mathbb{E}[g(\hat{\mathbf{f}}_k(Z),Z) - g(f(Z),Z)] = O(1)$. This implies that,

$$
\begin{aligned}
II &= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}-\mathcal{S}'\}}g(\hat{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z}) - g(f(\mathbf{Z}),\mathbf{Z})] \\
&= \mathbb{E}\left[\mathbb{E}[g(\hat{\mathbf{f}}_k(Z),Z) - g(f(Z),Z)] \mid 1_{\{\mathbf{Z}\in\mathcal{S}-\mathcal{S}'\}}\right] \times Pr(\mathbf{Z} \notin \mathcal{S}') \\
&= O(1) \times O((k/M)^{1/d}) = O((k/M)^{1/d}). \qquad (2.9)
\end{aligned}
$$

This concludes the proof. □

Theorem II.11. *The variance of the plug-in estimator* $\hat{\mathbf{G}}_{k,BC}(f)$ *is given by*

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

Proof. Under the logarithmic growth condition $k = O((\log(M))^{2/(1-\delta)})$, $g_2(k,M) = o(1)$ and $g_1(k,M) = 1+o(1)$ by assumption (2.4). Theorem II.11 follows by observing that $\hat{\mathbf{G}}_{k,BC}(f) = (\hat{\mathbf{G}}_k(f) - g_1(k,M))/g_2(k,M)$ and invoking Theorem II.2 . □

2.6.1.2 CLT

Theorem II.12. *The asymptotic distribution of the plug-in estimator* $\hat{\mathbf{G}}_{k,BC}(f)$ *is given by*

$$\lim_{\Delta\to 0} Pr\left(\frac{\hat{\mathbf{G}}_{k,BC}(f) - \mathbb{E}[\hat{\mathbf{G}}_{k,BC}(f)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_{k,BC}(f)]}} \leq \alpha\right) = Pr(\mathbf{S}\leq\alpha),$$

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

Proof. Under the logarithmic growth condition $k = O((\log(M))^{2/(1-\delta)})$, $g_2(k,M) = o(1)$ and $g_1(k,M) = 1+o(1)$ by assumption (2.4). Theorem II.12 follows by observing that $\hat{\mathbf{G}}_{k,BC}(f) = (\hat{\mathbf{G}}_k(f) - g_1(k,M))/g_2(k,M)$ and invoking Theorem II.3. □

2.6.1.3 MSE

Theorem IV. 1 specifies the bias of the plug-in estimator $\hat{\mathbf{G}}_{k,BC}(f)$ as $\Theta((k/M)^{2/d})$. Theorem IV. 2 specifies the variance as $\Theta(1/N + 1/M)$. By making k increase logarithmically in M, specifically, $k = O((\log(M))^{2/(1-\delta)})$ for any value $\delta \in (2/3, 1)$, the MSE is given by the rate $\Theta(((\log(T))^{2/(1-\delta)}/T)^{4/d})$. This estimator therefore has a faster rate of convergence in comparison to both Baryshnikov*et al.*'s estimators $\hat{\mathbf{H}}_S$ and $\hat{\mathbf{I}}_{\alpha,S}$ (MSE $= \Theta(T^{-2/(1+d)})$) and Leonenko*et al.*'s and Goria *et al.*'s estimators $\tilde{\mathbf{H}}_S$

and $\tilde{\mathbf{I}}_{\alpha,S}$ (MSE $= \Theta(T^{-2/d})$). Experimental MSE comparison of Leonenko's estimator against this estimator in Section V shows the MSE of this estimator to be significantly lower. Finally, note that such bias correction cannot be applied for general entropy functionals, and the bias correction factors cannot in general be incorporated. In the next section, the application of bias correction factors for estimation of Shannon and Rényi entropies is illustrated.

2.6.2 Shannon and Rényi entropy estimation

For the case of Shannon entropy ($g(u) = -\log(u)$), it can be verified that $g_1(k, M) = 1$, $g_2(k, M) = \psi(k) - \log(k-1)$ satisfy (2.4). Similarly, for the case of Rényi entropy ($g(u) = u^{\alpha-1}$), $g_1(k, M) = (\Gamma(k)/\Gamma(k+1-\alpha))(1/(k-1)^{\alpha-1})$, $g_2(k, M) = 0$ satisfy (2.4).

Let $\hat{\mathbf{H}}_k$ be the Shannon entropy estimate $\hat{G}_k(f)$ with the choice of functional $g(x) = -\log(x)$. Let $\hat{\mathbf{I}}_k^{(\alpha)}$ be the estimate of the Rényi α-integral estimate $\hat{G}_k(f)$ with the choice of functional $g(x) = x^{\alpha-1}$. Define $\check{\mathbf{H}}_k = \hat{\mathbf{H}}_k + [\log(k-1) - \Psi(k-1)]$ and $\check{\mathbf{I}}_k^{(\alpha)} = [(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}]^{-1}\hat{\mathbf{I}}_k^{(\alpha)}$. Also define the Rényi entropy estimator to be $\check{\mathbf{H}}_k^{(\alpha)} = (1-\alpha)^{-1}\log(\check{\mathbf{I}}_k^{(\alpha)})$. We note that the estimators $\check{\mathbf{H}}_k$ and $\check{\mathbf{H}}_k^{(\alpha)}$ correspond to data-split versions of the Shannon and Rényi entropy estimators of Goria *et al.* [32] and Leonenko *et al.* [49] respectively.

Because $[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}] \to 1$ and $\Psi(k-1) - \log(k-1) \to 0$ as $k \to \infty$, the estimators $\check{\mathbf{H}}_k$ and $\check{\mathbf{H}}_k^{(\alpha)}$ will have identical variance up to leading terms as $\hat{\mathbf{H}}_k$ and $\hat{\mathbf{H}}_k^{(\alpha)}$ respectively. Likewise, $\check{\mathbf{H}}_k$ and $\check{\mathbf{H}}_k^{(\alpha)}$, when suitably normalized, will converge to the same distribution as the estimators $\hat{\mathbf{H}}_k$ and $\hat{\mathbf{H}}_k^{(\alpha)}$ respectively.

From Theorem 2.1, it immediately follows that the bias of the estimators $\hat{\mathbf{H}}_k$ and $\hat{\mathbf{H}}_k^{(\alpha)}$is $O((k/M)^{1/d} + (1/k))$. On the other hand, from Theorem II.10, it is clear that

the bias of the estimators $\check{\mathbf{H}}_k$ and $\check{\mathbf{H}}_k^{(\alpha)}$ is given by

$$\mathbb{B}(\check{\mathbf{H}}_k) = c_0 \left(\frac{k}{M}\right)^{1/d} + c_1 \left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{1/d}\right),$$

and

$$\mathbb{B}(\check{\mathbf{H}}_k^{(\alpha)}) = c_0 \left(\frac{k}{M}\right)^{1/d} + c_1 \left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{1/d}\right),$$

In this case, the optimal choice of k needs to only grow logarithmically with M (necessary condition for the result on variance to go through). The minimum bias for optimal choice of k therefore reduces from $O(M^{-1/(1+d)})$ to $O(M^{-1/d})$. The optimal choice of N in this case is given by $N_{opt} = \Theta(M(\log(M)/M)^{(1/2-1/d)})$.

2.6.3 Estimation of K-L and Rényi divergence

We note that K-L and Rényi divergence are special cases of f-divergence with the choice of functionals $g_2(x) = -\log(x)$ and $g_2(x) = x^{\alpha-1}$ respectively. Let $\hat{\mathbf{D}}_k$ be the K-L divergence estimate $\hat{G}_k(f_1, f_2)$ with the choice of functional $g_2(x) = -\log(x)$. Similarly, let $\hat{\mathbf{I}}_{\mathbf{D}k}^{(\alpha)}$ be the estimate of the Rényi α divergence integral estimate $\hat{G}_k(f_1, f_2)$ with the choice of functional $g_2(x) = x^{\alpha-1}$.

As in the case of Shannon and Rényi entropy estimation, it is possible to define corrected estimators which reduce bias from $O(M^{-2/(2+d)})$ to $O(M^{-2/d})$. We describe these corrections next. Define $\check{\mathbf{D}}_k = \hat{\mathbf{D}}_k + [\log(k-1) - \Psi(k-1)]$ and $\check{\mathbf{I}}_{\mathbf{D}k}^{(\alpha)} = [(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}]^{-1}\hat{\mathbf{I}}_{\mathbf{D}k}^{(\alpha)}$. Also define the Rényi divergence estimator to be $\check{\mathbf{D}}_k^{(\alpha)} = (1-\alpha)^{-1}\log(\check{\mathbf{I}}_{\mathbf{D}k}^{(\alpha)})$. We note that the estimators $\check{\mathbf{D}}_k$ and $\check{\mathbf{D}}_k^{(\alpha)}$ correspond to data-split versions of the Shannon and Rényi divergence estimators of Wang *et al.* [87] and Poczos *et al.* [12] respectively.

Because $[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}] \to 1$ and $\Psi(k-1) - \log(k-1) \to 0$ as $k \to \infty$, the estimators $\check{\mathbf{D}}_k$ and $\check{\mathbf{D}}_k^{(\alpha)}$ will have identical variance up to leading terms

as $\hat{\mathbf{D}}_k$ and $\hat{\mathbf{D}}_k^{(\alpha)}$ respectively. Likewise, $\check{\mathbf{D}}_k$ and $\check{\mathbf{D}}_k^{(\alpha)}$, when suitably normalized, will converge to the same distribution as the estimators $\hat{\mathbf{D}}_k$ and $\hat{\mathbf{D}}_k^{(\alpha)}$ respectively.

From Theorem 2.3, it immediately follows that the bias of the estimators $\hat{\mathbf{D}}_k$ and $\hat{\mathbf{D}}_k^{(\alpha)}$ is $O((k/M)^{1/d} + (1/k))$. On the other hand, from Theorem II.10 it is clear that the bias of the estimators $\check{\mathbf{D}}$ and $\check{\mathbf{D}}_\alpha$ is $O((k/M)^{1/d})$. The minimum bias for optimal choice of k therefore reduces from $O(M^{-2/(1+d)})$ to $O(M^{-2/d})$.

2.6.4 Estimation of Shannon mutual information

As in the case of Shannon and Rényi entropy and divergence estimation, it is straightforward to define corrected MI estimators in an identical manner by using correction factors described above to reduce estimator bias from $O(M^{-1/(1+d)})$ to $O(M^{-1/d})$.

2.7 Comparison with existing results

Recently, Baryshnikov *et al.* [6] have developed asymptotic convergence results for estimators of f-divergence $G(f_0, f) = \int f(x)\phi(f_0(x)/f(x))dx$ for the case where f_0 is known. Their estimators are based on sums of functionals of k-NN distances. They assume that they have T i.i.d realizations from the unknown density f, and that f and f_0 are bounded away from 0 and ∞ on their support. The general form of the estimator of Baryshnikov *et al.* is given by

$$\hat{\mathbf{G}}_b(f) = \frac{1}{T}\sum_{i=1}^{T} g(\hat{\mathbf{f}}_{kS}(\mathbf{X}_i)),$$

where $\hat{\mathbf{f}}_{kS}(\mathbf{X}_i)$ is the standard k-NN density estimator [55] estimated using the $T-1$ samples $\{\mathbf{X}_1, .., \mathbf{X}_T\} - \{\mathbf{X}_i\}$.

We note that the standard k-NN density estimates $\hat{\mathbf{f}}_{kS}(.)$ at $\{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ can be determined by constructing a k-nearest neighbor graph with the k-the nearest

neighbor edge from each $\mathbf{X}_i \in \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ linking the vertices.

Baryshnikov *et al.* do not analyze the bias of their estimator. They show that the leading term in the variance is given by c_k/T for some constant c_k which is a function of the number of nearest neighbors k. Finally they show that their estimator is asymptotically normal.

Wang *et al.* [87] propose k-NN based estimators to estimate f-divergence for the case when we have i.i.d samples from the unknown densities f_0, f. The general form of the estimator of Wang *et al.* is given by

$$\hat{\mathbf{G}}_w(f) = \frac{1}{T}\sum_{i=1}^{T} g(\hat{\mathbf{f}}_{1kS}(\mathbf{X}_i)/\hat{\mathbf{f}}_{2kS}(\mathbf{X}_i)),$$

where $\hat{\mathbf{f}}_{ikS}(\mathbf{X}_i)$, $i = 1, 2$ are the standard k-NN density estimators estimated using the $T_1 - 1$ and T_2 samples from f_0 and f respectively.

Wang *et al.* do not require the assumption that f and f_0 are bounded away from 0 and ∞ on their support. However, they only show that their estimator is asymptotically consistent and do not provide rates of convergence or results on asymptotic normality.

In contrast, we assume higher order conditions on continuity of the density f and the functional g (see Section 3) as compared to Baryshnikov *et al.* and provide results on bias, variance and asymptotic distribution of data-split k-NN functional estimators of entropies of the form $G(f) = \int 1_{x\in\mathcal{S}'} g(f(x))f(x)dx$. In addition, we have been able to extend our estimator and corresponding results on bias, variance and asymptotic distribution to estimate f-divergence for both cases (i) f_0 is known and we have i.i.d. samples from the unknown density f, or (ii) we have i.i.d samples from the unknown densities f_0, f. Note that we also require the assumption that f and f_0 are bounded away from 0 and ∞ on their support. Because we are able to establish both the bias and variance of our estimator, in turn, we are able to specify

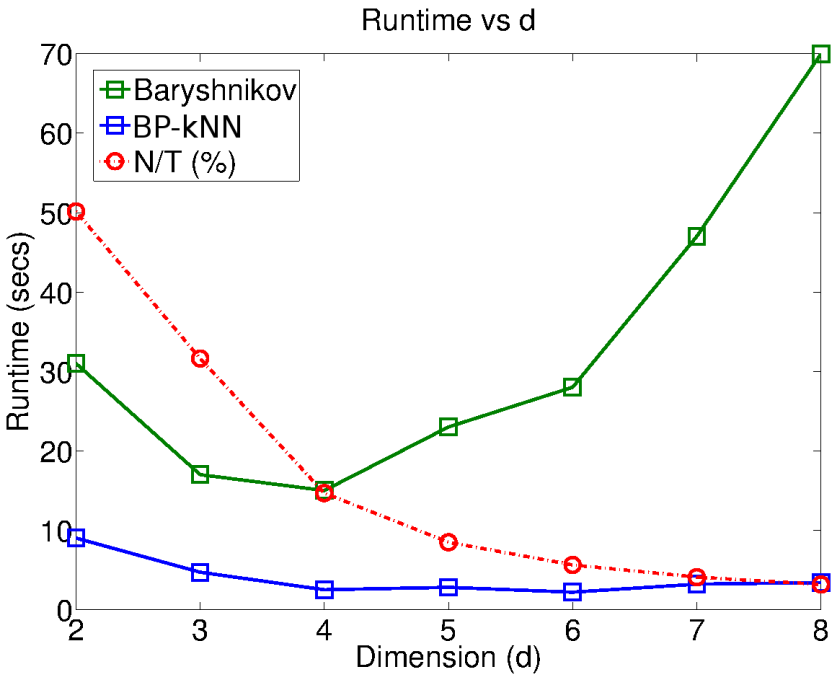


Figure 2.4: Comparison of average runtime of BP-kNN and Baryshnikov's estimator to estimate entropy as a function of dimension d. The runtime of BP-kNN, due to its bipartite nature, is superior to Baryshnikov's estimator.

optimal choice of free parameters k, N, M for minimum MSE. Finally, we are able to establish bias, variance and asymptotic distribution of MI estimators.

For estimating the functional $G(f) = \int g(f(x))f(x)dx$, we can use the estimator of Baryshnikov by restricting f_0 to be uniform. From our analysis in this chapter, under the assumption that f is two times continuously differentiable, we can establish the bias of $\hat{\mathbf{G}}_b(f)$ to be $\Theta((k/T)^{1/d} + 1/k)$. It is clear from our expression for the bias that the estimator of Baryshnikov will be unbiased iff $k \to \infty$ as $T \to \infty$. Furthermore, the optimal rate of growth of k is given by $k = T^{1/(1+d)}$. Furthermore, we can show that $c_k = \Theta(1)$ and therefore the overall optimal bias and variance of $\hat{\mathbf{G}}_b(f)$ is given by $\Theta(T^{-1/(1+d)})$ and $\Theta(T^{-1})$ respectively.

We also note that our estimator requires construction of bipartite k-NN graphs on the data split samples whereas the methods of Baryshnikov *et al.* require construction of directed k-NN graphs on all the T samples. Computationally, our method requires

$\Theta(dNM)$ time as compared to $\Theta(dT^2)$ for the estimators of Baryshnikov *et al.*.

We have shown that for optimal MSE, $N_{opt} = O(T^{(3+d)/(2(1+d))})$ which in turn implies that $M_{opt} = \Theta(T)$. Therefore the overall computation time in our case is $\Theta(dT^{(5+3d)/2(1+d)})$, which is $o(dT^2)$. To illustrate this, we plot the average runtime of Baryshnikov's estimator and our estimator as a function of sample size d in Fig. 2.4. The simulations were run using Matlab 7.6 on a Intel Pentium II processor. We also plot the percentage ratio of the optimal $N = N_{opt}$ over the number of samples T. Because of the decreasing ratio of N_{opt}/T with increasing dimension, the run-time of our estimator decreases relative to the run-time of Baryshnikov's estimator. We also exploit the bipartite nature of our k-NN graph MV estimators (see Chapter 6) to increase computational efficiency by an order of magnitude in the training sample size as compared to standard k-NN graphs.

Additionally, the bipartite nature of our estimators enables us to employ boundary correction to reduce MSE. To our knowledge, it is not straight forward to propose boundary corrected estimators to improve bias on standard k-NN graphs defined over all samples. For details, please see Chapter 3. The estimators of Baryshnikov *et al.*, the entropic graph estimator of Hero *et al.* [38] and the k-nearest neighbor estimator of Leonenko *et al.* [32] are examples which fall under the category of estimators defined on standard k-NN graphs. For comparison of MSE performance of these estimators, please see section 3.5.4.

Finally, we note that methods of proof of Baryshnikov *et al.* use stabilization methods for establishing asymptotic distributions of sums of weakly dependent terms in geometric probability. On the other hand, our methods of proof for determining MSE and CLT are based on statistical properties of k-NN neighborhoods and exchangeability respectively. The generality of our method of proof (lemma B.1-B.6) makes it possible to extend results on MSE and asymptotic normality for kernel density plug-in estimators by using the results established in Appendix A for kernel

density estimators.

2.7.1 Experimental validation of theory for Shannon entropy

We validate the theory of section 2.6 using using the 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 $p = 0.8$. First we estimate Shannon entropy using the estimator $\check{\mathbf{H}}_k$. Constants $c_i; i = 0, 1..5$ are estimated using Monte-Carlo methods [69].

In Fig. 2.7.1 we, we plot experimentally obtained and theoretically computed bias for finite N,M with $N+M = 20,000$. In the next experiment, we plot experimentally obtained and theoretically computed variance for fixed T, as N is varied. The results are shown in Fig. 2.7.1.

Finally, we show the Q-Q plot of the normalized MI estimate and the standard normal distribution in Fig. 2.7.1. The linear Q-Q plot validates our theorem on asymptotic normality of the plug-in estimator. Finally, using the CLT, we plot the 95% confidence intervals for the entropy functional as a function of sample size in Fig. 2.7.1.

2.8 Anomaly detection in networks

We apply our theory to the problem of anomaly detection in wireless sensor networks. Our objective is not an extensive comparison with competing anomaly detection methods, but rather to demonstrate the applicability of our theory to a real world application. 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

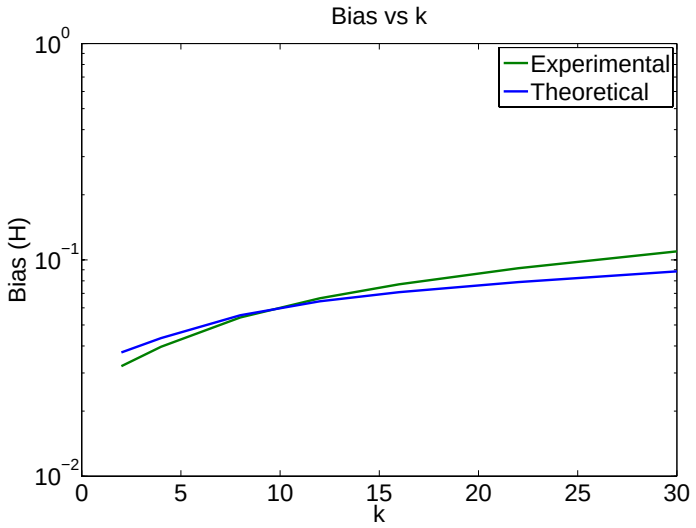


Figure 2.5: Comparison of theoretically predicted bias with experimentally observed bias for varying k. The experimentally observed bias agrees well with the theoretically predicted bias in Theorem II.10, which states that the bias is a monotonically increasing function of k.

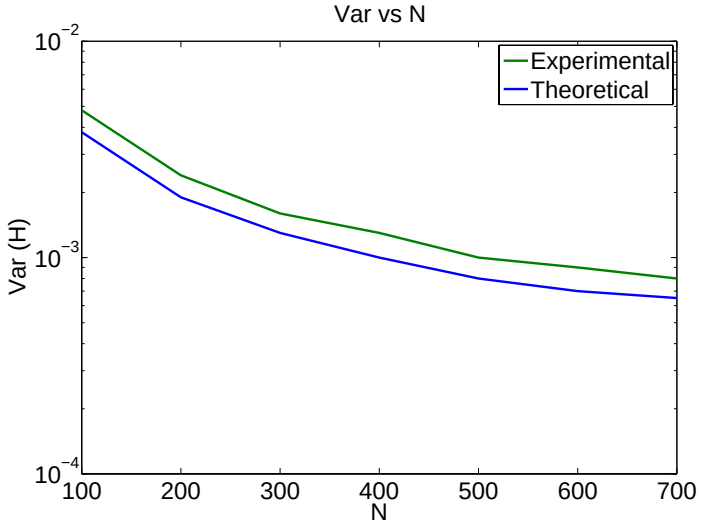


Figure 2.6: Comparison of theoretically predicted variance with experimentally observed variance for varying N. The experimentally observed variance agrees well with the theoretically predicted variance in Theorem II.11.

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.

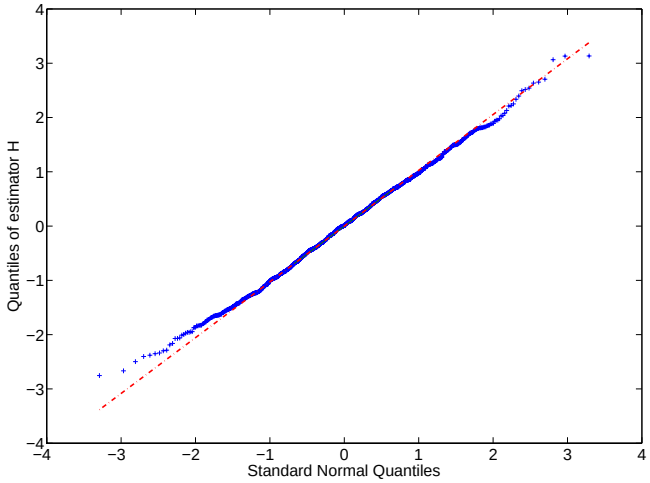


Figure 2.7: q-q comparing independent realizations of the normalized Shannon estimator (L.H.S. of Central limit theorem II.12) on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points validates the Central limit theorem.

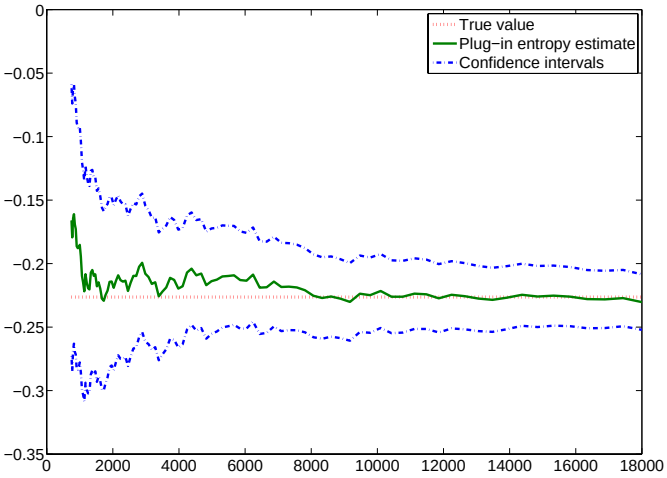


Figure 2.8: Predicted confidence intervals on Shannon entropy for varying sample size T using the Central limit theorem II.12. The confidence intervals decrease with sample size as expected.

The mission of this experiment is to use the 182 RSS sequences to detect any intruders (anomalies). To remove the temperature drifts of receivers we pre-process the data by removing their local mean values. Let $y_i[n]$ be the pre-processed n-th sample of the i-th signal and denote $y[n] = (y_1[n], \ldots, y_{182}[n])'$.

We now estimate the Shannon entropy for each 1-dimensional, 182 sample sequence $y[n]$ using the BP-kNN estimator $\check{\mathbf{H}}_k$. Denote the BP-kNN entropy estimate at each time instant n by $\tilde{H}[n]$. We detect anomalies by thresholding the entropy estimate $\tilde{H}[n]$. A time sample n is regarded to be anomalous if the entropy estimate $\check{H}_k[n]$ exceeds a specified threshold. We seek to choose the threshold appropriately for achieving a desired false alarm rate.

To this end, we estimate the entropies $\check{H}_k[n]$ for the time instants $n = 1, \ldots, 50$ when no anomalies were known to have occurred and subsequently estimate the mean μ and variance σ^2 of the entropy estimates for this nominal time interval $n \in [1, 50]$. Using these estimates of the mean and variance, we use the central limit theorem II.3 to set the threshold t_α for a given false alarm rate α as $t_\alpha = \mu + z_{\alpha/2}\sigma$ where $z_{\alpha/2}$ is the z-score corresponding to coverage $1 - \alpha$. This threshold t_α is then used to detect anomalies at time instants $n > 50$.

We note that the data in this experiment is not i.i.d. due to dependence between successive time samples, and therefore does not conform to assumptions of our theory. This dependence results in marginally higher entropy estimates at non-anomalous time instants immediately preceding and succeeding anomalous time intervals as compared to entropy estimates at nominal time instants farther away from anomalous activity. This is corroborated by Fig. 2.9, which shows the ground truth and the normalized entropy estimator response ($\check{H}_k[n] - t_\alpha$ with false alarm rate $\alpha = 0.05$) as a function of time.

Desired and observed false alarm rates						
Desired	.20	.10	.05	.02	.01	.005
Observed	.269	.111	.062	.026	.015	.009

The desired and corresponding observed false alarm rates in this experiment are shown in the table above. Despite the non i.i.d. nature of the data due to dependence between time samples, the observed false alarm rates is only marginally higher than

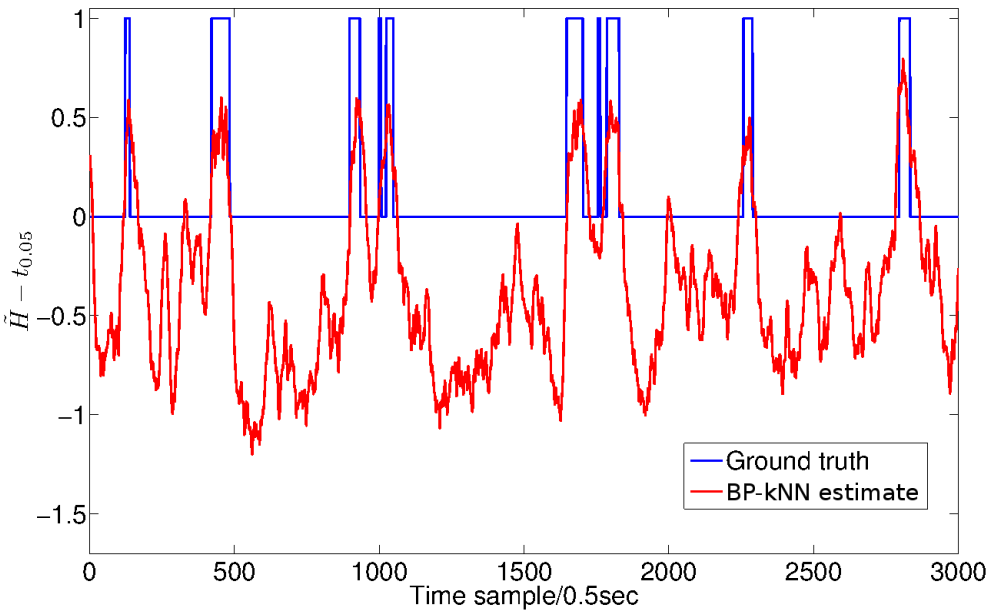


Figure 2.9: Entropy estimate $\tilde{H}[n]$ evaluated using BP-kNN estimator $\check{\mathbf{H}}_k$, implemented as a scan statistic over time n for anomaly detection in wireless ad hoc sensor network experiment. Ground truth indicator function (in blue) indicates when anomalous activity occurred. The entropy estimator detects these anomalies whenever the entropy estimate crosses the level $\alpha = 0.05$ threshold $t_{0.05}$ analytically determined by the CLT in Theorem II.3.

the desired false alarm rate. This result suggests that our theory can be applied to problems where there is dependence in the data.

ROC curves corresponding to the BP-kNN entropy estimator are shown in Fig. 2.10 in addition to the ROC curves using the subspace method of Lakhina *et al.* [45] and the covariance based estimator of Chen *et al.* [17]. It is clear that the detection performance using the entropy estimator is marginally better than the subspace and covariance based methods of Lakhina *et al.* and Chen *et al.* respectively. The Area under the ROC curves were found to be 0.9784, 0.9722 and 0.9645 for the entropy, covariance and subspace based anomaly detection methods respectively.

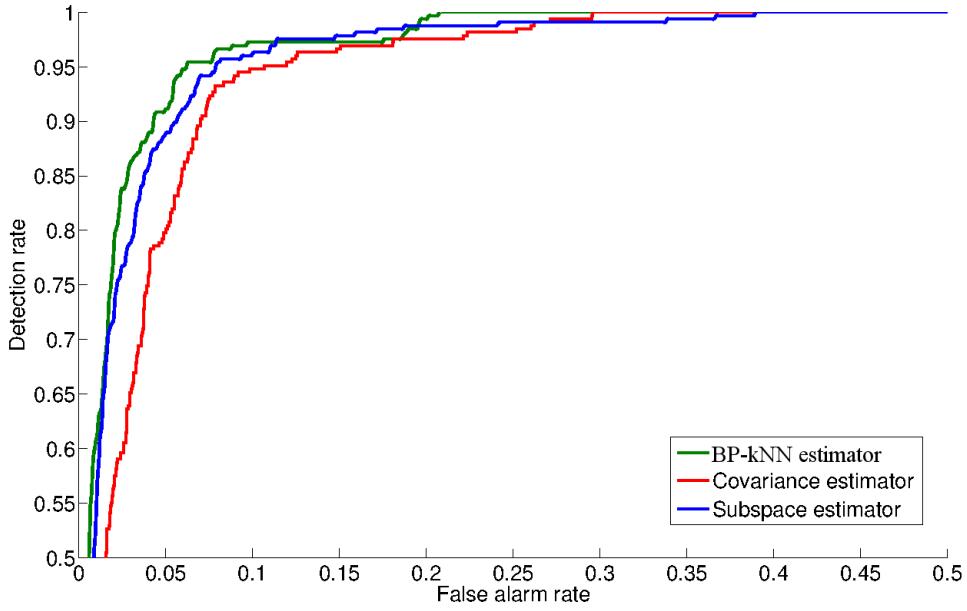


Figure 2.10: ROC curves for BP-kNN entropy estimate, covariance and subspace based anomaly detection. The performance of the BP-kNN entropy based method is the best as measured by area under the curve (0.9784 and compared to 0.9722 and 0.9645).

2.9 Discussion

We proposed a class of data-split k-NN density 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 central limit theorem for these estimators. We verified the validity of our theorems through simulations and established that the theory can be used to specify optimal estimator tuning parameters such as bandwidth and optimal partitioning of data samples.

Using the theory presented in this chapter one can tune the parameters of the plug-in estimator to achieve minimum asymptotic estimation MSE. Furthermore, it can be used to specify the minimum necessary sample size required to obtain requisite accuracy. This in turn can be used to predict and optimize performance in

applications like structure discovery in graphical models and dimension estimation for support sets of low intrinsic dimension. These applications are described in detail in future chapters.

CHAPTER III

Boundary compensation

3.1 Introduction

In chapter 2, we analyzed data-split bipartite k-NN graph estimators of entropy and divergence functionals of densities. The generic expression for the bias of these estimators was given by $c_0(k/M)^{1/d} + c_1(k/M)^{2/d}$. We showed that the source of the leading term $c_0(k/M)^{1/d}$ is due to the fact that if a probability density function has bounded support, the k-NN balls centered at points close to the boundary are often truncated at the the boundary. As a consequence of this truncation, the k-NN density estimates near the boundaries of the support suffer from significant bias. In this chapter, we will explore the source of this bias term due to truncation in further detail and present an modification of bipartite k-NN graphs which will reduce the leading term in the bias to $c_1(k/M)^{2/d}$.

Consider a large random sample from a continuous multivariate density that is zero outside a bounded region, which is the support of the density. When one constructs the k-NN graph on such a sample the local neighborhoods of the graph behave differently near the boundary of the support. For points well inside the boundary, the k-NN neighbors will be spread almost uniformly around the point. On the other hand, for points close to the boundary of the support, the k-NN neighbors are disproportionately distributed away from the boundary. This phenomenon becomes more

striking as the dimension of the multivariate density increases. As a result, the radius of the k-NN neighborhoods tend to be disproportionately larger near the boundary as compared to neighborhoods in the interior. These ideas will be formalized in Section 2 using analysis of the bias of k-NN density estimates.

The bias of finite supported density estimator performance has been previously studied in [42, 44] for kernel density estimates. Corrections have been suggested, primarily for the univariate case. These corrections also assume that the support is known apriori. In this chapter, we propose a method for compensating for the bias of k-NN density estimates for general multivariate data without any prior knowledge of the support of the density. Motivated by our analysis of k-NN density estimators, we suggest a corrected version of data-split k-NN plug-in estimators which compensates for k-NN graph behavior near the boundary of the support.

Throughout this chapter, we focus on the regime where the radius of the k-NN ball (which is $O((k/M)^{1/d})$ [55]) is small. This regime is equivalent to having a large number of samples relative to the dimension d. We note that k-NN methods will work poorly in high-dimensional spaces under small sample sizes and the above operating regime is necessary for k-NN methods to be effective. The consistency of k-NN estimation depends on the assumption that the size of the k-NN neighborhood becomes small relative to the modulus of continuity of the underlying probability density that generates the points. Thus one generally requires a large number of samples before the small estimation error behavior of a consistent estimator kicks-in. Specifically, as compared to low dimension sample space, for high dimensional samples one needs an exponentially greater number of samples to achieve equivalent bias. This follows from the fact that k-NN methods [55] require that $(k/M)^{1/d} \to 0$ and $k \to \infty$ for consistency and that the optimal rate is obtained by equalizing $(k/M)^{2/d}$ and $1/k$.

3.2 k-NN density estimators

In this section, we briefly review properties which have been established in the appendices of k-NN density estimators in the interior of the support. We then contrast this behavior to k-NN density estimators near the boundary of the support.

3.2.1 Concentration inequality for coverage probability

Define the coverage function as $\mathbf{P}(X) = \int_{\mathbf{S}_k(X)} f(Z)dZ$. Define spherical regions $S_r(X) = \{Y \in \mathcal{S} : d(X,Y) \leq r\}$. It has been previously established that $\mathbf{P}(X)$ has a beta distribution with parameters k, $M-k+1$. [55]. Using Chernoff inequalities, we can then establish the following concentration inequality. For $0 < p < 1/2$,

$$Pr(|\mathbf{P}(X) - k/M| < pk/M) = O(e^{-p^2k/2}). \tag{3.1}$$

Let $\natural(X)$ denote the event $(1-p_k)k/M < \mathbf{P}(X) < (p_k+1)k/M$ where $p_k = 1/(k^{\delta/2})$ with $\delta \in (2/3, 1)$. Then, $1 - Pr(\natural(X)) = O(e^{-p_k^2 k/2}) = o(1/k^a)$ for arbitrarily large values of a. Using the logarithmic growth condition on k which specifies $k = O(\log M)$, we have $1 - Pr(\natural(X)) = \mathbb{E}[1_{\natural^c(X)}] = o(1/M^a)$ for arbitrarily large values of a.

3.2.2 Interior points

Let $\mathcal{S}'' = \mathcal{S}_{\mathcal{I}}$ and observe that $Pr(\mathbf{X} \notin \mathcal{S}') = o(1)$, where X is random variable with density f. This implies that given the event $\natural(X)$, the k-NN neighborhoods of points $X \in \mathcal{S}''$ will lie completely inside the domain $\mathcal{S}$. Therefore 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 $2r(1-\alpha)/d > 1$.

3.2.3 Taylor series expansion of coverage probability

Let $X \in \mathcal{S}''$. Conditioned on the event $\natural(X)$, the k-NN region $\mathbf{S}_k(X)$ is a subset of $\mathcal{S}$. The coverage function $\mathbf{P}(X)$ can then be represented 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 [55].

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

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 $h(X) = c(X)f^{-2/d}(X)$. Note that r satisfies the condition $2r(1-\alpha)/d > 1$. Rearrange terms to obtain the following representation of $1/\mathbf{V_k}(X)$ [55]

$$\begin{aligned}\frac{1}{\mathbf{V}_k(X)} &= \frac{f(X)}{\mathbf{P}(X)} + \frac{h(X)}{\mathbf{P}^{1-2/d}(X)} \\ &+ \sum_{t\in\mathcal{T}} \frac{h_t(X)}{\mathbf{P}^{1-t}(X)} + \mathbf{h_r}(X),\end{aligned} \tag{3.2}$$

where $\mathcal{T}$ is some countable set with $\inf\{\mathcal{T}\} = 4/d$ and $\mathbf{h_r}(X) = o(1/\mathbf{P}^{1-2r/d}(X))$.

3.2.4 Bias of k-NN density estimates in interior

Finally, we analyze the bias of the k-NN density estimate which, unlike other central moments, cannot be obtained using (B.26). Let $X \in \mathcal{S}''$. Using the Taylor series expansion (B.13), it is shown in [55] that the bias of the k-NN density estimate is given by

$$\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) = h(X)\left(\frac{k}{M}\right)^{\frac{2}{d}} + o\left(\frac{k}{M}\right)^{\frac{2}{d}}. \tag{3.3}$$

Using these properties, we had showed that for any set $\mathcal{S}'' \subset \mathcal{S}_{\mathcal{I}}$, which is in the

interior of the support, the following properties hold:

$$\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}''\}}\hat{\mathbf{f}}_k(\mathbf{X})] - f(X) = \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}h(\mathbf{X})]\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d}. \tag{3.4}$$

3.2.5 Bias of k-NN density estimator near boundary

If a probability density function has bounded support, the k-NN balls centered at points close to the boundary are often truncated at the the boundary as shown in Fig. 3.1. Let

$$\alpha_k(X) = \frac{\int_{\mathbf{S}_k(X)\cap\mathcal{S}} dZ}{\int_{\mathbf{S}_k(X)} dZ}$$

be the fraction of the volume of the k-NN ball inside the boundary of the support. Also define $\mathbf{V}_{k,M}(X)$ to be the k-NN ball volume in a sample of size M. For interior points $X \in \mathcal{S}''$, with high probability, $\alpha_k(X) = 1$, 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

$$\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) = (1 - \alpha_k(X))f(X) + o(1). \tag{3.5}$$

Therefore the bias is much higher at the boundary of the support ($O(1)$) as compared to its interior ($O((k/M)^{2/d})$) (B.32). Furthermore, the bias at the support boundary does not decay to 0 as $k/M \to 0$.

As a result, if the set $\mathcal{S}'$ has non-empty intersection with the boundary of the support, we have

$$\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}(\hat{\mathbf{f}}_k(\mathbf{X}) - f(\mathbf{X}))] \;=\; h_0\left(\frac{k}{M}\right)^{1/d} + \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\cap\mathcal{S}_{\mathcal{I}}\}}h(\mathbf{X})]\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{1/d},$$

for some constant h_0 which depends on the density f and the support $\mathcal{S}$.

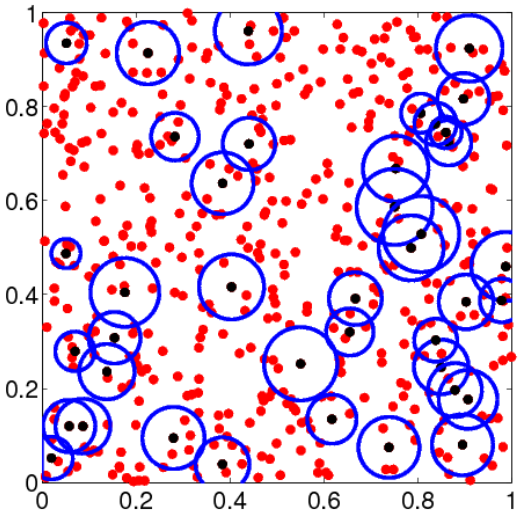


Figure 3.1: k-NN balls centered around a subsample of 2D uniformly distributed points. Note that the k-NN balls centered at points close to boundary are truncated by the boundary.

3.3 Boundary corrected k-NN density estimates

We have shown that the bias of k-NN density estimates $\hat{f}_k(.)$ in the interior of the density is of order $O((k/M)^{2/d})$ while the bias near the boundary is of order $O(1)$. We now describe a boundary corrected density estimator $\tilde{f}_k(.)$ which has bias of order $O((k/M)^{2/d})$ everywhere.

As a first step, we will identify interior points $\mathcal{I}_{\mathcal{N}}$ among the set of points among $\mathcal{X}_{\mathcal{N}} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ which are guaranteed to lie within the set $\mathcal{S}''$ with high probability. Define the set by $\mathcal{B}_{\mathcal{N}} = \mathcal{X}_{\mathcal{N}} - \mathcal{I}_{\mathcal{N}}$. Note that the bias of the standard k-NN density estimate is of order $O((k/M)^{2/d})$ for points $X \in \mathcal{I}_{\mathcal{N}}$.

Next, to compensate for the bias due to truncation, we modify the density estimator at the boundary points $\mathcal{B}_{\mathcal{N}}$ by using boundary corrected k-NN density estimates. We describe these steps in detail.

3.3.1 Boundary point detection

In this section, we will identify interior points $\mathcal{I}_{\mathcal{N}}$ while exclusively using the set $\mathcal{X}_{\mathcal{N}} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, with high probability. The fact that the detection of these points is independent of $\mathcal{X}_{\mathcal{M}} = \{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ is crucial for proving consistency of the proposed method.

Define $V_{k,M}(X) := \frac{k}{M\alpha_k(X)f(X)}$. Let $p(k, M)$ be any positive function satisfying $p(k, M) = \Theta((k/M)^{2/d}) + 1/k^{\delta/2}$. From the concentration inequality (B.1) and Taylor series expansion of the coverage function (B.13), for small values of k/M, we have

$$1 - Pr\left(\left|\frac{\mathbf{V}_{k,M}(X)}{V_{k,M}(X)} - 1\right| \leq p(k, M)\right) \leq o(M^{-a}).$$

To detect the interior points $\mathcal{I}_{\mathcal{N}}$, we construct a standard K-NN graph on $\mathcal{X}_{\mathcal{N}}$ where $K = \lfloor k \times (N/M) \rfloor$. By the concentration inequality (B.1), this choice of K guarantees that the size of the $2k$-NN ball in the partitioned sample is approximately the same as the the size of the K-NN ball in the N sample with high probability $1 - o(1/N^a)$.

Using the K-NN graph, for each sample $\mathbf{X} \in \mathcal{X}_{\mathcal{N}}$, we compute the number of points in $\mathcal{X}_{\mathcal{N}}$ that have $\mathbf{X}$ as a l-th nearest neighbor (l-NN), $l = \{1, \ldots, K\}$. Denote this count as $count(\mathbf{X})$. Then, for any $X \in \mathcal{X}_{\mathcal{N}}$,

$$1 - Pr\left(\left|\frac{\mathbf{V}_{K,N}(X)}{V_{K,N}(X)} - 1\right| \leq p(K, N)\right) \leq o(N^{-a}). \tag{3.6}$$

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

For X to be one of the K-NN of Y it is necessary that $R_K(Y) \geq ||Y - X||$ or equivalently, $R_K(Y)/R_K(X) \geq ||u||$. Using the concentration inequality (3.6) for

$R_K(X)$ and $R_K(Y)$, a sufficient condition for this is

$$\frac{\alpha_K(X)f(X)}{\alpha_K(Y)f(Y)}(1-2p(K,N)) \geq ||u||. \tag{3.7}$$

For $X \in \mathcal{B}_{\mathcal{T}}$, $\alpha_K(X) < 1$ with probability $1 - o(1/M^a)$. On the other hand, for a majority of the interior points $X \in \mathcal{I}_{\mathcal{N}}$ (of fraction $(K/N)^{1/d}$) $\alpha_K(X) = 1$ with high probability $1-o(1/M^a)$. This implies that X will be one of the K-NN of Y provided $||u|| \leq 1 - 2p(K,N)$. This implies that $count(\mathbf{X}) \geq K(1-2p(K,N))$.

It is therefore clear that the measure $count(X)$ is higher for points in the interior $X \in \mathcal{S}''$ as compared to points close to the boundary $\mathcal{B}$. For the specific choice $\mathcal{S}'' = \mathcal{S}_{\mathcal{I}}$, $\mathcal{I}_N$ and $\mathcal{B}_N$ can then be detected as follows. Sort the array $\{count(X); X \in \mathcal{X}_N\}$ in ascending order and assign the top κN points in this sorted array to be the set $\mathcal{B}_N$, where κ is the fraction $\kappa = \epsilon_\infty^2(k/M)^{1/d}$. Algorithm 1, shown below, codifies this sketch into a precise procedure. Having detected the boundary points,

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

1. Construct $K = k \times N/M$-NN tree on $\mathcal{X}_{\mathcal{N}}$
2. Compute $count(\mathbf{X})$ for each $\mathbf{X} \in \mathcal{X}_{\mathcal{N}}$
3. Sort array $\{count(X); X \in \mathcal{X}_N\}$ in ascending order
4. Detect boundary points $\mathcal{B}_{\mathcal{N}}$: Assign top κN points in sorted array to the set $\mathcal{B}_N$

we suggest replacing the standard k-NN density estimates at these points which suffer from bias of $O(1)$, with k-NN density estimates in the interior. Fig. 3.3.1 illustrates an instantiation of Algorithm 1 applied to detection of boundary points and pairing of boundary points with interior points in a uniform sample over a square support region.

3.3.2 Boundary corrected density estimator

Here the boundary corrected k-NN density estimator is defined and its asymptotic rates are computed. The proposed density estimator corrects the k-NN ball volumes

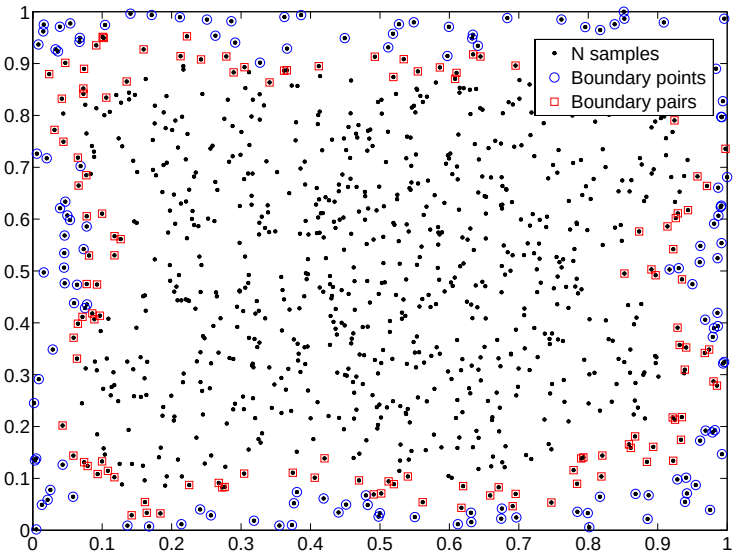


Figure 3.2: Detection of boundary points, and their closest interior neighbors, for realizations drawn from and 2d beta distribution. Clearly, the algorithm 1 identifies the boundary points in this example.

for points that are close to the boundary. To estimate the density at a boundary point $\mathbf{X} \in \mathcal{B}_{\mathcal{N}}$, we find a point $\mathbf{Y} \in \mathcal{I}_{\mathcal{N}}$ 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 $\mathbf{Y}$ instead and use this as an estimate of $f(\mathbf{Y})$. This informal argument is made more precise in what follows.

Consider the corrected density estimator $\tilde{\mathbf{f}}_k$ defined in (3.8). For each boundary point $\mathbf{X}_i \in \mathcal{B}_N$, let $\mathbf{X}_{n(i)} \in \mathcal{I}_{\mathcal{N}}$ be the interior sample point that is closest to $\mathbf{X}_i$. The corrected density estimator $\tilde{\mathbf{f}}_k$ is defined as follows.

$$\tilde{\mathbf{f}}_k(\mathbf{X}_i) = \begin{cases} \hat{\mathbf{f}}_k(\mathbf{X}_i) & \{\mathbf{X}_i \in \mathcal{I}_{\mathcal{N}}\} \\ \hat{\mathbf{f}}_k(\mathbf{X}_{n(i)}) & \{\mathbf{X}_i \in \mathcal{B}_N\} \end{cases} \tag{3.8}$$

This estimator has bias of order $O((k/M)^{1/d})$, which can be shown as follows. Let $\mathbf{X}$ denote $\mathbf{X}_i$ for some fixed $i \in \{1, .., N\}$. Also, let $\mathbf{X}_{-1} = \arg\min_{x \in \mathcal{S}'} d(x, \mathbf{X})$.

Given $\mathcal{X}_N$, if $X \in \mathcal{I}_N$, then by (B.32),

$$\mathbb{E}[\tilde{\mathbf{f}}_k(X)] \;=\; \mathbb{E}[\hat{\mathbf{f}}_k(X)] = f(X) + O((k/M)^{2/d}) + O(\mathcal{C}(k)).$$

Next consider the alternative case $X \in \mathcal{B}_N$. Let $X_n \in \mathcal{I}_N$ be the closest interior point to X. Define $h = X - X_n$. h can be rewritten as $h = h_1 + h_2$, where $h_1 = X - X_{-1}$ and $h_2 = X_{-1} - X_n$. Since $X \in \mathcal{B}_N$ implies that $X \in \mathcal{S} - \mathcal{S}'$ with probability $1 - o(1/M)$, consequently $||h_1|| = ||X - X_{-1}|| = O((k/M)^{1/d})$ with probability $1 - o(1/M)$. Again with probability $1 - o(1/M)$, $X_n \in \mathcal{S}'$ and consequently $||h_2|| = ||X_{-1} - X_n|| = o((k/M)^{1/d})$. This implies that $||h|| = O((k/M)^{1/d})$. Now,

$$f(X) = f(X_n) + O(||h||).$$

If X_n is located in the interior $\mathcal{S}'$, by (B.32),

$$\mathbb{E}[\hat{\mathbf{f}}_k(X_n)] \;=\; f(X_n) + O((k/M)^{2/d}) + o(1/M), \tag{3.9}$$

and therefore

$$\begin{aligned}
\mathbb{E}[\tilde{\mathbf{f}}_k(X)] &= \mathbb{E}[\hat{f}_k(\mathbf{X}_n)] + o(1/M) \\
&= f(X_n) + O((k/M)^{2/d}) + o(1/M) \\
&= f(X) + O(||h||) + O((k/M)^{2/d}) + o(1/M) \\
&= f(X) + O((k/M)^{1/d}) + o(1/M),
\end{aligned} \tag{3.10}$$

where the $o(1/M)$ accounts for error in the case of the event that $X_{n(i)} \notin \mathcal{S}'$. This implies that the corrected density estimate has lower bias as compared to the standard k-NN density estimate (compare to (B.32) and (C.1)). In particular, boundary compensation has reduced the bias of the estimator at points near the boundary from

$O(1)$ to $O((k/M)^{1/d}) + o(1/M)$.

3.3.3 Properties of boundary corrected density estimator

By section 3.3.1, $\mathcal{I}_{\mathcal{N}} \in \mathcal{S}'$ with probability $1 - o(1/M)$. The results on bias, variance and cross-moments of the standard k-NN density estimator $\hat{\mathbf{f}}_k$ derived in the previous Appendix for points $X \in \mathcal{S}'$ therefore carry over to the corrected density estimator $\tilde{\mathbf{f}}_k$ for points $\mathcal{I}_{\mathcal{N}}$ with error of order $o(1/M)$.

In the definition of the corrected estimator $\tilde{\mathbf{f}}_k$ in (3.8), $\hat{\mathbf{f}}_k(\mathbf{X}_{n(i)})$ is the standard k-NN density estimates and $\mathbf{X}_{n(i)} \in \mathcal{S}'$. It therefore follows that the variance and other central and cross moments of the corrected density estimator $\tilde{\mathbf{f}}_k$ will continue to decay at the same rate as the standard k-NN density estimator in the interior, as given by (B.36) and (B.37).

Given these identical rates and that the probability of a point being in the boundary region $\mathcal{S} - \mathcal{S}'$ is $O((k/M)^{1/d}) = o(1)$, the contribution of the boundary region to the overall variance and other cross moments of the boundary corrected density estimator $\tilde{\mathbf{f}}_k$ are asymptotically negligible compared to the contribution from the interior. As a result we can now generalize the results from Appendix A on the central moments and cross moments to include the boundary regions as follows. Denote $\tilde{\mathbf{f}}_k(X) - \mathbb{E}_X[\tilde{\mathbf{f}}_k(X) \mid X]$ by $\mathbf{e}(X)$.

3.3.3.1 Central and cross moments

For positive integers $q, r < k$

$$\mathbb{E}[\gamma(\mathbf{X})\mathbf{e}^q(\mathbf{X})] = 1_{\{q=2\}}\mathbb{E}\big[\gamma(\mathbf{X})f^2(\mathbf{X})\big]\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right), \tag{3.11}$$

$$
\begin{aligned}
&Cov[\gamma_1(\mathbf{X})\mathbf{e}^q(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}^r(\mathbf{Y})] \\
&= 1_{\{q,r=1\}} Cov[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})f(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})f(\mathbf{Y})] \left(\frac{1}{M} + o(1/M)\right) \\
&+ 1_{\{q+r>2\}} \left(O\left(\frac{1}{k^{((q+r)\delta/2-1)}M}\right) + O(k_M^{2/d}/M) + O(1/M^2)\right). \qquad (3.12)
\end{aligned}
$$

Next, we derive the following result on the bias of boundary corrected estimators.

3.3.3.2 Bias

For $k > 2$,

$$
\begin{aligned}
&\mathbb{E}[\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(\mathbf{X}) \mid \mathbf{X}]) - \gamma(f(\mathbf{X})))] = \mathbb{E}\left[\mathbb{E}\left[(\gamma(\tilde{\mathbf{f}}_k(\mathbf{X})) - \gamma(f(\mathbf{X}))) \mid \mathcal{X}_N\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{I}_N\}}(\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(X)]) - \gamma(f(X))) \mid \mathcal{X}_N\right]\right] \\
&+\mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(X)]) - \gamma(f(X))) \mid \mathcal{X}_N\right]\right] \\
&= I + II. \qquad (3.13)
\end{aligned}
$$

From (B.32), and $Pr(\mathbf{X} \in \mathcal{B}_N) = O((k/M)^{1/d})$, we have

$$
I = \mathbb{E}\left[\gamma'(f(\mathbf{X}))h(\mathbf{X})\right]\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d}. \qquad (3.14)
$$

Next, we will now derive II.

$$
\begin{aligned}
II &= \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(X)]) - \gamma(f(X))) \mid \mathcal{X}_N\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(f(X_n)) - \gamma(f(X))) + O\left(\frac{k}{M}\right)^{2/d} \mid \mathcal{X}_N\right]\right], \qquad (3.15)
\end{aligned}
$$

where the last step follows by (3.9). Let us concentrate on the inner expectation now. By section 3.3.1, we know that with probability $1 - o(1/M)$, if $X \in \mathcal{B}_N$, then $X \in \mathcal{S} - \mathcal{S}'$ and if $X_n \in \mathcal{I}_N$, then $X_n \in \mathcal{S}'$. Furthermore, $||X - X_{-1}|| = O(k/M)^{1/d}$

and $||X_{-1} - X_n|| = o(k/M)^{1/d}$ with probability $1 - o(1/M)$. This implies that

$$\begin{aligned}
&\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(f(X_n)) - \gamma(f(X))) + O\left(\frac{k}{M}\right)^{2/d} \mid \mathcal{X}_N\right] \\
&= \mathbb{E}\left[1_{\{X\in\mathcal{S}-\mathcal{S}'\}}(\gamma(f(X_{-1})) - \gamma(f(X))) \mid \mathcal{X}_N\right] + o\left(\frac{k}{M}\right)^{1/d} + o(1/M).
\end{aligned}$$

Since $Pr(\mathbf{X} \in \mathcal{S} - \mathcal{S}') = O((k/M)^{1/d})$, this in turn implies that

$$\begin{aligned}
II &= \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(X)]) - \gamma(f(X))) \mid \mathcal{X}_N\right]\right] \\
&= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}-\mathcal{S}'\}}(\gamma(f(\mathbf{X}_{-1})) - \gamma(f(\mathbf{X})))] + o\left(\frac{k}{M}\right)^{2/d} + o(1/M). \quad (3.16)
\end{aligned}$$

We therefore finally get,

$$\begin{aligned}
&\mathbb{E}[\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(\mathbf{X}) \mid \mathbf{X}]) - \gamma(f(\mathbf{X})))] = I + II \\
&= \mathbb{E}\left[\gamma'(f(\mathbf{X}))h(\mathbf{X})\right]\left(\frac{k}{M}\right)^{2/d} + \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}-\mathcal{S}'\}}(\gamma(f(\mathbf{X}_{-1})) - \gamma(f(\mathbf{X})))] \\
&+ o\left(\frac{k}{M}\right)^{2/d} + o(1/M). \quad (3.17)
\end{aligned}$$

Note that $||\mathbf{X} - \mathbf{X}_{-1}|| = O((k/M)^{1/d})$ with probability $1 - o(1/M)$. This therefore implies that

$$\begin{aligned}
c_3 &= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}-\mathcal{S}'\}}(\gamma(f(\mathbf{X}_{-1})) - \gamma(f(\mathbf{X})))] \\
&= O((k/M)^{1/d}) \times O((k/M)^{1/d}) + o(1/M) = O((k/M)^{2/d}) + o(1/M). \quad (3.18)
\end{aligned}$$

3.4 Functional estimation using boundary corrected density estimates

In this section, we use the boundary compensated k-NN density estimates to estimate functionals of the density. We are interested in estimating

$$G(f) = \int 1_{\{x \in \mathcal{S}'\}} g(f(x), x) f(x) d\mu(x) = \mathbb{E}[1_{\{x \in \mathcal{S}'\}} g(f(x), x)],$$

for some smooth function $g(f(x), x)$ and any subset $\mathcal{S}' \subset \mathcal{S}$ of the support $\mathcal{S}$. Define the plug-in estimators using boundary corrected k-NN density estimates as

$$\tilde{\mathbf{G}}_k(f) \;=\; \frac{1}{N} \sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i). \tag{3.19}$$

Let $\mathbf{Z}$ denote an independent realization drawn from f. Also, define $\mathbf{Z}_{-1} \in \mathcal{S}_I$ to be $\mathbf{Z}_{-1} = \arg\min_{x \in \mathcal{S}_I} d(x, \mathbf{Z})$. Under the assumptions stated in Section 2.1, with the additional assumption that M, N and T are linearly related through the proportionality constant α_{frac} with: $0 < \alpha_{frac} < 1$, $M = \alpha_{frac} T$ and $N = (1 - \alpha_{frac})T$, the following theorems hold.

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

$$\begin{aligned} \mathbb{B}(\tilde{\mathbf{G}}_k(f)) \;=\; & c_1 \left(\frac{k}{M}\right)^{2/d} + c_2 \left(\frac{1}{k}\right) \\ & + c_3(k, M, N) + o\left(\frac{1}{k} + \frac{k}{M}\right), \end{aligned}$$

where $c_3(k, M, N) = \mathbb{E}[1_{\{\mathbf{Z} \in \mathcal{S} - \mathcal{S}_I\}}(g(f(\mathbf{Z}_{-1}), \mathbf{Z}_{-1}) - g(f(\mathbf{Z}), \mathbf{Z}))] = O(k/M)^{2/d}$, $c_1 = \mathbb{E}[1_{\{\mathbf{Y} \in \mathcal{S}'\}} g'(f(\mathbf{Y}), \mathbf{Y}) c(\mathbf{Y})]$ and $c_2 = \mathbb{E}[1_{\{\mathbf{Y} \in \mathcal{S}'\}} f^2(\mathbf{Y}) g''(f(\mathbf{Y}), \mathbf{Y})/2]$.

Proof. We have shown in section 3.3.3 that the boundary corrected k-NN density estimate satisfies assumptions $\mathscr{A}.1$ and $\mathscr{A}.2$ listed in Appendix D, which in turn

implies that lemma D.1 holds. This gives:

$$\begin{aligned}\mathbb{E}[\tilde{\mathbf{G}}_k(f)] - G(f) &= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\ &\quad + c_2\left(\frac{1}{k}\right) + o(1/k).\end{aligned}$$

Using the properties of boundary corrected k-NN density estimates (section 3.3.3), we can then show

$$\begin{aligned}&\mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\ &= c_1\left(\frac{k}{M}\right)^{2/d} + c_3(k, M, N) + o\left(\left(\frac{k}{M}\right)^{2/d}\right).\end{aligned}$$

This concludes the proof.

□

The leading terms $c_1(k/M)^{2/d} + c_2/k$ arise due to the bias and variance of k-NN density estimates respectively (see Appendix A), while the term $c_3(k, M, N)$ arises due to boundary correction (see Appendix B). Henceforth, we will refer to $c_3(k, M, N)$ by c_3. Observe that $c_3 = O((k/M)^{2/d})$ (3.16).

In comparison to $\hat{\mathbf{G}}_k(f)$, we note that the bias reduces from $O(k/M)^{1/d}$ to $O(k/M)^{2/d}$ by using boundary correction. Because $\hat{\mathbf{G}}_k(f)$ and $\tilde{\mathbf{G}}_k(f)$ differ only near the boundaries of the support, the expressions for the variance and asymptotic distribution of $\tilde{\mathbf{G}}_k(f)$ are identical to the results of $\hat{\mathbf{G}}_k(f)$.

3.4.1 Optimized parameter tuning

From Theorem II.1 we see that it is required that $k \to \infty$ and $k/M \to 0$ for the estimator to be asymptotically unbiased. Likewise from Theorem II.2 we see that it is required that $N \to \infty$ and $M \to \infty$ for the variance of the estimator to converge to 0. We can now optimize the choice of density estimator tuning parameters to minimize

MSE. These tuning parameters are the number of nearest neighbors k, and the data splitting proportions $N/(N+M)$, $M/(N+M)$.

Throughout this section, we assume that the constant $c_0 = 0$. This is true if $\mathcal{S}' \cap \mathcal{B} = \phi$ when using bipartite k-NN estimators proposed in Chapter 2 and for arbitrary subsets $\mathcal{S}'$ when using boundary corrected estimators proposed in this chapter.

3.4.1.1 Optimal choice of k

Minimizing the MSE 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

$$k_{opt} \quad = \quad \arg\min_{k} |\mathbb{B}(f)| = \lfloor k_0 M^{\frac{2}{2+d}} \rfloor, \tag{3.20}$$

where $\lfloor x \rfloor$ is the closest integer to x and 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 to specify the bias at the optimal value of k (see Page 10, [80]). The bias evaluated at k_{opt} in this case is given by $b_0^- M^{\frac{-4}{2+d}}(1+o(1))$ where b_0^- is a constant which depends on the underlying density f. In practice, the constants c_1 and c_2 have to be estimated with error of at least order $o(1/k + (k/M)^{2/d})$ for the leading terms to cancel using the optimal choice $k_{opt} = \lfloor k_0 M^{\frac{2}{2+d}} \rfloor$, where k_0 depends on the estimated values of c_1 and c_2.

3.4.1.2 Choice of $\alpha_{frac} = M/T$

Observe that the MSE of $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ is dominated by the squared bias ($O(M^{-4/(2+d)})$) as contrasted to the variance ($O(1/N+1/M)$). This implies that the MSE rate of convergence is invariant to the choice of α_{frac}. This is corroborated by the experimental results shown in Fig. 6.6.

3.4.1.3 Discussion on optimal parameter choices

The optimal choice of k grows at a smaller rate as compared to the total number of samples M used in the first stage, which is the density estimation step. Furthermore, the rate at which k/M grows 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 estimates of the functional $G(f)$) to increase as the dimension increases. For increasing dimension an increasing number of the M points will be near the boundary of the support set. This in turn requires choosing a smaller k relative to M as the dimension d grows.

3.4.1.4 Optimal rate of convergence

We note that for high dimensions ($d > 6$), $N_{opt} = o(M_{opt})$, which 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 $c_5(1/T)(1 + o(1))$.

3.4.2 Extension to divergence and MI estimation

Finally, we note that boundary corrected plug-in estimators can be used in an identical manner for divergence and mutual information estimation as well. The bias again is reduced from $O(k/M)^{1/d}$ to $O(k/M)^{2/d}$. The variance and CLT are identical

to Theorems 2.5, 2.6, 2.8 and 2.9 because they differ on a set of probability $o(1)$. We do not repeat the results here.

3.4.3 Bias correction factors

In addition to using boundary corrected BP-kNN estimators, we can once again apply bias correction factors when estimating entropy, divergence and MI. Denote the BP-kNN plug-in estimator with bias correction as

$$\tilde{\mathbf{G}}_{k,BC}(f) \quad = \quad \frac{\tilde{\mathbf{G}}_k(f) - g_2(k,M)}{g_1(k,M)}. \tag{3.21}$$

In addition to the assumptions listed in section 2.3.1, assume the growth condition that $k = \Theta((\log(M))^{2/(1-\delta)})$ instead of the condition that $k = \Theta(M^\beta)$. Below the asymptotic bias and variance of the BP-kNN estimator with boundary and bias correction are specified.

Theorem III.2. *The bias of the BPI estimator $\tilde{\mathbf{G}}_{k,BC}(f)$ is given by*

$$\mathbb{B}[\tilde{\mathbf{G}}_{k,BC}(f)] \quad = \quad c_1 \left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{1/d}\right). \tag{3.22}$$

Proof. The proof trivially follows by combining the proofs of Theorems 2.1 and 3.1. □

Denote the boundary corrected Shannon and Rényi entropy estimators with correction factors by $\breve{H}_k$ and $\breve{H}_k^{(\alpha)}$. Let $\tilde{\mathbf{H}}_k$ be the boundary corrected Shannon entropy estimate $\tilde{G}_k(f)$ with the choice of functional $g(x) = -\log(x)$. Let $\tilde{\mathbf{I}}_k^{(\alpha)}$ be the boundary corrected estimate of the Rényi α-integral estimate $\tilde{G}_k(f)$ with the choice of functional $g(x) = x^{\alpha-1}$.

Define $\breve{\mathbf{H}}_k = \tilde{\mathbf{H}}_k + [\log(k-1) - \Psi(k-1)]$ and $\breve{\mathbf{I}}_k^{(\alpha)} = [(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}]^{-1}\tilde{\mathbf{I}}_k^{(\alpha)}$. Also define the Rényi entropy estimator to be $\breve{\mathbf{H}}_k^{(\alpha)} = (1-\alpha)^{-1}\log(\breve{\mathbf{I}}_k^{(\alpha)})$.

We note that the estimators $\breve{\mathbf{H}}_k$ and $\breve{\mathbf{H}}_k^{(\alpha)}$ correspond to data-split boundary corrected versions of the Shannon and Rényi entropy estimators of Goria *et al.* [32] and Leonenko *et al.* [49] respectively.

Because $[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}] \to 1$ and $\Psi(k-1) - \log(k-1) \to 0$ as $k \to \infty$, the estimators $\breve{\mathbf{H}}_k$ and $\breve{\mathbf{H}}_k^{(\alpha)}$ will have identical variance up to leading terms as $\tilde{\mathbf{H}}_k$ and $\tilde{\mathbf{H}}_k^{(\alpha)}$ respectively. Likewise, $\breve{\mathbf{H}}_k$ and $\breve{\mathbf{H}}_k^{(\alpha)}$, when suitably normalized, will converge to the same distribution as the estimators $\tilde{\mathbf{H}}_k$ and $\tilde{\mathbf{H}}_k^{(\alpha)}$ respectively.

From Theorem 3.1, it immediately follows that the bias of the estimators $\tilde{\mathbf{H}}_k$ and $\tilde{\mathbf{H}}_k^{(\alpha)}$ is $O((k/M)^{2/d} + (1/k))$. On the other hand, from the results we have established in Chapter 3 in conjunction with the results of Liitiäinen *et al.* (Theorem 2.1, [51]), it is clear that the bias of the estimators $\breve{\mathbf{H}}_k$ and $\breve{\mathbf{H}}_k^{(\alpha)}$ is given by

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

and

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

In this case, the optimal choice of k needs to only grow logarithmically with M (necessary condition for the result on variance to go through). The minimum bias for optimal choice of k therefore reduces from $O(M^{-2/(2+d)})$ to $O(M^{-2/d})$. The optimal choice of N in this case is given by $N_{opt} = \Theta(M(\log(M)/M)^{(1-1/d)})$.

As in the case of Shannon and Rényi entropy estimation, it is straightforward to define boundary corrected divergence and MI estimators with correction factors in an identical manner to reduce estimator bias from $O(M^{-2/(2+d)})$ to $O(M^{-2/d})$.

3.5 Experiments

We consider three sets of experiments. The first set verifies our boundary correction algorithm. The second set of experiments verifies the theoretical results on the bias, variance and central limit theorem for Shannon entropy and MI estimation. The final set compares the performance of the proposed boundary corrected estimator with other estimators in literature.

3.5.1 Boundary correction

We consider the problem of Shannon entropy estimation (with the choice of functional $g(u, v) = -\log u$) for a 2-dimensional distribution. To estimate the Shannon entropy, we use the standard BP-kNN estimator $\hat{\mathbf{G}}_k(f)$ (2.1) and the boundary corrected BP-kNN estimator $\tilde{\mathbf{G}}_k(f)$ (3.19). We consider two different types of densities: (i) Uniform distribution and (ii) a 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 $p = 0.8$. 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 for both the uncorrected and the boundary corrected plug-in estimator. This is shown in Fig. 3.3.

From the figure, it is clear that the bias corrected entropy estimator $\tilde{\mathbf{G}}_k(f)$(3.19) has significantly lower bias in the case of the uniform distribution, while 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 this mixture density, the fraction of boundary points is very small, thereby minimizing the influence of the boundary regions on the entropy estimate.

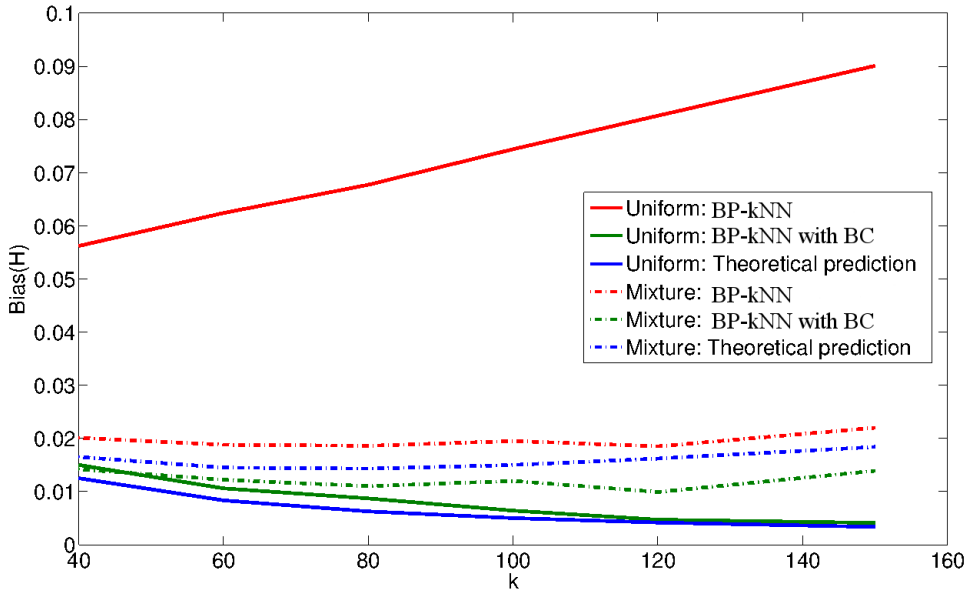


Figure 3.3: Variation of bias of estimated entropy vs bandwidth k using standard BP-kNN estimator $\hat{\mathbf{G}}_k(f)$ (2.1) and the boundary corrected BP-kNN estimator $\tilde{\mathbf{G}}_k(f)$ (3.19), denoted as 'BP-kNN' and 'BP-kNN with BC' respectively. The boundary corrected BP-kNN estimator clearly reduces bias in the entropy estimate in comparison to the uncorrected estimator for the uniform density. The boundary effects are negligible for the mixture density because of the small fraction of points at the boundary for the mixture density.

3.5.2 Experimental validation of theory for Shannon entropy

Here the theory established in Section 3 and Section 4 is validated. A three dimensional vector $\underline{X} = [X_1, X_2, X_3]^T$ was generated on the unit cube according to the i.i.d. Beta plus i.i.d. uniform mixture model:

$$f(x_1, x_2, x_3) = (1 - \epsilon) \prod_{i=1}^{3} f_{a,b}(x_i) + \epsilon, \tag{3.23}$$

where $f_{a,b}(x)$ is a univariate Beta density with shape parameters a and b. For the experiments the parameters were set to $a = 4, b = 4$, and $\epsilon = 0.2$. The Shannon entropy $(g(u) = -\log(u))$ is estimated using the BP-kNN estimator with boundary correc-

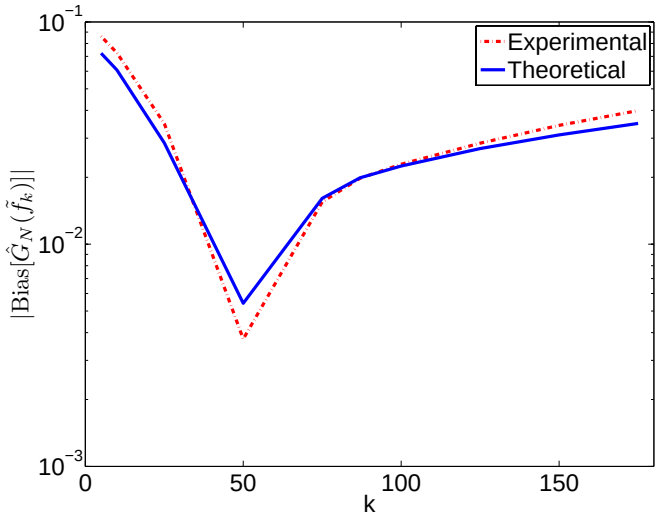


Figure 3.4: Comparison of theoretically predicted bias of plug-in estimator $\tilde{\mathbf{G}}_k(f)$ (3.19) against experimentally observed bias as a function of k. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the BP-kNN estimator $\tilde{\mathbf{G}}_k(f)$ on $T = 10^4$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (3.23). N, M were fixed as $N = 3000$, $M = 7000$ respectively. The theoretically predicted bias agrees well with experimental observations. The predictions of our asymptotic theory therefore extend to the finite sample regime. The theoretically predicted optimal choice of $k_{opt} = 52$ also minimizes the empirical bias.

tion: $\tilde{\mathbf{G}}_k(f)$ and BP-kNN estimator with boundary correction and bias correction: $\tilde{\mathbf{G}}_{k,BC}(f)$.

In Fig. 3.4, the bias approximations of Theorem III. 1 are compared to the empirically determined estimator bias of $\tilde{\mathbf{G}}_k(f)$. N and M are fixed as $N = 3000$, $M = 7000$. Note that the theoretically predicted optimal choice of $k_{opt} = 52$ minimizes the experimentally obtained bias curve. Thus, even though our theory is asymptotic it provides useful predictions for the case of finite sample size, specifying bandwidth parameters that achieve minimum bias. Further note that by matching rates, i.e. choosing $k = \bar{k} = M^{2/(2+d)} = 83$ also results in significantly lower MSE when compared to choosing k arbitrarily ($k < 10$ or $k > 150$). In Fig. 3.5, the bias approximations of Theorem IV. 1 are compared to the empirically determined estimator bias of $\tilde{\mathbf{G}}_{k,BC}(f)$. Observe that the empirical bias, in agreement with the bias

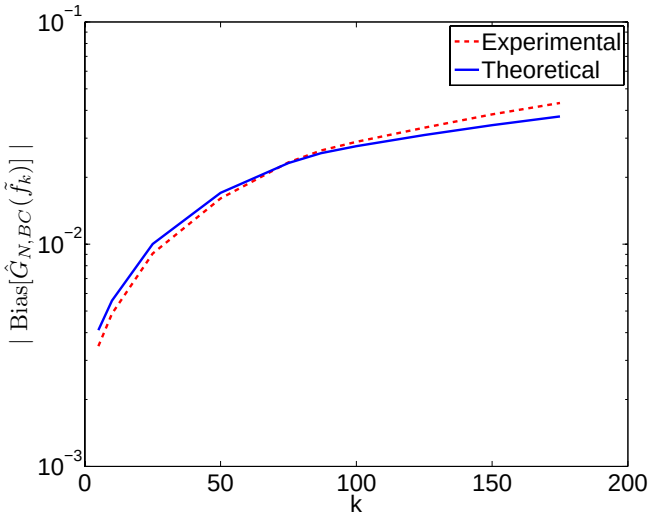


Figure 3.5: Comparison of theoretically predicted bias of the bias corrected estimator $\tilde{\mathbf{G}}_{k,BC}(f)$ (3.21) against experimentally observed bias as a function of k. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the proposed BP-kNN estimator with Bias correction $\tilde{\mathbf{G}}_{k,BC}(f)$ on $T = 10^4$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (3.23). N, M were fixed as $N = 3000$, $M = 7000$ respectively. The empirical bias is in agreement with the bias approximations of Theorem 3.2 and monotonically increases with k.

approximations of Theorem IV. 1, monotonically increases with k.

In Fig. 3.6, the empirically determined variance of $\tilde{\mathbf{G}}_k(f)$ is compared with the variance expressed by Theorem III. 2 for varying choices of N and M, with fixed $N + M = 10,000$. The theoretically predicted variance agrees well with experimental observations. A Q-Q plot of the normalized BP-kNN estimate $\tilde{\mathbf{G}}_k(f)$ and the standard normal distribution is shown in Fig. 3.7. The linear Q-Q plot validates the Central Limit Theorem 3.3. For Shannon entropy ($g(u) = -\log(u)$), the uncompensated and compensated BP-kNN estimators are related by

$$\tilde{\mathbf{G}}_{k,BC}(f) = \tilde{\mathbf{G}}_k(f) + \log(k-1) - \psi(k).$$

The variance and normalized distribution of these estimators are therefore identical.

Finally, using the CLT, the 95% coverage intervals of the BP-kNN estimator

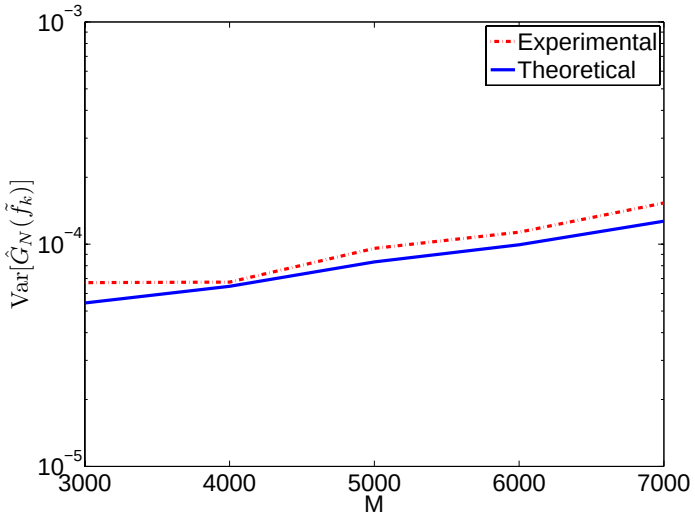


Figure 3.6: Comparison of theoretically predicted variance of BP-kNN estimator $\tilde{\mathbf{G}}_k(f)$ against experimentally observed variance as a function of M. The Shannon entropy $(g(u) = -\log(u))$ is estimated using the proposed BP-kNN estimator $\tilde{\mathbf{G}}_k(f)$ on $T = 10^4$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (3.23). k is chosen to be $k_{opt} = k_0 M^{2/(2+d)}$. The theoretically predicted variance agrees well with experimental observations.

$\tilde{\mathbf{G}}_{k,BC}(f)$ are shown as a function of sample size T in Fig. 3.8. The lengths of the predicted confidence intervals are accurate to within 12% of the true confidence intervals (determined by simulation over the range of 80% to 100% coverage - data not shown). These coverage intervals can be interpreted as confidence intervals on the true entropy, provided that the constants $c_1, .., c_5$ can be accurately estimated.

3.5.3 Experimental validation of theory for Shannon MI

We estimated the Shannon MI of a 2 dimensional beta distribution f_{12} with parameters $\alpha = 2$, $\beta = 2$ using the BP-kNN estimator $\tilde{\mathbf{G}}_k(f_{12})$ and compared our theoretical predictions with the observed bias and variance. In the first experiment, we fixed N to be 1000 and varied M. For each value of M, we optimized the kernel width k according to Eq.3.20. The variation of the bias of the estimator with changing M is shown in Fig. 3.9. In the next experiment, we fixed M to be 10000, chose

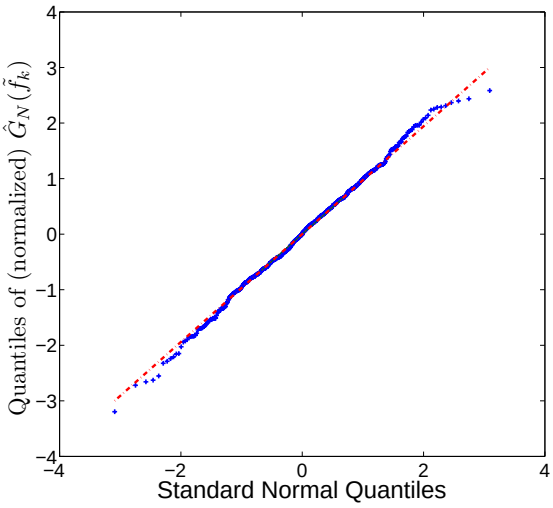


Figure 3.7: Q-Q plot comparing the quantiles of the BP-kNN estimator $\tilde{\mathbf{G}}_k(f)$ (with $g(u) = -\log(u)$) on the vertical axis to a standard normal population on the horizontal axis. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the proposed BP-kNN estimator $\tilde{\mathbf{G}}_k(f)$ on $T = 10^4$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (3.23). k, N, M are fixed as $k = k_{opt} = 52$, $N = 3000$ and $M = 7000$ respectively. The approximate linearity of the points validates our central limit theorem II.3.

the corresponding optimal value of k and varied N. The variation of the variance of the estimator against N is shown in Fig. 3.10. The proximity of the theoretical and empirical curves in these experiments validates our theory.

We performed the Kolmogorov-Smirnov test on the estimated MI, which resulted in the null hypothesis that the MI estimate could have the normal distribution. We generated a Q-Q plot of the MI estimate against the normal distribution. The resulting plot shown in Fig. 3.11 is linear, validating our theory on the asymptotic normal distribution of the plug-in estimates.

In the final experiment, we consider a mixture density $f_m = pf_\beta + (1-p)f_u$, where f_β is a beta distribution with parameters $\alpha = 2$, $\beta = 2$, f_u is a uniform density and p is the mixing ratio. We vary the mixing ratio p and evaluate the MI. The variation of the true MI and estimated MI with p is shown in Fig. 3.12 along with the 95%

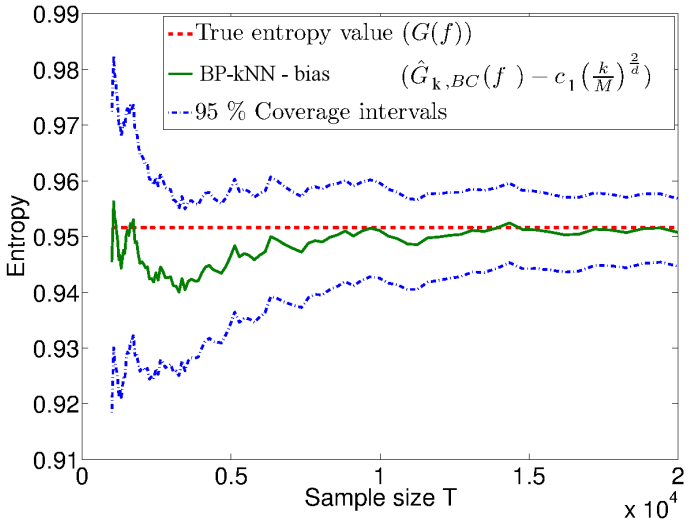


Figure 3.8: 95% coverage intervals of BP-kNN estimator $\tilde{\mathbf{G}}_k(f)$, predicted using the Central limit theorem II.3, as a function of sample size T. The Shannon entropy $(g(u) = -\log(u))$ is estimated using the proposed BP-kNN estimator $\tilde{\mathbf{G}}_{k,BC}(f)$ on T i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (3.23). The lengths of the coverage intervals are accurate to within 12% of the empirical confidence intervals obtained from the empirical distribution of the BP-kNN estimator.

confidence intervals using Theorem II.9. We find the estimated MI to lie within the confidence interval predicted by our theory.

3.5.4 Comparison to existing results

By using boundary correction, we are able to reduce the optimal bias of our estimator $\tilde{\mathbf{G}}(f)$ from $\Theta(T^{-1/(1+d)})$ to $\Theta(T^{-2/(2+d)})$. The overall optimal bias and variance of $\tilde{\mathbf{G}}(f)$ is therefore given by $\Theta(T^{-2/(2+d)})$ and $\Theta(T^{-1})$ respectively. Our estimator therefore has a faster rate of convergence in MSE $\Theta(T^{-4/(2+d)})$ as compared to the estimator of Baryshnikov *et al.* $(\Theta(T^{-2/(1+d)}))$.

Furthermore, when estimating Shannon and Rényi entropy, we can use correction factors to define estimators $\check{\mathbf{H}}_k$ and $\check{\mathbf{H}}_k^{(\alpha)}$ (see section 3.4.3) with bias given by $\Theta(T^{-2/(d)})$ instead of $\Theta(T^{-2/(2+d)})$. It is clear that the estimators $\tilde{\mathbf{H}}_k$ and $\tilde{\mathbf{H}}_k^{(\alpha)}$ have a faster rate of convergence as compared to Shannon and Rényi entropy estimators

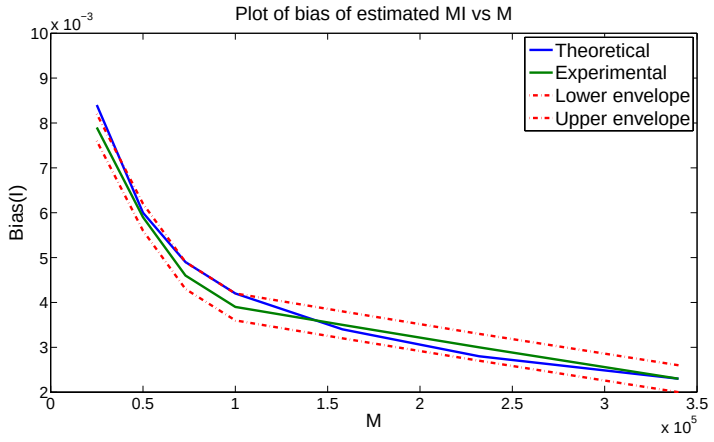


Figure 3.9: Variation of bias of BP-kNN estimator $\tilde{\mathbf{G}}_k(f_{12})$ vs M for fixed N = 1000 with ±95% confidence envelopes. The theoretically predicted bias agrees well with experimental observations.

of Goria *et al.* [32] and Leonenko *et al.* [49], which have bias of order $\Theta(T^{-1/(d)})$ due to bias from the boundary [51].

This is illustrated by the following experiment. We estimate the Rényi α-entropy for the choice $\alpha = 0.5$ for the 5-dimensional 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 $p = 0.8$ using Baryshnikov's estimator $\hat{\mathbf{G}}_b(f)$ with the choice of functional $g(u) = u^{\alpha-1}$, our data-split boundary-corrected estimator with correction factor $\breve{\mathbf{H}}_k^{(\alpha)}$, the entropic graph estimator of Hero *et al.* [38] and the k-nearest neighbor estimator of Leonenko *et al.* [32] with correction factors.

The results are shown in Fig. 3.13. It is clear from the figure that our data-split boundary-corrected estimator $\tilde{\mathbf{G}}(f)$ has a faster rate of convergence as predicted by our theory.

In the next section, we apply our thinned k-NN graphs to the problem of classification and compare the resulting performance with the standard k-NN classification algorithm.

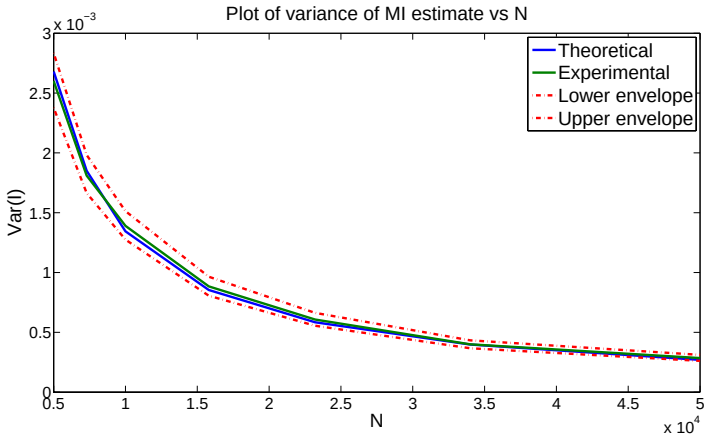


Figure 3.10: Variation of variance of BP-kNN estimator $\tilde{\mathbf{G}}_k(f_{12})$ vs N for fixed M = 10000 and bandwidth k = 411 with $\pm 95\%$ confidence envelopes. The theoretically predicted variance again agrees well with experimental observations.

3.6 Boundary compensated graphs

Our compensated k-NN density estimates can be extended to modify bipartite k-NN graphs as follows. This general k-NN graph compensation method is then illustrated for k-NN classification.

3.6.1 Relation between k-NN density estimate and k-NN graphs

Let $\mathbf{X}_1, .., \mathbf{X}_M$ denote M i.i.d realizations of the density f. Consider a k-NN graph constructed on these M samples. We therefore have an *equivalence* relation between a k-NN graph, and the k-NN density estimates constructed using the graph. To correct for boundary effects in the graph, we first analyze boundary effects in the k-NN density estimates and then use this equivalence to specify corrections to the graph.

3.6.2 Thinning k-NN graphs

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

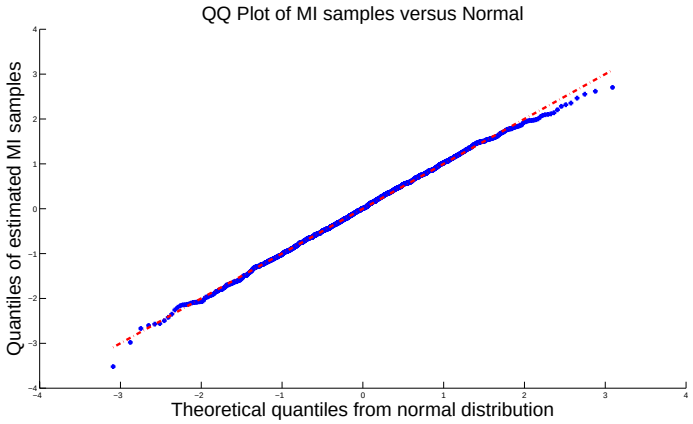


Figure 3.11: Q-Q plot of normalized BP-kNN estimate $\tilde{\mathbf{G}}_k(f_{12})$ and standard normal distribution. The approximate linearity of the points validates our central limit theorem.

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

3.6.3 k-NN classification

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

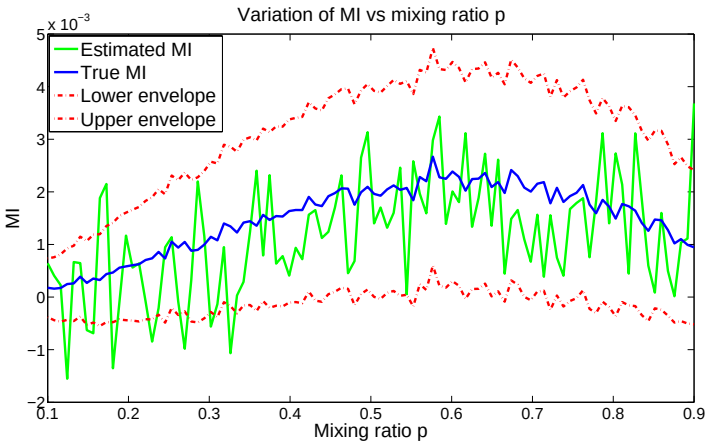


Figure 3.12: Variation of BP-kNN MI estimator $\tilde{\mathbf{G}}_k(f_{12})$ with mixing ratio p with $\pm 95\%$ confidence envelopes. Observe that the the estimated MI lies within the confidence interval predicted by our theory.

A simple example

We consider a simple example where 4 concentric 2D rings constitute 4 different classes of data. Each class consists of 400 samples. The confusion matrix (using the leave-one-out criteria) for the uncompensated and the compensated classifier ($k = 100$) is shown in Table 3.1.

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

Optical digit recognition

The 'Optical Recognition of Handwritten Digits Data Set' [3] consists of normalized bitmaps of handwritten digits from a preprinted form. This data set has 562 instances of each digit from $0 - 9$. Each instance is characterized by 64 dimensional pixel intensity values. As a first step, we use standard PCA embedding to reduce the dimension to 10. We then normalize these 10 dimensional vectors to unit length.

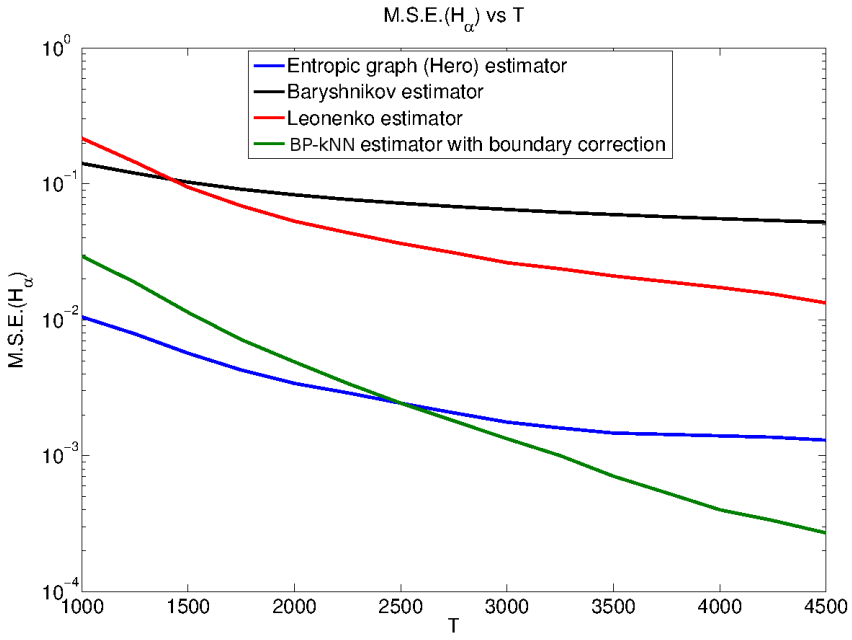


Figure 3.13: Variation of MSE of entropic graph estimator of Hero *et al.* [38], the k-nearest neighbor estimator of Leonenko *et al.* [32] and the k-nearest neighbor estimator of Baryshnikov *et al.* [6] and boundary-corrected BP-kNN estimator with correction factor $\breve{\mathbf{H}}_k^{(\alpha)}$ as a function of sample size T. From the figure we see that our estimator, in agreement with theory, has the fastest rate of convergence.

We treat the first 9 dimensions of each normalized vector as our feature vectors f_i. We note that the feature vectors f_i live in a unit hypercube in $\mathbf{R}^9$. A significant fraction of the feature vectors f_i will lie close to the surface of the hypercube, thereby behaving as boundary points. We apply the standard and boundary compensated k-NN classifiers ($k = 25$) to this data. The confusion matrix for the uncompensated and the compensated classifier is shown in Table 3.2. The leave-one-out classification error for the uncompensated classifier was found to be 4.59% and improved to 3.59% for the compensated classifier. Using a paired t-test, the p-value for this result was found to be well within a siginificance level of 1%, implying that the improvement in performance is indeed statistically significant.

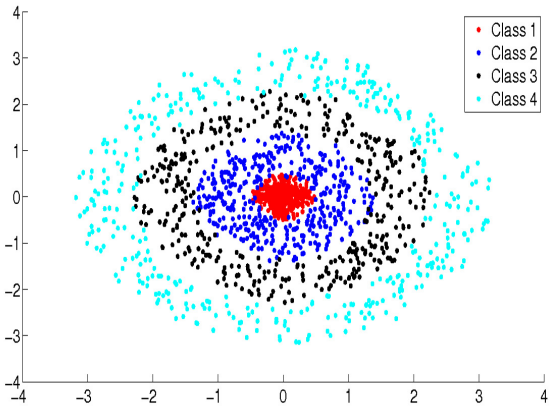


Figure 3.14: Data with four different classes, which lie on concentric circular discs.

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

Table 3.1: Confusion matrix for concentric circle data (Black: Standard k-NN graph; Blue: Boundary compensated k-NN graph).

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

Table 3.2: Confusion matrix for 'Handwritten Digits' dataset (Black: Standard k-NN graph; Blue: Boundary compensated k-NN graph).

3.7 Discussion

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

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

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

Finally, we note that in Section 6.5, another variant of boundary compensation based on spherical sector k-NN neighborhoods is proposed. Estimators with this *angular* variant of boundary compensation, in contrast to the extrapolation based compensation proposed in this Chapter, have a slower MSE rate of convergence (($O(T^{-1/(1+d)})$) vs ($O(T^{-2/(2+d)})$)). However, under higher order smoothness conditions on the density, we show in Chapter 6 that these angular estimators can be aggregated to produce ensemble estimators with much faster MSE rates of convergence ($O(1/T)$).

CHAPTER IV

Functional estimation on Manifolds

4.1 Introduction

In this chapter, we extend the results of Chapter 2 on entropy and divergence estimation from data which lies in $\mathbb{R}^d$ to data which lies on a possible non-linear manifold embedded in $\mathbb{R}^d$ having with smaller intrinsic dimension $d < D$. We first motivate the necessity of analyzing data on manifolds. Next, we extend the estimators defined in Chapter 2 for entropy and divergence estimation to the manifold setting and again analyze the MSE and obtain an asymptotic distribution. Finally, we extend our results on entropy estimation to propose a MSE optimized intrinsic dimension estimator.

Recent technological advancements in sensing and data storage have facilitated acquisition of large datasets which are very high dimensional in nature. High dimensional data by default is difficult to model and analyze because of the curse of dimensionality. However, often, there is underlying structure in this high dimensional data. This structure corresponds to redundancy. For example, consider the abilene network data where the extrinsic dimension = #(Routers). However, the data itself lies on a lower dimensional manifold. Practically, the complexity of capturing an high dimensional manifold is not possible. Often, all that is available are a finite number of representative samples of the manifold. We therefore analyze the manifolds via a

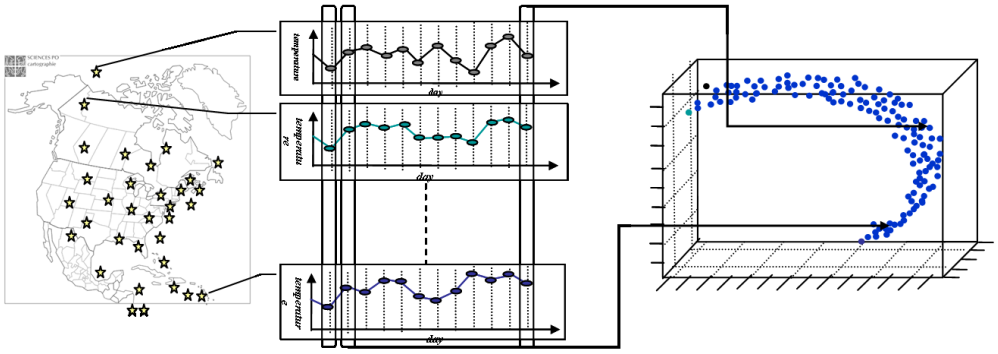


Figure 4.1: Illustration of the Abilene router network. The extrinsic dimension of this system at each time point is equal to the number of routers.

finite number of (possibly random) samples drawn from the manifold.

Specifically, we are interested in determining characteristics of the manifold such as its intrinsic dimension and entropy of the densities defined on manifolds.

4.2 Definition of a manifold

In this chapter, we are interested in manifolds $\mathcal{M}$ embedded in $\mathbb{R}^D$, which are topological spaces locally homeomorphic to the Euclidean space. A manifold is locally homeomorphic to Euclidean space if every point on the manifold has a neighborhood homeomorphic to an open Euclidean n-ball. We will assume throughout this chapter that the manifold does not have boundaries, i.e., for every for point $x \in \mathcal{M}$ there exists an open ball $U(x)$ centered at x and $U(x) \subset \mathcal{M}$.

Given a smooth compact manifold $\mathcal{M}$, a Riemann metric m is a mapping which associates to each point $x \in \mathcal{M}$, an inner product $m(.,.)$ between vectors tangent to $\mathcal{M}$ at x. A Riemann manifold $(\mathcal{M}, m)$ is the ordered pair of the manifold $\mathcal{M}$ with the metric m. When the underlying space in which the manifold $\mathcal{M}$ is embedded is $\mathbb{R}^D$, the naturally induced Riemann metric is just the usual dot product. A Riemann metric m endows the manifold $\mathcal{M}$ with a distance $d_m(.,.)$ via geodesics and a measure μ_m via the volume element. Throughout this chapter, we will assume that the manifold is embedded in $\mathbb{R}^D$ with the Riemann metric m being the dot product. For

a comprehensive explanation of the theory of Riemann manifolds, we refer the reader to [48].

We can now define k-nearest neighbors in a given sample in the following usual way via the geodesic distance $d_m(.,.)$: Given a sample $\{X_1,..,X_M\} \in \mathcal{M}$, the k-th nearest neighbor of a point $X \in \mathcal{M}$ to the sample $\{X_1,..,X_M\}$ is defined to be the point $X_{k(i)}$ which satisfies the following condition $\#\{d_m(X,X_i) < d_m(X,X_{k(i)}), i = 1,..,M\} = k-1$. Note that in the case where the manifold $\mathcal{M}$ is the extrinsic space $\mathbb{R}^D$, $X_{k(i)}$ is the standard k-th nearest neighbor with respect to the Euclidean distance.

We can now extend results established in Chapter 2 on entropy and divergence estimation to the case of manifolds. In order to extend our results, we use the fact that a Riemann manifold $\mathcal{M}$ with associated distance d_m and measure μ_m, looks locally like $\mathbb{R}^d$ with euclidean distance $||.||$ and Lebesgue measure λ. This fact is formalized by the following lemma:

Lemma IV.1. *([62], Lemma 5.1) Let $(\mathcal{M}, m)$ be a smooth Riemann d-dimensional manifold. For any $x \in \mathcal{M}$ and $\epsilon > 0$, there exists a chart $(\mathcal{U}, \phi)$ for $\mathcal{M}$, with $x \in \mathcal{U}$, such that*

$$(1+\epsilon)^{-1}|\phi(y)-\phi(z)| \leq d_m(y,z) \leq (1+\epsilon)|\phi(y)-\phi(z)| \forall y,z \in \mathcal{U}$$

and for any measurable subset $B \in \mathcal{U}$

$$(1-\epsilon)\lambda(\phi(B)) < \mu_m(B) < (1+\epsilon)\lambda(\phi(B)).$$

Recall that a chart $(\mathcal{U}, \phi)$ consists of a neighborhood $\mathcal{U}$ in $\mathcal{M}$ and a mapping $\phi : \mathcal{M} \cap \mathcal{U} \to \mathbb{R}^d$ that represents points in $\mathcal{M} \cap \mathcal{U}$ as points in the Euclidean d-dimensional space, i.e., for $y \in \mathcal{M} \cap \mathcal{U}$, $\phi(y)$ represents y in an Euclidean d-dimensional

coordinate system. For a chart $(\mathcal{U}, \phi)$, define the radial distance function r by

$$r(x, y) := \left(\sum_{i=1}^{n} (x^i - y^{(i)})^2 \right)^{1/2}$$

where $x, y \in \mathcal{M} \cap \mathcal{U}$ are two points in the neighborhood $\mathcal{U}$ and $\{x^{(i)}\}$ (respectively $\{y^{(i)}\}$) represent the d-dimensional coordinates of x (receptively y) in the mapped space $\phi(\mathcal{U})$.

4.2.1 Normal coordinate chart

We now further restrict our attention to normal coordinate charts, which we describe next. For any neighborhood $\mathcal{U}$ of a point $p \in \mathcal{M}$, it is possible to obtain a *normal coordinate chart* $\phi : \mathcal{U} \to \mathbb{R}^d$ which is a local coordinate system in a neighborhood of p obtained by applying the exponential map to the tangent space at p. The normal coordinate chart $\phi(.)$ satisfies the following properties : (i) (Proposition 5.11, [48]) the coordinates of the point p in the chart are $(0, \ldots, 0)$, i.e., $\phi(p) = (0, \ldots, 0)$ and (ii) (Corollary 6.11 [48]) for any other point $x \in \mathcal{U}$, the radial distance $r(p, x)$ *equals* the Riemannian distance from p to x.

We will exploit the fact that the radial distance on the mapped space is equal to the Riemannian distance on the manifold in order to extend the convergence results established in Chapter 2 to manifolds.

4.3 Functional estimation on manifolds

Let $f(X)$ be a density on the manifold, i.e. $f(X)$ is a non-negative function defined on $\mathcal{M}$ which satisfies $\int_{\mathcal{M}} f(x)dx = 1$. As in Chapter 2, we are interested in estimating entropy functionals of the form $G(f) = \int_{\mathcal{M}} g(f(x))f(x)dx$ and divergence functionals of the form $G(f_1, f_2) = \int_{\mathcal{M}} g(f_1(x)/f_2(x))f_2(x)dx$, for some smooth functions $g(f(x), x)$.

In this chapter, we restrict our attention to entropy estimation of functionals $G(f) = \int_{\mathcal{M}} g(f(x))f(x)dx$. We require that the density f be uniformly bounded away from 0 and finite on the set $\mathcal{S}' \in \mathcal{S}$ where $\mathcal{S}$ is the support of the density, i.e., there exist constants ϵ_0, ϵ_∞ such that $0 < \epsilon_0 < \epsilon_\infty < \infty$, with $\epsilon_0 \leq f(x) \leq \epsilon_\infty \; \forall x \in \mathcal{S}$. 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. Let $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ be T independent and identically distributed sample realizations in $\mathbb{R}^D$ distributed according to density f. Any realization $\mathbf{X}_i$ is constrained to lie on the d-dimensional Riemannian submanifold $\mathcal{M}$ of $\mathbb{R}^D$ $(d < D)$.

4.3.1 k-NN density estimation on manifolds

Let $d_m(X, Y)$ denote the Riemannian geodesic distance between points X and Y and $\mathbf{d}_{k,g}(X)$ denote the geodesic distance between a point X and its k-th nearest neighbor $\mathcal{N}_k(X)$ amongst $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$. The standard k-NN density estimator [53] on the manifold using the geodesic distance is defined as

$$\hat{\mathbf{f}}_{k,g}(X) = \frac{k-1}{Mc_d \mathbf{d}_{k,g}^d(X)},$$

where c_d is the unit Euclidean ball volume in d-dimensions.

4.3.2 Properties of k-NN density estimates on manifolds

Let $\phi : \mathcal{U} \to \phi(\mathcal{U})$ be the normal chart map at X and note that $\phi(X) = 0$. Let us map all random variables that fall in $\mathcal{U}$ via the normal chart $\phi(.)$, and map the rest of the points to some arbitrary points outside $\phi(\mathcal{U})$. Then the probability density function induced in $\mathcal{V} := \phi\mathcal{U}$ is given by $p(v) = \phi(f(u)) = \sqrt{|g(u)|} f(\phi^{-1}(u))$, where $g(u)$ is the local representation of the Riemannian metric at u, and $|g(u)|$ is its determinant. Let $\hat{\mathbf{p}}_k(0)$ be the k-NN density estimator in $\mathcal{V}$ at the point X, in terms of Euclidean distance.

In this normal coordinate system, the Euclidean distance from 0 to u, $||u||$, is equal to the Riemannian distance from p to $\phi^{-1}(u)$. Therefore as long as we have at least k random points in $\mathcal{U}$, the two estimators are equal, i.e.,

$$\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X) = \hat{\mathbf{p}}_k(0).$$

In section 3.3.1, we showed that for any positive function $q(k, M)$ satisfying $q(k, M) = \Theta((k/M)^{2/d}) + 1/k^{(1-\log\log k/\log k)/2}$, the following concentration inequality holds

$$1 - Pr\left(\left|\frac{\mathbf{V}_{k,M}(0)}{V_{k,M}(0)} - 1\right| \leq q(k, M)\right) \leq o(M^{-a}),$$

where $V_{k,M}(0)$ is the k-NN ball centered at $\phi(X) = 0$ in the mapped space $\mathcal{V}$. This implies that for sufficiently small values of k/M and $1/k$, the probability that $\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X) \neq \hat{\mathbf{p}}_k(0)$ decays exponentially fast in M.

Using this result, we can now extend results on the moments of k-NN density estimates on the manifolds by determining the corresponding moments of the Euclidean k-NN density estimates on the mapped space $\mathcal{V}$. We will assume throughout this chapter that the manifold does not have boundaries.

4.3.3 Moments of k-NN density estimate

By the concentration inequality, we know that with with high probability, the density estimates are equal. This in turn implies that the bias and variance of the k-NN density estimates on the manifold are 'close' to equal.

4.3.3.1 Bias

$$\begin{aligned}
\mathbb{B}[\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X)] &= \int_{\mathcal{M}^M} (\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X) - f(X)) \prod_{i=1}^{M} (f(\mathbf{x}_i) dx_i) \\
&= \int_{\mathcal{U}^M} (\hat{\mathbf{p}}_k(0) - p(0)) \prod_{i=1}^{M} (p(\phi^{-1}(\mathbf{x}_i)) dy_i) + o(1/M^a) \\
&= \mathbb{B}[\hat{\mathbf{p}}_k(0)] + o(1/M^a) \\
&= h(X)(k/M)^{2/d} + o((k/M)^{2/d}).
\end{aligned}$$

where $h(X)$ is given by

$$h(X) = |g(0)|^{-1/d} p^{-2/d}(0) \sum_i \frac{\partial^2}{\partial u_i^2}|_{u=0} \left(f(\phi^{-1}(u))) \right).$$

The second step in the derivation follows from the concentration inequality (4.1) and the last step follows from the properties of standard k-NN density estimates (see section B.3.3).

The determinant $|g(u)| = 1 - \sum_{i,j}(R_{ij}/3)u_i u_j + O(|u|^3)$ where R_{ij} is the Ricci curvature. We also have,

$$\frac{\partial}{\partial u_i}|g(u)| \,|_{u=0} = 0$$

for all i and

$$\frac{\partial^2}{\partial u_i^2}|g(u)| \,|_{u=0} = -\frac{2}{3} R_{ii}$$

which gives c_1 to be

$$\begin{aligned}
h(X) &= |g(0)|^{-1/d} p^{-2/d}(0) \sum_i \frac{\partial^2}{\partial u_i^2}|_{u=0} \sqrt{|g(u)|} \left(f(\phi^{-1}(u))) \right) \\
&= f^{-2/d}(p) \left(\frac{\partial^2}{\partial u_i^2}|_{u=0} \left(f(\phi^{-1}(u))) \right) - \frac{1}{3} R_{ii} \right) \\
&= f^{-2/d}(p)(\nabla(f(p) - S(p)/3)
\end{aligned}$$

where the last line follows from the fact that (i) $\sum_i \partial^2(f\phi^{-1})(0) is \nabla f(p)$ where $\nabla(.)$ is the Laplace-Beltrami operator since $\phi(.)$ is the normal chart map of p and (ii) since the standard basis of U maps back to an orthonormal basis at p, the sum $\sum_i R_{ii}$ is the scalar curvature $S(p)$.

4.3.3.2 Variance

In an identical manner, we have

$$\begin{aligned}
\mathbb{V}[\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X)] &= \int_{\mathcal{M}^M} (\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X) - f(X))^2 \prod_{i=1}^{M} (f(\mathbf{x}_i) dx_i) \\
&= \int_{\mathcal{U}^M} (\hat{\mathbf{p}}_k(0) - p(0))^2 \prod_{i=1}^{M} (p(\phi^{-1}(\mathbf{x}_i)) dy_i) + o(1/M^a) \\
&= \mathbb{V}[\hat{\mathbf{p}}_k(0)] + o(1/M^a) \\
&= p^2(0)(1/k) + o(1/k) \\
&= f^2(X)(1/k) + o(1/k),
\end{aligned}$$

where the last but one step follows from the properties of Euclidean k-NN density estimates (see appendix B.3.1).

4.3.3.3 Covariance

We now analyze covariance properties of k-NN density estimates on manifolds. Let X and Y be two fixed points on the support $\mathcal{S}$ of f. We first seek to answer the following question: for which set of pair of points $\{X, Y\}$ are the k-NN balls disjoint? Define $\mathbf{e}_{k,g,\mathcal{M}}(X) = \hat{\mathbf{f}}_{k,g,\mathcal{M}}(X) - \mathbb{E}[\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X)]$.

4.3.3.4 Intersecting and disjoint balls

Define spherical regions $S_{k,\mathcal{M}}(X) = \{Y \in \mathcal{M} : d_m(X, Y) \leq r\}$. Let $R_{k,\mathcal{M}}(X)$ correspond to the coverage value $2k/M$, i.e. $R_{k,\mathcal{M}}(X) = inf\{r : \int_{S_{k,\mathcal{M}}(X)} f(Z)dZ =$

$2k/M\}$. Let $P_{\mathcal{M}}(X)$ be the coverage function at X, i.e. $P_{\mathcal{M}}(X) = Pr(\mathbf{Z} \in \{Y \in \mathcal{M} : d_m(X,Y) \leq d_m(X, \mathcal{N}_k(X))\})$. Using the same arguments as in [55], it follows that $P_{\mathcal{M}}(X)$ has a beta distribution with parameters k, $M-k+1$.

Define $\Psi_k := \{X,Y\} \in \mathcal{S} : d_m(X,Y) \geq R_{k,\mathcal{M}}(X) + R_{k,\mathcal{M}}(Y)$. Let $\boldsymbol{\Upsilon}$ denote the event that the k-NN balls on the manifold intersect.

For $\{X,Y\} \in \Psi_k$,

$$\begin{aligned}
Pr(\boldsymbol{\Upsilon}) &= Pr(d_m(X, \mathcal{N}_k(X)) + d_m(Y, \mathcal{N}_k(Y)) \geq ||X-Y||) \\
&\leq Pr(d_m(X, \mathcal{N}_k(X)) + d_m(Y, \mathcal{N}_k(Y)) \geq R_k(X) + R_k(Y)). \\
&\leq Pr(d_m(X, \mathcal{N}_k(X)) \geq R_k(X)) + Pr(d_m(Y, \mathcal{N}_k(Y)) \geq R_k(Y)) \\
&= Pr(\mathbf{P}_{\mathcal{M}}(X) \geq 2k/M) + Pr(\mathbf{P}_{\mathcal{M}}(Y) \geq 2k/M) \\
&= Pr(\mathbf{P}_{\mathcal{V}}(X) \geq 2k/M) + Pr(\mathbf{P}_{\mathcal{V}}(Y) \geq 2k/M) + o(1/M^a) \\
&= o(1/M^a).
\end{aligned}$$

where the last step follows from the chernoff concentration inequality for the coverage function established in section B.1.1.1. We conclude that for $\{X,Y\} \in \Psi_k$, the probability of intersection of k-NN balls centered at X and Y decays exponentially in M.

Disjoint balls For the case where $\{X,Y\} \in \Psi_k$, we then know that with probability $1 - o(1/M^a)$, the k-NN balls are dis-joint. Let $\mathcal{U}_X$ be a geodesic ball around X and $\mathcal{U}_Y$ be a geodesic ball around Y. When the k-NN balls are dis-joint, we can then use two normal chart maps centered at X and Y as $\phi_X : \mathcal{U}_X \to \phi(\mathcal{U})_X$ be the normal chart map at X (similarly for Y) with the distinction that $\phi_X(X) = 0$ and $\phi_Y(Y) = p$ with p chosen to be large enough so that the maps $\phi(\mathcal{U})_X$ and $\phi(\mathcal{U})_Y$ are disjoint. Let us map all random variables that fall in $\mathcal{U}_X$ via the normal chart $\phi_X(.)$ (likewise for Y), and map the rest of the points to some arbitrary points outside $\phi(\mathcal{U})_X \cup \phi(\mathcal{U})_Y$.

Denote this mapping by a function $\phi : \mathcal{M} \to \mathcal{V}$.

Then the probability density function induced in $\mathcal{V} := \phi\mathcal{M}$ is given by $p(v) = \phi(f(u)) = \sqrt{|g(u)|}f(\phi^{-1}(u))$, where $g(u)$ is the local representation of the Riemannian metric at u, and $|g(u)|$ is its determinant. Let $\hat{\mathbf{p}}_k(0)$ be the k-NN density estimator in $\mathcal{V}$ at 0, and $\hat{\mathbf{p}}_k(p)$ be the k-NN density estimator in $\mathcal{V}$ at p in terms of Euclidean distance. As before, we have that these euclidean k-NN density estimates are identical to the k-NN density estimates on the manifold with probability $1 - o(1/M^a)$. Define $\mathbf{o}_k(p) = \hat{\mathbf{p}}_k(p) - \mathbb{E}[\hat{\mathbf{p}}_k(p)]$. Then for $\{X, Y\} \in \Psi_k$, the cross-correlation between the coverage density estimates is given by

$$\begin{aligned} I &:= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta_k}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_k^q(X)\mathbf{e}_k^r(Y)] \\ &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{o}_k^q(0)\mathbf{o}_k^r(p)] + o(1/M^a) \\ &= -1_{\{q=1,r=1\}}\mathbb{E}[\gamma_1(\mathbf{X})f(\mathbf{X})]\mathbb{E}[\gamma_2(\mathbf{Y})f(\mathbf{Y})]\frac{1}{M} + o(1/M). \end{aligned} \tag{4.1}$$

where the last step follows from section on the analysis of cross moments for standard Euclidean k-NN density estimates.

Intersecting balls For the case where $\{X, Y\} \in \Psi_k^c$, we then know that w.h.p, the k-NN balls intersect. When the k-NN balls intersect, we can then let $\phi : \mathcal{U} \to \phi(\mathcal{U})$ be the normal chart map at X and note that $\phi(X) = 0$. Let us map all random variables that fall in $\mathcal{U}$ via the normal chart $\phi(.)$, and map the rest of the points to some arbitrary points outside $\phi(\mathcal{U})$. Then the probability density function induced in $\mathcal{V} := \phi\mathcal{U}$ is given by $p(v) = \phi(f(u)) = \sqrt{|g(u)|}f(\phi^{-1}(u))$, where $g(u)$ is the local representation of the Riemannian metric at u, and $|g(u)|$ is its determinant. Let $\hat{\mathbf{p}}_k(0)$ be the k-NN density estimator in V at 0, in terms of Euclidean distance.

We note that in this case where with high probability the k-NN balls intersect, the k-NN ball around Y is a subset of the normal neighborhood $\mathcal{U}$ at X. Thus the Euclidean k-NN density estimates are identical to the k-NN density estimates on the

manifold with probability $1 - o(1/M^a)$. This implies that

$$\begin{aligned} II &:= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}^q_{k,g,\mathcal{M}}(\mathbf{X})\mathbf{e}^r_{k,g,\mathcal{M}}(\mathbf{Y})] \\ &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})[\mathbf{o}^q{}_k(0)\mathbf{o}^r{}_k(p)]] + o(1/M) \\ &= \mathbb{E}[\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})]\left(\frac{1}{M} + o\left(\frac{1}{M}\right)\right). \end{aligned}$$

where the last step follows from section on the analysis of cross moments for strandard Euclidean k-NN density estimates. This implies that

$$\begin{aligned} &Cov\big[\gamma_1(\mathbf{X})\mathbf{e}^q_{k,g,\mathcal{M}}(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}^r_{k,g,\mathcal{M}}(\mathbf{Y})\big] \\ &= I + II = 1_{\{q=1,r=1\}}Cov[\gamma_1(\mathbf{X})f(\mathbf{X}), \gamma_2(\mathbf{Y})f(\mathbf{Y})]\left(\frac{1}{M} + o\left(\frac{1}{M}\right)\right). \end{aligned}$$

4.3.4 Error between geodesic and euclidean k-NN distances

In the preceding sections, we assumed that we had access to the k-nearest neighbor geodesic lengths on a manifold. From a practical standpoint, one can only compute the Euclidean k-NN points in the extrinsic space.

In this section, we relate and characterize the error between the Euclidean k-NN lengths and geodesic lengths as follows. To this end, we use the following Lemma 4 in [54].

Define the minimum radius of curvature of a manifold $r_0 = r_0(\mathcal{M})$ to be $r_0^{-1} = \max_{\gamma,t} ||\gamma(t)||$ where γ varies over all unit-speed geodesics in $\mathcal{M}$ and t is in the domain D of the geodesic arc γ. Also define the minimum branch separation $s_0 = s_0(\mathcal{M})$ as the largest positive number for which $||x - y|| < s_0$ implies $d_m(x, y) \leq \pi r_0$, for $x, y \in \mathcal{M}$. The existence of r_0 and s_0 is guaranteed by the compactness of $\mathcal{M}$.

Lemma IV.2. *([54], Lemma 4) Let $\lambda > 0$ be given. Suppose the points x, y in $\mathcal{M}$ satisfy the conditions: (i)$||x - y|| < s_0$ and (ii) $||x - y|| \leq (2/\pi)r_0\sqrt{24\lambda}$. Suppose*

that there is a geodesic arc of length $d_m(x,y)$ connecting to x, y. Then

$$(1-\lambda)d_m(x,y) \leq ||x-y|| \leq d_m(x,y).$$

We will use the above lemma along with the characterization of k-NN geodesic lengths to obtain the following lemma.

Lemma IV.3. *For any fixed point $x \in \mathcal{M}$, let $\mathcal{N}_k(x)$ be the k-NN of x among the M samples $\{X_{N+1}, \ldots, X_{N+M}\}$. Let $\mathbf{d}_{k,g}(X)$ and $\mathbf{d}_{k,e}(X)$ be the geodesic distance between x and $\mathcal{N}_k(x)$ and the Euclidean distance respectively. Then, there exists a approximating function $\mathcal{A}_x : \mathbb{R} \to \mathbb{R}$ which depends on x, such that $\mathbf{d}_{k,e}(X) = \mathcal{A}_x(\mathbf{d}_{k,g}(X))$. This function is characterized by the following condition*

$$\mathcal{A}_x(s) = s + \sum_{i=3}^{d} a_x(i)s^i + o(s^d).$$

Proof. For any pair of points x, y on the manifold $\mathcal{M}$, let γ be the unit speed geodesic from x to y. Furthermore, let $\gamma(t) = x$ and $\gamma(t+s) = y$. We then have the geodesic length $d_m(x,y) = s$. We will now compute the Euclidean length as follows.

By Taylor series expansion, we have,

$$\begin{aligned} y - x &= \gamma(t+s) - \gamma(t) \\ &= \sum_{i=1}^{d} \gamma^{(i)}(t)s^i/i! + o(s^d). \end{aligned} \tag{4.2}$$

Also, note that $||\gamma^{(1)}(t)|| = 1$ because $\gamma(.)$ is unit speed. This then gives us that

$$||x-y|| = \mathcal{A}_x(d_m(x,y)) = \mathcal{A}_x(s) = s + \sum_{i=2}^{d} a_x(i)s^i + o(s^d).$$

Finally, we know from 4.1 that with high probability $1 - o(1/M^a)$, the maximum geodesic separation $d_m(x, x_k)$ between k-NN distances is $O((k/M)^{1/d})$. For small

values of k/M, $d_m(x, x_k)$ satisfies condition (i). Setting λ in condition (ii) to be $O((k/M)^{2/d})$ gives us

$$\mathcal{A}_x(d) = 1 + a_x d^2 + o(d^2).$$

These two results give us the required result. □

Let us now define standard k-NN density estimates using the Euclidean lengths on the manifold as

$$\hat{\mathbf{f}}_{k,e,\mathcal{M}}(X) = \frac{k-1}{Mc_d \mathbf{d}_{k,e}^d(X)}. \tag{4.3}$$

Note that we can then write the following relation between the Euclidean and geodesic k-NN density estimates.

$$\hat{\mathbf{f}}_{k,e,\mathcal{M}}(X) = \mathcal{A}_{\|_X}(\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X)), \tag{4.4}$$

where the new approximating function $\mathcal{A}_{\|}$ is given by

$$\mathcal{A}_{\|_x}(f) = f + \sum_{i=2}^{d} \tilde{a}_x(i)(k/M)^{i/d} f^{1-i/d} + o(k/M).$$

This relates the approximate k-NN density estimate using Euclidean distance with the 'exact' k-NN density estimates defined using geodesic distances on the manifold.

4.3.5 Moment properties of Euclidean approximate k-NN density estimates

The moment properties of the k-NN density estimate defined using Euclidean distances $\hat{\mathbf{f}}_{k,e,\mathcal{M}}(X)$ have the same central and cross moment properties, up to the leading terms as the 'exact' k-NN density estimates $\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X)$ defined using geodesic distances on the manifold, while the bias has an additional term. Let $\tilde{h}(X) = h(X) +$

$\tilde{a}_X f^{-2/d}(X)$. Define $\mathbf{e}_{k,e,\mathcal{M}}(X) = \hat{\mathbf{f}}_{k,e,\mathcal{M}}(X) - \mathbb{E}[\hat{\mathbf{f}}_{k,e,\mathcal{M}}(X)]$. We can summarize the results on moment properties of the Euclidean approximate k-NN density estimate as follows:

$$\mathbb{E}[\hat{\mathbf{f}}_{k,e,\mathcal{M}}(X)] - f(X) = \tilde{h}(X)\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d}, \tag{4.5}$$

$$\begin{aligned}&\mathbb{E}\big[\gamma(\mathbf{X})\mathbf{e}^q_{k,e,\mathcal{M}}(\mathbf{X})\big]\\ &= 1_{\{q=2\}}\mathbb{E}\big[\gamma(\mathbf{X})f^2(\mathbf{X})\big]\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right),\end{aligned} \tag{4.6}$$

$$\begin{aligned}&Cov\big[\gamma_1(\mathbf{X})\mathbf{e}^q_{k,e,\mathcal{M}}(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}^r_{k,e,\mathcal{M}}(\mathbf{Y})\big]\\ &= 1_{\{q,r=1\}}Cov[\gamma_1(\mathbf{X})f(\mathbf{X}), \gamma_2(\mathbf{Y})f(\mathbf{Y})]\left(\frac{1}{M}\right)\\ &+ o\left(\frac{1}{M}\right).\end{aligned} \tag{4.7}$$

4.3.6 Main results

Once again, we are interested in estimating functionals of the form

$$G(f) = \int 1_{\{x\in\mathcal{S}'\}} g(f(x), x) f(x) d\mu(x) = \mathbb{E}[1_{\{x\in\mathcal{S}'\}} g(f(x), x)],$$

for some smooth function $g(f(x), x)$ and some subset $\mathcal{S}' \subset \mathcal{S}$ of the support $\mathcal{S}$. Define the plug-in estimators as

$$\hat{\mathbf{G}}_{k,e,\mathcal{M}}(f) \;=\; \left(\frac{1}{N}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{S}'\}} g(\hat{\mathbf{f}}_{k,e,\mathcal{M}}(\mathbf{X}_i), \mathbf{X}_i)\right). \tag{4.8}$$

4.3.6.1 Bias and variance

Theorem IV.4. *The bias of the plug-in estimator* $\hat{\mathbf{G}}_{k,e,\mathcal{M}}(f)$ *is given by*

$$\begin{aligned}\mathbb{B}(\hat{\mathbf{G}}_{k,e,\mathcal{M}}(f)) &= c_0\left(\frac{k}{M}\right)^{1/d} + c_1\left(\frac{k}{M}\right)^{2/d} + c_2\left(\frac{1}{k}\right) \\ &\quad + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{1/d}\right),\end{aligned}$$

where $c_1 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} g'(f(\mathbf{Y}),\mathbf{Y})\tilde{h}(\mathbf{Y})]$ *and* $c_2 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f^2(\mathbf{Y}) g''(f(\mathbf{Y}),\mathbf{Y})/2]$.

Proof. The Euclidean k-NN density estimator on the manifold $\hat{\mathbf{f}}_{k,e,\mathcal{M}}(.)$ satisfies assumptions $\mathscr{A}.1$ and $\mathscr{A}.2$ (see section 4.3.5) listed in Appendix D. This implies that lemma D.1 holds. The rest of the proof is identical to the proof of Theorem 2.1 for bias of entropy functionals for Euclidean data. □

Theorem IV.5. *The variance of the plug-in estimator* $\hat{\mathbf{G}}_{k,e,\mathcal{M}}(f)$ *is given by*

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

where $c_4 = \mathbb{V}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} g(f(\mathbf{Y}),\mathbf{Y})]$ *and* $c_5 = \mathbb{V}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f(\mathbf{Y}) g'(f(\mathbf{Y}),\mathbf{Y})]$.

Proof. The Euclidean k-NN density estimator on the manifold $\hat{\mathbf{f}}_{k,e,\mathcal{M}}(.)$ satisfies assumptions $\mathscr{A}.1$ and $\mathscr{A}.3$ (see section 4.3.5) listed in Appendix D. This implies that lemma D.2 holds. The rest of the proof is identical to the proof of Theorem 2.2 for variance for entropy functionals of Euclidean data. □

4.3.6.2 Central limit theorem

In addition to the results on bias and variance shown in the previous section, we show that our plug-in estimator, appropriately normalized, weakly converges to the normal distribution. We study the asymptotic behavior of the plug-in estimates

under the following limiting conditions: (a) $k/M \to 0$, (b) $k \to \infty$, and (c) $N \to \infty$. As shorthand, we will collectively denote the above limiting assumptions by $\Delta \to 0$.

Theorem IV.6. *The asymptotic distribution of the normalized plug-in estimator* $\hat{\mathbf{G}}(f)$ *is given by*

$$\lim_{\Delta \to 0} Pr\left(\frac{\hat{\mathbf{G}}_{k,e,\mathcal{M}}(f) - \mathbb{E}[\hat{\mathbf{G}}_{k,e,\mathcal{M}}(f)]}{\sqrt{\mathbb{V}[1_{\{\mathbf{Y} \in \mathcal{S}'\}} g(f(\mathbf{Y}), \mathbf{Y})]/N}} \leq \alpha \right) = Pr(\mathbf{Z} \leq \alpha),$$

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

Proof. Proof is identical to the proof of Theorem 2.3 which establishes CLT for entropy functionals on Euclidean data. □

We note that divergence functionals can be estimated in an identical manner to Section 2.4 by using k-NN density estimates on the manifold instead of the standard k-NN density estimates on Euclidean space. The bias, variance and CLT for these estimators can be similarly analyzed.

4.3.7 Discussion

We observe that the functional forms of Theorems 4.4, 4.5 and 4.6 on bias, variance and asymptotic distribution respectively for entropy estimation of data on manifolds are identical to Theorems 2.1, 2.2 and 2.3 on bias, variance and asymptotic distribution respectively for entropy estimation of Euclidean data. In addition, the constants $c_i, i = 2, 4, 5$ are identical in both cases. This can be attributed to the fact that the radial distance on the mapped Euclidean range of the coordinate chart is *e*qual to the Riemannian distance on the manifold. The constant c_1 differs in the two cases only because of the additional bias term due to the approximation of the geodesic distance on the manifold by the Euclidean distance as discussed in section 4.3.4.

4.4 Dimension estimation

Intrinsic dimensionality is an important concept in high dimensional datasets whose principal modes of variation lie on a subspace of substantially lower dimension, the intrinsic dimension d. In such cases dimensionality reduction can be accomplished without loss of information. An accurate estimator of intrinsic dimension is a prerequisite for setting the embedding dimension of DR algorithms such as principal components analysis (PCA), ISOMAP, and Laplacian eigenmaps. Until recently the most common method for selecting an embedding dimension for these algorithms was to detect a knee in a residual error curve, e.g., scree plots of sorted eigenvalues. In this section, 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 [20] and to Farahmand's dimension estimator based on nearest neighbor distances [28]. The estimator can also be related to the Leonenko's Rényi entropy estimator [49]. However, unlike these estimators, our new dimension estimator is derived directly from a mean squared error (MSE) 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.

The section is organized as follows. We first introduce the the general form of the new dimension estimator. We then show that the estimator is related to a general class of k-NN density estimators. We review results on the statistical properties of functionals of kNN density estimators and use this theory to obtain expressions for the asymptotic bias and variance of the new dimension estimator, in addition to establishing that it satisfies a central limit theorem. The analytical expressions for bias and variance allow us to optimize over the tuning parameters of the dimension

estimator. Next, motivated by this analysis, we propose a modified dimension estimator with reduced variance. Finally, we report empirical comparisons that illustrate the improved performance of the new dimensionality estimator relative to previous approaches.

4.4.1 k-NN dimension estimator

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

4.4.1.1 Log-length statistic

Define the k-log-length statistic to be

$$\mathbf{L_k}(\mathcal{X}) = \frac{1}{N} \sum_{i=1}^{N} \log \left(\mathbf{d}_{k,e}(\mathbf{X}_i) \right),$$

where $\mathbf{d}_{k,e}(\mathbf{X}_i)$ is the Euclidean k-nearest neighbor (k-NN) distance from target sample X_i to the M reference samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$.

4.4.1.2 Relation to Shannon entropy

Let $\breve{H}_k$ denote the negative of the Shannon entropy plug-in estimator (with functional $g(u) = -\log(u)$) with bias correction,

$$\breve{H}_k = -\left(\tilde{G}_{k,\mathcal{M}} + \log(k-1) - \psi(k-1) \right).$$

We can write the following relation

$$\begin{aligned}\breve{H}_k &= \frac{1}{N}\sum_{i=1}^{N}\psi(k) - \log(c_d M) - d\log(\mathbf{d}_{k,e}(\mathbf{X}_i)) \\ &= \psi(k) - \log(c_d M) - d\mathbf{L_k}(\mathcal{X}).\end{aligned}$$

From the theory established in Chapter 4(Theorem 4.1 and 4.2), we know $\breve{H}_k = -H + o_p(1)$, where H is the Shannon entropy. Using this relation, one can estimate the dimension consistently using the following idea of slope based estimation.

4.4.1.3 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 $\mathbf{L_{k_2}}(\mathcal{X})$ be the length statistics evaluated at bandwidths k_1 and k_2 respectively. Define the slope based inverse dimension estimator to be

$$\hat{\mathbf{d}}_s^{-1} = \frac{\mathbf{L_{k_2}}(\mathcal{X}) - \mathbf{L_{k_1}}(\mathcal{X})}{\psi(k_2 - 1) - \psi(k_1 - 1)}.$$

The dimension can then be consistently estimated using the estimate $\hat{d}_s = 1/\hat{\mathbf{d}}_s^{-1}$.

Theorem IV.7. *The bias of the slope based inverse dimension estimator* $\hat{\mathbf{d}}_s^{-1}$ *is given by*

$$\mathbb{E}[\hat{\mathbf{d}}_s^{-1}] - d^{-1} = \frac{c_0}{d}\frac{(k_2/M)^{1/d} - (k_1/M)^{1/d}}{\psi(k_2) - \psi(k_1)} + o(M^{-1/d}),$$

and the variance is given by

$$\mathbb{V}[\hat{\mathbf{d}}_s^{-1}] = o(1/N + 1/M).$$

Proof. From Theorem 2.3, [51], we have

$$\mathbb{E}[\breve{H}_k] = -H + c_0(k/M)^{1/d} + o((k/M)^{1/d}).$$

This implies that

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{d}}_s^{-1}] - d^{-1} &= \frac{1}{d}\left(\frac{\breve{H}_{k_2} - \breve{H}_{k_1}}{\psi(k_2) - \psi(k_1)}\right) \\ &= \frac{c_0}{d}\frac{(k_2/M)^{1/d} - (k_1/M)^{1/d}}{\psi(k_2) - \psi(k_1)} + o(M^{-1/d}). \end{aligned} \tag{4.9}$$

From Theorem 4.5, we have $\mathbb{V}[\breve{H}_k] = O(1/N + 1/M)$. This implies that,

$$\mathbb{V}[\hat{\mathbf{d}}_s^{-1}] = (d(\psi(k_2) - \psi(k_1)))^{-2}\mathbb{V}[\breve{H}_{k_2} - \breve{H}_{k_1}] = o(1/N + 1/M), \tag{4.10}$$

where the last step follows from Cauchy-Schwarz.

□

From this analysis, we note that the bias and variance of the dimension estimator $\hat{\mathbf{d}}_s = 1/\hat{\mathbf{d}}_s^{-1}$ are also of order $O((1/M)^{1/d})$ and $o(1/N + 1/M)$ respectively.

4.4.2 Mixture of manifolds

The notion of intrinsic dimension has been extended to data that lies on a mixture of manifolds of varying intrinsic dimensions, that are embedded in $\mathbb{R}^D$. In this case, Carter et. al. proposed a local dimension estimate for each point in the sample. They define the local dimension estimate of each point to be the intrinsic dimension of a small sample centered at the point.

The k-NN neighborhoods in the above illustration are depicted by the blue lines. It is clear from the illustration that the k-NN neighborhoods 'hug' the respective manifolds in each case. Carter *et al.* proposed that the local dimension of each sample be estimated by estimating the intrinsic dimension of a small subset of samples close to the sample of interest.

In the simulations that follow, we will illustrate the superior performance of the proposed weighted dimension estimator in estimating the local dimension of a sample.

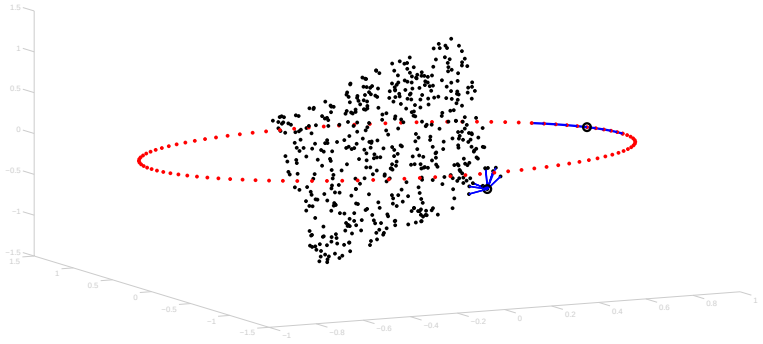


Figure 4.2: Illustration of data in a sample belonging to a mixture of manifolds. The black points on the plane have intrinsic dimension 2 while the red points on the circle have intrinsic dimension 1. The blue lines depict the k-NN edges.

4.4.3 Experimental results

4.4.3.1 Comparison of dimension estimation methods

We generate $T = 500$ samples $\mathcal{B}$ drawn from a $d = 3$ mixture density $f_m = .8f_\beta + .2f_u$, where f_β is the product of three 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 5-dimensional hyperplane in $\mathbb{R}^5$ by applying the transformation $\mathcal{Y} = U\mathcal{B}$ where U is a 5×3 random matrix whose columns are orthonormal. We apply our intrinsic dimension estimates on the samples $\mathcal{Y}$, and repeat the experiment a total of 100 times. The estimated dimension over these 100 trials is shown in Fig. 4.3.

We compare the performance of our proposed dimension estimator $\hat{\mathbf{d}}_s$ to the estimated proposed by Frahmand *et al.* [28] (denote as $\hat{\mathbf{d}}_f$), Levina and Bickel [50] (denote as $\hat{\mathbf{d}}_l$) and Costa *et al.* [20] (denote as $\hat{\mathbf{d}}_j$). We note that the performance of estimators $\hat{\mathbf{d}}_s$ and $\hat{\mathbf{d}}_l$ is perfect and outperform the other estimators.

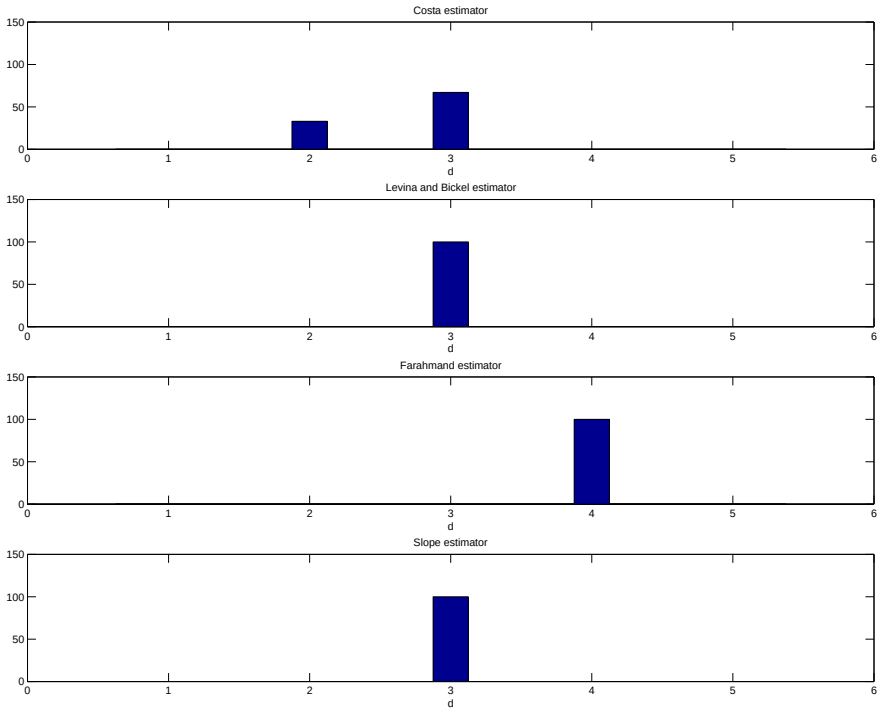


Figure 4.3: Comparison of dimension estimators. The proposed slope estimator $\hat{\mathbf{d}}_s$, and Levina and Bickel's estimator $\hat{\mathbf{d}}_l$, outperform the other estimators.

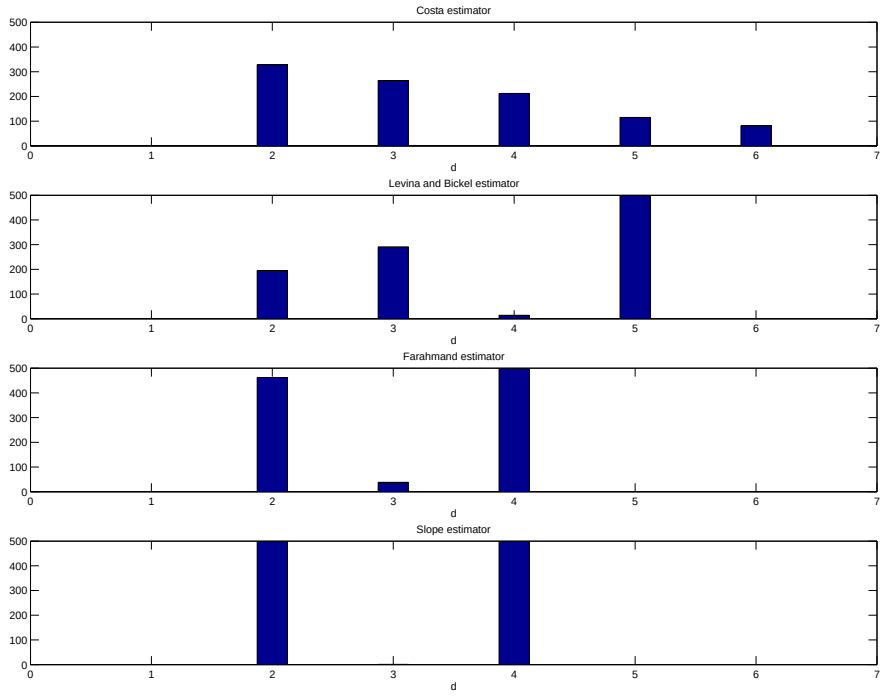


Figure 4.4: Comparison of local dimension estimation performance. The proposed slope estimator $\hat{\mathbf{d}}_s$ outperforms the other estimators.

4.4.3.2 Local dimension estimation: Toy example

In this experiment, we project data of intrinsic dimension 2 and 4 onto $\mathbb{R}^6$ (total of 1000 points), and then perform local dimension estimation using the dimension estimator. We compare our results to the dimension estimator of Frahmand *et al.* ($\hat{\mathbf{d}}_f$), Levina and Bickel ($\hat{\mathbf{d}}_l$) and Costa *et al.* ($\hat{\mathbf{d}}_j$). The results are shown in Fig. 4.4. From the histogram, it is clear that the slope estimator outperforms the other estimators.

4.5 Discussion

From our analysis, we note that the bias and variance of the slope dimension estimator $\hat{\mathbf{d}}_s$ are of order $(1/T)^{1/d}$ and $(1/T)$ respectively. The bias and therefore MSE can be large depending for large values of d is. In Chapter 6, we propose a

weighted dimension estimator which has a much lower MSE of order $O(1/T)$.

CHAPTER V

Minimum volume set testing

5.1 Introduction

In the previous chapters, we were concerned with the estimation of entropy, divergence, mutual information and intrinsic dimension. In this chapter, we consider the related problem of p-value estimation to test membership of individual samples in a level sets. We use the proposed p-value estimator to detect anomalies in data sets.

Minimum volume (MV) sets or equivalently level sets provide useful summaries of multi-dimensional functions for many applications including clustering [35, 82], anomaly detection [75, 81, 86], functional neuroimaging [64], bioinformatics [88] and digital elevation mapping [78].

Estimation of minimum volume sets is a difficult problem, especially for high dimensional data. There are two types of approaches to this problem: (1) transform the MV estimation problem to an equivalent density level set estimation problem, which requires estimation of the nominal density; and (2) directly identify the minimal set using function approximation and non-parametric estimation [75, 61, 73]. Both types of approaches involve explicit approximation of high dimensional quantities - the multivariate density function in the first case and the boundary of the minimum volume set in the second and are therefore not easily applied to high dimensional problems.

The GEM principle developed by Hero [36] for determining MV sets circumvents the above difficulties by using the asymptotic theory of random Euclidean graphs instead of function approximation. However, the GEM based K-kNNG anomaly detection scheme proposed in [36] is computationally difficult. To address this issue, a surrogate L1O-kNNG anomaly detection scheme was proposed in [36]. L1O-kNNG is computationally simpler than K-kNNG, but loses some desirable properties of the K-kNNG, including asymptotic consistency, as shown below.

In this chapter, we use the GEM principle to develop a bipartite k-nearest neighbor (k-NN) graph-based anomaly detection algorithm. BP-kNNG retains the desirable properties of the GEM principle and as a result inherits the following features: (i) it is not restricted to linear or even convex decision regions, (ii) it is completely non-parametric, (iii) it is optimal in that it converges to the uniformly most powerful (UMP) test when the anomalies are drawn from a mixture of the nominal density and the uniform density, (iv) it does not require knowledge of anomalies in the training sample, (v) it is asymptotically consistent in recovering the p-value of the test point and (vi) it produces estimated p-values, allowing for false positive rate control.

K-LPE [89] and RRS [66] are anomaly detection methods which are also based on k-NN graphs. BP-kNNG differs from L1O-kNNG, K-LPE and RRS in the following respects. L1O-kNNG, K-LPE and RRS do not use bipartite graphs. We will show that the bipartite nature of BP-kNNG results in significant computational savings. In addition, the K-LPE and RRS test statistics involve only the k-th nearest neighbor distance, while the statistic in BP-kNNG, like the L1O-kNNG, involves summation of the power weighted distance of all the edges in the k-NN graph. This will result in increased robustness to outliers in the training sample. Finally, we will show that the mean square rate of convergence of p-values in BP-kNNG ($O(T^{-2/(2+d)})$) is faster as compared to the convergence rate of K-LPE ($O(T^{-2/5} + T^{-6/5d})$), where T is the size of the nominal training sample and d is the dimension of the data.

The rest of this chapter is organized as follows. In Section 2, we outline the statistical framework for minimum volume set detection. In Section 3, we describe the GEM principle and the K-kNNG and L1O-kNNG detection schemes proposed in [36]. Next, in Section 4, we develop our bipartite k-NN graph (BP-kNNG) method. We show consistency of the method and compare its computational complexity with that of the K-kNNG, L1O-kNNG and K-LPE algorithms. In Section 5, we show simulation results that illustrate the superior performance of BP-kNNG over L1O-kNNG in the context of anomaly detection. We also show that our method compares favorably to other state of the art anomaly detection schemes when applied to real world data from the UCI repository [3]. We conclude with a short discussion in Section 6.

5.2 Statistical novelty detection

Given a training set of normal events, the anomaly detection problem aims to identify unknown, anomalous events that deviate from the normal set. This novelty detection problem arises in applications where failure to detect anomalous activity could lead to catastrophic outcomes, for example, detection of faults in mission-critical systems, quality control in manufacturing and medical diagnosis.

Several approaches have been proposed for anomaly detection. One class of algorithms assumes a family of parametrically defined nominal distributions. Examples include Hotelling's T test and the Fisher F-test, which are both based on a Gaussian distribution assumption. The drawback of these algorithms is model mismatch: the supposed distribution need not be a correct representation of the nominal data, which can then lead to poor false alarm rates. More recently, several non-parametric methods based on minimum volume (MV) set estimation have been proposed. These methods aim to find the minimum volume set that recovers a certain probability mass α with respect to the unknown probability density of the nominal events. If a new

event falls within the MV set, it is classified as normal and otherwise as anomalous.

The problem setup is as follows. We assume that a training sample $\mathcal{X}_T = \{X_1, \ldots, X_T\}$ of d-dimensional vectors is available. Given a new sample X, the objective is to declare X to either be a 'nominal' event consistent with $\mathcal{X}_T$ or an 'anomalous' event which deviates from $\mathcal{X}_T$. Fig. 5.1 shows a collection of training samples and two test samples - one of which is nominal and the other anomalous. We seek to find a functional D and corresponding detection rule $D(x) > 0$ so that X is declared to be nominal if $D(x) > 0$ holds and anomalous otherwise. The acceptance region is given by $A = \{x : D(x) > 0\}$. We seek to further constrain the choice of D to allow as few false negatives as possible for a fixed allowance of false positives.

To formulate this problem, we adopt the standard statistical framework for testing composite hypotheses. We assume that the training sample $\mathcal{X}_T$ is an i.i.d sample draw from an unknown d-dimensional probability distribution $f_0(x)$ on $[0,1]^d$. Let X have density f on $[0,1]^d$. The anomaly detection problem can be formulated as testing the hypotheses $H_0 : f = f_0$ versus $H_1 : f \neq f_0$.

5.2.1 Minimum volume set detection

For a given $\alpha \in (0,1)$, we seek an acceptance region A that satisfies $Pr(X \in A|H_0) \geq 1 - \alpha$. This requirement maintains the false positive rate at a level no greater than α. Let $\mathcal{A} = \{A : \int_A f_0(x)dx \geq 1-\alpha\}$ denote the collection of acceptance regions of level α. The most suitable acceptance region from the collection $\mathcal{A}$ would be the set which minimizes the false negative rate. Assume that the density f is bounded above by some constant C. In this case the false negative rate is bounded by $C\lambda(A)$ where $\lambda(.)$ is the Lebesgue measure in $\mathbb{R}^d$. Consider the relaxed problem of minimizing the upper bound $C\lambda(A)$ or equivalently the volume $\lambda(A)$ of A. The optimal acceptance region with a maximum false alarm rate α is therefore given by the minimum volume set of level α: $\Lambda_\alpha = min\{\lambda(A) : \int_A f_0(x)dx \geq \alpha\}$.

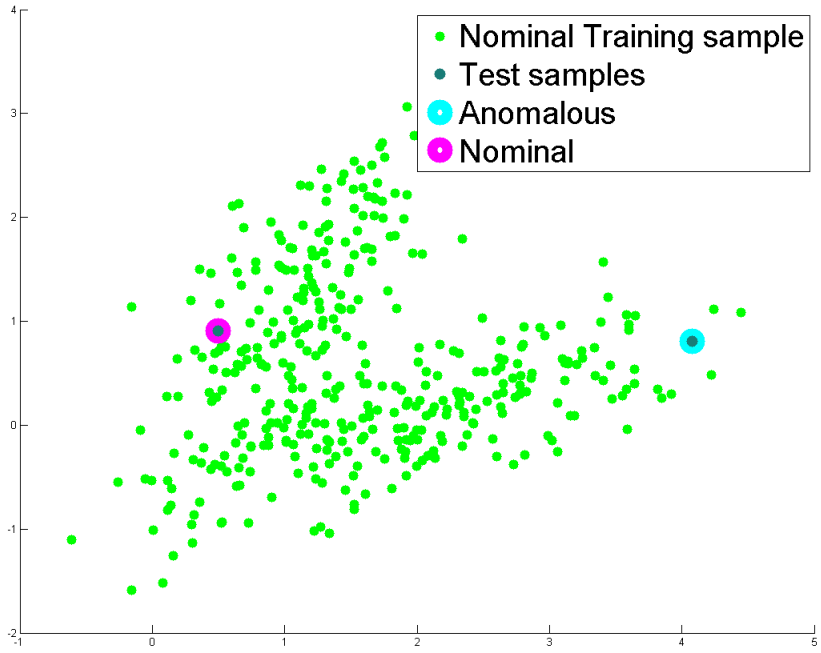


Figure 5.1: Illustration of a collection of training samples and two test samples - one of which is nominal and the other anomalous.

Define the minimum entropy set of level α to be $\Omega_\alpha = \min\{H_\nu(A) : \int_A f_0(x)dx \geq 1-\alpha\}$ where $H_\nu(A) = (1-\nu)^{-1}\int_A f_0^\nu(x)dx$ is the Rényi ν-entropy of the density f_0 over the set A. It can be shown that when f_0 is a Lebesgue density in $\mathbb{R}^d$, the minimum volume set and the minimum entropy set are equivalent, i.e. Λ_α and Ω_α are identical. Therefore, the optimal decision rule for a given level of false alarm α is to declare an anomaly if $X \notin \Omega_\alpha$.

This decision rule has a strong optimality property [36]: when f_0 is Lebesgue continuous and has no 'flat' regions over its support, this decision rule is a *uniformly most powerful* (UMP) test at level $1-\alpha$ for the null hypothesis that the test point has density $f(x)$ equal to the nominal $f_0(x)$ versus the alternative hypothesis that $f(x) = (1-\epsilon)f_0(x) + \epsilon U(x)$, where $U(x)$ is the uniform density over $[0,1]^d$ and $\epsilon \in [0,1]$. Furthermore, the power function is given by $\beta = Pr(X \notin \Omega_\alpha | H_1) = (1-\epsilon)\alpha + \epsilon(1-\lambda(\Omega_\alpha))$.

5.3 GEM principle

In this section, we briefly review the geometric entropy minimization (GEM) principle method [36] for determining minimum entropy sets Ω_α of level α. The GEM method directly estimates the critical region Ω_α for detecting anomalies using minimum coverings of subsets of points in a nominal training sample. These coverings are obtained by constructing minimal graphs, e.g., the k-minimal spanning tree or the k-nearest neighbor graph, covering a K-point subset that is a given proportion of the training sample. Points in the training sample that are not covered by the K-point minimal graphs are identified as tail events.

In particular, let $\mathcal{X}_{K,T}$ denote one of the $\binom{T}{K}$ K point subsets of $\mathcal{X}_T$. The k-nearest neighbors (k-NN) of a point $X_i \in \mathcal{X}_{K,T}$ are the k closest points to X_i among $\mathcal{X}_{K,T} - X_i$. Denote the corresponding set of edges between X_i and its k-NN by $\{e_{i(1)}, \ldots, e_{i(k)}\}$. For any subset $\mathcal{X}_{K,T}$, define the total power weighted edge length of the k-NN graph on $\mathcal{X}_{K,T}$ with power weighting γ $(0 < \gamma < d)$, as

$$L_{kNN}(\mathcal{X}_{K,T}) = \sum_{i=1}^{K} \sum_{l=1}^{k} |e_{t_i(l)}|^\gamma,$$

where $\{t_1, \ldots, t_K\}$ are the indices of $X_i \in \mathcal{X}_{K,T}$. Define the K-kNNG graph to be the K-point k-NN graph having minimal length $\min_{\mathcal{X}_{T,K} \in \mathcal{X}_T} L_{kNN}(\mathcal{X}_{T,K})$ over all $\binom{T}{K}$ subsets $\mathcal{X}_{K,T}$. Denote the corresponding length minimizing subset of K points by $\mathcal{X}^*_{T,K} = \underset{\mathcal{X}_{T,K} \in \mathcal{X}}{\operatorname{argmin}} L_{kNN}(\mathcal{X}_{K,T})$.

The K-kNNG thus specifies a minimal graph covering $\mathcal{X}^*_{K,T}$ of size K. This graph can be viewed as capturing the densest regions of $\mathcal{X}_T$. If $\mathcal{X}_T$ is an i.i.d. sample from a multivariate density $f_0(x)$ and if $\lim_{K,T\to\infty} K/T = \rho$, then the set $\mathcal{X}^*_{K,T}$ converges a.s. to the minimum ν-entropy set containing a proportion of at least ρ of the mass of $f_0(x)$, where $\nu = 1 - \gamma/d$ [36]. This set can be used to perform anomaly detection.

5.3.1 K-kNNG anomaly detection

Given a test sample X, denote the pooled sample $\mathcal{X}_{T+1} = \mathcal{X}_T \cup \{X\}$ and determine the K-kNNG graph over $\mathcal{X}_{T+1}$. Declare X to be an anomaly if $X \notin \mathcal{X}^*_{K,T+1}$ and nominal otherwise. When the density f_0 is Lebesgue continuous, it follows from [36] that as $K, T \to \infty$, this anomaly detection algorithm has false alarm rate that converges to $\alpha = 1 - K/T$ and power that converges to that of the minimum volume set test of level α. An identical detection scheme based on the K-minimal spanning tree has also been developed in [36].

The K-kNNG anomaly detection scheme therefore offers a direct approach to detecting outliers while bypassing the more difficult problems of density estimation and level set estimation in high dimensions. However, this algorithm requires construction of k-nearest neighbor graphs (or k-minimal spanning trees) over $\binom{T}{K}$ different subsets. For each input test point, the runtime of this algorithm is therefore $O(dK^2\binom{T}{K})$. As a result, the K-kNNG method is not well suited for anomaly detection for large sample sizes. The output of K-kNNG algorithm is illustrated in Fig. 5.2.

5.3.2 L1O-kNNG

To address the computational problems of K-kNNG, Hero [36] proposed implementing the K-kNNG for the simplest case $K = T - 1$. The runtime of this algorithm for each input test point is $O(dT^2)$. Clearly, the L1O-kNNG is of much lower complexity that the K-kNNG scheme. However, the L1O-kNNG detects anomalies at a fixed false alarm rate $1/(T+1)$, where T is the training sample size. To detect anomalies at a higher false alarm rate α^*, one would have to subsample the training set and only use $T^* = 1/\alpha^* - 1$ training samples. This destroys any hope for asymptotic consistency of the L1O-kNNG.

In the next section, we propose a different GEM based algorithm that uses bipartite graphs. The algorithm has algorithm has a much faster runtime than the

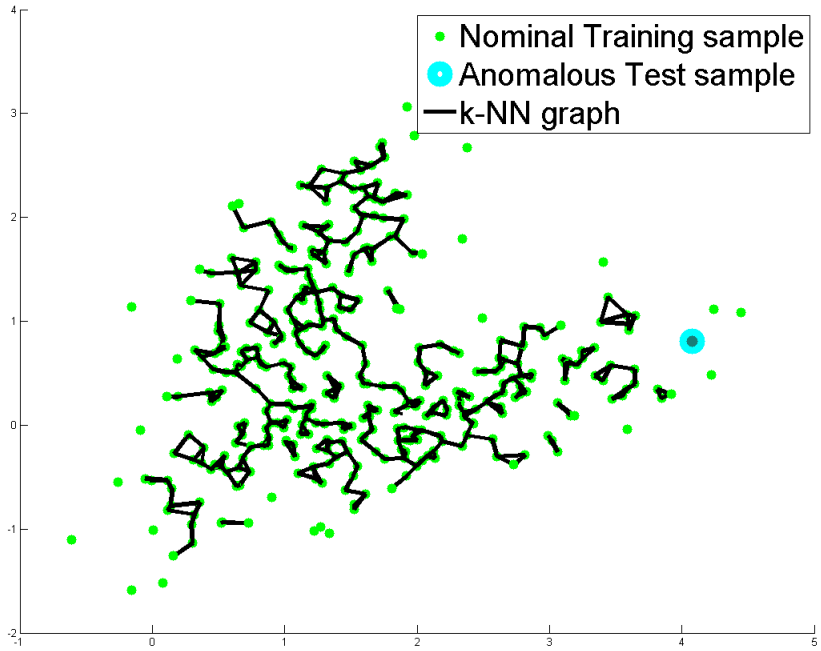


Figure 5.2: Output of K-kNNG algorithm

L1O-kNNG, and unlike the L1O-kNNG, is asymptotically consistent and can operate at any specified alarm rate α. We describe our algorithm below.

5.4 BP-kNNG

Let $\{\mathcal{X}_N, \mathcal{X}_M\}$ be a partition of $\mathcal{X}_T$ with $card\{\mathcal{X}_N\} = N$ and $card\{\mathcal{X}_M\} = M = T - N$ respectively. This partitioning is illustrated in Fig. 5.3.

As above, let $\mathcal{X}_{K,N}$ denote one of the $\binom{N}{K}$ subsets of K distinct points from $\mathcal{X}_N$. Define the bipartite k-NN graph on $\{\mathcal{X}_{K,N}, \mathcal{X}_M\}$ to be the set of edges linking each $X_i \in \mathcal{X}_{K,N}$ to its k nearest neighbors in $\mathcal{X}_M$. Define the total power weighted edge length of this bipartite k-NN graph with power weighting γ $(0 < \gamma < d)$ and a fixed

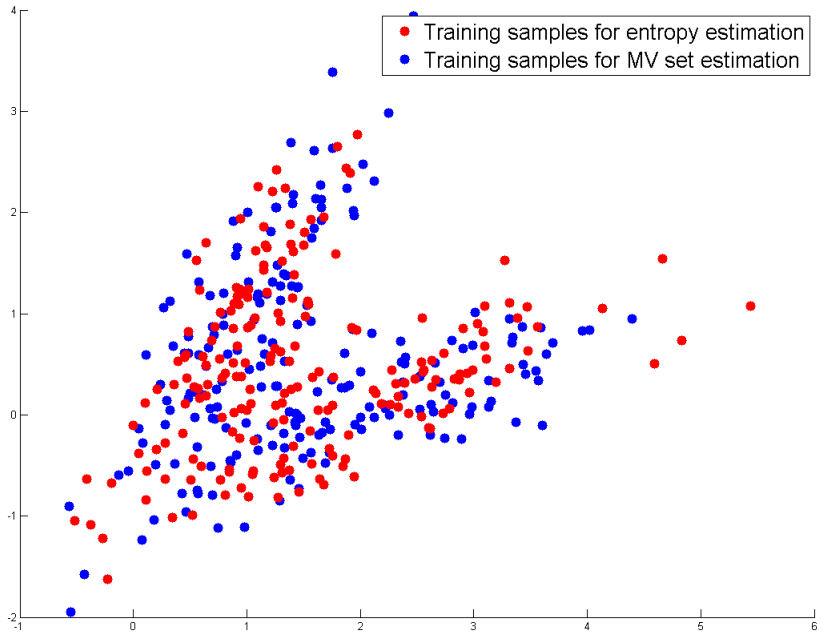


Figure 5.3: Illustration of the first step in the BP-kNNG anomaly detection algorithm: partitioning of data points into disjoint sets which are subsequently used for entropy and MV set estimation respectively.

number of edges s $(1 \leq s \leq k)$ corresponding to each vertex $X_i \in \mathcal{X}_{K,N}$ to be

$$L_{s,k}(\mathcal{X}_{K,N}, \mathcal{X}_M) = \sum_{i=1}^{K} \sum_{l=k-s+1}^{k} |e_{t_i(l)}|^{\gamma},$$

where $\{t_1, \ldots, t_K\}$ are the indices of $X_i \in \mathcal{X}_{K,N}$ and $\{e_{t_i(1)}, \ldots, e_{t_i(k)}\}$ are the k-NN edges in the bipartite graph originating from $X_{t_i} \in \mathcal{X}_{K,N}$. Define the bipartite K-kNNG graph to be the one having minimal weighted length $\min_{\mathcal{X}_{N,K} \in \mathcal{X}_N} L_{s,k}(\mathcal{X}_{N,K}, \mathcal{X}_M)$ over all $\binom{N}{K}$ subsets $\mathcal{X}_{K,N}$. Define the corresponding minimizing subset of K points of $\mathcal{X}_{K,N}$ by $\mathcal{X}^*_{K,N} = \underset{\mathcal{X}_{K,N} \in \mathcal{X}}{\operatorname{argmin}} L_{s,k}(\mathcal{X}_{K,N}, \mathcal{X}_M)$.

Using the theory of partitioned k-NN graph entropy estimators, we now show that as $k/M \to 0$, $k, N \to \infty$ and for fixed s, the set $\mathcal{X}^*_{K,N}$ converges a.s. to the minimum ν-entropy set $\Omega_{1-\rho}$ containing a proportion of at least ρ of the mass of

$f_0(x)$, where $\rho = \lim_{K,N\to\infty} K/N$ and $\nu = 1 - \gamma/d$. Assume without loss of generality that $\{X_1, \ldots, X_N\} \in \mathcal{X}_N$ and $\{X_{N+1}, \ldots, X_T\} \in \mathcal{X}_M$.

Denote the support of the density f_0 to be $\mathcal{S}$. Let $\mathcal{S}' \subset \mathcal{S}$ be any arbitrary subset of $\mathcal{S}$. Denote the collective behavior $k/M \to 0$, $k, N \to \infty$ by $\Delta \to 0$. Note that the distance $e_i(l)$ of a point $X_i \in \mathcal{X}_N$ to its l-th nearest neighbor in $\mathcal{X}_M$ is related to the bipartite l-nearest neighbor density estimate $\hat{f}_l(X_i) = \frac{l-1}{Mc_d e_i^d(l)}$ where c_d is the unit ball volume in d dimensions. From Theorem 2.1 and 2.2 it therefore immediately follows that for some fixed s

$$\lim_{\Delta\to 0} \mathbb{E}\left[\frac{1}{sN}\sum_{i=1}^{N} 1_{\{X_i\in\mathcal{S}'\}}(k/c_d M)^{\nu-1} d_{s,k}(X_i) - \int_{z\in\mathcal{S}'} f_0^{\nu}(z)\right]^2 = 0.$$

Because

$$\mathcal{X}^*_{K,N} = \underset{\mathcal{X}_{K,N}\in\mathcal{X}}{\operatorname{argmin}} L_{s,k}(\mathcal{X}_{K,N}, \mathcal{X}_M),$$

it follows that the set $\mathcal{X}^*_{K,N}$ converges to the minimum entropy set $\Omega_{1-\rho}$ containing a proportion of at least ρ of the mass of $f_0(x)$ where $\rho = \lim_{K,N\to\infty} K/N$.

This suggests using the bipartite k-NN graph to detect anomalies in the following way. Given a test point X, denote the pooled sample $\mathcal{X}_{N+1} = \mathcal{X}_N \cup \{X\}$ and determine the optimal bipartite K-kNNG graph $\mathcal{X}^*_{K,N+1}$ over $\{\mathcal{X}_{K,N+1}, \mathcal{X}_M\}$. Now declare X to be an anomaly if $X \notin \mathcal{X}^*_{K,N+1}$ and nominal otherwise. It is clear that by the GEM principle, this algorithm detects false alarms at a rate that converges to $\alpha = 1 - K/T$ and power that converges to that of the minimum volume set test of level α. The bipartite k-NN graph is shown in Fig. 5.4.

We can equivalently determine $\mathcal{X}^*_{K,N+1}$ as follows. For each $X_i \in \mathcal{X}_N$, construct $d_{s,k}(X_i) = \sum_{l=k-s+1}^{k} |e_{i(l)}|^\gamma$. For each test point X, define $d_{s,k}(X) = \sum_{l=s-k+1}^{k} |e_{X(l)}|^\gamma$, where $\{e_{X(1)}, \ldots, e_{X(k)}\}$ are the k-NN edges from X to $\mathcal{X}_M$. Now, choose the K points among $\mathcal{X}_N \cup X$ with the K smallest of the $N+1$ edge lengths $\{d_{s,k}(X_i), X_i \in \mathcal{X}_N\} \cup \{d_{s,k}(X)\}$. Because of the bipartite nature of the construction, this is equivalent to

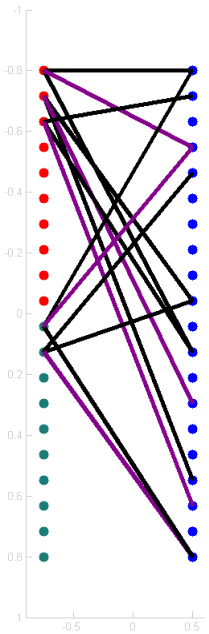


Figure 5.4: Bipartite k-NN graph on training and test data (red = N training samples for MV set estimation, blue = M training samples for MV set estimation, green = test samples)

choosing $\mathcal{X}^*_{K,N+1}$. This leads to the proposed BP-kNNG anomaly detection algorithm described by Algorithm 2. The output of Algorithm 2 is illustrated in Fig. 5.5.

5.4.1 BP-kNNG p-value estimates

The p-value is a score between 0 and 1 that is associated with the likelihood that a given point X_0 comes from a specified nominal distribution. The BP-kNNG generates an estimate of the p-value that is asymptotically consistent, guaranteeing that the BP-kNNG detector is a consistent novelty detector.

Specifically, for a given test point X_0, the true p-value associated with a point X_0 in a minimum volume set test is given by $p_{true}(X_0) = \int_{S(X_0)} f_0(z)dz$ where $S(X_0) = \{z : f_0(z) \leq f_0(X_0)\}$ and $E(X_0) = \{z : f_0(z) = f_0(X_0)\}$. $p_{true}(X_0)$ is the minimal level α at which X_0 would be rejected. The empirical p-value associated with the

Algorithm 2 Anomaly detection scheme using bipartite k-NN graphs

1. Input: Training samples $\mathcal{X}_T$, test samples X, false alarm rate α
2. Training phase
a. Create partition $\{\mathcal{X}_N, \mathcal{X}_M\}$
b. Construct k-NN bipartite graph on partition
c. Compute k-NN lengths $d_{s,k}(X_i)$ for each $X_i \in \mathcal{X}_N$: $d_{s,k}(X_i) = \sum_{l=k-s+1}^{k} |e_{i(l)}|^\gamma$
3. Test phase: detect anomalous points

for each input test sample X **do**
Compute k-NN length $d_{s,k}(X) = \sum_{l=k-s+1}^{k} |e_{X(l)}|^\gamma$
if

$$(1/N) \sum_{X_i \in \mathcal{X}_N} 1(d_{s,k}(X_i) < d_{s,k}(X)) \geq 1 - \alpha$$

then
Declare X to be anomalous
else
Declare X to be non-anomalous
end if
end for

BP-kNNG is defined as

$$p_{bp}(X_0) \;=\; \frac{\sum_{X_i \in \mathcal{X}_N} 1(d_{s,k}(X_i) \geq d_{s,k}(X_0))}{N}. \tag{5.1}$$

5.4.2 Asymptotic consistency and optimal convergence rates

Here we prove that the BP-kNNG detector is asymptotically consistent by showing that for a fixed number of edges s, $\mathbb{E}[(p_{bp}(X_0) - p_{true}(X_0))^2] \to 0$ as $k/M \to 0$, $k, N \to \infty$. In the process, we also obtain rates of convergence of this mean-squared error. These rates depend on k, N and M and result in the specification of an optimal number of neighbors k and an optimal partition ratio N/M that achieve the best trade-off between bias and variance of the p-value estimates $p_{bp}(X_0)$. We assume that the density f_0 (i) is bounded away from 0 and ∞ and is continuous on its support $\mathcal{S}$, (ii) has no flat spots over its support set and (iii) has a finite number of modes. Let $\mathbb{E}$ denote the expectation w.r.t. the density f_0, and $\mathbb{B}$, $\mathbb{V}$ denote the bias and variance operators. Throughout this section, assume without loss of generality that

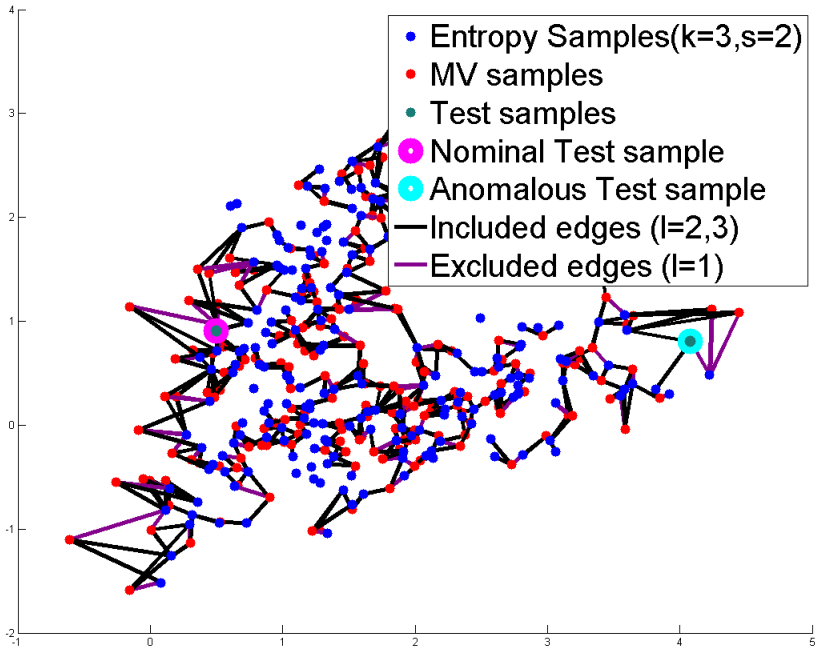


Figure 5.5: Output of BP-kNNG algorithm

$\{X_1, \ldots, X_N\} \in \mathcal{X}_N$ and $\{X_{N+1}, \ldots, X_T\} \in \mathcal{X}_M$.

Bias We first introduce the oracle p-value

$$p_{orac}(X_0) = (1/N) \sum_{X_i \in \mathcal{X}_N} 1(f_0(X_i) \leq f_0(X_0))$$

and note that $\mathbb{E}[p_{orac}(X_0)] = p_{true}(X_0)$. Let

$$e(X) = \left(\sum_{l=k-s+1}^{k} \left(\frac{k-1}{l-1} \hat{f}_l(X) \right)^{\nu-1} \right) - s(f(X))^{\nu-1}$$

and

$$\delta(X_i, X_0) = \delta_i = (f(X_i))^{\nu-1} - (f(X_0))^{\nu-1}.$$

We then have

$$\begin{aligned}\mathbb{B}[p_{bp}(X_0)] &= \mathbb{E}[p_{bp}(X_0)] - p_{true}(X_0) = \mathbb{E}[p_{bp}(X_0) - p_{orac}(X_0)] \\ &= \mathbb{E}[1(d_{s,k}(X_1) \geq d_{s,k}(X_0))] - \mathbb{E}[1(f(X_1) \leq f(X_0))] \\ &= \mathbb{E}[1(e(X_1) - e(X_0) + \delta_1 \leq 0) - 1(\delta_1 \leq 0)].\end{aligned}$$

This bias will be non-zero when $1(e(X_1) - e(X_0) + \delta_1 \leq 0) \neq 1(\delta_1 \leq 0)$. First we investigate this condition when $\delta_1 > 0$. In this case, for $1(e(X_1) - e(X_0) + \delta_1 \leq 0) \neq 1(\delta_1 \leq 0)$, we need $-e(X_1)+e(X_0) \geq \delta_1$. Likewise, when $\delta_1 \leq 0$, $1(e(X_1)-e(X_0)+\delta_1 \leq 0) \neq 1(\delta_1 \leq 0)$ occurs when $e(X_1) - e(X_0) > |\delta_1|$.

From the theory developed in Appendix C in [80], for any fixed s, $|e(X)| = O(k/M)^{1/d} + O(1/\sqrt{k})$ with probability greater than $1 - o(1/M)$. This implies that

$$\begin{aligned}\mathbb{B}[p_{bp}(X_0)] &= \mathbb{E}[1(e(X_1) - e(X_0) + \delta_1 \leq 0) - 1(\delta_1 \leq 0)] \\ &= Pr\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} + o(1/M).\end{aligned}$$

We first analyze the case where f_0 is monotonic. By the continuity of f_0, we then have $||X_1 - X_0||^d = O(\delta_1)$. Because we assume the density f_0 is bounded above by some constant C on its support, we have

$$\begin{aligned}Pr\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} &= Pr(||X_1 - X_0||^d = O(k/M)^{1/d} + O(1/\sqrt{k})) \\ &= O(k/M)^{1/d} + O(1/\sqrt{k}).\end{aligned}$$

We now extend this analysis to the general case where f_0 is assumed to have a finite number of modes. Let $S_{X_0}(\delta) = \{X \in \mathcal{S} : |(f_0(X))^{\nu-1} - (f_0(X_0))^{\nu-1}| < \delta\}$. By the

continuity of f_0, the volume $V_{X_0}(\delta) = \int_{S_{X_0}(\delta)} dx = O(\delta)$. We then have

$$\begin{aligned}
\mathbb{B}[p_{bp}(X_0)] &= \mathbb{E}[1(e(X_1) - e(X_0) + \delta_1 \leq 0) - 1(\delta_1 \leq 0)] \\
&= Pr\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} + o(1/M) \\
&= Pr\{X_1 \in S_{X_0}(O(k/M)^{1/d} + O(1/\sqrt{k}))\} + o(1/M) \\
&= O(V_{X_0}(O(k/M)^{1/d} + O(1/\sqrt{k})) \\
&= O((k/M)^{1/d} + 1/\sqrt{k}).
\end{aligned}$$

Variance Define $b_i = 1(e(X_i) - e(X_0) + \delta_i \leq 0) - 1(\delta_i \leq 0)$. We can compute the variance in a similar manner to the bias as follows

$$\begin{aligned}
&\mathbb{V}[p_{bp}(X_0)] \\
&= \frac{1}{N}\mathbb{V}[1(d_{s,k}(X_1) \geq d_{s,k}(X_0))] \\
&+ \frac{N-1}{N} Cov[1(d_{s,k}(X_1) \geq d_{s,k}(X_0)), 1(d_{s,k}(X_2) \geq d_{s,k}(X_0))] \\
&= \frac{1}{N}\mathbb{V}[1(e(X_1) - e(X_0) + \delta_1 \leq 0)] + \frac{N-1}{N} Cov[b_1, b_2] \\
&= O(1/N) + \mathbb{E}[b_1 b_2] - (\mathbb{E}[b_1]\mathbb{E}[b_2]) \\
&= O(1/N) \\
&+ Pr\{\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} \cap \{|\delta_2| = O(k/M)^{1/d} + O(1/\sqrt{k})\}\} \\
&- Pr\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} Pr\{|\delta_2| = O(k/M)^{1/d} + O(1/\sqrt{k})\} \\
&= O(1/N + (k/M)^{2/d} + 1/k).
\end{aligned}$$

Consistency of p-values: From (5.2) and (5.2), we obtain an asymptotic representation of the estimated p-value $\mathbb{E}[(p_{bp}(X_0) - p_{true}(X_0))^2] = O((k/M)^{2/d}) + O(1/k) + O(1/N)$. This implies that p_{bp} converges in mean square to p_{true}, for a fixed number of edges s, as $k/M \to 0$, $k, N \to \infty$.

Optimal choice of parameters: The optimal choice of k to minimize the MSE is given by $k = \Theta(M^{2/(2+d)})$. For fixed $M + N = T$, to minimize MSE, N should then be chosen to be of the order $O(M^{(4+d)/(4+2d)})$, which implies that $M = \Theta(T)$. The mean square convergence rate for this optimal choice of k and partition ratio N/M is given by $O(T^{-2/(2+d)})$. In comparison, the K-LPE method requires that k grows with the sample size at rate $k = \Theta(T^{2/5})$. The mean square rate of convergence of the p-values in K-LPE is then given by $O(T^{-2/5} + T^{-6/5d})$. The rate of convergence of the p-values is therefore faster in the case of BP-kNNG as compared to K-LPE.

5.4.3 Comparison of run time complexity

Here we compare complexity of BP-kNNG with that of K-kNNG, L1O-kNNG and K-LPE. For a single query point X, the runtime of K-kNNG is $O(dK^2\binom{T}{K})$, while the complexity of the surrogate L1O-kNN algorithm and the K-LPE is $O(dT^2)$. On the other hand, the complexity of the proposed BP-kNNG algorithm is dominated by the computation of $d_k(X_i)$ for each $X_i \in \mathcal{X}_N$ and $d_k(X)$, which is $O(dNM) = O(dT^{(8+3d)/(4+2d)}) = o(dT^2)$.

For the K-kNNG, L1O-kNNG and K-LPE, a new k-NN graph has to be constructed on $\{\mathcal{X}_N \cup \{X\}\}$ for every new query point X. On the other hand, because of the bipartite construction of our k-NN graph, $d_k(X_i)$ for each $X_i \in \mathcal{X}_N$ needs to be computed and stored only once. For every new query X that comes in, the cost to compute $d_k(X)$ is only $O(dM) = O(dT)$. For a total of L query points, the overall runtime complexity of our algorithm is therefore much smaller than the L1O-kNNG, K-LPE and K-kNNG anomaly detection schemes ($O(dT(T^{(4+d)/(4+2d)} + L))$ compared to $O(dLT^2)$, $O(dLT^2)$ and $O(dLK^2\binom{T}{K})$ respectively).

5.5 Simulation comparisons

We compare the L1O-kNNG and the bipartite K-kNNG schemes on a simulated data set. The training set contains 1000 realizations drawn from a 2-dimensional Gaussian density f_0 with mean 0 and diagonal covariance with identical component variances of $\sigma = 0.1$. The test set contains 500 realizations drawn from $0.8f_0 + 0.2U$, where U is the uniform density on $[0, 1]^2$. Samples from the uniform distribution are classified to be anomalies. The percentage of anomalies in the test set is therefore 20%.

The distribution f_0 has essential support on the unit square. For this simple case the minimum volume set of level α is a disk centered at the origin with radius $\sqrt{2\sigma^2}\log(1/\alpha)$. The power of the uniformly most powerful (UMP) test is $1 - 2\pi\sigma^2\log(1/\alpha)$.

L1O-kNNG and BP-kNNG were implemented in Matlab 7.6 on an 2 GHz Intel processor with 3 GB of RAM. The value of k was set to 5. For the BP-kNNG, we set $s = 1$, $N = 100$ and $M = 900$. In Fig. 5.6, we compare the detection performance of L1O-kNNG and BP-kNNG against the 'clairvoyant' UMP detector in terms of the ROC. We note that the proposed BP-kNNG is closer to the optimal UMP test as compared to the L1O-kNNG. In Fig. 5.7 we note the close agreement between desired and observed false alarm rates for BP-kNNG. Note that the L1O-kNNG significantly underestimates its false alarm rate for higher levels of true false alarm.

In the case of the L1O-kNNG, it took an average of 60ms to test each instance for possible anomaly. The total run-time was therefore 60x500 = 3000ms. For the BP-kNNG, for a single instance, it took an average of 57ms. When all the instances were processed together, the total run time was only 97ms. This significant savings in runtime is due to the fact that the bipartite graph does not have to be constructed separately for each new test instance; it suffices to construct it once on the entire data set.

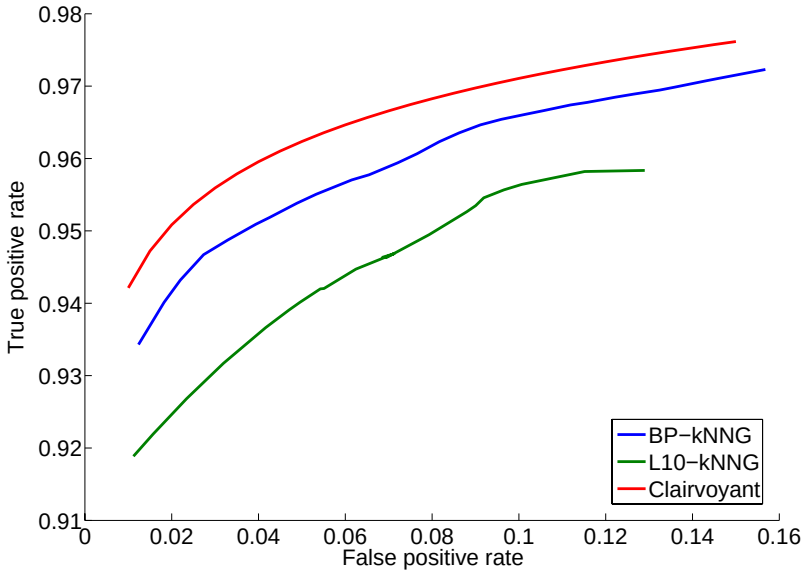


Figure 5.6: ROC curves for L1O-kNNG and BP-kNNG. The labeled 'clairvoyant' curve is the ROC of the UMP anomaly detector. The ROC curve for the BP-kNNG estimator is closer to the performance of the 'clairvoyant' UMP detector.

5.5.1 Experimental comparisons

In this section, we compare our algorithm to several other state of the art anomaly detection algorithms, namely: MassAD [83], isolation forest (or iForest) [52], two distance-based methods ORCA [7] and K-LPE [89], a density-based method LOF [13], and the one-class support vector machine (or 1-SVM) [73]. All the methods are tested on the five largest data sets used in [52]. The data characteristics are summarized in Table 5.1. One of the anomaly data generators is Mulcross [71] and the other four

Data set	Sample size	Dimension	Anomaly class
HTTP (KDD'99)	567497	3	attack (0.4%)
Forest	286048	10	class 4 vs class 2 (0.9%)
Mulcross	262144	4	2 clusters (10%)
SMTP (KDD'99)	95156	3	attack (0.03%)
Shuttle	49097	9	class 2,3,5,6,7 vs class 1 (7%)

Table 5.1: Description of data used in anomaly detection experiments.

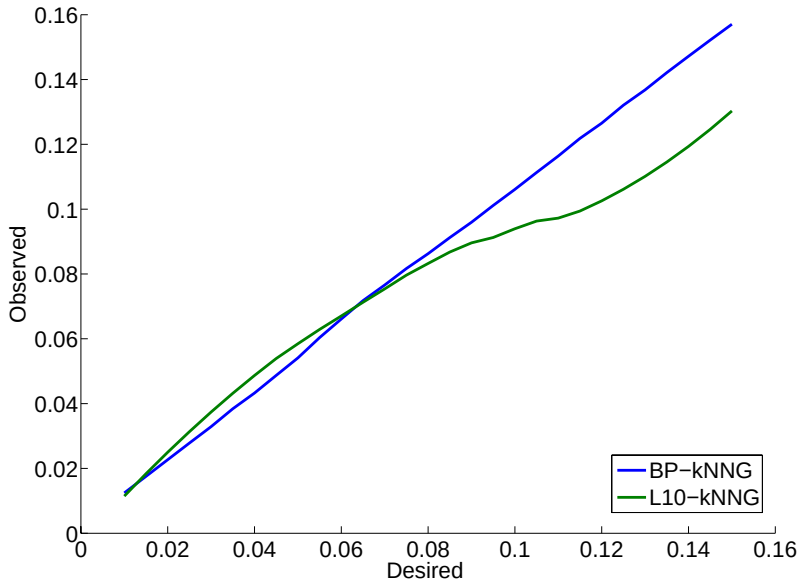


Figure 5.7: Comparison of observed false alarm rates for L1O-kNNG and BP-kNNG with the desired false alarm rates. The observed false alarm rates agree well with the desired false alarm rates.

are from the UCI repository [3]. Full details about the data can be found in [52].

The comparison performance is evaluated in terms of averaged AUC (area under ROC curve) in Table 5.3 and processing time (a total of training and test time) in Table 5.2. Results for BP-kNNG are compared with results for L1O-kNNG, K-LPE, MassAD, iForest and ORCA in Table 5.3 and Table 5.2. The results for MassAD, iForest and ORCA are reproduced from [83]. MassAD and iForest were implemented in Matlab and tested on an AMD Opteron machine with a 1.8 GHz processor and 4 GB memory. The results for ORCA, LOF and 1-SVM were conducted using the same experimental setting but on a faster 2.3 GHz machine. We exclude the results for LOF and 1-SVM because MassAD, iForest and ORCA have been shown to outperform LOF and 1-SVM in [83].

We implemented BP-kNNG, L1O-kNNG and K-LPE in Matlab on an Intel 2 GHz processor with 3 GB RAM. We note that this machine is comparable to the AMD Opteron machine with a 1.8 GHz processor. We choose $T = 10^4$ training samples

Data sets	BP	L10	K-LPE	1-SVM	Kernel	Mass	iF	ORCA
HTTP	3.81	.10/i	.19/i	35872	33	34	147	9487
Forest	7.54	.18/i	.18/i	9738	18	18	79	6995
Mulcross	4.68	.26/i	.17/i	7343	17	17	75	2512
SMTP	0.74	.11/i	.17/i	987	7	7	26	267
Shuttle	1.54	.45/i	.16/i	333	4	4	15	157

Table 5.2: Comparison of anomaly detection schemes in terms of run-time for BP-kNNG (BP) against other state-of-the-art anomaly detection methods. When reporting results for L1O-kNNG and K-LPE, we report the processing time per test instance (/i). We note that BP-kNNG algorithm requires the least run-time.

and fix $k = 50$ in all three cases. For BP-kNNG, we fix $s = 5$ and $N = 10^3$. When reporting results for L1O-kNNG and K-LPE, we report the processing time per test instance (/i). We are unable to report the AUC for K-LPE because of the large processing time and for L1O-kNNG because it cannot operate at high false alarm rates.

From the results in Table 5.3 and Table 5.2, we see that BP-kNNG performs comparably in terms of AUC to the other algorithms, while having the least processing time across all algorithms (implemented on different, but comparable machines). In addition, BP-kNNG allows the specification of a threshold for anomaly detection at a desired false alarm rate. This is corroborated by the results in Table 5.4, where we see that the observed false alarm rates across the different data sets are close to the desired false alarm rate.

5.6 Discussion

The geometric entropy minimization (GEM) principle was introduced in [36] to extract minimal set coverings that can be used to detect anomalies from a set of training samples. In this paper we propose a bipartite k-nearest neighbor graph (BP-kNNG) algorithm based on the GEM principle. BP-kNNG inherits the theoretical

Data sets	BP	L10	K-LPE	1-SVM	Kernel	Mass	iF	ORCA
HTTP	0.994	NA	NA	0.90	0.99	1.00	1.00	0.36
Forest	0.862	NA	NA	0.90	0.69	0.91	0.87	0.83
Mulcross	1.00	NA	NA	0.58	1.00	0.99	0.96	0.33
SMTP	0.924	NA	NA	0.78	0.60	0.86	0.88	0.87
Shuttle	0.992	NA	NA	0.79	0.92	0.99	1.00	0.60

Table 5.3: Comparison of anomaly detection schemes in terms of AUC against other state-of the-art anomaly detection methods. We are unable to report the AUC for K-LPE and L1O-kNNG because of the large processing time. We note that BP-kNNG compares favorably in terms of AUC.

Data sets	Desired false alarm				
	0.01	0.02	0.05	0.1	0.2
HTTP (KDD'99)	0.007	0.015	0.063	0.136	0.216
Forest	0.009	0.015	0.035	0.071	0.150
Mulcross	0.008	0.014	0.040	0.096	0.186
SMTP (KDD'99)	0.006	0.017	0.046	0.099	0.204
Shuttle	0.026	0.030	0.045	0.079	0.179

Table 5.4: Comparison of desired and observed false alarm rates for BP-kNNG. There is good agreement between the desired and observed rates.

optimality properties of GEM methods including consistency, while being an order of magnitude faster than the methods proposed in [36].

We compared BP-kNNG against state of the art anomaly detection algorithms and showed that BP-kNNG compares favorably in terms of both ROC performance and computation time. In addition, BP-kNNG enjoys several other advantages including the ability to detect anomalies at a desired false alarm rate. In BP-kNNG, the p-values of each test point can also be easily computed (5.1), making BP-kNNG easily extendable to incorporating false discovery rate constraints.

CHAPTER VI

Ensemble methods

6.1 Introduction

We showed in the previous chapters that for d-dimensional data, the variance of the proposed k-NN estimators decay as $O(T^{-1})$, where T is the sample size, while the bias, because of the curse of dimensionality, decays as $O(T^{-1/(1+d)})$. One can use the boundary compensated k-NN graphs described in Chapter 3 to reduce the bias to $O(T^{-2/(2+d)})$. However, the bias $O(T^{-2/(2+d)})$ still dominates the mean square error (MSE) in high dimensions.

To accelerate the slow rate of convergence of the bias in high dimensions, we propose a weighted ensemble estimator for ensembles of estimators that satisfy conditions $\mathscr{C}.1(6.1)$ and $\mathscr{C}.2(6.2)$ defined below. Optimal weights, which serve to lower the bias of the ensemble estimator to $O(T^{-1/2})$, can be determined by solving a sparsity-inducing convex optimization problem. Remarkably, this optimization problem does not involve any density-dependent parameters and can therefore be performed off-line. This then ensures MSE convergence of the weighted estimator at the parametric rate of $O(T^{-1})$.

We will first explain the proposed ensmeble method in a general setting, and then apply the method to specific examples including entropy estimation for anomaly detection, intrinsic dimension estimation and minimum volume set estimation.

6.1.1 Previous work

For Shannon and Rényi entropy estimators proposed by Leonenko *et al.*, Liitiäinen *et al.* [51] showed that the bias is of order $O(T^{-1/d})$ while the variance is of order $O(T^{-1})$. For moderate to large dimensions d, the contribution of the bias therefore dominates the MSE. To partially address this problem, Liitiäinen *et al.* considered a weighted k-NN estimator with reduced bias of $o(T^{-1/d})$ and variance of $O(T^{-1})$. In this paper, we extend Liitiäinen *et al.*'s work by determining weights which will reduce the bias of the weighted estimator to $O(T^{-1/2})$. Furthermore, we will extend the application of weighted estimators to general entropy, divergence and MI estimation and intrinsic dimension estimation.

Birge and Massart show that for density f in a Holder smoothness class with s derivatives, the minimax MSE rate for estimation of a smooth functional is $T^{-2\gamma}$, where $\gamma = \min\{1/2, 4s/(4s+d)\}$. This means that for $s > 4/d$, parametric rates are achievable. When the density f is $s > d/4$ times differentiable, certain estimators of functionals of the form of the form $\int g(f(x), x)f(x)dx$, proposed by Birge and Massart [10], Laurent [47] and Giné and Mason [31], achieve the parametric MSE convergence rate of $O(T^{-1})$. The key ideas in [10, 47, 31] are: (i) estimation of quadratic functionals $\int f^2(x)dx$ with MSE convergence rate $O(T^{-1})$; (ii) use of kernel density estimators with kernels that satisfy the following symmetry constraints:

$$\int K(x)dx = 1, \int x^r K(x)dx = 0$$

for $r = 1, .., s$; and finally (iii) truncating the kernel density estimate so that it is bounded away from 0. By using these ideas, the estimators proposed by [10, 47, 31] are able to avoid the curse of dimensionality.

In contrast, the ensemble estimators proposed in this chapter require higher order smoothness conditions on the density, i. e. the density must be $s > 2d$ times

differentiable. However, our estimators are much simpler to implement in contrast to the estimators proposed in [10, 47, 31]. In particular, the estimators in [10, 47, 31] require separately estimating quadratic functionals of the form $\int f^2(x)dx$, and using truncated kernel density estimators with symmetric kernels, conditions that are not required in this paper. Our estimator is a simple convex combination of an ensemble of estimators, where the ensemble satisfies conditions $\mathscr{C}.1$ and $\mathscr{C}.2$, and is therefore trivial to implement. Observe that Birge and Massart showed that for $s > 2d$ (which is effectively the assumption that the density f has $2d$ derivatives), it is indeed possible to achieve parametric sqrt MSE rate of $1/\sqrt{T}$. The proposed ensemble estimators therefore do not violate the minimax rate results of Birge and Massart.

Ensemble based methods have been previously proposed in the context of classification. For example, in both boosting [72] and multiple kernel learning [46] algorithms, lower complexity weak learners are combined to produce classifiers with higher accuracy. Our work differs from these methods in several ways. First and foremost, our proposed method performs estimation rather than classification. An important consequence of this is that the weights we use are *data independent*, while the weights in boosting and multiple kernel learning must be estimated from training data since they depend on the unknown distribution.

6.2 General methodology

Let $\bar{l} = \{l_1, .., l_L\}$ denote a vector of indices. For an ensemble of estimators $\{\hat{\mathbf{E}}_l\}_{l\in\bar{l}}$ of E, define the weighted ensemble estimator with respect to weights $w = \{w(l_1), \ldots, w(l_L)\}$ as

$$\hat{\mathbf{E}}_w = \sum_{l\in\bar{l}} w(l)\hat{\mathbf{E}}_l$$

where the weights satisfy $\sum_{l\in\bar{l}} w(l) = 1$. Assume that this ensemble of estimators $\{\hat{\mathbf{E}}_l\}_{l\in\bar{l}}$ satisfy the following two conditions:

- $\mathscr{C}.1$ The bias is given by

$$\mathbb{B}(\hat{\mathbf{E}}_l) \;=\; \sum_{i\in\mathcal{I}} c_i \psi_i(l) T^{-i/2d} + O(1/\sqrt{T}), \tag{6.1}$$

where c_i are constants that depend on the underlying density, $\mathcal{I}$ is a finite index set with cardinality $I < L$, $\min(\mathcal{I}) = i_0 > 0$ and $\max(\mathcal{I}) = i_d <= d$, and $\psi_i(l)$ are basis functions that depend only on l.

- $\mathscr{C}.2$ The variance is given by

$$\mathbb{V}(\hat{\mathbf{E}}_l) \;=\; c_v \left(\frac{1}{T}\right) + o\left(\frac{1}{T}\right). \tag{6.2}$$

Theorem VI.1. *For an ensemble of estimators $\{\hat{\mathbf{E}}_k\}_{k\in\bar{k}}$, assume that the conditions $\mathcal{C}.1$ and $\mathcal{C}.2$ hold. Then, it is possible to choose a weight vector w such that*

$$\mathbb{E}[(\hat{\mathbf{E}}_w - E)^2] = \Theta(1/T).$$

Furthermore, the optimal weight w

(i) is independent of the underlying density, the T sample realizations and the functional E of interest, and

(ii) can be determined as the solution to an off-line convex optimization problem.

Proof. For each $i \in \mathcal{I}$, define $\gamma_w(i) = \sum_{l\in\bar{l}} w(l)\psi_i(l)$. The bias of the ensemble estimator is given by

$$\mathbb{B}(\hat{\mathbf{E}}_w) \;=\; \sum_{i\in\mathcal{I}} c_i \gamma_w(i) T^{-i/2d} + O\left(\frac{1}{\sqrt{T}}\right). \tag{6.3}$$

Denote the covariance matrix of $\{\hat{E}_l; l \in \bar{l}\}$ by Σ_L. Let $\bar{\Sigma}_L = \Sigma_L/T$. Observe that by (6.2) and Cauchy-Schwarz, the entries of $\bar{\Sigma}_L$ are $O(1)$. The variance of the

weighted estimator $\hat{E}_w$ can then be bound as follows:

$$\begin{aligned}\mathbb{V}(\hat{E}_w) &= \mathbb{V}(\sum_{l=1}^{d} w_l \hat{E}_l) = w'\Sigma_L w = \frac{w'\bar{\Sigma}_L w}{T} \\ &\leq \frac{\lambda_{\max}(\bar{\Sigma}_L)||w||_2^2}{T}. \end{aligned} \tag{6.4}$$

We seek a weight vector w that (i) ensures that the bias of the weighted estimator is $O(T^{-1/2})$ and (ii) has low ℓ_2 norm $||w||_2$ in order to limit the contribution of the variance of the weighted estimator. To this end, let w_o be the solution to the convex optimization problem

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & ||w||_2 \\ \text{subject to} \quad & \sum_{l \in \bar{l}} w(l) = 1, \\ & \gamma_w(i) = 0, \ i \in \mathcal{I}. \end{aligned}$$

This problem can be equivalently expressed as

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & ||w||_2 \\ \text{subject to} \quad & A_0 w = b, \end{aligned}$$

where A_0 and b are defined below. Let $f_{\mathcal{IN}} : \mathcal{I} \to \{1, .., I\}$ be a bijective mapping. Let a_0 be the vector of ones: $[1, 1..., 1]_{1\times L}$; and let $a_{f_{\mathcal{IN}}(i)}$, for $i \in \mathcal{I}$ be given by $a_{f_{\mathcal{IN}}(i)} = [\psi_i(l_1), .., \psi_i(l_L)]$. Define $A_0 = [a_0', a_1', ..., a_I']'$, $A_1 = [a_1', ..., a_I']$ and $b = [1; 0; 0; ..; 0]_{(I+1)\times 1}$. Then, the optimal minimum $\eta(d) := ||w_o||_2$ is given by

$$\eta(d) = \sqrt{\frac{\det(A_1 A_1')}{\det(A_0 A_0')}}.$$

Consequently, by (6.3), the bias $\mathbb{B}[\hat{E}_{w_o}] = O(1/\sqrt{T})$. By (6.4), the estimator variance $\mathbb{V}[\hat{E}_{w_0}]$ is of order $O(1/T)$. This concludes the proof.

□

In practice, for stability reasons, we solve for w_o using the alternative optimization problem defined below:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \epsilon \\ \text{subject to} \quad & \gamma_w(0) = 1, \\ & |\gamma_w(i) T^{-i/d}| \leq \epsilon, \ i \in \mathcal{I}, \\ & ||w||_2 \leq \eta(d). \end{aligned} \tag{6.5}$$

By design, for this choice of $\eta(d)$, the solution ϵ to (II.5) will be of order $O(1/\sqrt{T})$ for every T and insure that the weighted estimator bias behaves as $\mathbb{B}[\hat{E}_{w_o}] = O(1/\sqrt{T})$ while the estimator variance $\mathbb{V}[\hat{E}_{w_o}]$, by (6.4), is of order $O(1/T)$. The optimization problem (6.5) is convex.

Also observe that if we choose to minimizing the ℓ_1 norm of w, for moderately large values of the length of the weight vector w, the solution to the optimization problem would have been sparse [23]. The relative pros and cons of using ℓ_1 norm in place of the ℓ_2 norm is a topic of future work.

Next, we will verify conditions $\mathscr{C}.1$(6.1) and $\mathscr{C}.2$(6.2) for (i) density estimation, (ii) entropy, divergence and MI estimation and finally, (iii) we will extend the idea of ensemble estimators to dimension estimation.

6.3 Ensemble estimators for density estimation

In this section, we apply the theory of ensemble estimation to density estimation problems. Let X be any arbitrary point such that that the density f is d-times differentiable in a neighborhood of X. We estimate the density at X using the k-NN density estimator $\hat{\mathbf{f}}_k(X)$, which is evaluated using the T independent realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ drawn from f. Then,

6.3.1 Analysis of MSE

Theorem VI.2. *The bias of the bipartite k-NN density estimator is given by*

$$\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) \;=\; \sum_{i\in\mathcal{I}} h_i(X)\left(\frac{k}{T}\right)^{i/d} + O\left(\frac{1}{k} + \frac{k}{T}\right), \tag{6.6}$$

where h_i are constants and $\mathcal{I} = \{2, \ldots, d\}$.

Theorem VI.3. *The variance of the bipartite k-NN density estimator is given by*

$$\mathbb{V}[\hat{\mathbf{f}}_k(X)] \;=\; f^2(X)\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right). \tag{6.7}$$

Proof. These results are derived in Appendix B. In particular, please refer to section B.3.3. □

6.3.2 Optimal MSE rate

From Theorem VI.2, $k/T \to 0$ for the estimator to be unbiased. Likewise from Theorem VI.3 $k \to \infty$ for the variance of the estimator to converge to 0. We optimize the choice of number of nearest neighbors k for minimum M.S.E. Minimizing the M.S.E. over k is equivalent to minimizing the square of the bias over k. The optimal choice of k is given by

$$k_{opt} \;=\; \Theta(T^{\frac{4}{4+d}}), \tag{6.8}$$

and the MSE evaluated at k_{opt} is $\Theta(T^{\frac{-4}{4+d}})$.

6.3.2.1 Discussion

The optimal MSE for the estimator $\hat{\mathbf{f}}_k(X)$ is achieved for the choice of $k = \Theta(T^{4/(4+d)})$, and is given by $\Theta(T^{-4/(4+d)})$. Our goal is to reduce the MSE to $O(T^{-1/2})$. We do so by applying the method of weighted ensembles.

6.3.3 Weighted ensemble entropy estimator

For a positive integer $L > d$, choose $\bar{l} = \{l_1, \ldots, l_L\}$ to be positive real numbers. Define the mapping $k(l) = \lfloor l\sqrt{T} \rfloor$ and let $\bar{k} = \{k(l); l \in \bar{l}\}$. Define the weighted ensemble estimator

$$\hat{\mathbf{f}}_w(X) = \sum_{l \in \bar{l}} w(l) \hat{\mathbf{f}}_k(X).$$

From theorems II.1 and II.2, we see that the bias of the ensemble of estimators $\{\hat{\mathbf{f}}_k(X); l \in \bar{l}\}$ satisfy a modified form of $\mathscr{C}.1$ [79], with T replaced by $\sqrt{T}$, when we set $\psi_i(l) = l^{i/d}$ and $\mathcal{I} = \{2, .., d\}$. Furthermore, the general form of the variance of $\hat{\mathbf{f}}_k(X)$ follows a modified version of $\mathscr{C}.2$ [79] with T again replaced by $\sqrt{T}$. This implies that we can use the weighted ensemble estimator $\hat{\mathbf{f}}_w(X)$ to estimate entropy at $O(T^{-1/2})$ convergence rate by setting w equal to the optimal weight w_o given by (II.5). In other words, the MSE reduces from $O(T^{-4/(4+d)})$ [30] to $O(1/\sqrt{T})$ using ensemble estimation. This result is stated as the following theorem.

Theorem VI.4. *The MSE of the density estimator* $\hat{\mathbf{f}}_w(X)$ *is given by*

$$MSE(\hat{\mathbf{f}}_w(X)) \;\; = \;\; \Theta(1/\sqrt{M}).$$

6.3.4 Experiments

In our simulations we consider density estimation using four different choices of density estimators: (i) the k-NN density estimator defined in Chapter 2, (ii) a weighted density estimator proposed by Biau *et al.* [8], (iii) an ensemble estimator

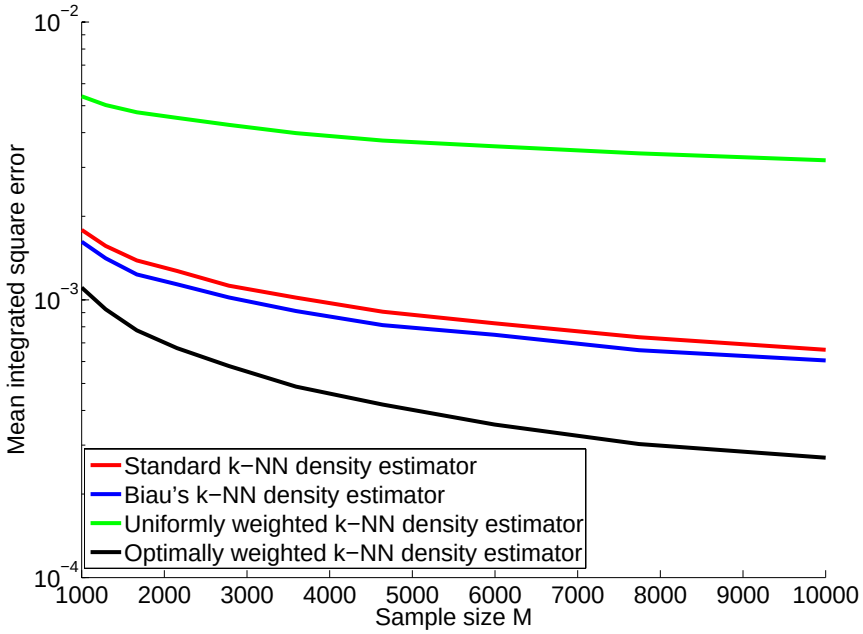


Figure 6.1: Variation of integrated mean square error of density estimates as a function of sample size T using samples drawn from 5-d standard normal distribution. From the figure, we see that our weighted estimator has the fastest rate of convergence.

with uniform weights and (iv) the optimal weighted ensemble estimator $\hat{\mathbf{f}}_w(X)$. We estimate density for the following class of densities: (i) 5 dimensional standard normal density, and (ii) 5 dimensional beta density with parameters $1.5, 2$. The MSE results of these different estimators for the two different densities f are shown in Fig. 6.1 and Fig. 6.2 as a function of sample size T. It is clear from the figures that the proposed ensemble estimator $\hat{\mathbf{f}}_w$ has significantly faster rate of convergence as predicted by our theory.

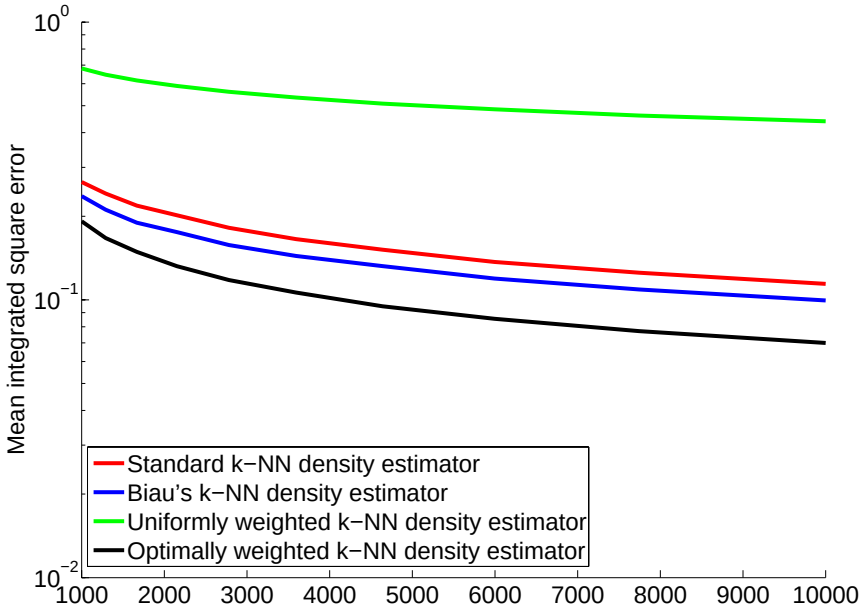


Figure 6.2: Variation of integrated mean square error of density estimates as a function of sample size T using samples drawn from 5-d beta distribution. From the figure, we see that our weighted estimator has the fastest rate of convergence.

6.4 Ensemble estimators for entropy and divergence estimation

We first deal with the case where $\mathcal{S}' \cap \mathcal{S}_{\mathcal{I}} = \phi$. If we assume that the density f is d-times differentiable, we show in section B.3.3 that,

$$\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) = \sum_{i=2}^{d} h_i(X) \left(\frac{k}{M}\right)^{i/d} + O\left(\frac{1}{k} + \frac{k}{M}\right), \tag{6.9}$$

for any $X \in \mathcal{S}'$ (see Appendix B.1).

Let $\hat{G}_k$ denote any one of any one of the following: the entropy estimator with $\hat{G}_k := \hat{G}_k(f)$, the divergence estimator with $\hat{G}_k := \hat{G}_k(f_1, f_2)$ and the MI estimator with $\hat{G}_k = \hat{G}_k(f_{12})$. Assume that the density f is $2d$-times differentiable and the functional $g(x, y)$ is d-times differentiable wrt x. Using the results from Chapter 2 in

combination with Eq. 6.9, we then have the following general form for the bias of $\hat{G}_k$:

Theorem VI.5. *The bias of the plug-in estimator $\hat{G}_k$ is given by*

$$\mathbb{B}(\hat{\mathbf{G}}_k) \;=\; \sum_{i\in\mathcal{I}} c_{1,i}\left(\frac{k}{M}\right)^{i/d} + c_2\left(\frac{1}{k}\right) + o\left(\frac{1}{k}+\frac{k}{M}\right),$$

where $c_{1,i}$ and c_2 are constants and $\mathcal{I} = \{1,\ldots,d\}$.

Proof. The proof follows exactly along the lines of the proof of Theorem 2.1, 2.4 and 2.7 and subsequently observing that

$$\begin{aligned}
&\mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}) - g(f(\mathbf{Z}),\mathbf{Z}))]\\
&= \sum_{i=1}^{d}\mathbb{E}\left[g^{(i)}(f(\mathbf{Z}),\mathbf{Z})\Big(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_k(\mathbf{Z})] - f(\mathbf{Z})\Big)^i\right]\\
&= \sum_{i=1}^{d} c_{1,i}(k/M)^{i/d} + o((k/M)). \qquad (6.10)
\end{aligned}$$

□

6.4.1 Weighted ensemble entropy estimator

For a positive integer $L > d$, choose $\bar{l} = \{l_1,\ldots,l_L\}$ to be positive real numbers. Define the mapping $k(l) = \lfloor l\sqrt{M}\rfloor$ and let $\bar{k} = \{k(l); l \in \bar{l}\}$. Define the weighted ensemble estimator

$$\hat{G}_w = \sum_{l\in\bar{l}} w(l)\hat{G}_{k(l)}.$$

From theorems VI.5 and Theorem 2.2, 2.5 and 2.8, we see that the bias of the ensemble of estimators $\{\hat{G}_{k(l)}; l \in \bar{l}\}$ satisfy $\mathscr{C}.1$ (Section 6.2) when we set $\psi_i(l) = l^{i/d}$ and $\mathcal{I} = \{1,..,d\}$. Furthermore, the general form of the variance of $\hat{\mathbf{G}}_{k(l)}$ follows $\mathscr{C}.2$ (Section 6.2) because $N = M = T/2$. This implies that we can use the weighted ensemble estimator $\hat{G}_w$ to estimate entropy, divergence or MI at $O(T^{-1})$ convergence rate by setting w equal to the optimal weight w_o (shown in Fig. 6.3) given by (II.5).

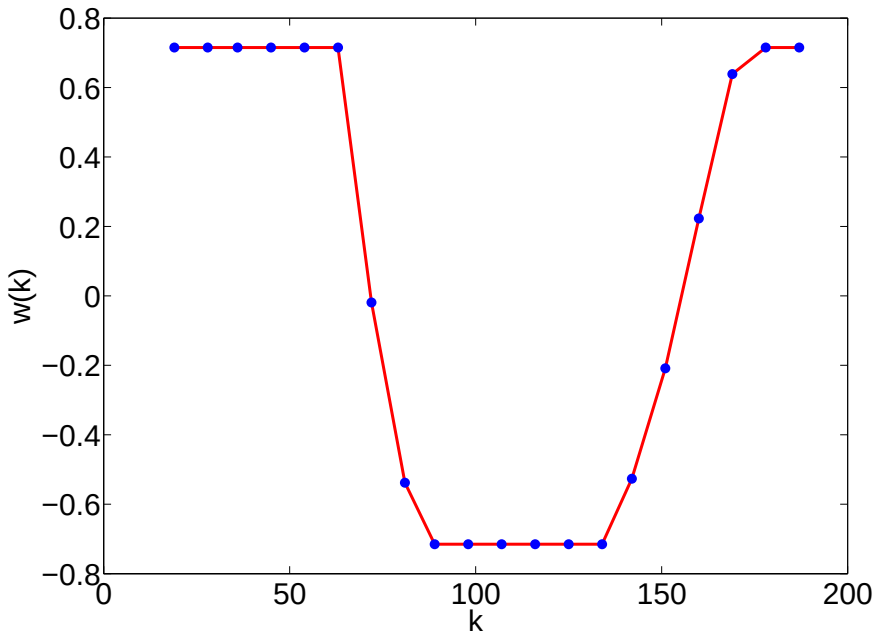


Figure 6.3: Variation of optimal weight w as a function of k (blue dots represent each entry $k \in \bar{k}$).

This result is stated as the theorem below.

Theorem VI.6. *The MSE of the ensemble estimator* $\hat{\mathbf{G}}_w$ *is given by*

$$MSE(\hat{\mathbf{G}}_w) \quad = \quad \Theta(1/T).$$

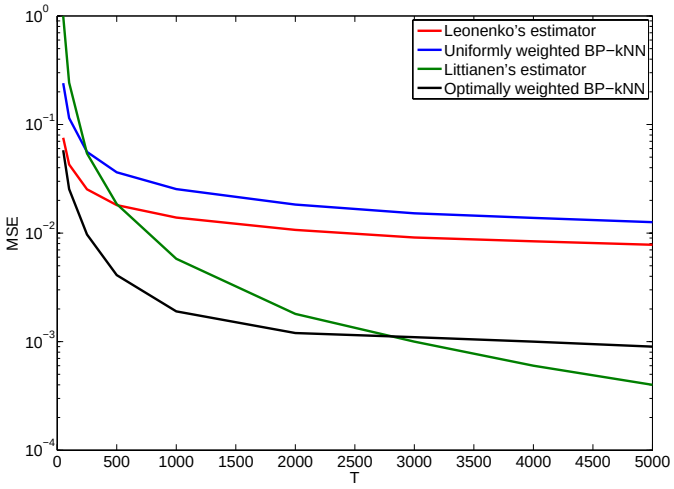
In the next section, we illustrate the performance of weighted ensemble estimators for estimation of Rényi entropy. Observe that in the case of estimation of Rényi entropy, the set $\mathcal{S}' = \mathcal{S}$ and therefore, the assumption $\mathcal{S}' \subset \mathcal{S}_{\mathcal{I}}$ is violated. Even so, we observe that the weighted ensemble estimator works well.

6.4.2 Simulations

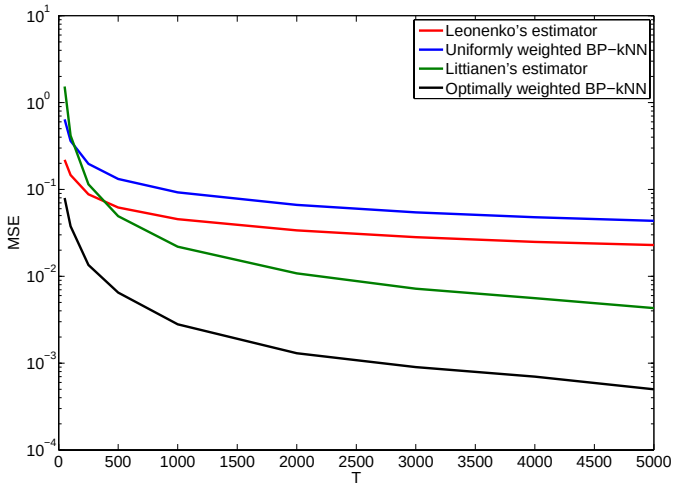
We will compare the MSE of the ensemble estimator $\hat{G}_w$ for estimating Rényi entropy with $\alpha = 0.95$. The ensemble is given by $\{\check{H}_l^{(\alpha)}\}, l = \{1, .., 50\}$ with $\alpha = 0.95$. We consider four different choices of weight vectors: The nearest neighbor estimator

of Leonenko *etal* with weight $w_s = [1, 0, \ldots, 0]$, the uniform weighted estimator with weight $w_u = (1/k)[1, \ldots, 1]$, the first-order correction estimator of Liitiäinen *etal* with weight w_f, and finally the optimized weighted estimator with weight w_o.

We estimate entropy for the following class of densities: 6 dimensional mixture density $f_m(p, a, b) = pf_\beta(a, b) + (1 - p)f_u$; f_β: Beta density with parameters a,b; f_u: Uniform density; Mixing ratio p. We show representative results obtained by simulating samples from $f_m(0.8, 1.5, 1.5)$. The MSE error performances for these densities are shown in Fig. 6.4(a) and Fig. 6.4(b) respectively.



(a) MSE comparison for density $f_m(.8, 2, 2)$. A lower order bias correction suffices for this density. The weighted BP-kNN estimator outperforms other estimators for small sample sizes, while the first-order correction estimator of Liitiäinen *etal* works better for larger sample sizes.



(b) MSE comparison for density $f_m(.8, 1.5, 1.5)$. This density requires higher order bias correction. The weighted BP-kNN estimator therefore has superior MSE performance for all sample size regimes.

Figure 6.4: Comparison of MSE of weighted estimators for different choices of weight vectors. The proposed optimal weight (6.5) outperforms the rest of the choice of weight vectors.

The observed MSE performance can be explained as follows. For the density $f_m(.8, 2, 2)$, we note that the higher order co-efficients in the bias expansion c_i, $i > 2$ are identically 0. The MSE performance for the optimized weighted entropy estimator is better than Liitiäinen *etal*'s first-order correction estimator for small sample sizes because the first-order correction estimator does not account for second order terms in the bias. With increasing sample size, the contribution of the second order bias terms become negligible. Because the first order bias term is explicitly set to 0 in the first-order correction estimator, it performs better with increasing sample size as compared to the optimized weighted estimator. On the other hand, for the density $f_m(.8, 1.5, 1.5)$, higher order co-efficients are non-zero and therefore contribute to bias. The optimized weighted estimator with higher order bias correction and lower norm therefore works better in this case.

6.4.2.1 Anomaly detection revisited

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). To capture the temporal dependency between successive measurements, for each time point we form a temporal dependency discriminant by con-

sidering vectors of $d = 3$ successive time samples at each sensor and estimating the entropy by averaging over $M = 182$ spatial samples. We note that the ground truth indicator is only for evaluating the detecting performance and the detection scheme presented here is conducted in a completely unsupervised manner.

In order to detect anomalies, we form a running estimate of the Rényi α-entropy with $\alpha = 0.95$, of the 3-dimensional time sequence using weighted k-NN estimators with first order correction weight w_f and optimized correction weight w_o. We perform anomaly detection by thresholding the entropy estimate. A time sample is regarded to be anomalous if the entropy estimate exceeds a specified threshold. ROC curves corresponding to first-order correction weight w_f and optimized correction weight w_o are shown in Fig. 2. The Area under the ROC curve (AUC) was found to be 0.9538 and 0.9821 for the first-order correction estimator and the optimized weighted estimator respectively. It is clear that the detection performance using the optimized weight w_o is superior to the performance using Liitiäinen *etal*'s first-order correction weight w_f.

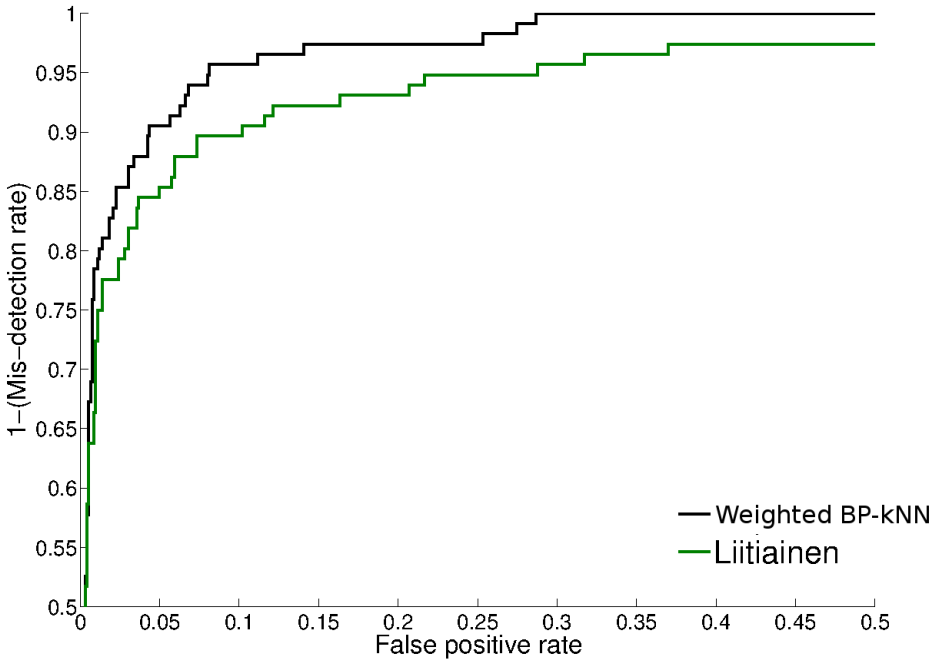


Figure 6.5: Comparison of ROC curves for anomaly detection. The weighted BP-kNN estimator outperforms Liitiäinen *etal*'s first-order correction estimator.

6.5 Angular plug-in estimators for entropy

We will now extend these ideas to the case where $\mathcal{S}' \cap \mathcal{S}_{\mathcal{I}} \neq \phi$. We do so by modifying the k-NN estimators as follows. We propose a modification of k-NN densities, called angular k-NN densities, with the property that the k-NN neighborhoods of these densities are sectors of hyper-spheres, as opposed to the complete spherical regions $\mathbf{S}_k(X)$ in the case of standard k-NN. These sectors are directed such that they always lie in the interior of the support $\mathcal{S}$ with high probability.

Observe that, because the sectors are directed to lie in the interior of the support $\mathcal{S}$, angular k-NN plug-in estimators also compensate for the bias due to the boundary. However, in contrast to the extrapolation based compensation proposed in Chapter 3, angular k-NN plug-in estimators have a slower MSE rate of convergence

$((O(T^{-1/(1+d)}))$ vs $(O(T^{-2/(2+d)})))$. This is because the sectors of hyper-spheres are asymmetric regions about the center of the sphere. However, under higher order smoothness conditions on the density, we show that these angular estimators satisfy the regularity conditions $\mathscr{C}.1$ and $\mathscr{C}.2$, and can therefore be aggregated to produce ensemble estimators with much faster MSE rates of convergence $O(1/T)$. The regularity conditions $\mathscr{C}.1$ and $\mathscr{C}.2$ are in general not satisfied by the extrapolation based estimators of Chapter 3, thereby necessitating the use of angular k-NN plug-in estimators.

6.5.1 Entropy estimation problem

Without loss of generality, set $\mathcal{S}' = \mathcal{S}$. We are therefore interested in estimating non-linear functionals $G(f)$ of d-dimensional multi-variate densities f with compact support $\mathcal{S}$, where $G(f)$ has the form

$$G(f) = \int g(f(x), x) f(x) d\mu(x),$$

for some smooth function $g(f(x), x)$. Let $\mathcal{B}$ denote the boundary of $\mathcal{S}$. Assume that the support $\mathcal{S}$ is a union of a finite set of convex regions. Let 2θ be a known upper bound on the curvature of the boundary of these regions.

6.5.2 Plug-in estimators of entropy

The plug-in estimator is constructed using a data splitting approach as follows. The data is randomly subdivided into two parts $\mathcal{X}_N = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\mathcal{X}_M = \{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ of N and M points respectively. Consider plug-in estimators of the form

$$\hat{\mathbf{G}}_{k,\theta}(f) \;=\; \frac{1}{N} \sum_{i=1}^{N} g(\hat{\mathbf{f}}_{k,\theta}(\mathbf{X}_i), \mathbf{X}_i). \tag{6.11}$$

where $\hat{\mathbf{f}}_{k,\theta}(.)$ is the angular k-NN density estimator defined below.

6.5.3 Angular k-NN density estimates

Let $\{\mathbf{X}_{i1}, .., \mathbf{X}_{iK}\}$ be the $K = \lfloor k \times N/M \rfloor$-nearest neighbors of the point $\mathbf{X}_i$ among $\{\mathbf{X}_1, .., \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, .., \mathbf{X}_N\}$. For each $\mathbf{X}_i \in \mathcal{X}_N$, Define $N(\mathbf{X}_i)$ to be the normalized average direction of the K-NN of $\mathbf{X}_i$ with respect to $\mathcal{X}_N$, i. e. ,

$$N(\mathbf{X}_i) = \mathbf{X}_i + \sum_{j=1}^{K} \frac{\mathbf{X}_{ij} - X}{||\mathbf{X}_{ij} - X||}.$$

Conditioned on $\mathcal{X}_N$, for each $X \in \mathcal{X}_N$, let $\mathbf{d}_{k,\theta}(X) = \inf_{\tau>0}\{\tau : \#\{i : i \in \{N + 1, .., N+M\}; |X-\mathbf{X}_i| < \tau; \langle X - \mathbf{X}_i, X - N(X)\rangle < \theta\} \geq k\}$. Define the corresponding k-NN region to be the arc $\mathbf{S}_{k,\theta}(X) = \{Z : d(X, Z) \leq \mathbf{d}_{k,\theta}(X); \langle X - Z, X - N(X)\rangle < \theta\}$. Let the volume of this region be $\mathbf{V}_{k,\theta}(X)$. Define the angular k-NN density estimate at X as

$$\hat{\mathbf{f}}_{k,\theta}(X) = \frac{k-1}{M\mathbf{V}_{k,\theta}(X)}. \tag{6.12}$$

The moment properties of angular k-NN densities are discussed in detail in Appendix B.6.5.

6.5.4 Analysis of MSE

Under the assumptions stated in Section 2.3.1, in addition to the assumption that the density f has $2d$ continuous partial derivatives, we show that the following theorems hold:

Theorem VI.7. *The bias of the plug-in estimator $\hat{\mathbf{G}}_{k,\theta}(f)$ is given by*

$$\mathbb{B}(\hat{\mathbf{G}}_{k,\theta}(f)) \;=\; \sum_{i\in\mathcal{I}} c_{1,i} \left(\frac{k}{M}\right)^{i/d} + c_2 \left(\frac{1}{k}\right) + o\left(\frac{1}{k} + \frac{k}{M}\right),$$

where $c_{1,i}$ and c_2 are constants and $\mathcal{I} = \{1, \ldots, d\}$.

Proof. The proof of the theorem is identical to the proof of Theorem II.1 and follows by the use of Lemma B.6 and Lemma B.7. □

Theorem VI.8. *The variance of the plug-in estimator* $\hat{\mathbf{G}}_{k,\theta}(f)$ *is given by*

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

where c_4 *and* c_5 *are constants.*

Proof. The proof of the theorem is identical to the proof of Theorem II.2 and follows by the use of Lemma B.7 and Lemma B.8. □

6.5.5 Optimal MSE rate

From Theorem VI.7, $k \to \infty$ and $k/M \to 0$ for the estimator $\hat{\mathbf{G}}_k$ to be unbiased. Likewise from Theorem VI.8 $N \to \infty$ and $M \to \infty$ for the variance of the estimator to converge to 0. We can optimize the choice of number of nearest neighbors k, and the data splitting proportions $N/(N+M)$, $M/(N+M)$ for minimum M.S.E.

6.5.5.1 Optimal choice of k

Minimizing the M.S.E. over k is equivalent to minimizing the square of the bias over k. The optimal choice of k is given by

$$k_{opt} \quad = \quad \Theta(M^{\frac{1}{1+d}}), \tag{6.13}$$

and the bias evaluated at k_{opt} is $\Theta(M^{\frac{-1}{1+d}})$.

6.5.5.2 Optimal choice of α_{frac}

Observe that the MSE of $\hat{\mathbf{G}}_{k,\theta}$ is dominated by the squared bias ($\Theta(M^{-2/(1+d)})$) as contrasted to the variance ($\Theta(1/N + 1/M)$). This implies that the asymptotic MSE rate of convergence is invariant to the choice of α_{frac}.

6.5.5.3 Discussion

The optimal MSE for the estimator $\hat{\mathbf{G}}_{k,\theta}$ is therefore achieved for the choice of $k = \Theta(M^{1/(1+d)})$, and is given by $\Theta(T^{-2/(1+d)})$. Our goal is to reduce the MSE to $O(T^{-1})$. We do so by applying the method of weighted ensembles described in Section 6.2.

6.5.6 Weighted ensemble entropy estimator

For a positive integer $L > d$, choose $\bar{l} = \{l_1, \ldots, l_L\}$ to be positive real numbers. Define the mapping $k(l) = \lfloor l\sqrt{M} \rfloor$ and let $\bar{k} = \{k(l); l \in \bar{l}\}$. Define the weighted ensemble estimator

$$\hat{G}_{w,\theta} = \sum_{l \in \bar{l}} w(l) \hat{G}_{k(l),\theta}.$$

From theorems II.1 and II.2, we see that the bias of the ensemble of estimators $\{\hat{G}_{k(l),\theta}; l \in \bar{l}\}$ satisfy $\mathscr{C}.1$ (Section 6.2) when we set $\psi_i(l) = l^{i/d}$ and $\mathcal{I} = \{1, .., d\}$. Furthermore, the general form of the variance of $\hat{\mathbf{G}}_{k(l),\theta}$ follows $\mathscr{C}.2$ (Section 6.2) because $N = M = T/2$. This implies that we can use the weighted ensemble estimator $\hat{G}_{w,\theta}$ to estimate entropy at $O(T^{-1})$ convergence rate by setting w equal to the optimal weight w_o given by (II.5). This result is stated as the theorem below.

Theorem VI.9. *The MSE of the plug-in estimator $\hat{\mathbf{G}}_{w,\theta}$ is given by*

$$MSE(\hat{\mathbf{G}}_{w,\theta}) \;\; = \;\; \Theta(1/T).$$

6.5.7 Experiments

In our simulations we consider the estimation of Shannon entropy using five different choices of functional estimators: (i) the k-NN plug-in estimator defined in Chapter 2, (ii) the bias corrected k-NN estimator [32], (iii) the angular plug-in estimator $\hat{G}_{k,\theta}$, (iv) a weighted ensemble of the bias corrected k-NN estimator [32], and

(v) the weighted ensemble estimator $\hat{\mathbf{G}}_{w,\theta}$. We estimate entropy for the following class of densities: 6 dimensional mixture density $f_m(p,a,b) = pf_\beta(a,b) + (1-p)f_u$; f_β: Beta density with parameters a,b; f_u: Uniform density; Mixing ratio p.

6.5.7.1 Variation with sample size T

The MSE results of these different estimators are shown in Fig. 6.6 as a function of sample size T. In this experiment, we fixed $a = 3, b = 3, p = 0.75$ and set $d = 8$. It is clear from the figure that the proposed ensemble estimator $\hat{\mathbf{G}}_{w,\theta}$ has significantly faster rate of convergence as predicted by our theory.

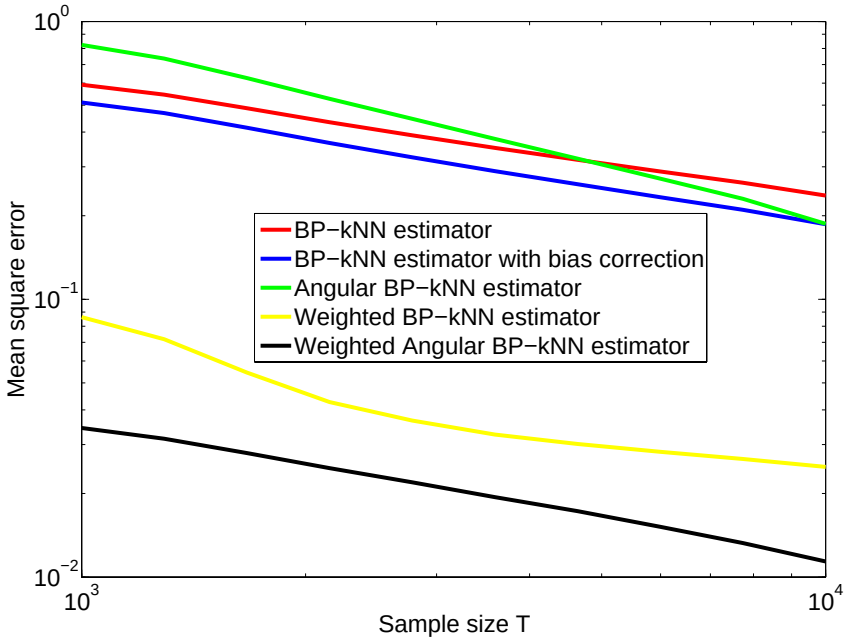


Figure 6.6: Variation of MSE of Shannon entropy estimates as a function of sample size T. From the figure, we see that our proposed angular weighted BP-kNN estimator has the fastest rate of convergence.

6.5.7.2 Variation with dimension d

The MSE results of these different estimators are shown in Fig. 6.7 as a function of dimension d, for fixed sample size $T = 10^3$. In this experiment, we once again fixed $a = 3, b = 3, p = 0.75$. Observe that the MSE of the weighted estimator $\hat{\mathbf{G}}_{w,\theta}$ is significantly smaller than the MSE of the other estimators, and furthermore, the MSE of the weighted estimator $\hat{\mathbf{G}}_{w,\theta}$ increases at a much slower rate as a function of d. This is in agreement with our theory that the MSE of the weighted angular estimator is $O(\eta(d)/T)$, where the function $\eta(d)$ is a parameter in the optimization problem (II.5)(Section 6.2) which increases as a function of d.

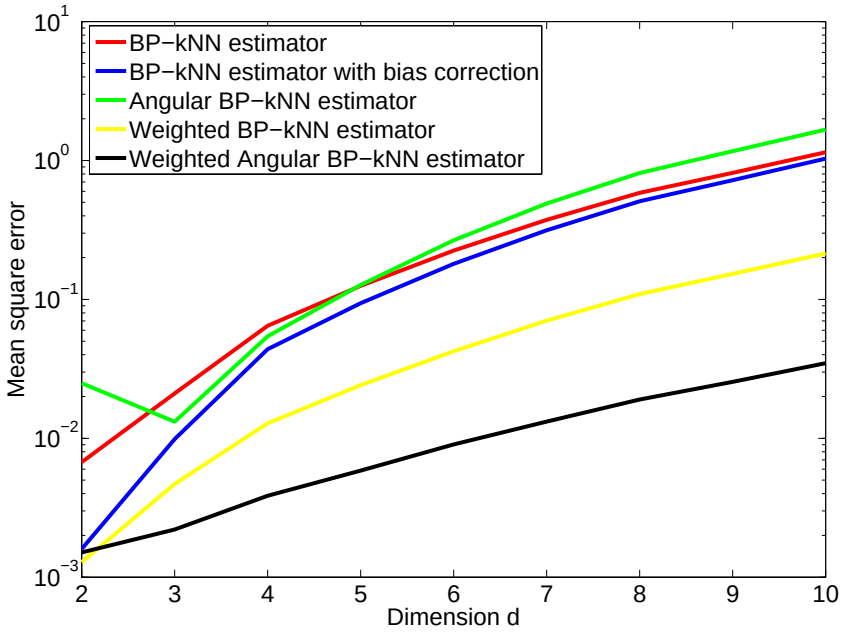


Figure 6.7: Variation of MSE of Shannon entropy estimates as a function of dimension d. From the figure, we see that our proposed angular weighted BP-kNN estimator has the fastest rate of convergence for all dimensions $d > 2$.

6.6 Extension of ensemble estimators to manifolds

In this section, we extend the use of weighted ensemble estimators to data on manifolds.

6.6.1 Bias expansion for density estimation on manifolds

The bias of the k-NN density estimate on the manifold $\hat{f}_{k,g,\mathcal{M}}(X)$ has the following expansion:

$$\begin{aligned}
\mathbb{B}[\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X)] &= \int_{\mathcal{M}^M} (\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X) - f(X)) \prod_{i=1}^{M} (f(\mathbf{x}_i) dx_i) \\
&= \int_{\mathcal{U}^M} (\hat{\mathbf{p}}_k(0) - p(0)) \prod_{i=1}^{M} (p(\phi^{-1}(\mathbf{x}_i)) dy_i) + o(1/M^a) \\
&= \mathbb{B}[\hat{\mathbf{p}}_k(0)] + o(1/M^a) \\
&= \sum_{i \in \mathcal{I}} h_i(X)(k/M)^{i/d} + o((k/M).
\end{aligned}$$

Using lemma IV.3, we have shown in section 4.3.4 that we can write the following relation between the Euclidean and geodesic k-NN density estimators on the manifold:

$$\hat{\mathbf{f}}_{k,e,\mathcal{M}}(X) = \mathcal{A}_{\|_X}(\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X)), \tag{6.14}$$

where the new approximating function $\mathcal{A}_{\|}$ is given by

$$\mathcal{A}_{\|_x}(f) = f + \sum_{i=2}^{d} \tilde{a}_x(i)(k/M)^{i/d} f^{1-i/d} + o(k/M).$$

This implies that the bias of the Euclidean k-NN density estimate $\hat{\mathbf{f}}_{k,e,\mathcal{M}}(X)$ is

given by

$$\begin{aligned}\mathbb{B}[\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X)] &= \mathbb{E}[\hat{\mathbf{f}}_{k,e,\mathcal{M}}(X)] - f(X) \\ &= \mathbb{E}[\mathcal{A}_{\|X}(\hat{\mathbf{f}}_{k,g,\mathcal{M}}(X))] - f(X) \\ &= \sum_{i\in\mathcal{I}} \tilde{h}_i(X)(k/M)^{i/d} + o((k/M).\end{aligned}$$

6.6.2 Bias expansion for entropy estimation on manifolds

The bias of the entropy estimator $\hat{G}_{k,e,\mathcal{M}}$ has the following expansion.

Theorem VI.10. *The bias of the plug-in estimator $\hat{E}_k$ is given by*

$$\mathbb{B}[\hat{G}_{k,e,\mathcal{M}}] = \sum_{i\in\mathcal{I}} c_{1,i}\left(\frac{k}{M}\right)^{i/d} + c_2\left(\frac{1}{k}\right) + o\left(\frac{1}{k} + \frac{k}{M}\right),$$

where $c_{1,i}$ and $c_2 = \mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}} f^2(\mathbf{Y}) g''(f(\mathbf{Y}), \mathbf{Y})/2]$ are constants and $\mathcal{I} = \{1, \ldots, d\}$.

Proof. The proof follows exactly along the lines of the proof of Theorem 2.1 by plugging in (6.15) in the proofs of Lemma D.1. □

Theorem VI.10 verifies condition $\mathscr{C}.1(6.1)$ for the ensemble $\hat{G}_{l,e,\mathcal{M}}$. Theorem IV.5 and therefore verifies condition $\mathscr{C}.2(6.2)$. This implies that we can use ensemble weighted estimators to estimate entropy on the manifold at MSE rate of $O(1/T)$. Next, we will extend these ensemble entropy estimators on the manifold to estimate dimension.

6.7 Ensemble weighted dimension estimator

Recall that we had previously defined the log-length statistic to be

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

in section 4.4.1.1 and showed that

$$\begin{aligned}\breve{H}_k &= \frac{1}{N}\sum_{i=1}^{N}\psi(k) - \log(c_d M) - d\log(\mathbf{d}_{k,e}(\mathbf{X}_i)) \\ &= \psi(k) - \log(c_d M) - d\mathbf{L_k}(\mathcal{X}).\end{aligned}$$

where

$$\breve{H}_k = -\left(\tilde{G}_{k,\mathcal{M}} + \log(k-1) - \psi(k-1)\right),$$

denotes the negative of the Shannon entropy estimator with bias correction.

In this section, we improve on the previous slope based estimator (see section 4.4.1.3) by proposing the following inverse weighted slope estimator

$$\hat{\mathbf{d}}_w^{-1} = \frac{\mathbf{L}_w(\mathcal{X})}{\sum_{k\in\bar{k}} w_k\psi(k)}$$

where

$$\mathbf{L}_w(\mathcal{X}) = \sum_{k\in\bar{k}} w_k\mathbf{L_k}(\mathcal{X}).$$

with the weights w_k satisfying the condition $\sum w_k = 0$ and $\sum w_k\psi(k) = 1$. Now, when we rewrite the dimension estimator, we get the following relation:

$$\hat{\mathbf{d}}_w^{-1} = \mathbf{L}_w(\mathcal{X}) = (1/d)\sum_{k\in\bar{k}} w_k(1-\breve{H}_k)$$

For the ensemble $\breve{H}_l$, $l \in \bar{k}$, we have previously established in section 6.6.2 that $\mathscr{C}.1(6.1)$ and $\mathscr{C}.2(6.2)$ are satisfied. In order to reduce the MSE of the inverse dimension estimator $\hat{\mathbf{d}}_w^{-1}$ to $O(1/T)$, we can therefore use the concept of weighted estimators and solve the convex optimization problem (6.5), but with the constraints $\sum w_k = 0$ and $\sum w_k\psi(k) = 1$ in place of the constraint $\sum w_k = 1$.

The modified optimization problem is therefore given by

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & ||w||_2 \\ \text{subject to} \quad & \sum_{k\in\bar{k}} w_k = 0, \\ & \sum_{k\in\bar{k}} w_k \psi(k) = 1, \\ & \sum_{k\in\bar{k}} w(k) k^{i/d} = 0,\ i \in 1,..,d. \end{aligned} \tag{6.15}$$

Observe that the constraints

$$\sum_{k\in\bar{k}} w(k) k^{i/d} = 0,\ i \in 1,..,d$$

depend on the unknown dimension intrinsic d. We therefore adopt the following iterative matching solution to determine the inverse optimally weighted dimension estimate $\hat{\mathbf{d}}_o^{-1}$:

Algorithm 3 Iterative matching algorithm for weighted dimension estimation

1. Determine initial dimension estimate $d_0 = \lfloor \hat{\mathbf{d}}_s \rfloor$
2. Initialize diff $= \infty$
3. Initialize weighted inverse dimension estimate $\hat{\mathbf{d}}_o^{-1} = 1/d_0$
4. Do:

for Each d_{in} in $[1, D]$ **do**

a. Determine weight $w(d_{in})$ using (6.15) with $d = d_{in}$

b. Determine inverse intrinsic dimension $\hat{\mathbf{d}}_{out}^{-1} = \hat{\mathbf{d}}_{w(d_{in})}^{-1}$

c. Check:

if $\mid \hat{\mathbf{d}}_{out}^{-1} - 1/d_{in} \mid <$ diff **then**

(i) diff $=\mid \hat{\mathbf{d}}_{out}^{-1} - 1/d_{in} \mid$

(ii) $\hat{\mathbf{d}}_o^{-1} = \hat{\mathbf{d}}_{w(d_{in})}^{-1}$

end if

end for

5. Output $\hat{\mathbf{d}}_o^{-1}$

With very high probability $1 - O(1/T)$, the difference $\mid \hat{\mathbf{d}}_{w(d_{in})}^{-1} - 1/d_{in} \mid$ will be minimized for the case when d_{in} is equal to the true intrinsic dimension d. Subse-

quently, when using these optimized weights w evaluated at d, by lemma VI.1, the bias of $\hat{\mathbf{d}}_o^{-1}$ is given by

$$\begin{aligned}\mathbb{B}(\hat{\mathbf{d}}_o^{-1}) &= \sum_{i\in\mathcal{I}} c_{1,i}\left(\frac{k}{M}\right)^{i/d} + c_2\left(\frac{1}{k}\right) + o\left(\frac{1}{k}+\frac{k}{M}\right), \\ &= O(1/\sqrt{M}),\end{aligned}$$

and the variance is given by $\mathbb{V}[\hat{\mathbf{d}}_o^{-1}] = O(1/M + 1/N)$. Because $N = M = T/2$, the overall MSE is order $O(1/T)$. This in turn implies that the weighted dimension estimator given by

$$\hat{\mathbf{d}}_w = 1/\hat{\mathbf{d}}_o^{-1}$$

also converges at the MSE rate of $O(1/T)$ and is therefore an significant improvement over the estimators of Farahmand *et al.* ($\hat{\mathbf{d}}_f$), Levina and Bickel ($\hat{\mathbf{d}}_l$), Costa and Hero ($\hat{\mathbf{d}}_j$) and the uniform slope estimator ($\hat{\mathbf{d}}_s$), which converge at the rate of $(1/T)^{1/d}$.

6.7.1 Simulations

First, we repeat the experiments in section 4.4.3, but with higher dimensional data and lower sample sizes, using the proposed dimension estimator $\hat{\mathbf{d}}_w$ in addition to the estimators proposed by Farahmand *et al.etal* ($\hat{\mathbf{d}}_f$), Levina and Bickel ($\hat{\mathbf{d}}_l$), Costa and Hero *etal* ($\hat{\mathbf{d}}_j$) and the uniform slope estimator ($\hat{\mathbf{d}}_s$) proposed in Chapter 4.. Next, we apply the weighted dimension estimator to detect anomalies in the Abilene router data and to fuse AVIRIS hyperspectral images. Note that in practice, we vary d_{in} in the smaller range $[\min(1, d_0 - 2), \max(d_0 + 2, D)]$ to speed up run-time with the assumption that the true intrinsic dimension d will be contained in the interval $[\min(1, d_0 - 2), \max(d_0 + 2, D)]$.

6.7.1.1 Comparison of dimension estimation methods

We generate $T = 300$ samples $\mathcal{B}$ drawn from a $d = 6$ mixture density $f_m = .8f_\beta + .2f_u$, where f_β is the product of six 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 10-dimensional hyperplane in $\mathbb{R}^10$ by applying the transformation $\mathcal{Y} = U\mathcal{B}$ where U is a 10×6 random matrix whose columns are orthonormal. We apply our intrinsic dimension estimates on the samples $\mathcal{Y}$, and repeat the experiment a total of 100 times. The estimated dimension over these 100 trials is shown in Fig. 6.8.

We compare the performance of our proposed dimension estimator $\hat{\mathbf{d}}_w$ to the estimators proposed by Farahmand *et al.etal* ($\hat{\mathbf{d}}_f$), Levina and Bickel ($\hat{\mathbf{d}}_l$), Costa and Hero *etal* ($\hat{\mathbf{d}}_j$) and the uniform slope estimator ($\hat{\mathbf{d}}_s$) proposed in Chapter 4. We note that the estimator $\hat{\mathbf{d}}_w$ outperforms the other estimators.

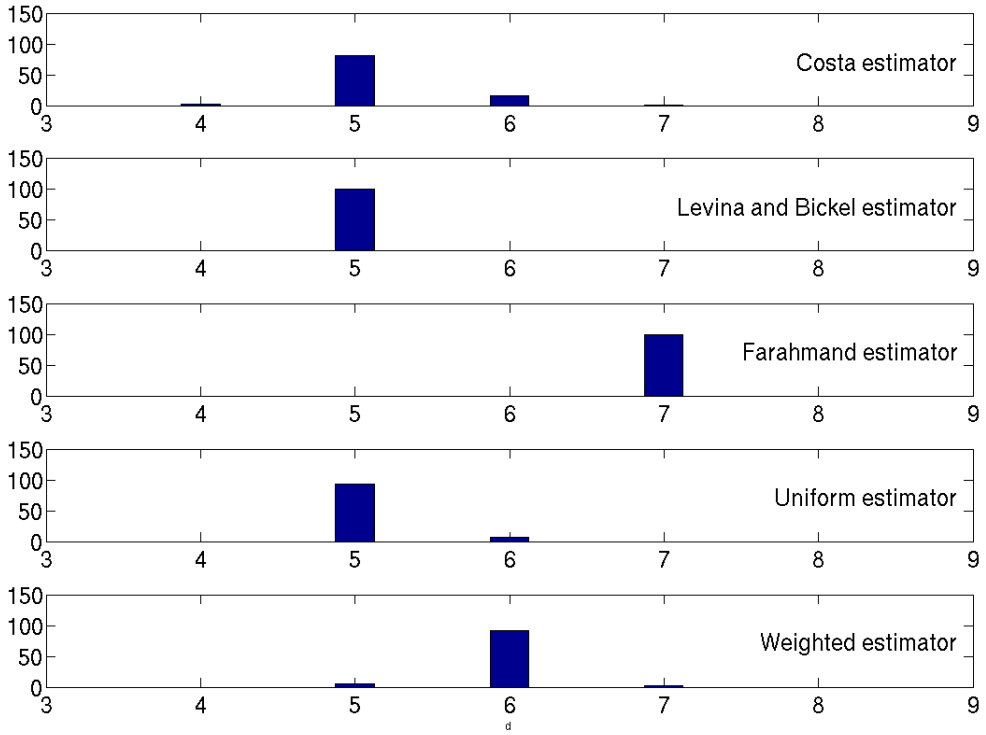


Figure 6.8: Comparison of dimension estimators. The proposed weighted estimator $\hat{\mathbf{d}}_w$ outperforms the other estimators.

6.7.1.2 Local dimension estimation: Toy example

In this experiment, we project data of intrinsic dimension 3 and 5 onto $\mathbb{R}^7$ (total of 600 points), and then perform local dimension estimation using the dimension estimator. We compare our results to the dimension estimator of Farahmand *et al.* ($\hat{\mathbf{d}}_f$), Levina and Bickel ($\hat{\mathbf{d}}_l$), Costa and Hero ($\hat{\mathbf{d}}_j$) and the uniform slope estimator ($\hat{\mathbf{d}}_s$). The results are shown in Fig. 6.9. From the histogram, it is clear that the weighted estimator outperforms the other estimators.

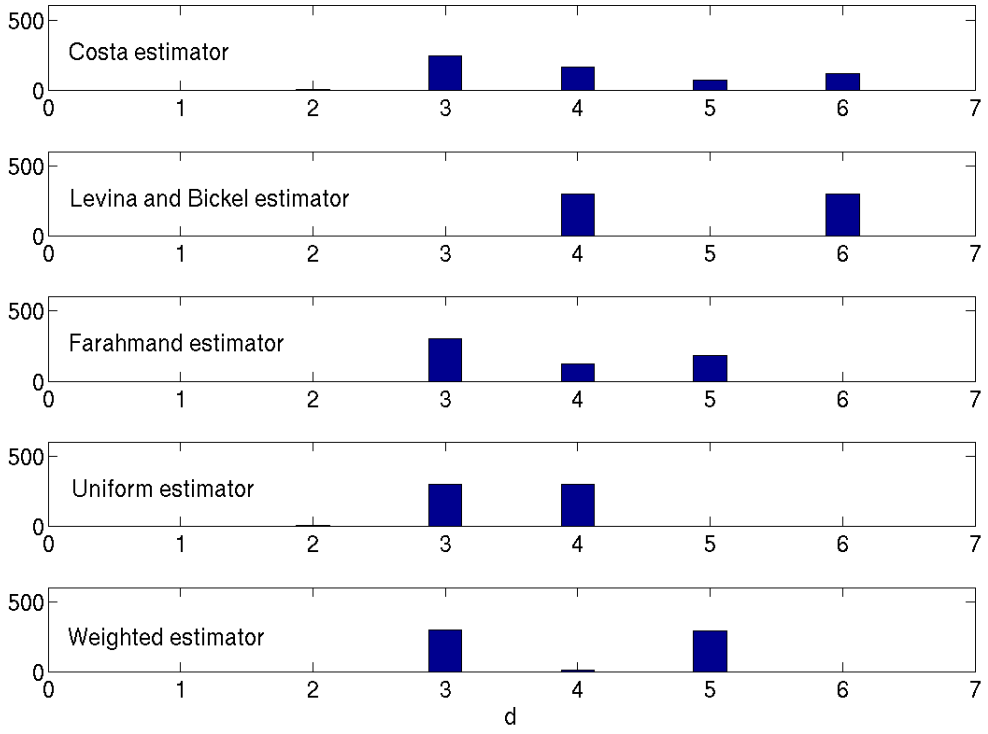


Figure 6.9: Comparison of local dimension estimation performance. The proposed weighted estimator $\hat{\mathbf{d}}_w$ outperforms the other estimators.

6.7.1.3 Anomaly detection in Abilene network data

Anomalies can be detected in router networks by estimating the local dimension at each time point and by monitoring change in dimension. The data used is the number of packets sent by each of the 11 routers on the Abilene network between January 1-2, 2005. A sample is taken every 5 minutes, leading to 576 samples with an extrinsic dimension of 11. We treat each time point as a single sample realization. We seek to estimate the local dimension of the support set of the distribution at each time point.

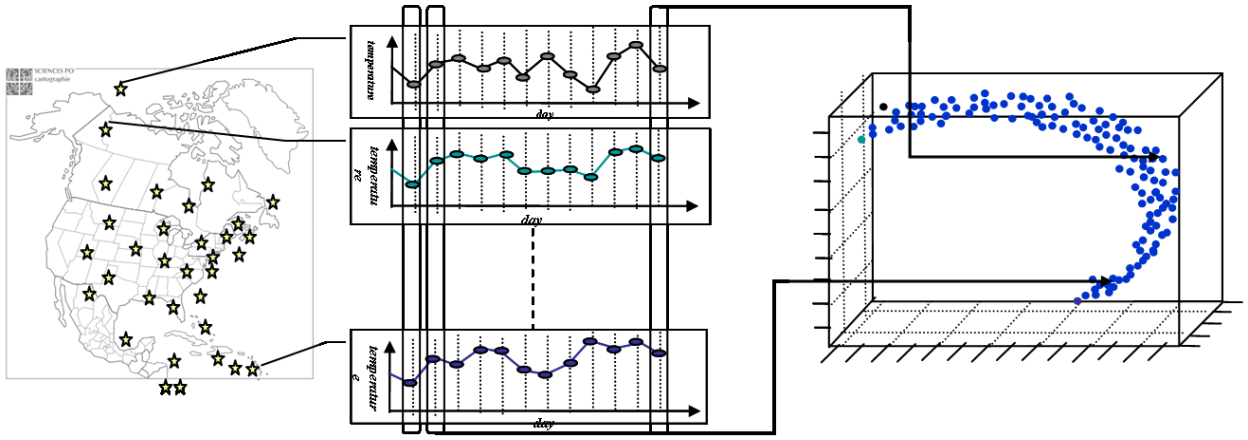


Figure 6.10: Illustration of the Abilene router network. The extrinsic dimension of this system at each time point is equal to the number of routers.

We can use the local dimension estimate as a statistic for doing network anomaly detection. Simultaneous peaks in router traffic imply strong correlation between router traffic and should correspond to lower dimension. This is reflected better by the weighted estimator relative to the other estimators in comparison in Fig. 6.11. In particular, the weighted estimator produces a smoother estimate, and is able to pick-up less obvious correlated time instants - for eg, see time instant 480. This is easier to see in the zoomed in Fig. 6.12.

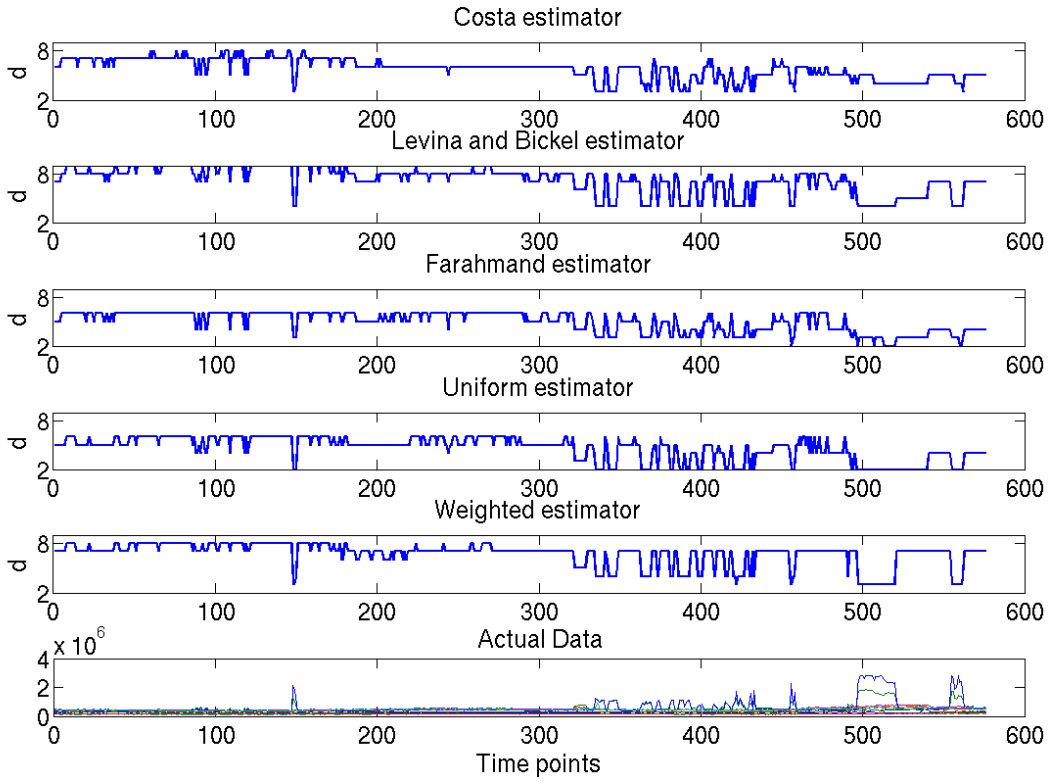


Figure 6.11: Comparison of performance of dimension estimates on Abilene network traffic data. The weighted estimator performs better relative to the other estimators in comparison. In particular, the weighted estimator produces a smoother estimate, and is able to pick-up less obvious correlated time instants - for eg, see time instant 480.

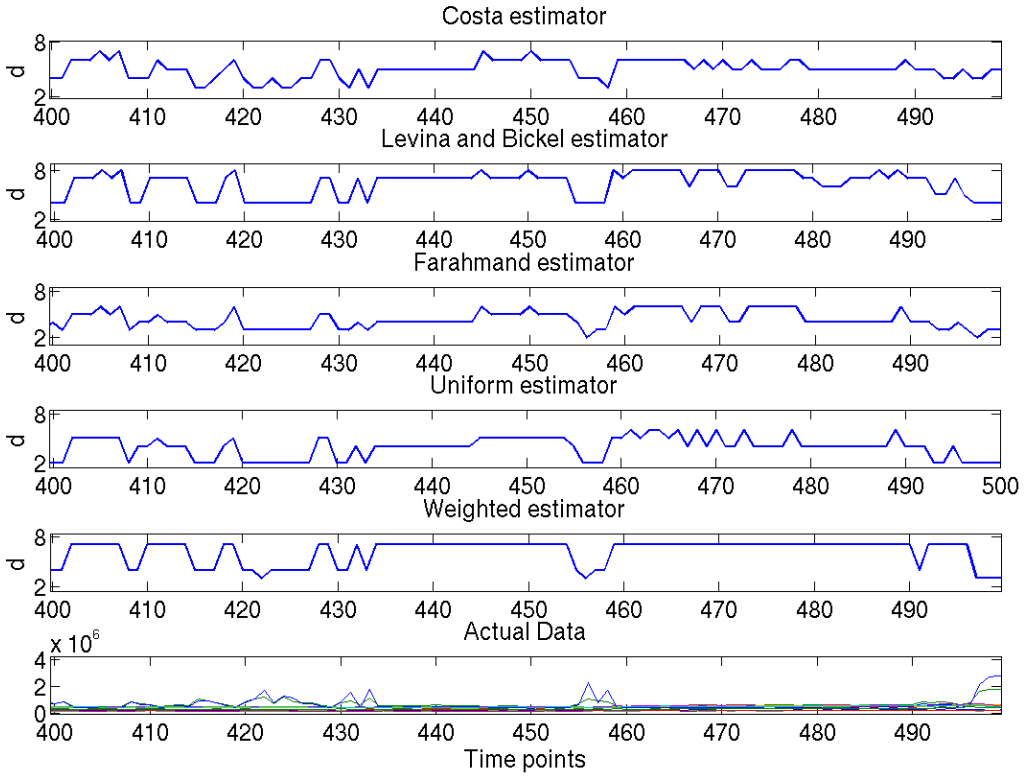


Figure 6.12: Comparison of performance of dimension estimates on Abilene network traffic data (Zoomed into time instants between 400 and 500). The weighted estimator performs better relative to the other estimators in comparison. In particular, the weighted estimator produces a smoother estimate, and is able to pick-up less obvious correlated time instants - for eg, see time instant 480.

6.7.1.4 Dimension based image fusion

In our final experiment, we use the local dimension estimate to perform dimension based image fusion. The data consists of hyperspectral radiance images of Moffett Field [4]. A significant portion - including the entire left portion - of the Moffett Field image (shown in the visible band in Fig. 6.13) is comprised of water bodies, while the

rest corresponds to vegetation and urban areas.

Figure 6.13: Picture of Moffett field in the visible band. A significant portion - including the entire left portion - is comprised of water bodies, while the rest corresponds to vegetation and urban areas.

The AVIRIS hyperspectral response of each image is in the visible to near-infrared range (400 to 2500 nm), of 224 contiguous channels which are approximately 10 nm wide. The scanner type is nadir-viewing, whiskbroom. The hyperspectral images at channels 10, 50, 100, 160 of Moffett field are shown in Fig. 6.14.

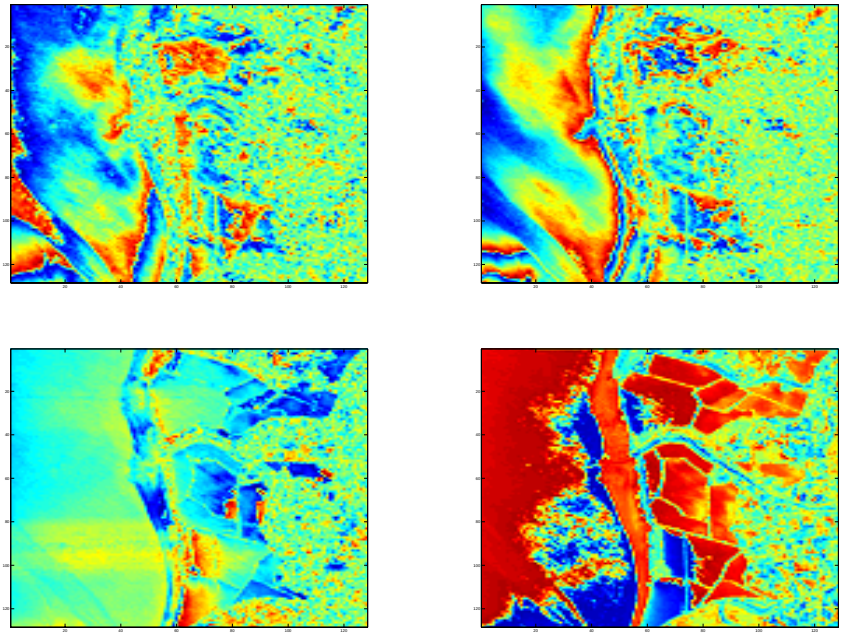


Figure 6.14: AVIRIS hyperspectral radiance images at channels 10, 50, 100, 160 of Moffett field.

Hyperspectral characteristics of different surfaces including water, vegetation and soil has been previously studied. The reflectance (1-radiance) characteristics of water, vegetation and soil are shown in Fig. 6.15. From Fig. 6.15, we can infer that the bandwidth of radiance response of water is much smaller in comparison to the radiance response of soil and vegetation.

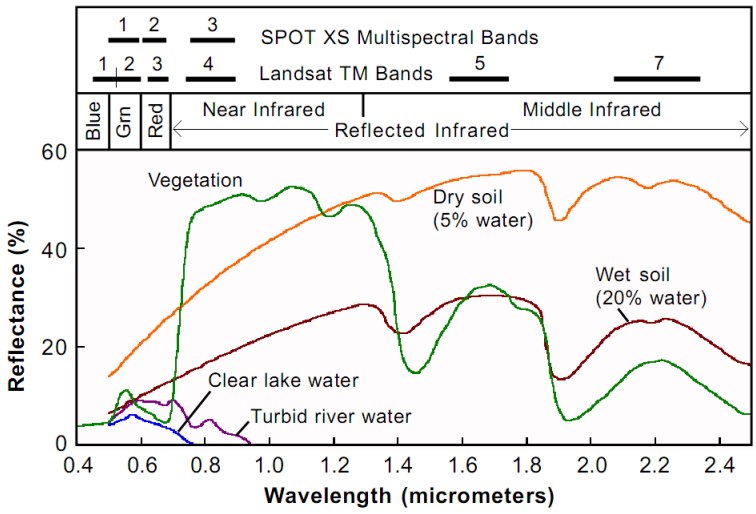


Figure 6.15: Hyperspectral reflectance response as a function of wavelength for different materials - water, vegetation and soil.

The data matrix is therefore of dimension 128x128 (pixels) x 224 wavelengths. We estimate the local dimension at each pixel location and then perform image segmentation using the local dimension estimates. The local dimension estimate, and for the sake of comparison, the standard deviation of the hyperspectral response at each pixel location, are shown in Fig. 6.16.

Observe that the local dimension bandwidth at each pixel location is in complete contrast to the standard deviation at each pixel location - when the local dimension is high at a pixel location, the standard deviation is low and vice versa. However, from Fig. 6.15, we know that the bandwidth of the hyperspectral response of water is low in contrast to the bandwidth of the hyperspectral response of vegetation and soil. The local dimension estimate image (Fig. 6.16) is in complete agreement with Fig. 6.15 in that the dimension estimate of the regions of Moffett field corresponding to water bodies have a much lower estimated dimension value. The standard deviation image, in contrast, is unable to capture this behavior.

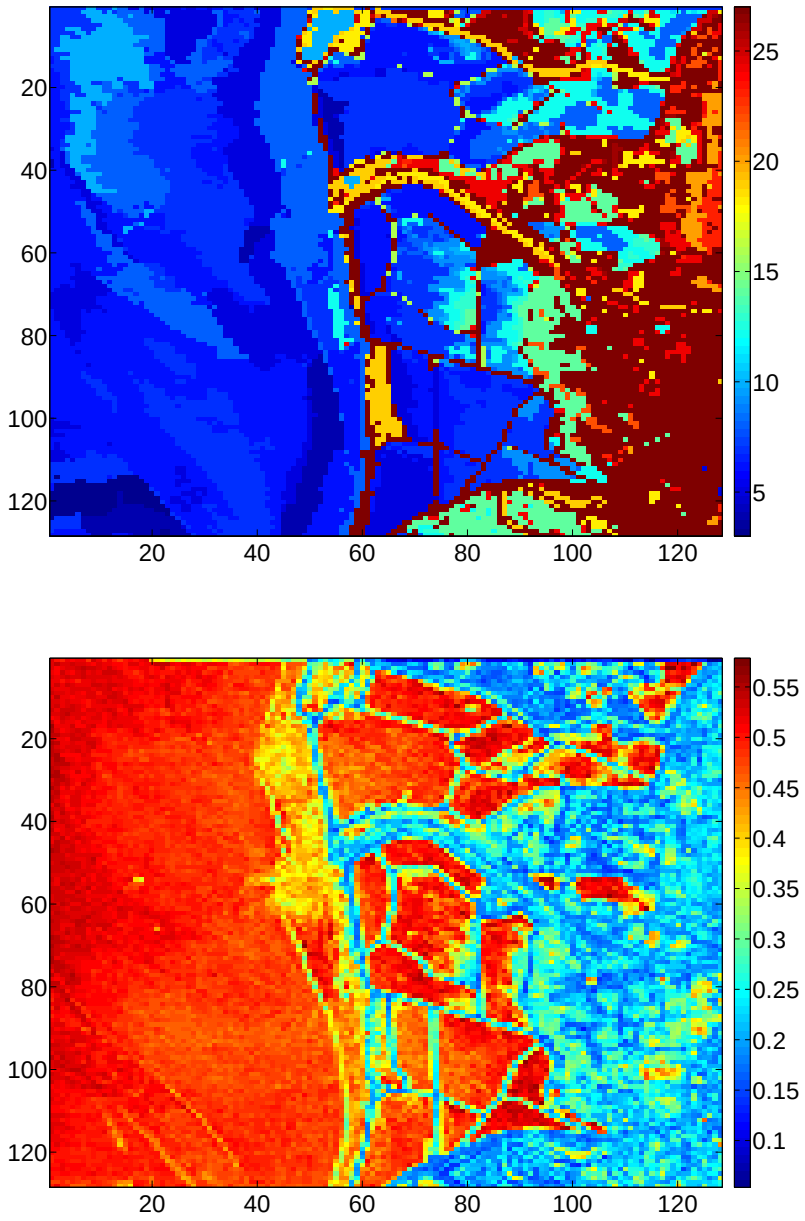


Figure 6.16: Local dimension estimate image (top) and standard deviation image (bottom) of AVIRIS data of Moffett field. The local dimension estimate image is in complete agreement with Fig. 6.15 in that the dimension estimate of the regions of Moffett field corresponding to water bodies have a much lower estimated dimension value. The standard deviation image, in contrast, is unable to capture this behavior.

The local dimension estimate therefore accurately captures the variability in hyperspectral response at each pixel location, and can be used to characterize the hyperspectral response properties of the material at each location. Equivalently, the local dimension estimate accurately summarizes the hyperspectral response characteristics of each pixel, and can in turn be used to differentiate between regions with different hyperspectral response characteristics. This can be achieved, for example, by segmenting by the local dimension estimate image. This is shown in Fig. 6.17.

Figure 6.17: Segmented image of Moffett field. The local dimension estimate image accurately summarizes the hyperspectral response characteristics of each pixel, and can in turn be used to differentiate between regions with different hyperspectral response characteristics. This can be achieved, for example, by segmenting by the local dimension estimate image.

Data set	Sample size	Dimension	Anomaly class
HTTP (KDD'99)	567497	3	attack (0.4%)
Forest	286048	10	class 4 vs class 2 (0.9%)
Mulcross	262144	4	2 clusters (10%)
SMTP (KDD'99)	95156	3	attack (0.03%)
Shuttle	49097	9	class 2,3,5,6,7 vs class 1 (7%)

Table 6.1: Description of data used in anomaly detection experiments.

6.7.2 Discussion

Estimating intrinsic dimension is fundamental to analyzing high dimensional data. Estimators defined in literature suffer from high bias due to curse of dimensionality. We have proposed a new intrinsic dimension estimator based on weighted k-NN graphs. Optimal weights are derived from higher order analysis of bias. The resulting estimator has parametric convergence rate of $O(1/T)$.

6.8 Extension of ensemble estimators to anomaly detection

In this section, we apply the weighted ensembles method to the anomaly detection algorithm described in Chapter 5. In particular, define the weighted statistic

$$d_{w,s,k}(X) = \sum_{l=k-s+1}^{k} w_o(l)|e_{X(l)}|^\gamma, \tag{6.16}$$

where w_o is the optimal weight defined in Section 6.3.3. Define the anomaly detection algorithm WBP-kNNG identically to the BP-kNNG algorithm 2, but using the weighted statistic $d_{w,s,k}(X)$ in place of $d_{s,k}(X)$.

6.8.1 Experimental comparisons

We apply the WBP-kNNG algorithm to the data sets described in Table 6.1. Observe that the run-time of the WBP-kNNG algorithm is equal to the run-time of BP-kNNG plus the additional off-line time required to solve for the optimal weight

Data sets	BP	WBP	L10	K-LPE	1-SVM	Kernel	Mass	iF	ORCA
HTTP	0.994	0.995	NA	NA	0.90	0.99	1.00	1.00	0.36
Forest	0.862	0.941	NA	NA	0.90	0.69	0.91	0.87	0.83
Mulcross	1.00	1.00	NA	NA	0.58	1.00	0.99	0.96	0.33
SMTP	0.924	0.935	NA	NA	0.78	0.60	0.86	0.88	0.87
Shuttle	0.992	0.992	NA	NA	0.79	0.92	0.99	1.00	0.60

Table 6.2: Comparison of anomaly detection schemes in terms of AUC for WBP-kNNG (WBP) against BP-kNNG (BP), L1O-kNNG (L10), K-LPE, MassAD (Mass), iForest (iF) and ORCA. We are unable to report the AUC for K-LPE and L1O-kNNG because of the large processing time. We note that WBP-kNNG outperforms the other algorithm in terms of AUC.

w_o.

We compare the performance of the WBP-kNNG algorithm in terms of the AUC in Table 6.2. The WBP-kNNG algorithm outperforms all other algorithms in comparison, and in particular works better than the BP-kNNG algorithm (which corresponds to uniform weights). The superior performance of WBP-kNNG can simply be explained by observing that the statistics $d_{s,k}(X)$ and $d_{w,s,k}(X)$ correspond to uniformly weighted and optimally weighted estimators of a quantity directly proportional to the Rényi-$(1-\gamma/d)$ entropy.

In particular, observe that the improvement in performance of WBP-kNNG relative to BP-kNNG, is most significant in the case of the high dimensional ($d = 10$) Forest data set. On the other hand, the performance of the two algorithms are comparable wrt the lower dimensional SMTP data set. This is in agreement with our theory that the performance of weighted ensemble estimators is significantly better in comparison to the individual or uniformly weighted estimators for higher values of dimension d.

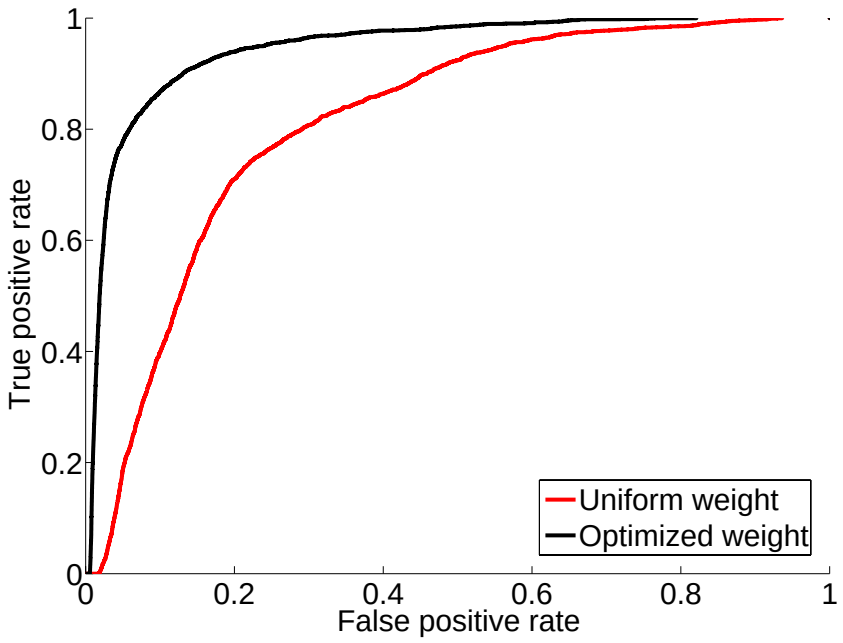


Figure 6.18: ROC performance curves for the BP-kNNG and WBP-kNNG algorithm on the Forest data set. The WBP-kNNG algorithm uniformly outperforms the BP-kNNG algorithm.

The ROC performance curves for the BP-kNNG and WBP-kNNG algorithm on the Forest data set are compared in Fig. 6.18. Observe that the WBP-kNNG algorithm uniformly outperforms the BP-kNNG algorithm.

CHAPTER VII

Conclusion and Future Work

7.1 Summary

This thesis was motivated by the need for analysis of finite sample performance of k-NN estimators of functionals of densities. This thesis (Chapter 2 and 4) filled this void by (i) introducing a a new class of bipartite k-NN estimators of density functionals and subsequently (ii) providing finite sample analysis of the mean square error and establishing a central limit theorem for these estimators. These results characterize the finite sample performance of these k-NN estimators in terms of sample size T, the number of nearest neighbors k, the dimension of the data d, and the underlying density f. A direct consequence of these results is that the estimator performance can be optimized over free parameters (for eg, the number of neighbors k).

Of greater consequence of this analysis of k-NN estimators was the fact that the analysis lent significant insight in to how bipartite k-NN estimators can be modified to significantly improve rate of convergence of the estimators. In particular, a first order bias compensation was proposed in Chapter 3 and Section 6.2 to reduce MSE from $O(T^{-2/(1+d)})$ to $O(T^{-4/(2+d)})$. Under higher order smoothness conditions, a general ensemble estimation method was later proposed in Chapter 6 to further reduce the MSE to $O(T^{-1})$.

This statistical analysis of bipartite k-NN estimators contributed to the development of several performance driven applications in this thesis. In particular, entropy estimates were used as test statistics for anomaly detection in wireless sensor networks at desired false alarm rates (Chapter 2). Analysis of bipartite k-NN functional estimators for data lying on manifolds led to a new class of MSE optimal estimators of intrinsic dimension. Estimates of intrinsic dimension were used as measures of signal complexity to drive applications including image segmentation and data fusion (Chapter 4). Finally, the bipartite nature of the proposed k-NN estimators was used to develop an extremely quick procedure to determine membership in minimum volume sets (Chapter 5).

7.2 Future work

There are several promising directions in which the research presented in this thesis can be extended. These are highlighted below.

Extension to kernel density estimators The results in this thesis concern MSE analysis and asymptotic distributions of estimators defined on bipartite k-NN graphs. Our method of analysis for determining MSE and asymptotic distribution are based on statistical properties of k-NN neighborhoods and exchangeability respectively. The generality of our method of proof (lemma B.1-B.6) makes it possible to extend results on MSE and asymptotic normality to kernel density plug-in estimators. We have made preliminary progress in this direction by developing such results for kernel density estimators with uniform kernels (see Appendix A for details).

Applications of k-NN graphs k-NN graphs enjoy ubiquitous presence in several applications. The theory presented in this thesis can be applied to improve the performance of k-NN graph driven applications. In particular, weighted k-NN graphs

can be potentially used for improving regression and classification tasks.

Applications of divergence functionals Divergence functionals are widely used as similarity measures in signal processing and machine learning applications. The estimators introduced in this thesis can be used to estimate divergence measures more accurately in these applications, potentially leading to improved end results. For example, factor graph structure discovery and image registration performance could be significantly improved by using weighted ensemble estimators.

Comparison with minimax rates Birge and Massart [10] have shown that for density f in a Holder smoothness class with s derivatives, the minimax MSE rate for estimation of a smooth functional is $T^{-2\gamma}$, where $\gamma = \min\{1/2, 4s/(4s+d)\}$. In this thesis, several bipartite k-NN estimators of functionals have been proposed and the MSE performance of these estimators has been studied. A comparison study of the MSE performance of these estimators with the minimax MSE rates derived by Birge and Massart would be very useful in determining how close the performance of the various proposed estimators are to the minimax bounds.

Extension to family of densities In this thesis, we introduced a theory for entropy and divergence estimation of fixed densities with confidence. To predict performance in applications like image registration, this theory has to be extended to characterize divergence estimation for a parameterized family of densities $\{f_\theta, \theta \in \Theta\}$.

Specifically, the goal to estimate the divergence function $I(\theta)$ between a fixed density f and a parameterized density f_θ, $\theta \in \Theta$. If the divergence is estimated using i.i.d. realizations from f and f_0, the goal of this theory is to characterize the joint covariance matrix for the estimator $\hat{\mathbf{I}}_\Theta$ and to determine a joint asymptotic distribution. This theory can then subsequently be extended to the estimator of $I_s = \inf_{\theta\in\Theta} \hat{\mathbf{I}}_\Theta$ and subsequently to characterize the estimator $\theta_s = \arg\min_{\theta\in\Theta} \hat{\mathbf{I}}_\Theta$.

This theory will facilitate the development of performance driven algorithms for tasks such as image registration which involve optimization of divergence measures over parameterized densities f_θ, $\theta \in \Theta$.

APPENDICES

APPENDIX A

Uniform kernels

A.1 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.2 Uniform kernel density estimator

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

$$V_u(X) = \frac{k}{M}, \tag{A.1}$$

and the kernel region is given by

$$S_u(X) = \{Y : c_d||X - Y||^d \leq V_u\}. \tag{A.2}$$

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

$$\mathbf{l_u}(X) = \Sigma_{i=1}^{M} 1_{X_i \in S_u(X)}, \tag{A.3}$$

and the *uniform kernel* density estimator is defined by

$$\hat{\mathbf{f}}_\mathbf{u}(X) = \frac{\mathbf{l_u}(X)}{MV_u(X)}. \tag{A.4}$$

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

$$U(X) = \int_{S_u(X)} f(z)dz = \mathbb{E}[1_{\mathbf{Z} \in S_u(X)}]. \tag{A.5}$$

We observe that $\mathbf{l_u}(X)$ is a binomial random variable with parameters M and $U(X)$.

A.2.1 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 [55].

$$\begin{aligned} U(X) &= \int_{S_u(X)} f(Z)dZ \\ &= f(X)V_u(X) + c(X)V_u^{1+2/d}(X) + o(V_u^{1+2/d}(X)) \\ &= f(X)\frac{k}{M} + c(X)\left(\frac{k}{M}\right)^{1+2/d} + o\left(\left(\frac{k}{M}\right)^{1+2/d}\right), \end{aligned} \tag{A.6}$$

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

A.2.2 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.2.3 Concentration around true density

For $0 < p < 1/2$,

$$Pr(\mathbf{l_u}(X) > (1+p)MU(X)) \leq e^{-MU(X)p^2/4}, \tag{A.7}$$

and

$$Pr(\mathbf{l_u}(X) < (1-p)MU(X)) \leq e^{-MU(X)p^2/4}. \tag{A.8}$$

Using the Taylor expansion of coverage, we then have

$$Pr(\hat{\mathbf{f}}_\mathbf{u}(X) > (1+p)(f(X) + O((k/M)^{2/d}))) \leq\sim e^{-p^2kf(X)/4}, \tag{A.9}$$

and

$$Pr(\hat{\mathbf{f}}_{\mathbf{u}}(X) < (1-p)(f(X) + O((k/M)^{2/d}))) \leq\sim e^{-p^2 k f(X)/4}. \tag{A.10}$$

This then implies that

$$Pr(\hat{\mathbf{f}}_{\mathbf{u}}(X) > (1+p)f(X)) \leq\sim e^{-p^2 k f(X)/4}, \tag{A.11}$$

and

$$Pr(\hat{\mathbf{f}}_{\mathbf{u}}(X) < (1-p)f(X)) \leq\sim e^{-p^2 k f(X)/4}. \tag{A.12}$$

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,

$$\begin{aligned} Pr(\hat{\mathbf{f}}_{\mathbf{u}}(\mathbf{X}) > (1+p)f(\mathbf{X})) &= \mathbb{E}_{\mathbf{X}}[Pr(\hat{\mathbf{f}}_{\mathbf{u}}(\mathbf{X}) > (1+p)f(\mathbf{X}))] \\ &\leq \mathbb{E}[\sim (e^{-p^2 k f(\mathbf{X})/4})] \\ &= \sim e^{-p^2 k/4}, \end{aligned} \tag{A.13}$$

and

$$\begin{aligned} Pr(\hat{\mathbf{f}}_{\mathbf{u}}(\mathbf{X}) < (1-p)f(\mathbf{X})) &= \mathbb{E}_{\mathbf{X}}[Pr(\hat{\mathbf{f}}_{\mathbf{u}}(\mathbf{X}) < (1-p)f(\mathbf{X}))] \\ &\leq \mathbb{E}[\sim (e^{-p^2 k f(\mathbf{X})/4})] \\ &= \sim e^{-p^2 k/4}. \end{aligned} \tag{A.14}$$

A.3 Central Moments

Define the error function of the uniform kernel density,

$$\mathbf{e}_{\mathbf{u}}(X) = \hat{\mathbf{f}}_{\mathbf{u}}(X) - \mathbb{E}[\hat{\mathbf{f}}_{\mathbf{u}}(X)]. \tag{A.15}$$

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

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

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

First Moment:

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{f}}_\mathbf{u}(X)] - f(X) &= \frac{M}{k} U(X) - f(X) \\ &= c(X) \left(\frac{k}{M}\right)^{2/d} + o\left(\left(\frac{k}{M}\right)^{2/d}\right). \end{aligned} \tag{A.16}$$

Second Moment:

$$\begin{aligned} \mathbb{V}[\hat{\mathbf{f}}_\mathbf{u}(X)] &= \mathbb{E}[\mathbf{e}_\mathbf{u}^2(X)] \\ &= \frac{M}{k^2} U(X)(1 - U(X)) \\ &= f(X)\frac{1}{k} + o\left(\frac{1}{k}\right). \end{aligned} \tag{A.17}$$

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

$$\mathbb{E}[\mathbf{e}_u^r(X)] = O\left(\frac{1}{k^{r/2}}\right). \tag{A.18}$$

A.4 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.

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 A.1.

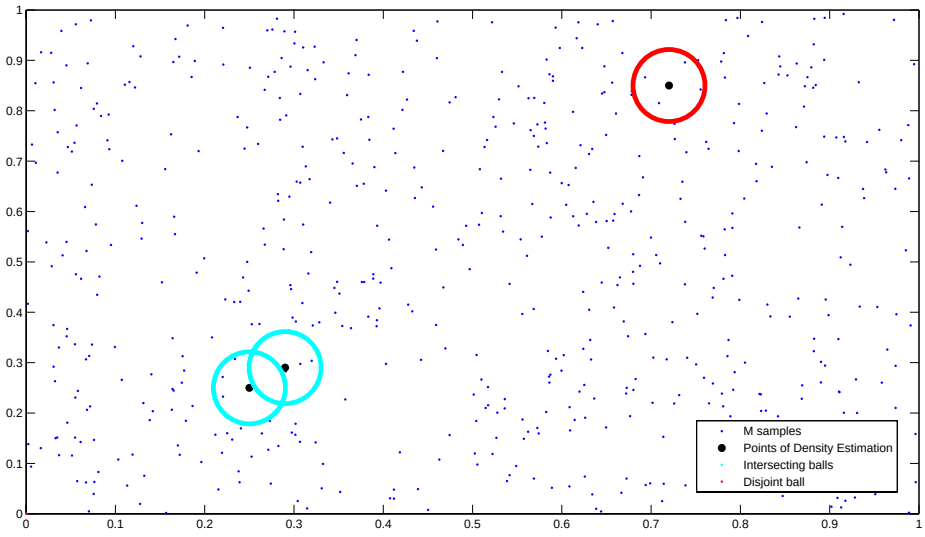


Figure A.1: Intersecting and disjoint uniform kernel neighborhoods centered at the two points X and Y.

Define,

$$U(X,Y) := \mathbb{E}[1_{\mathbf{Z}\in S_u(X)} 1_{\mathbf{Z}\in S_u(Y)}]. \tag{A.19}$$

Intersecting balls

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

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

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,

$$
\begin{aligned}
\aleph(||X - Y||/R_u(X)) &= V(S_u(X) \cap S_u(Y))/V_u(X) \\
&= \frac{V[1_{\mathbf{Z}\in B(0,R_u(X))}1_{\mathbf{Z}\in B(||Y-X||,R_u(Y))}]}{V_u(X)} \\
&= \frac{V[1_{\mathbf{Z}\in B(0,1)}1_{\mathbf{Z}\in B(||Y-X||/R_u(X),1)}]}{V[1_{\mathbf{Z}\in B(0,1)}]} \\
&= O(1). \qquad \text{(A.20)}
\end{aligned}
$$

Because f is assumed to be continuous, we have

$$
U(X,Y) = \mathbb{E}[1_{\mathbf{Z}\in S_u(X)}1_{\mathbf{Z}\in S_u(Y)}] = [f(X) + o(1)]V(S_u(X) \cap S_u(Y)). \qquad \text{(A.21)}
$$

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

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

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

$$
\begin{aligned}
& Cov(\hat{\mathbf{f}}^q_{\mathbf{u}}(X), \hat{\mathbf{f}}^r_{\mathbf{u}}(Y)) = \frac{1}{k^{q+r}} Cov(\mathbf{l}^q_{\mathbf{u}}(X), \mathbf{l}^r_{\mathbf{u}}(Y)) \\
& = \frac{1}{k^{q+r}} \sum l^q_x l^r_y Pr(\mathbf{l}_{\mathbf{u}}(X) = l_x, \mathbf{l}_{\mathbf{u}}(Y) = l_y) \\
& - \frac{1}{k^{q+r}} \sum l^q_x l^r_y Pr(\mathbf{l}_{\mathbf{u}}(X) = l_x) Pr(\mathbf{l}_{\mathbf{u}}(Y) = l_y) \\
& = \sum_{\chi} \frac{l^q_x l^r_y}{k^{q+r}} \left[Pr(\mathbf{l}_{\mathbf{u}}(X) = l_x, \mathbf{l}_{\mathbf{u}}(Y) = l_y) - Pr(\mathbf{l}_{\mathbf{u}}(X) = l_x) Pr(\mathbf{l}_{\mathbf{u}}(Y) = l_y) \right] \\
& + o(1/M) \\
& = \sum_{\chi} \frac{f^q(X) f^r(Y) l^q_x l^r_y U^q(X) U^r(Y)}{k^{q+r}(l_x \times \ldots \times l_x - q + 1)(l_y \times \ldots \times l_y - r + 1)} \times \\
& [\prod_{l=0}^{q+r-1} (M - l) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(X) = l_x, \hat{\mathbf{l}}_{\mathbf{u}}(Y) = l_y) \\
& - \prod_{l=0}^{q-1} (M - l) \prod_{l=0}^{r-1} (M - l) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(X) = l_x) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(Y) = l_y)] \\
& + o(1/M) \\
& = \left(\frac{f^q(X) f^r(Y)}{M^{q+r}} + O\left(\frac{1}{k M^{q+r}} \right) \right) \times \\
& \sum_{\chi} [\prod_{l=0}^{q+r-1} (M - l) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(X) = l_x, \hat{\mathbf{l}}_{\mathbf{u}}(Y) = l_y) \\
& - \prod_{l=0}^{q-1} (M - l) \prod_{l=0}^{r-1} (M - l) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(X) = l_x) Pr(\hat{\mathbf{l}}_{\mathbf{u}}(Y) = l_y)] \\
& + o(1/M) \\
& = \left(\frac{f^q(X) f^r(Y)}{M^{q+r}} + O\left(\frac{1}{k M^{q+r}} \right) \right) \times \\
& [\prod_{l=0}^{q+r-1} (M - l) - \prod_{l=0}^{q-1} (M - l) \prod_{l=0}^{r-1} (M - l)] \\
+ \quad & o(1/M) \\
= \quad & \frac{-qr f^q(X) f^r(Y)}{M} + o\left(\frac{1}{M} \right).
\end{aligned}
$$

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 $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density f. Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. 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 $q + r > 1$,*

$$
\mathbb{E}[\mathbf{1}_{\mathbf{\Delta_u}}(\mathbf{X}, \mathbf{Y})\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_{\mathbf{u}}^{q}(\mathbf{X})\mathbf{e}_{\mathbf{u}}^{r}(\mathbf{Y})] \;=\; o\left(\frac{1}{M}\right), \tag{A.22}
$$

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.A.18.

□

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

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

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

We have already established in the previous lemma that

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

Now,

$$\begin{aligned} D &= \mathbb{E}[(1-\mathbf{1}_{\mathbf{\Delta_u}}(\mathbf{X},\mathbf{Y}))\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\mathbf{X},\mathbf{Y}}[Cov(\mathbf{e}^q_{\mathbf{u}}(X),\mathbf{e}^r_{\mathbf{u}}(Y))]] \qquad \text{(A.23)} \\ &= \mathbb{E}[(1-\mathbf{1}_{\mathbf{\Delta_u}}(\mathbf{X},\mathbf{Y}))\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})o(1/M)] \\ &= o\left(\frac{1}{M}\right). \end{aligned}$$

This concludes the proof. □

APPENDIX B

k-NN density estimates

B.1 Introduction

In this appendix, moment properties of the standard k-NN density estimate $\hat{\mathbf{f}}_k(X)$ are derived conditioned on $X_1, \ldots, X_N$. As the samples $X_1, \ldots, X_N, X_{N+1}, \ldots, X_T$, $T = M + N$ are i.i.d., these conditional moments are independent of the N samples $\mathbf{X}_1, .., \mathbf{X}_N$.

B.1.1 Preliminaries

Let $d(X, Y)$ denote the Euclidean distance between points X and Y and $\mathbf{d}_X^{(k)}$ denote the Euclidean distance between a point X and its k-th nearest neighbor amongst $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$. Let c_d denote the unit ball volume in d dimensions. 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 [53] is defined as

$$\hat{\mathbf{f}}_k(X) = \frac{k-1}{M\mathbf{V}_k(X)}.$$

Define the coverage function as

$$\mathbf{P}(X) = \int_{\mathbf{S}_k(X)} f(Z)dZ.$$

Define spherical regions

$$S_r(X) = \{Y \in \mathbb{R}^d : d(X,Y) \leq r\}.$$

B.1.1.1 Concentration inequality for coverage probability

It has been previously established that $\mathbf{P}(X)$ has a beta distribution with parameters k, $M-k+1$ [30]. Using Chernoff inequalities, we can then establish the following concentration inequality (Section B.1, [80]). For some $0 < p < 1/2$,

$$\begin{aligned} Pr(\mathbf{P}(X) > (1+p)(k-1)/M) &= O(e^{-p^2k/2(1+p)}) \\ Pr(\mathbf{P}(X) < (1-p)(k-1)/M) &= O(e^{-p^2k/2(1-p)}). \end{aligned} \tag{B.1}$$

Define

$$k_M = (k-1)/M.$$

Let $\natural(X)$ denote the event

$$\mathbf{P}(X) < (p_k+1)k_M, \tag{B.2}$$

where $p_k = \sqrt{6}/(k^{\delta/2})$. Then, $1 - Pr(\natural(X)) = O(e^{-p_k^2k/2}) = O(e^{-3k^{(1-\delta)}})$. Equivalently,

$$1 - Pr(\natural(X)) = O(\mathcal{C}(k)), \tag{B.3}$$

where $\mathcal{C}(k)$ is a function which satisfies the rate of decay condition $\mathcal{C}(k) = O(e^{-3k^{(1-\delta)}})$. Similarly, let $\natural_{-1}(X)$ denote the event

$$\mathbf{P}(X) > (1 - p_k)k_M, \tag{B.4}$$

Then

$$1 - Pr(\natural_{-1}(X)) = O(\mathcal{C}(k)), \tag{B.5}$$

Also let $\natural\natural(X) = \natural(X) \cap \natural_{-1}(X)$. Then

$$1 - Pr(\natural\natural(X)) = O(\mathcal{C}(k)), \tag{B.6}$$

Finally, we note that $\Gamma(x + a)/\Gamma(x) = x^a + o(x^a)$. Then for any $a < k$, $\mathbb{E}[\mathbf{P}^{-a}(X)]$ exists and is given by

$$\mathbb{E}[\mathbf{P}^{-a}(X)] = \frac{\Gamma(k - a)\Gamma(M + 1)}{\Gamma(k)\Gamma(M + 1 - a)} = \Theta((k_M)^{-a}). \tag{B.7}$$

B.1.1.2 Interior points

Let $\mathcal{S}'$ to be any arbitrary subset of $\mathcal{S}_I$ (Section 2.3.1) satisfying the condition $Pr(\mathbf{Y} \notin \mathcal{S}') = o(1)$ where $\mathbf{Y}$ is random variable with density f. This implies that given the event $\natural(X)$, the k-NN neighborhoods $\mathbf{S}_k(X)$ of points $X \in \mathcal{S}'$ will lie completely inside the domain $\mathcal{S}$. Therefore the density f has continuous partial derivatives of order 2ν in the k-NN ball neighborhood $\mathbf{S}_k(X)$ for each $X \in \mathcal{S}'$ (assumption $(\mathcal{A}.2)$). We will now derive moments for the interior set of points $X \in \mathcal{S}'$. This excludes the set of points X close to the boundary of the support whose k-NN neighborhoods $\mathbf{S}_k(X)$ intersect with the boundary of the support. We will deal with these points in Appendix B.

B.1.1.3 Taylor series expansion of coverage probability

Let $X \in \mathcal{S}'$. Given the event $\natural(X)$, the coverage function $\mathbf{P}(X)$ can be represented 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 as follows. In particular, for some fixed $x \in \mathcal{S}'$, let

$$p(u) = \int_{S_u(x)} f(z)dz.$$

Using ($\mathcal{A}$.2), we can write, by a Taylor series expansion of f around x using multi-index notation [70]

$$f(z) = \sum_{0 \leq |\alpha| \leq 2\nu} \frac{(z-x)^\alpha}{\alpha!}(\partial^\alpha f)(x) + o(||z-x||^{2\nu}) \tag{B.8}$$

Assuming $S_u(x) \subset \mathcal{S}$, we can then write

$$\begin{aligned} p(u) &= \int_{S_u(x)} f(z)dz \\ &= \int_{S_u(x)} \left(\sum_{|0 \leq \alpha \leq 2\nu|} \frac{(z-x)^\alpha}{\alpha!}(\partial^\alpha f)(x) \right) dz + o(u^{d+2\nu}) \\ &= f(x)c_d u^d + \sum_{i=1}^{\nu-1} c_i(x) c_d^{1+2i/d} u^{d+2i} + o(u^{d+2\nu}). \end{aligned} \tag{B.9}$$

where $c_i(x)$ are functionals of the derivatives of f. Now, denote $v(u) = \int_{S_u(x)} dz$ to be the volume of $S_u(x)$. Let $u^{inv}(v)$ be the inverse function of $v(u)$. Note that this inverse is well-defined since $v(u)$ is monotonic in u. Since $S_u(x) \subset \mathcal{S}$, $v(u) = c_d u^d$. This gives $u^{inv}(v) = (v/c_d)^{1/d}$. Define

$$P(v) = \int_{S_{u^{inv}(v)}(x)} f(z)dz.$$

Using (B.9),

$$P(v) \;=\; f(X)v + \sum_{i=1}^{\nu-1} c_i(X) v^{1+2i/d} + o(v^{1+2\nu/d}). \tag{B.10}$$

Now denote $V(p) = P^{inv}(p)$ to be the inverse of $P(.)$. Note that this inverse is well-defined since $P(v)$ is monotonic in v. Dividing (B.10) by $vP(v)$ on both sides, we get

$$\frac{1}{v} \;=\; \frac{f(X)}{P(v)} + \sum_{i=1}^{\nu-1} \frac{c_i(X)}{P(v)} v^{2i/d} + o(v^{2\nu/d} P^{-1}(v)) \tag{B.11}$$

By repeatedly substituting the LHS of (B.11) in the RHS of (B.11), we can obtain (B.12):

$$\frac{1}{V(p)} \;=\; \frac{f(X)}{p} + \sum_{i=1}^{\nu-1} \frac{h_i(X)}{p^{1-2i/d}} + o(p^{2\nu/d-1}), \tag{B.12}$$

From our derivation of (B.12) using (B.10), it is clear that $h_i(X)$ are of the form

$$h_i(X) = \sum_{\{a_i\}=A; A\in\mathcal{A}} \frac{\prod_{i=1}^{\nu-1} c_i^{a_i}}{f^{a_0}(X)}$$

where A is a ν-tuple of positive real numbers $a_0, .., a_{\nu-1}$ and the cardinality of $\mathcal{A}$ is finite. By assumptions ($\mathcal{A}$.1) and ($\mathcal{A}$.2), this implies that the constants $h_i(X)$ are *bounded*. Also, we note that $h(X) = h_1(X) = c(X) f^{-2/d}(X)$ [30], where $c(X) := c_1(X) = \Gamma^{(2/d)}(\frac{d+2}{2}) tr[\nabla^2(f(X))]$. This then implies that under the event $\natural(X)$

$$\frac{1}{\mathbf{V}_k(X)} = \frac{f(X)}{\mathbf{P}(X)} + \sum_{t\in\mathcal{T}} \frac{h_t(X)}{\mathbf{P}^{1-t}(X)} + \mathbf{h_r}(X), \tag{B.13}$$

where $\mathcal{T} = \{2/d, 4/d, 6/d.., 2\nu/d\}$ and $\mathbf{h_r}(X) = o(\mathbf{P}^{2\nu/d-1}(X))$. Now, by ($\mathcal{A}$.2), we have $(k/M)^{2\nu/d} = o(1/M)$. This implies that $2\nu/d > 1$. Under the event $\natural(X)$, we

have $\mathbf{P}(X) \leq (p_k+1)k/M$, which, in conjunction with the condition $2\nu/d > 1$ implies that

$$\mathbf{h_r}(X) \;=\; o(\mathbf{P}^{2\nu/d-1}(X)) = o((k/M)^{2\nu/d-1}) = o(1/k_M M). \tag{B.14}$$

On the other hand, under the event, $\natural^c(X)$, $(p_k+1)k/M \leq \mathbf{P}(X) \leq 1$, which gives

$$\mathbf{h_r}(X) \;=\; O(1). \tag{B.15}$$

B.1.1.4 Approximation to the k-NN density estimator

Define the *coverage* density estimate to be,

$$\hat{\mathbf{f}}_c(X) = f(X)\frac{k-1}{M}\frac{1}{\mathbf{P}(X)}.$$

The estimate $\hat{\mathbf{f}}_c(X)$ is clearly not implementable. Note also that the two estimates - $\hat{\mathbf{f}}_c(X)$ and $\hat{\mathbf{f}}_k(X)$ - are identical in the case of the uniform density.

$$\frac{1}{\mathbf{V}_k(X)} = \frac{f(X)}{\mathbf{P}(X)} + \frac{h(X)}{\mathbf{P}^{1-2/d}(X)} + \mathbf{h_s}(X), \tag{B.16}$$

where $\mathbf{h_s}(X) = o(1/\mathbf{P}^{1-2/d}(X))$. This gives,

$$\hat{\mathbf{f}}_k(X) \;=\; \hat{\mathbf{f}}_c(X) + \left(\frac{k-1}{M}\right)\frac{h(X)}{\mathbf{P}^{1-2/d}(X)} + \frac{k-1}{M}\mathbf{h_s}(X). \tag{B.17}$$

whenever $\natural(X)$ is true.

B.1.2 Bounds on k-NN density estimates

Let X be a Lebesgue point of f, i.e., an X for which

$$\lim_{r\to 0} \frac{\int_{S_r(X)} f(y)dy}{\int_{S_r(x)} dy} = f(X).$$

Because f is an density, we know that almost all $X \in \mathcal{S}$ satisfy the above property. Now, fix $\epsilon \in (0,1)$ and find $\delta > 0$ such that

$$\sup_{0<r\leq\delta} \frac{\int_{S_r(X)} f(y)dy}{\int_{S_r(x)} dy} - f(X) \leq \epsilon f(X).$$

This in turn implies that, for $\mathbf{P}(X) \leq P(\delta)$,

$$\frac{\mathbf{P}(X)}{(1+\epsilon)f(X)} \leq \mathbf{V}_k(X) \leq \frac{\mathbf{P}(X)}{(1-\epsilon)f(X)} \tag{B.18}$$

and in turn implies

$$(1-\epsilon)\hat{\mathbf{f}}_c(X) \leq \quad \hat{\mathbf{f}}_k(X) \quad \leq (1+\epsilon)\hat{\mathbf{f}}_c(X). \tag{B.19}$$

Also, because $\delta > 0$ is fixed, we note that the event $\mathbf{P}(X) \leq P(\delta)$ is a subset of $\natural(X)$ and therefore (B.18) holds under $\natural(X)$.

Under the event $\natural^c(X)$, we can bound $\mathbf{V}_k(X)$ from above by $c_d\mathcal{D}^d$. Also, since $\mathbf{V}_k(X)$ is monotone in $\mathbf{P}(X)$, under the event $\natural^c(X)$, we can bound $\mathbf{V}_k(X)$ from below by $(1+p_k)(k-1)/M(1-\epsilon)f(X)$ and therefore by $(k-1)/M(1-\epsilon)f(X)$. Written explicitly,

$$\frac{(k-1)}{M(1-\epsilon)f(X)} \leq \mathbf{V}_k(X) \leq c_d\mathcal{D}^d \tag{B.20}$$

and in turn implies

$$(k-1)/(Mc_d\mathcal{D}^d) \leq \quad \hat{\mathbf{f}}_k(X) \quad \leq (1-\epsilon)f(X). \tag{B.21}$$

Finally, note that $k_M/\mathbf{P}(X)$ is bounded above by $O(1)$ under the event $\natural(X)$. This implies that for any $a < k$,

$$\mathbb{E}[\natural^c(X)]k_M^a\mathbf{P}^{-a}(X) \leq O(1)Pr(\natural^c(X)) = O(\mathcal{C}(k)). \tag{B.22}$$

B.2 Approximation to the kNN density estimator

Define the *coverage* density estimate to be,

$$\hat{\mathbf{f}}_c(X) = f(X)\frac{k-1}{M}\frac{1}{\mathbf{P}(X)}.$$

The estimate $\hat{\mathbf{f}}_c(X)$ is clearly not implementable. Note also that the two estimates - $\hat{\mathbf{f}}_c(X)$ and $\hat{\mathbf{f}}_k(X)$ - are identical in the case of the uniform density. Define the error functions $\mathbf{e}_c(X) = \hat{\mathbf{f}}_c(X) - \mathbb{E}[\hat{\mathbf{f}}_c(X)]$ and $\mathbf{e}_k(X) = \hat{\mathbf{f}}_k(X) - \mathbb{E}[\hat{\mathbf{f}}_k(X)]$. Note that the coverage density estimate corresponds to the leading term in the Taylor series expansion of the volume. Therefore

$$\begin{aligned}\hat{\mathbf{f}}_k(X) \quad &= \quad \hat{\mathbf{f}}_c(X) + \sum_t \left(\frac{k-1}{M}\right) h_t(X)(1/\mathbf{P}^{1-t}(X)) \\ &\quad + \frac{k-1}{M}\mathbf{h_r}(X).\end{aligned}$$

B.2.1 Key idea

We note that the coverage random variable is independent of the underlying density and therefore the representation is a sufficient statistic to determine properties of kNN density estimates.

B.2.2 Similarity between k-NN and coverage density estimates

We first establish the following two lemmas which relate the moments of the k-NN density estimate to the moments of the coverage density estimate. Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions satisfying the condition: $\mathbb{E}[\gamma_i^2(\mathbf{X})]$ is finite, $i = 1, 2$. Also let $\gamma(X) = \gamma_1(X)$. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density f. Let q, r be arbitrary positive integers.

Lemma B.1.

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

Lemma B.2.

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

As a consequence of these lemma, for $X \in \mathcal{S}'$, we can compute all central and cross moments of the k-NN density $\hat{\mathbf{f}}_k(X)$ 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 $\mathbf{e}_r(X) = \mathcal{M}(((k-1)/M)\mathbf{h_r}(X))$. Define the event $\{X \in \mathcal{S}'\} \cap \{\natural(X)\}$ by $\dagger(X)$. Note that under the event $\dagger(X)$, $\mathbf{e}_k(X) = \mathbf{e}_c(X) + \mathbf{e}_t(X) + \mathbf{e}_r(X) =: \mathbf{e}_o(X)$. Finally, let β be any real number and define $\mathbf{E}_\beta(X) = ((k-1)/M)^\beta(\mathcal{M}(\mathbf{P}^{-\beta}(X)))$.

Proof. of Lemma B.1.

$$\begin{aligned}
&\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] \\
&= \mathbb{E}\big[1_{\dagger(\mathbf{X})}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] + o(1/M) \\
&= \mathbb{E}\big[1_{\dagger(\mathbf{X})}\gamma(\mathbf{X})\mathbf{e}_o^q(\mathbf{X})\big] + o(1/M) \\
&= \mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_o^q(\mathbf{X})\big] + o(1/M) \\
&= \mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbb{E}_{\mathbf{X}}(\mathbf{e}_c(X)+\mathbf{e}_t(X)+\mathbf{e}_r(X))^q\right] \\
&+ o(1/M).
\end{aligned}$$

Let us focus on the inner expectation first. Since $\mathbf{P}(X)$ is a beta random variable, the probability density function of $\mathbf{P}(X)$ is given by

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

We have $\mathbb{E}[1_{\natural(X)}\mathbf{P}^{-\beta}(X)] = \Theta((k/M)^{-\beta})$. For large enough k, M, $\mathbb{E}[\mathbf{P}^{-2\beta}(X)]$ is bounded between 0 and 1, which implies that $\mathbb{E}[1_{\natural^c(X)}\mathbf{P}^{-\beta}(X)] = o(1/M^{a/2})$ using Cauchy-Schwarz and the concentration inequality (B.1). This then gives $\mathbb{E}[\mathbf{P}^{-\beta}(X)] = \Theta((k/M)^{-\beta})$. This yields $\mathbb{E}[1_{\natural(X)}\mathbf{E}_\beta^q(X)] = O\left(k^{-(\delta_k q/2)}\right)$. We can again bound $\mathbb{E}[1_{\natural^c(X)}\mathbf{E}_\beta^q(X)]$ by $o(1/M^{a/2})$ using Cauchy-Schwarz inequality and the concentration bound obtaining $\mathbb{E}[\mathbf{E}_\beta^q(X)] = O\left(k^{-(\delta_k q/2)}\right)$. Noting that $\delta_k \to 1$ as $k \to \infty$ gives

$$\mathbb{E}[\mathbf{E}_\beta^q(X)] = O(k^{-q/2}). \tag{B.23}$$

From this analysis on $\mathbf{E}_\beta(X)$, it follows $\mathbb{E}[\mathbf{e}_r^l(X)] = O((k/M)^{2rl/d}) = o(1/M)$ for any $l > 1$. Similarly, $\mathbb{E}[\mathbf{e}_c^l(X)] = O(k^{-l/2})$. Now, $\mathbf{e}_t^l(X)$ can be expressed 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 repeated application of Cauchy-Schwarz, the expectation of each of these terms, and

therefore $\mathbb{E}[\mathbf{e}_t^l(X)]$, is $o(k^{-l/2})$.

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

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

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 $q \neq m$,

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

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

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

□

Before proving Lemma B.2, we seek to answer the following question: for which set of pair of points $\{X, Y\}$ are the k-NN balls disjoint?

B.2.3 Intersecting and disjoint balls

Define $\Psi_\epsilon := \{X, Y\} \in \mathcal{S}' : ||X - Y|| \geq R_\epsilon(X) + R_\epsilon(Y)$ where $R_\epsilon(X)$ and $R_\epsilon(Y)$ are the ball radii of the spherical regions $S_u(X)$ and $S_u(Y)$. We will now show that

for $\{X, Y\} \in \Psi_\epsilon$, the k-NN balls will be disjoint with exponentially high probability. Let $\mathbf{d_X^{(k)}}$ and $\mathbf{d_Y^{(k)}}$ denote the k-NN distances from X and Y and let $\boldsymbol{\Upsilon}$ denote the event that the k-NN balls intersect. For $\{X, Y\} \in \Psi_\epsilon$,

$$\begin{aligned} Pr(\boldsymbol{\Upsilon}) &= Pr(\mathbf{d_X^{(k)}} + \mathbf{d_Y^{(k)}} \geq ||X - Y||) \\ &\leq Pr(\mathbf{d_X^{(k)}} + \mathbf{d_Y^{(k)}} \geq R_\epsilon(X) + R_\epsilon(Y)). \\ &\leq Pr(\mathbf{d_X^{(k)}} \geq R_\epsilon(X)) + Pr(\mathbf{d_Y^{(k)}} \geq R_\epsilon(Y)) \\ &= Pr(\mathbf{P}(X) \geq (p_k + 1)((k-1)/M)) \\ &\quad + Pr(\mathbf{P}(Y) \geq (p_k + 1)((k-1)/M)) \\ &= o(1/M^a), \end{aligned}$$

where the last inequality follows from the concentration inequality (B.1). 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 $\{X, Y\} \in \Psi_\epsilon^c$. From the definition of the region Ψ_ϵ, we have $Pr(\{\mathbf{X}, \mathbf{Y}\} \in \Psi_\epsilon^c) = O(k/M)$.

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

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

Define

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

and note that

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

Figure B.1 shows the distribution of the M samples when the k-NN balls are disjoint.

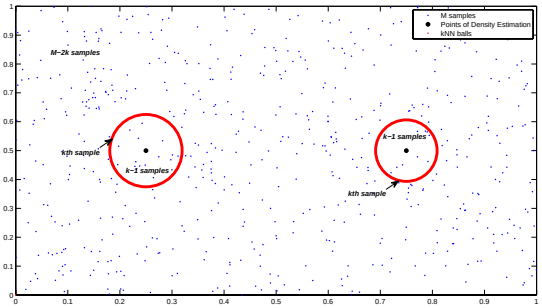


Figure B.1: Distribution of random samples when k-NN balls centered at X and Y are disjoint.

Now note that $i(p_X,p_Y)$ corresponds to the density function $f_{\boldsymbol{\Upsilon}^c}(p_X,p_Y)$ for the choices $t=k$, $u=k$ and $v=M-2k+1$. 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,

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

From the joint distribution representation, it follows that

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

Now observe that

$$\begin{aligned}
&\left(\frac{k-1}{M}\right)^{t+u} Cov(1/\mathbf{P}^t(X), 1/\mathbf{P}^u(Y)) \\
&= \left(\frac{k-1}{M}\right)^{t+u} [\mathbb{E}[1/\mathbf{P}^t(X)\mathbf{P}^u(Y)] - \mathbb{E}[1/\mathbf{P}^t(X)]\mathbb{E}[1/\mathbf{P}^u(Y)]] \\
&= \left(\frac{k-1}{M}\right)^{t+u} \mathbb{E}[1/\mathbf{P}^t(X)]\mathbb{E}[1/\mathbf{P}^u(Y)] \left[\frac{\mathbb{E}[1/\mathbf{P}^t(X)\mathbf{P}^u(Y)]}{\mathbb{E}[1/\mathbf{P}^t(X)]\mathbb{E}[1/\mathbf{P}^u(Y)]} - 1\right] \\
&= (1 + o(1/k)) \left[1 - \frac{tu}{M} + o(1/M) - 1\right] \\
&= -\left(\frac{tu}{M}\right) + o(1/M). \qquad \text{(B.24)}
\end{aligned}$$

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

$$\begin{aligned}
&Cov(\mathbf{e}_t^q(X), \mathbf{e}_\beta^r(Y)) \\
&= k_M^{tq+\beta r} Cov\left(\left[\frac{1}{\mathbf{P}^t(X)} - \mathbb{E}\left[\frac{1}{\mathbf{P}^t(X)}\right]\right]^q, \left[\frac{1}{\mathbf{P}^\beta(Y)} - \mathbb{E}\left[\frac{1}{\mathbf{P}^\beta(Y)}\right]\right]^r\right) \\
&= \sum_{a=1}^{q}\sum_{b=1}^{r} \binom{q}{a}\binom{r}{b} [(-1)^{a+b} + o(1)] k_M^{ta+\beta b} Cov(1/\mathbf{P}^{ta}(X), 1/\mathbf{P}^{\beta b}(Y)) \\
&= -t\beta \sum_{a=1}^{q}\sum_{b=1}^{r} \binom{q}{a}\binom{r}{b} \frac{(-1)^a a (-1)^b b}{M} + o\left(\frac{1}{M}\right) \\
&= 1_{\{q=1,r=1\}} \left(\frac{-t\beta}{M}\right) + o(1/M) \\
&= 1_{\{q=1,r=1\}} \Theta(1/M) + o(1/M). \qquad \text{(B.25)}
\end{aligned}$$

Proof. of Lemma B.2. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density f. Then

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

Using the same arguments as in proof of Lemma A.1, we can show that the contribution of terms $\mathbf{e}_r(\mathbf{X})$,$\mathbf{e}_r(\mathbf{Y})$ to the R.H.S. of the above equation is $o(1/M)$. Define $\sharp(\mathbf{X},\mathbf{Y}) := \gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})Cov_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))^q,(\mathbf{e}_c(Y)+\mathbf{e}_t(Y))^r]$. We can then reduce,

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

Now note that $(\mathbf{e}_c(X)+\mathbf{e}_t(X))^q$ 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 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 (B.24), which follows after recalling 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 (B.24) and noting that the error $\mathbf{e}_c(X)/f(X)$ is a special instance of $\mathbf{E}_\beta(X)$ and subsequently invoking (B.25).

For $\{X,Y\}\in\Psi_\epsilon^c$, the 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 (B.23) and Cauchy-Schwarz can be shown to be $o(k^{-(q+r)/2})$. On the other hand, if $m=q$ and $n=r$, $Cov[(\mathbf{e}_c(X))^q,(\mathbf{e}_c(Y))^r]=O(k^{-(q+r)/2})$.

This yields expressions for terms I and II that are expressed as double sums over m,q and n,r of the aforementioned covariance terms. Asymptotically in k,M these sums are dominated by the components $m=q$ and $n=r$. This completes the proof. □

B.3 Moments of coverage function

We have therefore established that moments of the k-NN and covergae density estimates are identical up to leading terms. Next the central and cross moments of the coverage density estimate are derived.

B.3.1 Central moments for the Coverage density estimate

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

$$\begin{aligned}&\mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_c^q(\mathbf{X})\right]\\&= 1_{\{q=2\}}\mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})f^2(\mathbf{X})\right]\left(\frac{1}{k}\right)+o\left(\frac{1}{k}\right).\end{aligned}\tag{B.26}$$

B.3.2 Cross Moments for the Coverage density estimate

We will first show

$$\begin{aligned}&Cov\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_c^q(\mathbf{X}),1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_c^r(\mathbf{Y})\right]\\&=\left(1_{\{q=1,r=1\}}O(1/M)+o(1/M)\right).\end{aligned}\tag{B.27}$$

From Lemma B.2, it is clear that the contributions of the term $\mathbf{e_t}(X)$ are of a smaller asymptotic order than the contributions of the term $\mathbf{e_c}(X)$. Redefine $\sharp(\mathbf{X},\mathbf{Y}) := \gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})Cov_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e_c}(X))^q,(\mathbf{e_c}(Y))^r]$. By Lemma B.2, we can then write

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

From the results in the proof of Lemma B.2

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

where the last step follows from the fact that probability $Pr(\{\mathbf{X},\mathbf{Y}\} \in \Psi_\epsilon) = 1 - O(k/M) = O(1)$. Similarly,

$$
\begin{aligned}
II &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})} \sharp(\mathbf{X},\mathbf{Y})] \\
&= \mathbb{E}\big[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})} 1_{\{q=1,r=1\}} O(1/k) + o(1/k)\big] \\
&= 1_{\{q=1,r=1\}} O(1/M) + o(1/M).
\end{aligned}
$$

where the last but one step follows since the probability $Pr(\{\mathbf{X},\mathbf{Y}\} \in \Psi_\epsilon^c) = O(k/M)$. We focus on the case $\{q=1, r=1\}$ and seperately analyze disjoint balls and intersecting balls:

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

Disjoint balls For $\{X, Y\} \in \Psi_\epsilon$, the cross-correlation between the coverage density estimates is expressed using (B.24)

$$\begin{aligned} I &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})Cov[\mathbf{e}_c(X), \mathbf{e}_c(Y)]] \\ &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})]\,(-1/M + o(1/M)) \\ &= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})]\,(-1/M + o(1/M)) \\ &= -\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})]\mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})]\frac{1}{M} + o(1/M). \end{aligned} \tag{B.28}$$

where the second to last equality follows by applying the Cauchy-Schwarz inequality and subsequently using the fact that $\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}] = (1 - O(k/M))\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}1_{\{\mathbf{Y}\in\mathcal{S}'\}}]$.

Intersecting balls For $\{X, Y\} \in \Psi_\epsilon^c$, we will show that the cross-correlations of the coverage density estimator and an oracle uniform kernel density estimator (defined below) are identical up to leading terms (without explicitly evaluating the cross-correlation between the coverage density estimates) and then derive the correlation of the oracle density estimator to obtain corresponding results for the coverage estimate.

Oracle ϵ ball density estimate In order to estimate cross moments for the coverage (and thereby k-NN density estimates), we introduce the ϵ *ball* density estimator. The ϵ-ball density estimator is a kernel density estimator that uses a uniform kernel with bandwidth which depends on the unknown density f. Let the volume of the kernel be $V_\epsilon(X)$ and the corresponding kernel region be $S_\epsilon(X) = \{Y \in \mathcal{S} : c_d||X-Y||^d \leq V_\epsilon(X)\}$. The volume is chosen such that the coverage $Q_\epsilon(X) = \int_{S_\epsilon(X)} f(z)dz$ is set to $(1 + p_k)k/M$. 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

$$\hat{\mathbf{f}}_\epsilon(X) = \frac{\mathbf{l}_\epsilon(\mathbf{X})}{MV_\epsilon(X)}. \tag{B.29}$$

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 using results on the volumes of intersections of hyper spheres (refer Appendix A for details).

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

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

Next, the cross-correlations of the coverage density estimator and the ϵ ball density estimator are shown to be asymptotically equal.

Lemma B.4.

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

Proof. We begin by establishing the conditional density and expectation of $\hat{\mathbf{f}}_\epsilon(X)$ given $\hat{\mathbf{f}}_\mathbf{c}(X)$. We drop the dependence on X and denote $\mathbf{l}_\epsilon = \Sigma_{i=1}^M 1_{\{X_i\in S_\epsilon(X)\}}$, the k-NN coverage by $\mathbf{P}$ and the ϵ ball coverage by Q. Let $\mathbf{q} = Q/\mathbf{P}$ and $\mathbf{r} = (Q-\mathbf{P})/(1-\mathbf{P})$. The following expressions for conditional densities and expectations are derived in [57]

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

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

which implies

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

Using the above expressions for conditional expectations, the following marginal expectation are obtained. 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

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

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-Schwarz inequality and the concentration inequality to show $C \times III = o(1/M)$.

Therefore, we have

$$\begin{aligned}\mathbb{E}[\mathbf{e_c}(X)\mathbf{e}_\epsilon(X)] &= \frac{f^2(X)}{k}(1-Q) + O(e^{-p_k^2 k/6}). \\ &= \frac{f^2(X)}{k} - \frac{f^2(X)}{M} + o\left(\frac{1}{M}\right) \\ &= f^2(X)\left(\frac{1}{k} + o\left(\frac{1}{k}\right)\right). \end{aligned} \tag{B.30}$$

Now denote $\mathbf{E}(X) = (\mathbf{e_c}(X) - \mathbf{e}_\epsilon(X))$. Note that $\mathbb{E}[\mathbf{E}^2(X)] = \mathbb{E}[\mathbf{e_c}(X)^2] - 2E[\mathbf{e_c}(X)\mathbf{e}_\epsilon(X)] + \mathbb{E}[\mathbf{e}_\epsilon(X)^2]$. Since $E[\mathbf{e_c}(X)^2] = f^2(X)\frac{1}{k} + o(1/k)$ and $E[\mathbf{e}_\epsilon(X)^2] = f^2(X)(1/k + o(1/k))$ it follows from (B.30) that $\mathbb{E}[E(X)] = o(1/k)$. This result means $\mathbf{e_c}(X)$ and $\mathbf{e}_\epsilon(X)$ are almost perfectly correlated. We can now express the covariance between the coverage density estimates in terms of the covariance between the ϵ ball estimates as follows:

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

Using Cauchy-Schwarz, a bound on 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 $|IV| \leq \sqrt{\mathbb{E}[\mathbf{E}(X)]\mathbb{E}[\mathbf{E}(Y)]}$ can be obtained. Note that the above application of Cauchy-Schwarz 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). Therefore all the three terms II, III and IV are $o(1/k)$. This concludes the proof of Lemma B.4. □

For Lemma B.4 to be useful, we would want $\mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)]$ must be orders of magnitude larger than the error $o(1/k)$, which is indeed the case for $\{X, Y\} \in {\Psi_\epsilon}^c$ since $\mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)] = O(1/k)$ (Lemma A.2, Appendix .1) for such X and Y. We can then use this lemma and the previously established results on co-variance of ϵ-ball density estimates (lemma B.3) to obtain the corresponding result for coverage density estimates:

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

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

Proof.

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

In the second to last step, we obtain o(1/M) for the second term by recognizing that $Pr(\{\mathbf{X}, \mathbf{Y}\} \in \Psi_\epsilon^c) = O(k/M)$ and $O(k/M) \times o(1/k) = o(1/M)$. The above result in conjunction with (B.28) gives the required result. □

B.3.3 Bias of the kNN density estimates

Finally, we analyze the bias of the k-NN density estimate which, unlike other central moments, cannot be obtained using (B.26). Let $X \in \mathcal{S}'$. We can analyze the

bias of k-NN density estimates as follows by using (B.17)

$$\begin{aligned}
\mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_k(X)] &= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_c(X)] + \mathbb{E}\left[1_{\natural(X)}k_M\frac{h(X)}{\mathbf{P}^{1-2/d}(X)}\right] + \mathbb{E}\left[1_{\natural(X)}k_M\mathbf{h}_{\mathbf{s}}(X)\right] \\
&= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_c(X)] + \mathbb{E}\left[1_{\natural(X)}k_M\frac{h(X)}{\mathbf{P}^{1-2/d}(X)}\right] + o\left((k/M)^{2/d}\right) \\
&= \mathbb{E}[\hat{\mathbf{f}}_c(X)] + \mathbb{E}\left[k_M\frac{h(X)}{\mathbf{P}^{1-2/d}(X)}\right] + o\left(\frac{k}{M}\right)^{2/d} + o(1/M) \\
&= f(X) + h(X)\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d}, \qquad \text{(B.31)}
\end{aligned}$$

where we used the fact that under the event $\natural^c(X)$, $((k-1)/M)\mathbf{P}^{1-t}(X) = O(1)$ for any $t >= 0$, which in turn gives $\mathbb{E}[1_{\natural^c(X)}((k-1)/M)\mathbf{P}^{1-t}(X)] = O(Pr(\natural^c(X))) = o(1/M)$. This implies that

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) &= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_k(X)] + \mathbb{E}[1_{\natural^c(X)}\hat{\mathbf{f}}_k(X)] - f(X) \\
&= h(X)\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d} + o(1/M) + \mathbb{E}[1_{\natural^c(X)}\hat{\mathbf{f}}_k(X)] \\
&= h(X)\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d} + o(1/M), \qquad \text{(B.32)}
\end{aligned}$$

where the last step follows because , by (B.21), $1_{\natural^c(X)}\hat{\mathbf{f}}_k(X) = O(1)$. This expression is true for $k >= 3$ by (B.7).

If we assume that the density f is $d+2$-times differentiable, we have

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) &= \mathbb{E}\left[k_M\frac{f(X)}{\mathbf{P}(X)}\right] + k_M\sum_{i\in\mathcal{I}}\mathbb{E}\left[\frac{h_i(X)}{\mathbf{P}^{1-i/d}(X)}\right] + o(k_M) \\
&= \sum_{i\in\mathcal{I}} h_i(X)\left(\frac{k}{M}\right)^{i/d} + O\left(\frac{1}{k} + \frac{k}{M}\right), \qquad \text{(B.33)}
\end{aligned}$$

with $\mathcal{I} = \{2, \ldots, d\}$ and $h_2(X) = h(X)$. The first step follows using the Taylor series

expansion (B.13) and the second step follows from that

$$\frac{\Gamma(k+a)}{\Gamma(k)} = k^a(1+O(1/k)),$$

which implies that $\mathbb{E}[1/\mathbf{P}^{1-i/d}(X)] = \Theta((k/M)^{i/d-1}) + O(1/k)$.

Next, assuming that (2.4) holds, we evaluate $\mathbb{E}[g(\hat{\mathbf{f}}_k(X), X)]$ in an identical fashion to the derivation of (B.32). For $X \in \mathcal{S}_{\mathcal{I}}$,

$$\begin{aligned}
&\mathbb{E}[1_{\natural(X)} g(\hat{\mathbf{f}}_k(X), X)] \\
&= \mathbb{E}\left[1_{\natural}(X) g\left(\hat{\mathbf{f}}_c(X) + k_M h(X)(\mathbf{P}(X))^{2/d-1} + k_M \mathbf{h}_{\mathbf{s}}(X), X\right)\right] \\
&= \mathbb{E}\left[1_{\natural(X)} g\left(\hat{\mathbf{f}}_c(X) + k_M h(X)(\mathbf{P}(X))^{2/d-1} + k_M o((\mathbf{P}(X))^{2/d-1}), X\right)\right] \\
&= \mathbb{E}\left[g\left(\hat{\mathbf{f}}_c(X) + k_M h(X)(\mathbf{P}(X))^{2/d-1} + k_M o((\mathbf{P}(X))^{2/d-1}), X\right)\right] \\
&= \mathbb{E}\left[g(\hat{\mathbf{f}}_c(X), X) + g'(\hat{\mathbf{f}}_c(X), X) k_M h(X)(\mathbf{P}(X))^{2/d-1} + o(k_M \mathbf{P}(X))^{2/d-1})\right] \\
&= g(f(X), X) g_1(k, M) + g_2(k, M) + g'(f(X), X) h(X)(k/M)^{2/d} + o((k/M)^{2/d}).
\end{aligned}$$

This gives,

$$\begin{aligned}
&\mathbb{E}[g(\hat{\mathbf{f}}_k(X), X)] \\
&= \mathbb{E}[1_{\natural(X)} g(\hat{\mathbf{f}}_k(X), X)] + \mathbb{E}[1_{\natural^c(X)} g(\hat{\mathbf{f}}_k(X), X)] \\
&= g(f(X), X) g_1(k, M) + g_2(k, M) \\
&+ g'(f(X), X) h(X)(k/M)^{2/d} + o((k/M)^{2/d}).
\end{aligned} \tag{B.34}$$

B.4 Summary

We summarize the results derived in this appendix here. We then have,

$$\mathbb{E}[1_{\{\mathbf{X} \in \mathcal{S}'\}} \hat{\mathbf{f}}_k(\mathbf{X})] - f(X) = \mathbb{E}[1_{\{\mathbf{X} \in \mathcal{S}'\}} h(\mathbf{X})] \left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d}, \tag{B.35}$$

$$\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] = 1_{\{q=2\}}\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})f^2(\mathbf{X})\big]\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right), \tag{B.36}$$

$$\begin{aligned} &Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k^q(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k^r(\mathbf{Y})\big] \\ &= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})f(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})f(\mathbf{Y})]\left(\frac{1}{M}\right) \\ &+ o\left(\frac{1}{M}\right). \end{aligned} \tag{B.37}$$

B.5 k-NN moments for MI estimation

Define $\mathbf{e}_{ik}(Z) = \hat{f}_{ik}(Z) - \mathbb{E}[\hat{f}_{ik}(Z)]$, $i = 1, 2$ or 12. Also, let $\mathbf{e}_{i\epsilon}$ define the error in the corresponding oracle kernel estimates (refer lemma B.3). Define $\mathcal{S}'$ to be the intersection of the individual $\mathcal{S}'$. Rather than rederive the properties afresh, we will borrow from the theory previously established.

B.5.1 Central and Intra-Cross Moments

We note that since these are standard k-NN density estimates, all the central and intra-cross moments carry over from the previous appendix:

$$\mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\hat{\mathbf{f}}_{ik}(\mathbf{Z})] - f_i(\mathbf{Z}) = \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}h_i(\mathbf{Z})]\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d}, \tag{B.38}$$

$$\mathbb{E}\big[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma(\mathbf{Z})\mathbf{e}_{ik}^q(\mathbf{Z})\big] = 1_{\{q=2\}}\mathbb{E}\big[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma(\mathbf{Z})f_i^2(\mathbf{Z})\big]\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right), \tag{B.39}$$

$$
\begin{aligned}
&Cov\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_1(\mathbf{Z})\mathbf{e}_{ik}^q(\mathbf{Z}), 1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_2(\mathbf{Z})\mathbf{e}_{ik}^r(\mathbf{Z})\right] \\
&= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_1(\mathbf{Z})f_i(\mathbf{Z}), 1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_2(\mathbf{Z})f(\mathbf{Z})]\left(\frac{1}{M}\right) \\
&+ o\left(\frac{1}{M}\right). \qquad \text{(B.40)}
\end{aligned}
$$

B.5.2 Inter-cross moments

In this section, we are interested in evaluating inter-cross moments of the form $Cov[\mathbf{e}_{1k}^q(Z_1), \mathbf{e}_{2k}^r(Z_2)]$, $Cov[\mathbf{e}_{12k}^q(Z_1), \mathbf{e}_{2k}^r(Z_2)]$ and $Cov[\mathbf{e}_{12k}^q(Z_1), \mathbf{e}_{1k}^r(Z_2)]$.

B.5.2.1 Marginal - marginal moments

Because X and Y represent dis-joint sets of variables, it is clear that

$$
Cov[\mathbf{e}_{1k}^r(Z_1), \mathbf{e}_{2k}^q(Z_2)] = 0. \qquad \text{(B.41)}
$$

This in turn implies that for two random variables $\mathbf{Z}_1$ and $\mathbf{Z}_2$ which are (i) equal and drawn from f_{12} or (ii) drawn i.i.d. from f_{12}.

$$
Cov\left[1_{\{\mathbf{Z_1}\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)\mathbf{e}_{1k}^q(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)\mathbf{e}_{2k}^r(\mathbf{Z}_2)\right] = 0. \qquad \text{(B.42)}
$$

B.5.2.2 Joint - marginal moments

First consider the case where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ which are equal and drawn from f_{12}. In this case, when conditioned on $Z = (X, Y) = \mathbf{Z}_1 = \mathbf{Z}_2$,

$$
\begin{aligned}
Cov[\mathbf{e}_{12k}^r(X,Y), \mathbf{e}_{2k}^q(Y)] &= Cov[\mathbf{e}_{12\epsilon}^r(X,Y), \mathbf{e}_{2\epsilon}^q(Y)] + o(1/k) \\
&= o(1/k),
\end{aligned}
$$

where the first step follows from the similarity between oracle kernel and k-NN density estimates established in lemma B.4. The last step follows by recognizing that the

ratio of intersection of the uniform kernel region of the joint density estimate $\hat{\mathbf{f}}_{12u}$ and the marginal estimate $\hat{\mathbf{f}}_{1u}$ to either ball volume k/M is $o(1)$. This is because the intersecting volume is given by the intersection between a sphere and a cylinder whose axis collide and whose volumes are k/M each, which implies that the volume of the intersection is $o(k/M)$. This gives,

$$Cov\big[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_1(\mathbf{Z})\mathbf{e}_{12k}^q(\mathbf{Z}), 1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_2(\mathbf{Z})\mathbf{e}_{1k}^r(\mathbf{Z})\big] = o(1/k). \tag{B.43}$$

To analyze the case where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ which are drawn i.i.d. from f_{12}, we once again consider two cases: the k-NN balls at $\mathbf{Z}_1$ and $\mathbf{Z}_2$ corresponding to $\hat{\mathbf{f}}_{12k}$ and $\hat{\mathbf{f}}_{1k}$ (i) are disjoint and (ii) intersect. Denote these regions by $\Psi_{12\epsilon 1}$ and $\Psi_{12\epsilon 1}^c$ respectively. Let $\Delta_{12\epsilon 1}(X,Y)$ denote the event $\{Z_1, Z_2\} \in \Psi_{12\epsilon 1}^c$. Also define $\sharp_{12,1}(\mathbf{Z}_1,\mathbf{Z}_2) := \gamma_1(\mathbf{Z}_1)\gamma_2(\mathbf{Z}_2)Cov_{\{\mathbf{Z}_1,\mathbf{Z}_2\}}[(\mathbf{e}_{12k}(Z_1))^q, (\mathbf{e}_{1k}(Z_2))^r]$. We can then write

$$\begin{aligned}
&Cov\big[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_!)\mathbf{e}_{12k}(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)\mathbf{e}_{1k}(\mathbf{Z}_2)\big] \\
&= \mathbb{E}[1_{\{\mathbf{Z}_1,\mathbf{Z}_2\in\mathcal{S}'\}}\sharp(\mathbf{X},\mathbf{Y})] \\
&= \mathbb{E}[\mathbf{1}_{\boldsymbol{\Delta}_{\mathbf{12}\epsilon\mathbf{1}}{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\sharp(\mathbf{X},\mathbf{Y})] \\
&+\mathbb{E}[\mathbf{1}_{\boldsymbol{\Delta}_{\mathbf{12}\epsilon\mathbf{1}}(\mathbf{X},\mathbf{Y})}\sharp(\mathbf{X},\mathbf{Y})] \\
&= I + II.
\end{aligned} \tag{B.44}$$

For the disjoint case, using the exact methods shown in Appendix B, we show that

$$\begin{aligned}
I &= 1_{q,r=1}\mathbb{E}[\mathbf{1}_{\boldsymbol{\Delta}_{\mathbf{12}\epsilon\mathbf{1}}{}^{\mathbf{c}}(\mathbf{Z_1},\mathbf{Z_2})}\gamma_1(\mathbf{Z}_1)\gamma_2(\mathbf{Z}_2)Cov[\mathbf{e}_{12k}(\mathbf{Z}_1),\mathbf{e}_{1k}(Z_2)]] + o(1/M) \\
&= \mathbb{E}[\mathbf{1}_{\boldsymbol{\Delta}_{\mathbf{12}\epsilon\mathbf{1}}{}^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{Z}_1)\gamma_2(\mathbf{Z}_2)]\,(-1/M + o(1/M)) \\
&= \mathbb{E}[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_2)\gamma_2(\mathbf{Z}_2)]\,(-1/M + o(1/M)) \\
&= -\mathbb{E}[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_2)]\mathbb{E}[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)]\frac{1}{M} + o(1/M).
\end{aligned} \tag{B.45}$$

For the intersecting balls case, we can first derive that

$$\begin{aligned}&\mathbb{E}\big[\mathbf{1}_{\mathbf{\Delta_{12\epsilon 1}(Z_1,Z_2)}}\gamma_1(\mathbf{Z}_1)\mathbf{e_{12\epsilon}}(\mathbf{Z}_2)\gamma_2(\mathbf{Z}_2)\mathbf{e_{1\epsilon}}(\mathbf{Z}_1 2)\big]\\&= \mathbb{E}[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)\gamma_2(\mathbf{Z}_1)f^2(\mathbf{Z}_1)]\left(\frac{1}{M}+o\left(\frac{1}{M}\right)\right).\end{aligned}$$

Next, in conjunction with lemma B.4, we can show

$$\begin{aligned}I &= Cov\big[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)\mathbf{e}^q_{12k}(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)\mathbf{e}^r_{1k}(\mathbf{Z}_2)\big]\\&= 1_{q,r=1}Cov[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)f(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)f(\mathbf{Z}_2)]\left(\frac{1}{M}+o\left(\frac{1}{M}\right)\right)\end{aligned} \tag{B.46}$$

Plugging (B.45) and (B.46) in to (B.44), we get

$$\begin{aligned}&Cov\big[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)\mathbf{e}^q_{12k}(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)\mathbf{e}^r_{1k}(\mathbf{Z}_2)\big]\\&= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)f_{12}(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)f_1(\mathbf{Z}_2)]\left(\frac{1}{M}\right)+o\left(\frac{1}{M}\right).\end{aligned} \tag{B.47}$$

B.6 Moment properties of angular k-NN density estimates

Throughout this Appendix, let X, Y be distinct points in $\mathcal{X}_N$. We will now analyze properties of $\hat{\mathbf{f}}_{k,\theta}(X)$, conditioned on $\{\mathbf{X}_1, .., \mathbf{X}_N\}$. First, it is show that the region $\mathbf{S}_{k,\theta}(X)$ is contained in the support region $\mathcal{S}$ with high probability. This in turn guarantees the consistency of the angular k-NN density estimates. Define

$$\check{f}_{k,\theta}(X) = \mathbb{E}[\hat{\mathbf{f}}_{k,\theta}(X) \mid X]$$

and denote

$$\hat{\mathbf{e}}_{k,\theta}(X) = \hat{\mathbf{f}}_{k,\theta}(X) - \mathbb{E}[\hat{\mathbf{f}}_{k,\theta}(X) \mid X].$$

Define the angular coverage function to be $\mathbf{P}_{k,\theta}(X) = \int_{\mathbf{S}_{k,\theta}(X)} f(Z)dZ$. It is clear that the random variable $\mathbf{P}_{k,\theta}(X)$ is beta distributed with parameters $k, M-k+1$ [30].

B.6.1 Support subset guarantee

Let $\mathcal{P}(X)$ be the normal vector of X w.r.t the boundary $\mathcal{B}$. By the definition of $\mathcal{P}(X)$, it is clear that for any $Y \in \mathcal{S}$, the angle between $X - \mathcal{P}(X)$ and $X - Y$ lies in the interval $\{-\theta, \theta\}$. To show that $\mathbf{S}_{k,\theta}(X)$ is contained in $\mathcal{S}$, it therefore suffices to show that (i) the angle $\theta(X)$ between the vector $N(X) - X$ and $\mathcal{P}(X) - X$ is between $\{-\theta/2, \theta/2\}$, and (ii) the distance $\mathbf{d}_{k,\theta}(X)$ is small. Next, it is shown that (i) and (ii) hold with high probability.

B.6.1.1 Concentration inequality for angle $\theta(X)$

Let θ_j be the angle between $X - \mathcal{P}(X)$ and $X - X_j$, where X_j is the j-th nearest neighbor of X among $\mathcal{X}_N$. Consider the K-NN graph constructed on $\mathcal{X}_N$, and let $\mathbf{d}_K(X_i)$ be the $K+1$-th nearest neighbor distance from $X_i \in \mathcal{X}_N$. From the results in Section 2 [55], it directly follows that conditioned on the $K+1$-NN distance $\mathbf{d}_K(X_i)$, the K nearest neighbors of X_i are *uniformly* distributed. Using this result, it follows that the angles $\{\theta_j; j = 1, .., K\}$, conditioned on $\mathbf{d}_K(X_i)$, are uniformly distributed with mean 0. Using the Chernoff inequality for uniform distribution, and using the fact that $k = k_0 M^\beta$ and $K = k \times N/M$,

$$Pr\{|\theta(X)| > a\} = \mathbb{E}[Pr\{|\theta(X)| > a \mid \mathbf{d}_K(X_i)\}] \leq exp(-a^2 K) = o(1/T),$$

for any $a > 0$. Setting $a = \theta/2$, it immediately follows that $\theta(X)$ lies between $\{-\theta/2, \theta/2\}$ with high probability $1 - o(1/T)$. Denote the event $\theta(X) \in \{-\theta/2, \theta/2\}$ by $\sharp(X)$.

B.6.1.2 Concentration inequality for coverage probability

Define

$$k_M = (k-1)/M.$$

Let $\flat(X)$ denote the event

$$\mathbf{P}_{k,\theta}(X) < (p_k + 1)k_M, \tag{B.48}$$

where $p_k = \sqrt{6}/(k^{\delta/2})$. Using Chernoff inequalities, $1 - Pr(\flat(X)) = O(e^{-p_k^2 k/2}) = O(e^{-3k^{(1-\delta)}})$. Equivalently,

$$1 - Pr(\flat(X)) = O(\mathcal{C}(k)), \tag{B.49}$$

where $\mathcal{C}(k)$ is a function which satisfies the rate of decay condition $\mathcal{C}(k) = O(e^{-3k^{(1-\delta)}})$. Observe that for k/M sufficiently small, the event $\flat(X)$ translates to the event that $\mathbf{d}_{k,\theta}(X)$ is correspondingly small.

B.6.1.3 Support subset condition

Under the event $\natural(X) = \flat(X) \cap \sharp(X)$, it is clear that $\mathbf{V}_{k,\theta}(X)$ is a subset of $\mathcal{S}$. Also observe that

$$1 - Pr(\natural(X)) = O(\mathcal{C}(k)). \tag{B.50}$$

B.6.2 Taylor series expansion of coverage probability

Identical to the derivation of (70) in [80], when $\natural(X)$ is true, or equivalently when $\mathbf{V}_{k,\theta}(X)$ is a subset of $\mathcal{S}$, we can write

$$\begin{aligned}\mathbf{P}_{k,\theta}(X) &= \int\limits_{\mathbf{S}_{k,\theta}(X)} f(z)dz \\ &= f(X)\mathbf{V}_{k,\theta}(X) + \sum_{i=1}^{\nu-1} c_t(X)\mathbf{V}_{k,\theta}^{1+2i/d}(X) + o(\mathbf{V}_{k,\theta}^{1+2i/d}(X)),\end{aligned}$$

where $c_i(X)$ depend only on the derivatives of the underlying density f and $N(X)$, and therefore only on $\mathcal{X}_N$, which in turn implies that these terms are statistically independent of $\mathcal{X}_M$. This then implies that under the event $\natural(X)$

$$\frac{1}{\mathbf{V}_{k,\theta}(X)} = \frac{f(X)}{\mathbf{P}_{k,\theta}(X)} + \sum_{t\in\mathcal{T}} \frac{h_i(X)}{\mathbf{P}^{1-t}(X)} + \mathbf{h_r}(X), \tag{B.51}$$

where $\mathcal{T} = \{1/d, 2/d, .., 2\nu/d\}$ and $\mathbf{h_r}(X) = o(\mathbf{P}_{k,\theta}^{2\nu/d-1}(X))$. Again, the terms $h_i(X)$ and the reminder term $h_r(\tilde{X})$ depend only on $\mathcal{X}_N$, and are statistically independent of $\mathcal{X}_M$. Now, by $(\mathcal{A}.2)$, we have $(k/M)^{2\nu/d} = o(1/M)$. This implies that $2\nu/d > 1$. Under the event $\natural(X)$, we have $\mathbf{P}(X) \leq (p_k+1)k/M$, which, in conjunction with the condition $2\nu/d > 1$ implies that

$$\mathbf{h_r}(X) \;\; = \;\; o(\mathbf{P}_{k,\theta}^{2\nu/d-1}(X)) = o((k/M)^{2\nu/d-1}) = o(1/k_M M). \tag{B.52}$$

B.6.3 Bounds on k-NN density estimates

We have the following bounds under the events $\natural(X)$ and $\natural^c(X)$. Under the event $\natural(X)$,

$$(1-\epsilon)\epsilon_0 \frac{k_M}{\mathbf{P}_{k,\theta}(X)} \leq \;\; \hat{\mathbf{f}}_k(X) \;\; \leq (1+\epsilon)\epsilon_\infty \frac{k_M}{\mathbf{P}_{k,\theta}(X)}. \tag{B.53}$$

Under the event $\natural^c(X)$, we can bound $\mathbf{V}_{k,\theta}(X)$ from above by $c_d\mathcal{D}^d$. Also, since $\mathbf{V}_k(X)$ is monotone in $\mathbf{P}_{k,\theta}(X)$, under the event $\natural^c(X)$, we can bound $\mathbf{V}_{k,\theta}(X)$ from below by $(1+p_k)k_M/(1-\epsilon)$. Written explicitly,

$$(k-1)/(Mc_d\mathcal{D}^d) \leq \;\; \hat{\mathbf{f}}_k(X) \;\; \leq (1-\epsilon)\epsilon_\infty. \tag{B.54}$$

B.6.4 Bias of angular boundary corrected density estimates

We can analyze the bias of k-NN density estimates as follows. Define

$$\hat{\mathbf{f}}_{c,\theta}(X) = \frac{k_M f(X)}{\mathbf{P}_{k,\theta}(X)}.$$

When $\natural(X)$ holds, by using (B.13) and the applying the fact that $2\nu/d > 1$,

$$\frac{1}{\mathbf{V}_k(X)} = \frac{f(X)}{\mathbf{P}_{k,\theta}(X)} + \sum_{i\in\mathcal{I}} \frac{h_i(X)}{\mathbf{P}_{k,\theta}^{1-i/d}(X)} + \mathbf{h_s}(X), \tag{B.55}$$

where $\mathcal{I} = \{1, .., d\}$ and $\mathbf{h_s}(X) = o(1)$. This gives,

$$\hat{\mathbf{f}}_{k,\theta}(X) \;=\; \hat{\mathbf{f}}_{c,\theta}(X) + \sum_{i\in\mathcal{I}} k_M \frac{h(X)}{\mathbf{P}_{k,\theta}^{1-i/d}(X)} + k_M \mathbf{h_s}(X), \tag{B.56}$$

whenever $\natural(X)$ is true. Then,

$$\begin{aligned}
\mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_{k,\theta}(X)] &= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_{c,\theta}(X)] \\
&\quad + \sum_{i\in\mathcal{I}} \mathbb{E}\left[1_{\natural(X)} k_M \frac{h_i(X)}{\mathbf{P}_{k,\theta}^{1-i/d}(X)}\right] + \mathbb{E}\left[1_{\natural(X)} k_M \mathbf{h_s}(X)\right] \\
&= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_{c,\theta}(X)] + \sum_{i\in\mathcal{I}} \mathbb{E}\left[1_{\natural(X)} k_M \frac{h(X)}{\mathbf{P}_{k,\theta}^{1-i/d}(X)}\right] + o\left(\mathbb{E}\left[1_{\natural(X)} k_M\right]\right) \\
&= \mathbb{E}[\hat{\mathbf{f}}_{c,\theta}(X)] + \mathbb{E}\left[k_M \frac{h(X)}{\mathbf{P}_{k,\theta}^{1-i/d}(X)}\right] + o\left(\frac{k}{M}\right) \\
&= f(X) + \sum_{i\in\mathcal{I}} h_i(X) \left(\frac{k}{M}\right)^{i/d} + o\left(\frac{k}{M}\right),
\end{aligned} \tag{B.57}$$

where we used the fact that under the event $\natural^c(X)$, $((k-1)/M)\mathbf{P}_{k,\theta}^{1-t}(X) = O(1)$ for any $t >= 0$, which in turn gives $\mathbb{E}[1_{\natural^c(X)}((k-1)/M)\mathbf{P}_{k,\theta}^{1-t}(X)] = O(Pr(\natural^c(X))) = O(\mathcal{C}(k))$.

This implies that

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{f}}_{k,\theta}(X)] - f(X) &= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_{k,\theta}(X)] + \mathbb{E}[1_{\natural^c(X)}\hat{\mathbf{f}}_{k,\theta}(X)] - f(X) \\ &= \sum_{i\in\mathcal{I}} h_i(X)\left(\frac{k}{M}\right)^{i/d} + o\left(\frac{k}{M}\right) + \mathbb{E}[1_{\natural^c(X)}\hat{\mathbf{f}}_k(X)] \\ &= \sum_{i\in\mathcal{I}} h_i(X)\left(\frac{k}{M}\right)^{i/d} + o\left(\frac{k}{M}\right), \end{aligned} \tag{B.58}$$

where the last step follows because, by (B.21), $1_{\natural^c(X)}\hat{\mathbf{f}}_k(X) = O(1)$ and $O(\mathcal{C}(k)) = o(1/M)$. This gives the following lemma:

Lemma B.6. *Let $\gamma(x,y)$ be an arbitrary function with d partial derivatives wrt x and $\sup_{x,y}|\gamma(x,y)| < \infty$. Then,*

$$\mathbb{E}[\gamma(\check{f}_{k,\theta}(\mathbf{Z}),\mathbf{Z})] - \mathbb{E}[\gamma(f(\mathbf{Z}),\mathbf{Z})] = \sum_{i=1}^{d} c_{1,i}(\gamma(x,y))(k/M)^{i/d} + o((k/M)), \tag{B.59}$$

where $c_{1,i}(\gamma(x,y))$ are functionals of γ and f.

Proof.

$$\begin{aligned}\mathbb{E}[\gamma(\check{f}_{k,\theta}(\mathbf{Z}),\mathbf{Z})] - \mathbb{E}[\gamma(f(\mathbf{Z}),\mathbf{Z})] &= \mathbb{E}[\gamma(\check{f}_{k,\theta}(\mathbf{Z}),\mathbf{Z}) - \gamma(f(\mathbf{Z}),\mathbf{Z})] \\ &= \sum_{i=1}^{d}\mathbb{E}\left[\gamma^{(i)}(f(\mathbf{Z}),\mathbf{Z})\big(\check{f}_{k,\theta}(\mathbf{Z}) - f(\mathbf{Z})\big)^i\right] \\ &= \sum_{i=1}^{d} c_{1,i}(\gamma(x,y))(k/M)^{i/d} + o((k/M)), \end{aligned} \tag{B.60}$$

where $c_{1,i}(\gamma(x,y))$ are functionals of $\gamma(x,y)$ and its derivatives. □

B.6.5 Central and cross moments for angular k-NN density estimates

Lemma B.7. *Let $\gamma(x)$ be an arbitrary function satisfying $\sup_x|\gamma(x)| < \infty$. Then,*

$$\mathbb{E}\big[\gamma(\mathbf{X})\hat{\mathbf{e}}^q_{k,\theta}(\mathbf{X})\big] = 1_{\{q=2\}}c_2(\gamma(x))\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right), \tag{B.61}$$

where $c_2(\gamma(x))$ is a functional of γ and f.

Lemma B.8. *Let $\gamma_1(x)$, $\gamma_2(x)$ be arbitrary functions with 1 partial derivative wrt x and $\sup_x |\gamma_1(x)| < \infty$, $\sup_x |\gamma_2(x)| < \infty$. Then,*

$$Cov\left[\gamma_1(\mathbf{X})\hat{\mathbf{e}}^q_{k,\theta}(\mathbf{X}), \gamma_2(\mathbf{Y})\hat{\mathbf{e}}^r_{k,\theta}(\mathbf{Y})\right] = 1_{\{q=1,r=1\}} c_5(\gamma_1(x), \gamma_2(x)) \left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right) \quad \text{(B.62)}$$

where $c_5(\gamma_1(x), \gamma_2(x))$ is a functional of $\gamma_1(x)$, $\gamma_2(x)$ and f.

Proof. The derivation of these results are identical to the derivation of (116) and (117) in [80]. (116) and (117) are derived using (70). Observing that $\mathbf{P}(X)$ (section B.1 [80]) and $\mathbf{P}_{k,Y,\theta}(X)$ and both beta random variables with parameters k, $M - k + 1$, it follows that (70) and (B.13) are identical up to leading constants. This in turn implies that (C.3) and (C.4) follow from (B.13), in an identical manner to the derivation of (116) and (117) from (70) and finally using the fact that $O(\mathcal{C}(k)) = o(1/M)$ under the assumption $k = k_o M^\beta$. □

APPENDIX C

Boundary extension

C.1 k-NN density estimator moments for entropy estimation

In the previous section, we established results for points in any set $\mathcal{S}'$ satisfying the condition $\mathcal{S}'$ is a subset of $\mathcal{S}_{\mathcal{I}}$. In this section, we extend these results to the case when $\mathcal{S}'$ is not a subset of $\mathcal{S}_{\mathcal{I}}$.

We begin by observing that the volume of the set $\mathcal{S} - \mathcal{S}_{\mathcal{I}} = O(k/M)^{1/d}$ by virtue of the fact that the volume of the k-NN balls in d-dimensions is $O(k/M)$.

C.1.1 Bias

When $X \in \mathcal{S} - \mathcal{S}_{\mathcal{I}}$, the k-NN balls centered at X are often truncated at the the boundary. Let

$$\alpha_k(X) = \frac{\int_{\mathbf{S}_k(X) \cap \mathcal{S}} dZ}{\int_{\mathbf{S}_k(X)} dZ}$$

be the fraction of the volume of the k-NN ball inside the boundary of the support. Also define $\mathbf{V}_{k,M}(X)$ to be the k-NN ball volume in a sample of size M. For interior points $X \in \mathbf{S}''$, with high probability, $\alpha_k(X) = 1$, 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

$$\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) = (1 - \alpha_k(X))f(X) + o(1). \tag{C.1}$$

Therefore the bias is much higher at the boundary of the support ($O(1)$) as compared to its interior ($O((k/M)^{2/d})$) (B.32). Furthermore, the bias at the support boundary does not decay to 0 as $k/M \to 0$.

This implies that the overall bias is given by

$$\begin{aligned}
&\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}(\hat{\mathbf{f}}_k(\mathbf{X}) - f(\mathbf{X}))] \\
&= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\cap\mathcal{S}_{\mathcal{I}}{}^c\}}(\hat{\mathbf{f}}_k(\mathbf{X}) - f(\mathbf{X}))] + \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\cap\mathcal{S}_{\mathcal{I}}\}}(\hat{\mathbf{f}}_k(\mathbf{X}) - f(\mathbf{X}))] \\
&= h_0 \left(\frac{k}{M}\right)^{1/d} + \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\cap\mathcal{S}_{\mathcal{I}}\}}h(\mathbf{X})] \left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{1/d}.
\end{aligned} \tag{C.2}$$

for some constant h_0 which depends on the density f and the support $\mathcal{S}$.

C.1.2 Central moments

Observe that $Pr(\mathbf{Y} \in \mathcal{S}_{\mathcal{I}}{}^c) = O((k/M)^{1/d})$. Also, from the concentration inequality, we have $\mathbb{E}[e_k^q(X)] = O(1/k^{q/2})$ for any $X \in \mathcal{S}$. From this, it follows that

$$\begin{aligned}
&\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] \\
&= \mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\cap\mathcal{S}_{\mathcal{I}}{}^c\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] + \mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\cap\mathcal{S}_{\mathcal{I}}\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] \\
&= \mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\cap\mathcal{S}_{\mathcal{I}}{}^c\}}\gamma(\mathbf{X})O(1/k^{q/2})\big] + 1_{\{q=2\}}\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\cap\mathcal{S}_{\mathcal{I}}\}}\gamma(\mathbf{X})f^2(\mathbf{X})\big] \\
&= 1_{\{q=2\}}\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})f^2(\mathbf{X})\big] \left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right).
\end{aligned} \tag{C.3}$$

C.1.3 Cross moments

From the work done by Evans *etal* [26], $Cov[e_k^q(X), e_k^r(Y)] = O(k^5/M)$ for any $X, Y \in \mathcal{S}$. We can then write

$$\begin{aligned} &Cov\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k^q(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k^r(\mathbf{Y})\right] \\ &= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})f(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})f(\mathbf{Y})]\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right). \end{aligned} \tag{C.4}$$

C.2 k-NN density estimator moments for MI estimation

Given that the probability of a point being in the boundary region $\mathcal{S} - \mathcal{S}''$ is $O((k/M)^{1/d}) = o(1)$, the contribution of the boundary region to the overall bias, variance and other cross moments of the boundary corrected density estimator $\hat{\mathbf{f}}$ are asymptotically negligible compared to the contribution from the interior. As a result we can now generalize the results from section B.5 on the cross moments for arbitrary subsets $\mathcal{S}'$ as follows.

$$Cov\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_1(\mathbf{Z})\mathbf{e}_{12k}^q(\mathbf{Z}), 1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_2(\mathbf{Z})\mathbf{e}_{1k}^r(\mathbf{Z})\right] = o(1/k). \tag{C.5}$$

$$\begin{aligned} &Cov\left[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)\mathbf{e}_{12k}^q(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)\mathbf{e}_{1k}^r(\mathbf{Z}_2)\right] \\ &= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)f_{12}(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)f_1(\mathbf{Z}_2)]\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right). \end{aligned} \tag{C.6}$$

APPENDIX D

General results for Bias and Variance of plug-in estimators

D.1 General properties of k-NN density estimators

Let $\mathbf{Z}$ be a random variable with density f. Denote the conditional expected value $E[\hat{\mathbf{f}}(\mathbf{Z})|\mathbf{Z}]$ by $\mathbb{E}_Z[\hat{\mathbf{f}}(\mathbf{Z})]$ and define $\hat{\mathbf{f}}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]$ by $\mathbf{e}(\mathbf{Z})$.

In this section, we will assume that the density estimate $\hat{\mathbf{f}}(\mathbf{Z})$ satisfies the following properties:

- $\mathscr{A}.1$ Bounds: Assume that the following bounds hold under the event $\natural(X)$ and $\natural^c(X)$ respectively:

$$(1-\epsilon)k_M f(X)/\mathbf{P}(X) \leq \quad \hat{\mathbf{f}}(X) \quad \leq (1+\epsilon)k_M f(X)/\mathbf{P}(X). \tag{D.1}$$

and

$$(k-1)/(Mc_d\mathcal{D}^d) \leq \quad \hat{\mathbf{f}}(X) \quad \leq (1-\epsilon)f(X). \tag{D.2}$$

- $\mathscr{A}.2$ Higher order central moments: $\mathbb{E}[\mathbf{e}^r(\mathbf{Z})] = O(1/k^r)$ for any integer $r >= 2$. Also, for $r = 2$, assume that $\mathbb{E}[\mathbf{e}^2(\mathbf{Z})] = f^2(\mathbf{Z})(1/k + o(1/k))$.

- $\mathscr{A}.3$ Higher order cross moments: Assume that the following condition holds:

$$\begin{aligned} &Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k^q(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k^r(\mathbf{Y})\big] \\ &= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})f(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})f(\mathbf{Y})]\left(\frac{1}{M}\right) \\ &+ o\left(\frac{1}{M}\right). \end{aligned} \tag{D.3}$$

Observe that these properties are satisfied by any one of (i) standard k-NN density estimates $\hat{f}_k(.)$, (ii) boundary corrected density estimates $\tilde{f}_k(.)$ or (iii) angular weighted boundary corrected density estimates $\hat{f}_{k,K}(.)$ (please refer to Appendix B for details).

Lemma D.0. *Assume that $U(x, y)$ is any arbitrary functional which satisfies*

$$(i) \sup_{x\in(\epsilon_0,\epsilon_1)} |U(x, y)| = G_0 < \infty,$$

$$(ii) \sup_{x\in(q_l,q_u)} |U(x, y)|\mathcal{C}(k) = G_1 < \infty,$$

$$(iii)\mathbb{E}[\sup_{x\in(p_l,p_u)} |U(x/\mathbf{p}, y)|] = G_2 < \infty.$$

Let $\mathbf{Z}$ denote $\mathbf{X}_i$ for some fixed $i \in \{1, .., N\}$. Let $\zeta_{\mathbf{Z}}$ be any random variable which almost surely lies in the range $(f(\mathbf{Z}), \hat{\mathbf{f}}(\mathbf{Z}))$. If assumption $\mathscr{A}.1$ holds, then

$$\mathbb{E}[|U(\zeta_{\mathbf{Z}}, \mathbf{Z})|] < \infty.$$

Proof. We will show that the conditional expectation $\mathbb{E}[|U(\zeta_Z, Z)| \mid \mathcal{X}_N] < \infty$. Be-

cause $0 < \epsilon_0 < f(X) < \epsilon_\infty < \infty$ by $(\mathcal{A}.1)$, it immediately follows that

$$\mathbb{E}[|U(\zeta_{\mathbf{Z}}, \mathbf{Z})|] = \mathbb{E}[\mathbb{E}[|U(\zeta_Z, Z)| \mid \mathcal{X}_N]] < \infty.$$

By (D.1) and $(\mathcal{A}.1)$, we know that if $\natural(Z)$ holds, $p_l/\mathbf{P}(Z) < \hat{\mathbf{f}}(Z) < p_u/\mathbf{P}(Z)$. On the other hand, if $\natural^c(Z)$ holds, by (D.2) and $(\mathcal{A}.1)$, $q_l < \hat{\mathbf{f}}(Z) < q_u$. This therefore implies that if $\natural(Z)$ holds, $\min\{\epsilon_0, p_l/\mathbf{P}(Z)\} < \zeta_Z < \max\{\epsilon_\infty, p_u/\mathbf{P}(Z)\}$ and if $\natural^c(Z)$ holds, $\min\{\epsilon_0, q_l\} < \zeta_Z < \max\{\epsilon_\infty, q_u\}$. Then,

$$\begin{aligned}
&\mathbb{E}[|U(\zeta_Z, Z)| \mid \mathcal{X}_N] = \mathbb{E}[1_{\natural(Z)}|U(\zeta_Z, Z)| \mid \mathcal{X}_N] + \mathbb{E}[1_{\natural^c(Z)}|U(\zeta_Z, Z)| \mid \mathcal{X}_N] \\
&\leq G_0 + \mathbb{E}[1_{\natural(Z)} \sup_{x \in (p_l, p_u)} |U(x/\mathbf{P}(Z), Z)|] + \max\{G_0, G_1/\mathcal{C}(k)\} Pr(\natural^c(Z)) \\
&\leq G_0 + \mathbb{E}[\sup_{x \in (p_l, p_u)} |U(x/\mathbf{P}(Z), Z)|] + \max\{G_0, G_1/\mathcal{C}(k)\} Pr(\natural^c(Z)) \\
&= G_0 + G_2 + \max\{G_1/\mathcal{C}(M), G_0\}\mathcal{C}(k) \\
&= G_0 + G_2 + \max\{G_1, G_0\mathcal{C}(k)\} < \infty
\end{aligned} \tag{D.4}$$

where the final step follows from the fact that $\mathcal{C}(k) = o(1)$. This concludes the proof. □

D.2 Entropy estimators

The entropy estimators we have defined are of the general form

$$\hat{\mathbf{G}}(f) \;=\; \left(\frac{1}{N} \sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}(\mathbf{X}_i), \mathbf{X}_i) \right). \tag{D.5}$$

where the set $\mathcal{S}'$ is arbitrary and $\hat{f}(.)$ can be any one of (i) standard k-NN density estimates $\hat{f}_k(.)$, (ii) boundary corrected density estimates $\tilde{f}_k(.)$ or (iii) angular boundary corrected density estimates $\hat{f}_{k,\theta}(.)$. Under these assumptions, we prove the following lemmas:

D.2.1 Bias

Lemma D.1. *If assumptions $\mathscr{A}.1$ and $\mathscr{A}.2$ are satisfied, the bias of the entropy estimator is given by*

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{G}}(f)] - G(f) &= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z}) - g(f(\mathbf{Z}),\mathbf{Z}))] \\ &\quad + \frac{1}{2}\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g''(f(\mathbf{Z}),\mathbf{Z})f^2(\mathbf{Z})\right]\left(\frac{1}{k}\right) + o(1/k).\end{aligned}$$

Proof. Using the continuity of $g'''(x,y)$, construct the following third order Taylor series of $g(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})$ around the conditional expected value $\mathbb{E}_Z[\hat{\mathbf{f}}(\mathbf{Z})]$.

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

where $\zeta_{\mathbf{Z}} \in (g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z}), g(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})$ is defined by the mean value theorem. This gives

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

where the last but one step follows from (C.3), joint continuity of $g^{(3)}(x,y)$ (in the interval $x \in (\epsilon_0,\epsilon_\infty)$) and the fact $\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] = f(\mathbf{Z}) + o(1)$. Let $\Delta(\mathbf{Z}) = \frac{1}{6}1_{\{\mathbf{Z}\in\mathcal{S}'\}}g^{(3)}(\zeta_{\mathbf{Z}},\mathbf{Z})$. From Lemma D.0, it follows that $\Delta(\mathbf{Z})$ converges in probability to $\frac{1}{6}1_{\{\mathbf{Z}\in\mathcal{S}'\}}g^{(3)}(f(\mathbf{Z}),\mathbf{Z})$. This combined with the fact that $\Delta(\mathbf{Z})$ is uniformly bounded implies that $\mathbb{E}[\Delta^2(\mathbf{Z})] = O(1)$. By Cauchy-Schwarz,

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

By observing that the samples $\{\mathbf{X}_i\}, i = 1, \ldots, N$ are identical, we therefore have

$$\begin{aligned}
&\mathbb{E}[\hat{\mathbf{G}}(f)] - G(f) = \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\hat{f}(\mathbf{Z}), \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\
&= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] + \mathbb{E}\left[\frac{1}{2}1_{\{\mathbf{Z}\in\mathcal{S}'\}}g''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z})\mathbf{e}^2(\mathbf{Z})\right] \\
&+o(1/k).
\end{aligned}$$

As a final step, we note that $\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] = f(\mathbf{Z}) + o(1)$. This implies that

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{G}}(f)] - G(f) &= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\
&\quad +\frac{1}{2}\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g''(f(\mathbf{Z}), \mathbf{Z})\mathbf{e}^2(\mathbf{Z})\right] + o(1/k) \\
&= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\
&\quad +\frac{1}{2}\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g''(f(\mathbf{Z}), \mathbf{Z})f^2(\mathbf{Z})\right]\left(\frac{1}{k}\right) + o(1/k).
\end{aligned}$$

□

D.2.2 Variance

Lemma D.2. *Under assumptions $\mathscr{A}.2$ and $\mathscr{A}.3$ listed above, the variance of the entropy estimator is given by*

$$\begin{aligned}
\mathbb{V}(\hat{\mathbf{G}}(f)) &= \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z})]\left(\frac{1}{N}\right) \\
&\quad +\mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g'(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z})\hat{\mathbf{f}}(\mathbf{Z})]\left(\frac{1}{M}\right) \\
&\quad +o\left(\frac{1}{M} + \frac{1}{N}\right).
\end{aligned}$$

Proof. By the continuity of $g^{(\lambda)}(x, y)$, we can construct the following Taylor series of

$g(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z})$ around the conditional expected value $\mathbb{E}_Z[\hat{\mathbf{f}}(\mathbf{Z})]$.

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

where $\xi_{\mathbf{Z}} \in (g(f(\mathbf{Z})), g(\hat{\mathbf{f}}(\mathbf{Z})))$. Denote $(g'(\xi_{\mathbf{Z}}, \mathbf{Z}))/\lambda!$ by $\Psi(\mathbf{Z})$. Further define the operator $\mathcal{M}(\mathbf{Z}) = \mathbf{Z} - \mathbb{E}[\mathbf{Z}]$ and

$$\begin{aligned} p_i &= \mathcal{M}(1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)], \mathbf{X_i})), \\ q_i &= \mathcal{M}(1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g'(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)], \mathbf{X_i}) \mathbf{e}(\mathbf{X_i})), \\ r_i &= \mathcal{M}\left(\sum_{i=2}^{\lambda} \frac{1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g^{(i)}(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)], \mathbf{X_i})}{i!} \mathbf{e}^i(\mathbf{X_i}) \right) \\ s_i &= \mathcal{M}\left(1_{\{\mathbf{X}_i \in \mathcal{S}'\}} \Psi(\mathbf{X_i}) \mathbf{e}^{\lambda}(\mathbf{X_i}) \right) \end{aligned}$$

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

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

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

$$\begin{aligned} \mathbb{E}\left[(p_1 + q_1 + r_1 + s_1)^2 \right] &= \mathbb{E}[{p_1}^2] + o(1) \\ &= \mathbb{V}[1_{\{\mathbf{Z} \in \mathcal{S}'\}} g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z})] + o(1) \end{aligned}$$

From assumption 3, it follows that

- $\mathbb{E}[{p_1}^2] = \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})] + o(1)$
- $\mathbb{E}[q_1 q_2] = \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g'(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})\hat{\mathbf{f}}(\mathbf{Z})]\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right)$
- $\mathbb{E}[q_1 r_2] = o\left(\frac{1}{M}\right)$
- $\mathbb{E}[r_1 r_2] = o\left(\frac{1}{M}\right)$

Since q_1 and s_2 are 0 mean random variables

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

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

$$\begin{aligned}
\mathbb{V}(\hat{\mathbf{G}}(f)) &= \frac{1}{N}\mathbb{E}\left[{p_1}^2\right] + \frac{(N-1)}{N}\mathbb{E}[q_1 q_2] + o\left(\frac{1}{M}+\frac{1}{N}\right). \\
&= \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})]\left(\frac{1}{N}\right) \\
&\quad +\mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g'(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})\hat{\mathbf{f}}(\mathbf{Z})]\left(\frac{1}{M}\right) \\
&\quad +o\left(\frac{1}{M}+\frac{1}{N}\right).
\end{aligned}$$

□

D.3 Divergence estimators

The divergence estimators to estimate $G(f_1, f_2)$ that we have defined are of the general form

$$\hat{\mathbf{G}}(f_1, f_2) \;=\; \left(\frac{1}{N}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g_2(\hat{\mathbf{f}}_1(\mathbf{X}_i)/\hat{\mathbf{f}}_2(\mathbf{X}_i), \mathbf{X}_i)\right). \tag{D.6}$$

where the set $\mathcal{S}'$ is arbitrary and $\hat{f}_a(.)$ $(a = 1, 2)$ can be any one of (i) standard k-NN density estimates $\hat{f}_k(.)$, (ii) boundary corrected density estimates $\tilde{f}_k(.)$ or (iii) angular weighted boundary corrected density estimates $\hat{f}_{k,K}(.)$.

Let $\mathbf{Z}$ be a random variable with density f_2. Denote the conditional expected value $E[\hat{\mathbf{f}}_a(\mathbf{Z})|\mathbf{Z}]$ by $\mathbb{E}_Z[\hat{\mathbf{f}}_a(\mathbf{Z})]$ and define $\hat{\mathbf{f}}_a(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_a(\mathbf{Z})]$ by $\mathbf{e}(\mathbf{Z})$. Also define $\hat{\mathbf{f}} := \hat{\mathbf{f}}_1/\hat{\mathbf{f}}_2$, $f = f_1/f_2$ and $\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] = \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{Z})/\hat{\mathbf{f}}_2(\mathbf{Z})] = \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{Z})]/\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Z})]$. Also define $\hat{\mathbf{f}}_1(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{Z})]$ by $\mathbf{e}_1(\mathbf{Z})$ and $\hat{\mathbf{f}}_2(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Z})]$ by $\mathbf{e}_2(\mathbf{Z})$. Finally, define

$$\mathbf{e}_{1,2}(\mathbf{Z}) = \frac{\mathbf{e}_1(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{Z})]} - \frac{\mathbf{e}_2(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Z})]}.$$

Because $\mathbf{e}_1(\mathbf{Z})$ and $\mathbf{e}_2(\mathbf{Z})$ are conditionally independent on $\mathbf{Z}$, the variance $\mathbb{V}[\mathbf{e}_{12}(\mathbf{Z})|\mathbf{Z}] = \mathbb{V}[\frac{\mathbf{e}_1(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{Z})]}|\mathbf{Z}] + \mathbb{V}[\frac{\mathbf{e}_2(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Z})]}|\mathbf{Z}]$.

Uisng Taylor series expansion, we can write

$$\begin{aligned}\frac{\hat{\mathbf{f}}_1(x)}{\hat{\mathbf{f}}_2(x)} \;&=\; \frac{f_1(x)}{f_2(x)}\left(1 + \frac{\hat{\mathbf{f}}_1(x) - f_1(x)}{f_1(x)} - \frac{\hat{\mathbf{f}}_2(x) - f_2(x)}{f_2(x)} + \frac{(\hat{\mathbf{f}}_2(x) - f_2(x))^2}{f_2^2(x)}\right) \\ &\quad + O((\hat{\mathbf{f}}_2(x) - f_2(x))^3).\end{aligned}$$

D.3.1 Bias

Lemma D.3. *If assumptions $\mathscr{A}.1$ and $\mathscr{A}.2$ are satisfied, the bias of the divergence estimator is given by*

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{G}}(f_1,f_2)] - G(f_1,f_2) &= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z}) - g(f(\mathbf{Z}),\mathbf{Z}))] + o(1/k) \\ &\quad + \mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\left(g_2'(f(\mathbf{Z}),\mathbf{Z})f(\mathbf{Z}) + g_2''(f(\mathbf{Z}),\mathbf{Z})f(\mathbf{Z})^2\right)\right]\left(\frac{1}{k}\right).\end{aligned}$$

Proof. Using the continuity of $g_2'''(x,y)$, construct the following third order Taylor series of $g_2(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})$ around the conditional expected value $\mathbb{E}_Z[\hat{\mathbf{f}}(\mathbf{Z})]$.

$$\begin{aligned}&g_2(\hat{\mathbf{f}}_1(\mathbf{Z})/\hat{\mathbf{f}}_2(\mathbf{Z}),\mathbf{Z}) = g_2(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z}) \\ &+g_2'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z})\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]\left(\frac{\mathbf{e}_1(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{Z})]} - \frac{\mathbf{e}_2(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Z})]} + \frac{\mathbf{e}_2^2(\mathbf{Z})}{(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Z})])^2}\right) \\ &+g_2''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z})\frac{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]^2}{2}\left(\frac{\mathbf{e}_1^2(\mathbf{Z})}{(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{Z})])^2} + \frac{\mathbf{e}_2^2(\mathbf{Z})}{(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Z})])^2}\right) \\ &+o(\mathbf{e}_1^2(\mathbf{Z}) + \mathbf{e}_2^2(\mathbf{Z})).\end{aligned}$$

This gives

$$\begin{aligned}&\mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g_2(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z}) - g_2(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z}))] \\ &= \mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\left(g_2'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z})\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] + (1/2)g_2''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z})\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]^2\right)\right]\left(\frac{1}{k}\right) \\ &+ \mathbb{E}\big[o(\mathbf{e}_1^2(\mathbf{Z}) + \mathbf{e}_2^2(\mathbf{Z}))\big]\end{aligned}$$

Using Cauchy-Schwarz inequality as in the proof of Lemma D.1, we can show that $\mathbb{E}[o(\mathbf{e}_1^2(\mathbf{Z}) + \mathbf{e}_2^2(\mathbf{Z}))] = o(1/k)$. Once again, by observing that the samples $\{\mathbf{X}_i\}, i =$

$1, \ldots, N$ are identical, we therefore have

$$\begin{aligned}
&\mathbb{E}[\hat{\mathbf{G}}(f_1, f_2)] - G(f_1, f_2) = \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\hat{f}(\mathbf{Z}), \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\
&= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] + o(1/k) \\
&+\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\left(g_2'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z})\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] + g_2''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z})\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]^2/2\right)\right]\left(\frac{1}{k}\right).
\end{aligned}$$

As a final step, we note that $\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] = f(\mathbf{Z}) + o(1)$. This implies that

$$\begin{aligned}
&\mathbb{E}[\hat{\mathbf{G}}(f_1, f_2)] - G(f_1, f_2) = \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\
&+\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\left(g_2'(f(\mathbf{Z}), \mathbf{Z})f(\mathbf{Z}) + g_2''(f(\mathbf{Z}), \mathbf{Z})f(\mathbf{Z})^2\right)\right]\left(\frac{1}{k}\right) + o(1/k).
\end{aligned}$$

□

D.3.2 Variance

Lemma D.4. *Under assumptions $\mathscr{A}.2$ and $\mathscr{A}.3$ listed above, the variance of the divergence estimator is given by*

$$\begin{aligned}
\mathbb{V}(\hat{\mathbf{G}}(f_1, f_2)) &= \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z})]\left(\frac{1}{N}\right) \\
&\quad +\mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g'(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z})\hat{\mathbf{f}}(\mathbf{Z})]\left(\frac{1}{M}\right) \\
&\quad +o\left(\frac{1}{M} + \frac{1}{N}\right).
\end{aligned}$$

Proof. Using Taylor series identically to the proof of lemma D.2, we can show that it suffices to consider the following leading terms

$$\begin{aligned}
&g_2(\hat{\mathbf{f}}_1(\mathbf{Z})/\hat{\mathbf{f}}_2(\mathbf{Z}), \mathbf{Z}) = g_2(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) \\
&+g_2'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z})\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]\left(\mathbf{e}_{1,2}(\mathbf{Z})\right) \\
&+H.O.T
\end{aligned} \tag{D.7}$$

and show that the H.O.T terms will contribute order $o(1/M)$ to the variance. Then,

$$\begin{aligned}
&\mathbb{V}[\hat{\mathbf{G}}(f_1, f_2)] = \mathbb{V}\left[\left(\frac{1}{N}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g_2(\hat{\mathbf{f}}_1(\mathbf{X}_i)/\hat{\mathbf{f}}_2(\mathbf{X}_i), \mathbf{X}_i)\right)\right] \\
&= \mathbb{V}\left[\left(\sum_{i=1}^{N} \frac{1_{\{\mathbf{X}_i \in \mathcal{S}'\}}}{N}\left(g_2(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)], \mathbf{X}_i) + g_2'(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)], \mathbf{X}_i)\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)]\mathbf{e}_{1,2}(\mathbf{X}_i)\right)\right)\right] \\
&+ o(1/M) \\
&= \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}} g(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z})]\left(\frac{1}{N}\right) + \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}} g'(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z})\hat{\mathbf{f}}(\mathbf{Z})]\left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right).
\end{aligned}$$

□

D.4 MI estimators

The MI estimators we have defined are of the general form

$$\hat{\mathbf{G}}(f_{12}) \;=\; \left(\frac{1}{N}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}_1(\mathbf{X}_i)\hat{\mathbf{f}}_2(\mathbf{Y}_i)/\hat{\mathbf{f}}_{12}(\mathbf{Z}_i), \mathbf{Z}_i)\right). \tag{D.8}$$

where the set $\mathcal{S}'$ is arbitrary and $\hat{f}_a(.)$ $(a = 1, 2, 12)$ can be any one of (i) standard k-NN density estimates $\hat{f}_k(.)$, (ii) boundary corrected density estimates $\tilde{f}_k(.)$ or (iii) angular weighted boundary corrected density estimates $\hat{f}_{k,K}(.)$.

Let $\mathbf{Z}$ be a random variable with density f_{12}. Denote the conditional expected value $E[\hat{\mathbf{f}}_a(\mathbf{Z})|\mathbf{Z}]$ by $\mathbb{E}_Z[\hat{\mathbf{f}}_a(\mathbf{Z})]$ and define $\hat{\mathbf{f}}_a(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_a(\mathbf{Z})]$ by $\mathbf{e}(\mathbf{Z})$. Also define $\hat{\mathbf{f}} := \hat{\mathbf{f}}_1\hat{\mathbf{f}}_2/\hat{\mathbf{f}}_{12}$, $f = f_1 f_2/f_{12}$ and $\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] = \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{X})]\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Y})]/\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{12}(\mathbf{Z})]$. Also define $\hat{\mathbf{f}}_1(\mathbf{X}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{X})]$ by $\mathbf{e}_1(\mathbf{X})$, $\hat{\mathbf{f}}_2(\mathbf{Y}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Y})]$ by $\mathbf{e}_2(\mathbf{Y})$ and $\hat{\mathbf{f}}_2(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{12}(\mathbf{Z})]$ by $\mathbf{e}_{12}(\mathbf{Z})$. Finally, define

$$\mathbf{e}_{1,2,12}(\mathbf{Z}) = \frac{\mathbf{e}_1(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{X})]} + \frac{\mathbf{e}_2(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Y})]} - \frac{\mathbf{e}_{12}(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{12}(\mathbf{Z})]}.$$

Using Taylor series expansion, we can write

$$\frac{\hat{\mathbf{f}}_1(x)\hat{\mathbf{f}}_2(y)}{\hat{\mathbf{f}_{12}}(z)} = \frac{f_1(x)f_2(y)}{f_{12}(z)} \times$$
$$\left(1 + \frac{\hat{\mathbf{f}}_1(x) - f_1(x)}{f_1(x)} + \frac{\hat{\mathbf{f}}_2(y) - f_2(y)}{f_2(y)} - \frac{\hat{\mathbf{f}_{12}}(z) - f_{12}(z)}{f_{12}(z)} + \frac{(\hat{\mathbf{f}_{12}}(z) - f_{12}(z))^2}{f_{12}^2(z)}\right)$$
$$+O((\hat{\mathbf{f}}_2(x) - f_2(x))^3).$$

In this section, we will assume that the density estimate satisfies assumptions 1,2 and 3. In addition, we assume that the following marginal-joint cross moment conditions hold:

$\mathscr{A}.4$ Cross marginal-joint moments:

$$(a)Cov\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_1(\mathbf{Z})\mathbf{e}_{12k}^q(\mathbf{Z}), 1_{\{\mathbf{Z}\in\mathcal{S}'\}}\gamma_2(\mathbf{Z})\mathbf{e}_{1k}^r(\mathbf{Z})\right] = o(1/k).$$

$$(b)Cov\left[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)\mathbf{e}_{12k}^q(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)\mathbf{e}_{1k}^r(\mathbf{Z}_2)\right]$$
$$= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{Z}_1\in\mathcal{S}'\}}\gamma_1(\mathbf{Z}_1)f_{12}(\mathbf{Z}_1), 1_{\{\mathbf{Z}_2\in\mathcal{S}'\}}\gamma_2(\mathbf{Z}_2)f_1(\mathbf{Z}_2)]\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right).$$

D.4.1 Bias

Lemma D.5. *If assumptions $\mathscr{A}.1$, $\mathscr{A}.2$ and $\mathscr{A}.4(a)$ are satisfied, the bias of the MI estimator is given by*

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{G}}(f_{12})] - G(f_{12}) &= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}(g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] + o(1/k) \\ &\quad + \mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}'\}}\left(g'(f(\mathbf{Z}), \mathbf{Z})f(\mathbf{Z}) + g''(f(\mathbf{Z}), \mathbf{Z})f(\mathbf{Z})^2\right)\right]\left(\frac{1}{k}\right).\end{aligned}$$

Proof. Using the continuity of $g'''(x, y)$, construct the following third order Taylor

series of $g(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z})$ around the conditional expected value $\mathbb{E}_Z[\hat{\mathbf{f}}(\mathbf{Z})]$.

$$
\begin{aligned}
& g(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z}) = g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) \\
& + g'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] \left(\frac{\mathbf{e}_1(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{X})]} + \frac{\mathbf{e}_2(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Y})]} - \frac{\mathbf{e}_{12}(\mathbf{Z})}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{12}(\mathbf{Z})]} + \frac{\mathbf{e}_{12}^2(\mathbf{Z})}{(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{12}(\mathbf{Z})])^2} \right) \\
& + g''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) \frac{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]^2}{2} \left(\frac{\mathbf{e}_1^2(\mathbf{Z})}{(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_1(\mathbf{X})])^2} + \frac{\mathbf{e}_2^2(\mathbf{Z})}{(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Y})])^2} + \frac{\mathbf{e}_{12}^2(\mathbf{Z})}{(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_{12}(\mathbf{Z})])^2} \right) \\
& + o(\mathbf{e}_1^2(\mathbf{Z}) + \mathbf{e}_2^2(\mathbf{Z}) + \mathbf{e}_{12}^2(\mathbf{Z})).
\end{aligned}
$$

This gives

$$
\begin{aligned}
& \mathbb{E}[1_{\{\mathbf{Z} \in \mathcal{S}'\}} (g(\hat{\mathbf{f}}(\mathbf{Z}), \mathbf{Z}) - g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}))] \\
& = \mathbb{E}\left[1_{\{\mathbf{Z} \in \mathcal{S}'\}} \left(g'(f(\mathbf{Z}), \mathbf{Z}) f(\mathbf{Z}) + g''(f(\mathbf{Z}), \mathbf{Z}) f(\mathbf{Z})^2 \right) \right] \left(\frac{1}{k} \right) \\
& + \mathbb{E}\big[o(\mathbf{e}_1^2(\mathbf{Z}) + \mathbf{e}_2^2(\mathbf{Z}) + \mathbf{e}_{12}^2(\mathbf{Z})) \big]
\end{aligned}
$$

Using Cauchy-Schwarz inequality as in the proof of Lemma D.1, we can show that $\mathbb{E}[o(\mathbf{e}_1^2(\mathbf{Z}) + \mathbf{e}_2^2(\mathbf{Z}) + \mathbf{e}_{12}^2(\mathbf{Z}))] = o(1/k)$. Once again, by observing that the samples $\{\mathbf{X}_i\}, i = 1, \ldots, N$ are identical, we therefore have

$$
\begin{aligned}
& \mathbb{E}[\hat{\mathbf{G}}(f_1, f_2)] - G(f) = \mathbb{E}[1_{\{\mathbf{Z} \in \mathcal{S}'\}} (g(\hat{f}(\mathbf{Z}), \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] \\
& = \mathbb{E}[1_{\{\mathbf{Z} \in \mathcal{S}'\}} (g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] + o(1/k) \\
& + \mathbb{E}\left[1_{\{\mathbf{Z} \in \mathcal{S}'\}} \left(\frac{g_2'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]}{\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Y})]^2} + \frac{g_2''(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]^2}{2 \mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}_2(\mathbf{Y})]^2} \right) \mathbf{e}^2(\mathbf{Z}) \right].
\end{aligned}
$$

As a final step, we note that $\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})] = f(\mathbf{Z}) + o(1)$. This implies that

$$
\begin{aligned}
\mathbb{E}[\hat{\mathbf{G}}(f_{12})] - G(f_{12}) \quad = \quad & \mathbb{E}[1_{\{\mathbf{Z} \in \mathcal{S}'\}} (g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z}))] + o(1/k) \\
& + \mathbb{E}\left[1_{\{\mathbf{Z} \in \mathcal{S}'\}} \left(g'(f(\mathbf{Z}), \mathbf{Z}) f(\mathbf{Z}) + g''(f(\mathbf{Z}), \mathbf{Z}) f(\mathbf{Z})^2 \right) \right] \left(\frac{1}{k} \right).
\end{aligned}
$$

□

D.4.2 Variance

Lemma D.6. *If assumptions $\mathscr{A}.1$, $\mathscr{A}.2$ and $\mathscr{A}.4(a)$ are satisfied, the variance of the MI estimator is given by*

$$\begin{aligned}\mathbb{V}(\hat{\mathbf{G}}(f_{12})) &= \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})]\left(\frac{1}{N}\right)\\ &\quad+\mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g'(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})\hat{\mathbf{f}}(\mathbf{Z})]\left(\frac{1}{M}\right)\\ &\quad+o\left(\frac{1}{M}+\frac{1}{N}\right).\end{aligned}$$

Proof. Using Taylor series identically to the proof of lemma D.2, we can show that it suffices to consider the following leading terms

$$\begin{aligned}&g(\hat{\mathbf{f}}_1(\mathbf{Z})\hat{\mathbf{f}}_2(\mathbf{Z})/\hat{\mathbf{f}_{12}}(\mathbf{Z}),\mathbf{Z}) = g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z})\\ &+g'(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})],\mathbf{Z})\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})]\left(\mathbf{e}_{1,2,12}(\mathbf{Z})\right)\\ &+H.O.T\end{aligned} \tag{D.9}$$

and show that the H.O.T terms will contribute order $o(1/M)$ to the variance. Then,

$$\begin{aligned}&\mathbb{V}[\hat{\mathbf{G}}(f_{12})] = \mathbb{V}\left[\left(\frac{1}{N}\sum_{i=1}^{N}1_{\{\mathbf{X}_i\in\mathcal{S}'\}}g(\hat{\mathbf{f}}_1(\mathbf{X}_i)/\hat{\mathbf{f}}_2(\mathbf{X}_i),\mathbf{X}_i)\right)\right]\\ &= \mathbb{V}\left[\left(\sum_{i=1}^{N}\frac{1_{\{\mathbf{X}_i\in\mathcal{S}'\}}}{N}\left(g(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)],\mathbf{X}_i)+g'(\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)],\mathbf{X}_i)\mathbb{E}_{\mathbf{X}_i}[\hat{\mathbf{f}}(\mathbf{X}_i)\mathbf{e}_{1,2,12}(\mathbf{X}_i)\right)\right)\right]\\ &+o(1/M)\\ &= \mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})]\left(\frac{1}{N}\right)+\mathbb{V}[1_{\{\mathbf{Z}\in\mathcal{S}'\}}g'(\hat{\mathbf{f}}(\mathbf{Z}),\mathbf{Z})\hat{\mathbf{f}}(\mathbf{Z})]\left(\frac{1}{M}\right)+o\left(\frac{1}{M}+\frac{1}{N}\right).\end{aligned}$$

□

APPENDIX E

General result on CLT for interchangeable processes

E.1 CLT for Interchangeable Processes

Let $\{\mathbf{Z}_i; i = 1, 2, \ldots\}$ be an interchangeable stochastic process with 0 mean and variance 1. Blum et.al.[11] showed that the random variable $\mathbf{S}_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Z}_i$ converges in distribution to $N(0, 1)$ if and only if $Cov(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $Cov(\mathbf{Z}_1^2, \mathbf{Z}_2^2) = 0$.

E.1.1 De Finetti's Theorem

Let $\mathcal{F}$ be the class of one dimensional distribution functions and for each pair of real numbers x and y define $\mathcal{F}(x, y) = \{F \in \mathcal{F} | F(x) \leq y\}$. Let $\mathcal{B}$ 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 $\mathcal{B}$ such that

$$Pr\{\mathbf{B}\} = \int_{\mathcal{F}} Pr_F\{\mathbf{B}\} d\mu(F), \tag{E.1}$$

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.1.2 Necessary and Sufficient conditions for CLT

For each $F \in \mathcal{F}$ define $m(F)$ and $\sigma^2(F)$ as $m(F) = \int_{-\infty}^{\infty} x dF(x)$, $\sigma(F) = \int_{-\infty}^{\infty} x^2 dF(x) - 1$ and for all real numbers m and non-negative real numbers σ^2 let $\mathcal{F}_{m,\sigma^2}$ be the set of $F \in \mathcal{F}$ 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 $\mu(\mathcal{F}_{0,0}) = 1$. Furthermore, they show that the condition $\mu(\mathcal{F}_{0,0}) = 1$ is equivalent to the condition that $Cov(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $Cov(\mathbf{Z}_1^2, \mathbf{Z}_2^2) = 0$.

E.2 CLT for Asymptotically Uncorrelated processes

In this section, we establish the CLT for this type of asymptotically uncorrelated interchangeable processes. Define the sum $\mathbf{S_{N,M}}$

$$\mathbf{S}_{N,M} = \frac{\sum_{i=1}^{N} \mathbf{Y}_{M,i}}{\sqrt{\mathbb{V}[\sum_{i=1}^{N} \mathbf{Y}_{M,i}]}},$$

where the indices N and M explicitly stress the dependence of the sum $\mathbf{S}_{N,M}$ on the number of random variables $N + M$.

Lemma E.1. *Assume that the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ belong to an 0 mean, unit variance, interchangeable process [11] for all values of M. Further assume that $Cov(\mathbf{Y}_{M,1}, \mathbf{Y}_{M,2})$ and $Cov(\mathbf{Y}_{M,1}^2, \mathbf{Y}_{M,2}^2)$ are $O(1/M)$.*

Then, the random variables $\mathbf{S}_{N,M}$ converges in distribution to $N(0, 1)$.

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

such that $\delta_\mu(M) = o(1); \frac{1}{M\delta_\mu(M)} = o(1)$, $\delta_\sigma(M) = o(1); \frac{1}{M\delta_\sigma(M)} = o(1)$. Denote the set of $F \in \mathcal{F}$ with $\mathcal{F}_{m,\delta,M} := \{m^2(F) \geq \delta_\mu(M)\}$; $\mathcal{F}_{\sigma,\delta,M} := \{\sigma^2(F) \geq \delta_\sigma(M)\}$; $\mathcal{F}^*_{m,\delta,M} := \{m^2(F) \in (0, \delta_\mu(M))\}$ and $\mathcal{F}^*_{\sigma,\delta,M} := \{\sigma^2(F) \in (0, \delta_\sigma(M))\}$. Denote the measures of these sets by $\mu_{m,\delta,M}$, $\mu_{\sigma,\delta,M}$, $\mu^*_{m,\delta,M}$ and $\mu^*_{\sigma,\delta,M}$ respectively. We have from (E.1) that

$$\begin{aligned} \int_{\mathcal{F}} m^2(F) d\mu(F) &= Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) \\ \int_{\mathcal{F}} \sigma^2(F) d\mu(F) &= \int_{\mathcal{F}} [\mathbb{E}_F[\mathbf{Z}^2 - 1]]^2 d\mu(F) = Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}). \end{aligned} \tag{E.2}$$

Applying the Chebyshev inequality, we get

$$\begin{aligned} \delta_\mu(M)\mu_{m,\delta,M} &\leq Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}), \\ \delta_\sigma(M)\mu_{\sigma,\delta,M} &\leq Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}). \end{aligned}$$

Because the covariances decay at $O(1/M)$, $\mu_{m,\delta,M}$ and $\mu_{\sigma,\delta,M} \to 0$ as $M \to \infty$. From the definition of $\mathcal{F}^*_{m,\delta,M}$ and $\mathcal{F}^*_{\sigma,\delta,M}$, we also have that $\mu^*_{m,\delta,M}$ and $\mu^*_{\sigma,\delta,M} \to 0$ as $M \to \infty$. We also have

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

and therefore

$$\lim_{M\to\infty} \mu(\mathcal{F}_{0,0}) = 1. \tag{E.3}$$

Observe that

$$\lim_{\Delta\to 0} Pr\{\mathbf{S}_{N,M} \leq \alpha\} = \lim_{\Delta\to 0} \int_{\mathcal{F}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F)$$

$$= \lim_{\Delta\to 0} \int_{\mathcal{F}_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) + \lim_{\Delta\to 0} \int_{\mathcal{F}} 1_{\{F\in\mathcal{F}-\mathcal{F}_{0,0}\}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F)$$

$$= \lim_{\Delta\to 0} \int_{\mathcal{F}_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) + \int_{\mathcal{F}} \lim_{\Delta\to 0} \left(1_{\{F\in\mathcal{F}-\mathcal{F}_{0,0}\}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\}\right) d\mu(F) \quad \text{(E.4)}$$

$$= \lim_{\Delta\to 0} \int_{\mathcal{F}_{0,0}} Pr_F\{\mathbf{S}_{N,M} \leq \alpha\} d\mu(F) \quad \text{(E.5)}$$

$$= \lim_{\Delta\to 0} \int_{\mathcal{F}_{0,0}} Pr_F \left\{ \sum_{i=1}^{N} \left(\frac{\mathbf{Y}_{M,i}}{\sqrt{\mathbb{V}[\sum_{i=1}^{N} \mathbf{Y}_{M,i}]}} \right) \leq \alpha \right\} d\mu(F)$$

$$= \lim_{\Delta\to 0} \int_{\mathcal{F}_{0,0}} Pr_F \left\{ \sum_{i=1}^{N} \left(\frac{\mathbf{Y}_{M,i}}{\sqrt{N\mathbb{V}[\mathbf{Y}_{M,i}] + N(N-1)Cov[\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}]}} \right) \leq \alpha \right\} \int_{\mathcal{F}_{0,0}} d\mu(F)$$

$$= \lim_{\Delta\to 0} \int_{\mathcal{F}_{0,0}} Pr_F \left\{ \sum_{i=1}^{N} \left(\frac{\mathbf{Y}_{M,i}}{\sqrt{N\mathbb{V}[\mathbf{Y}_{M,i}]}} \right) \leq \alpha \right\} \int_{\mathcal{F}_{0,0}} d\mu(F) \quad \text{(E.6)}$$

$$= \lim_{\Delta\to 0} \int_{\mathcal{F}_{0,0}} Pr_F \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Y}_{M,i} \leq \alpha \right\} d\mu(F)$$

$$= \int_{\mathcal{F}} \lim_{\Delta\to 0} \left(1_{\{F\in\mathcal{F}_{0,0}\}} Pr_F \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Y}_{M,i} \leq \alpha \right\} \right) d\mu(F)$$

$$= \int_{\mathcal{F}} \phi(\alpha) d\mu(F) = \phi(\alpha), \quad \text{(E.7)}$$

where $\phi(.)$ is the distribution function of a Gaussian random variable with mean 0 and variance 1. Step (E.4) follows from the Dominated Convergence theorem. By (E.3), $\lim_{\Delta\to 0} 1_{\{F\in\mathcal{F}-\mathcal{F}_{0,0}\}} = 0$ almost surely. This gives Step (E.5). Step (E.6) is obtained by observing that, by (E.2), $Cov[\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}] = 0$ when $F \in \mathcal{F}_{0,0}$. The last step (E.7) follows from the CLT for sums of 0 mean, unit variance, i.i.d random variables and (E.3). This concludes the proof of Theorem E.1. □

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Ahmad, I., and Pi-Erh Lin (1976), A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.), *Information Theory, IEEE Transactions on*, *22*(3), 372 – 375, doi:10.1109/TIT.1976.1055550.

[2] Ali, S. M., and S. D. Silvey (1966), A general class of coefficients of divergence of one distribution from another, *Journal of the Royal Statistical Society B*, pp. 131–142.

[3] Asuncion, A., and D. Newman (2007), UCI machine learning repository.

[4] AVIRIS data (1999), Available online at http://aviris.jpl.nasa.gov/html/data.html.

[5] Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2003), Convexity, classification, and risk bounds, *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*.

[6] Baryshnikov, Y., M. D. Penrose, and J. Yukich (2009), Gaussian limits for generalized spacings, *Ann. Appl. Probab.*, *19*(1), 158–185.

[7] Bay, S. D., and M. Schwabacher (2003), Mining distance-based outliers in near linear time with randomization and a simple pruning rule, in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pp. 29–38, ACM, New York, NY, USA, doi:http://doi.acm.org/10.1145/956750.956758.

[8] Biau, G., F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodríguez (2011), A weighted k-nearest neighbor density estimate for geometric inference, *Electron. J. Stat.*, *5*, 204–237, doi:10.1214/11-EJS606.

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

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

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

[12] B.Poczos, and J. Schneider (2011), On the estimation of α-divergences, in *AISTATS 2011*.

[13] Breunig, M. M., H. Kriegel, R. T. Ng, and J. Sander (2000), Lof: identifying density-based local outliers, in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pp. 93–104, ACM, New York, NY, USA, doi:http://doi.acm.org/10.1145/342009.335388.

[14] Carter, K., R. Raich, and A. Hero (2010), On local intrinsic dimension estimation and its applications, *Signal Processing, IEEE Transactions on*, *58*(2), 650 –663, doi:10.1109/TSP.2009.2031722.

[15] Chatterjee, S. (2008), A new method of normal approximation, *ANNALS OF PROBABILITY*, *36*, 1584.

[16] Chaudhuri, B. B., and N. Sarkar (1995), Texture segmentation using fractal dimension, *IEEE Trans. Pattern Anal. Mach. Intell.*, *17*, 72–77, doi:10.1109/34.368149.

[17] Chen, Y., 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.

[18] Cheng, R. C. H., and N. A. K. Amin (1983), Estimating parameters in continuous univariate distributions with a shifted origin., *Journal of the Royal Statistical Society. Series B (Methodological)*, *11*, 394–403.

[19] Chitode, J. S. (2009), *Digital Communications*, Technical Publications.

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

[21] Cover, T., and J. Thomas (2006), *Elements of information theory*, Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience.

[22] Csiszar, I. (1967), Information-type measures of difference of probability distributions and indirect observations, *Studia Sci. Math. Hungarica*, pp. 299–318.

[23] Donoho, D. L. (2004), For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution, *Manuscript, Department of Statistics, Stanford University*.

[24] Dudewicz, E. J., and E. C. van der Meulen (1981), Entropy-based tests of uniformity., *Journal of the American Statistical Association*, *76*, 967–974.

[25] Eggermont, P. B., and V. N. LaRiccia (1999), Best asymptotic normality of the kernel density entropy estimator for smooth densities, *Information Theory, IEEE Transactions on*, *45*(4), 1321 –1326, doi:10.1109/18.761291.

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

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

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

[29] Fodor, I. (2002), A survey of dimension reduction techniques, *Tech. rep.*, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory.

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

[31] Giné, E., and D. M. Mason (2008), Uniform in bandwidth estimation of integral functionals of the density function, *Scandinavian Journal of Statistics*, *35*(4), 739–761.

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

[33] Gupta, R. (2001), Quantization strategies for low-power communications, Ph.D. thesis, University of Michigan, Ann Arbor.

[34] Hall, P., and J. S. Marron (1987), Estimation of integrated squared density derivatives, *Stat. Prob. Lett*, pp. 109–115.

[35] Hartigan, J. (1975), *Clustering algorithms*, xiii+351 pp., John Wiley & Sons, New York-London-Sydney, wiley Series in Probability and Mathematical Statistics.

[36] Hero, A. O. (2006), Geometric entropy minimization (gem) for anomaly detection and localization, in *Proc. Advances in Neural Information Processing Systems (NIPS*, pp. 585–592, MIT Press.

[37] Hero, A. O., B. Ma, O. Michel, and J. Gorman (2002), Applications of entropic spanning graphs, *Signal Processing Magazine, IEEE*, *19*(5), 85 – 95.

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

[39] Hulle, M. M. V. (2005), Edgeworth approximation of multivariate differential entropy, *Neural Computation*, *17*(9), 1903–1910.

[40] Jain, A. (1981), Image data compression: A review, *Proceedings of the IEEE*, *69*(3), 349 – 389, doi:10.1109/PROC.1981.11971.

[41] Jain, A. K., and R. C. Dubes (1988), *Algorithms for clustering data*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[42] Jones, M. C. (1993), Simple boundary correction for kernel density estimation, *Statistics and Computing*, *3*, 135–146.

[43] Kailath, T. (1967), The divergence and bhattacharyya distance measures in signal selection, *IEEE Transactions on Communications*, *15*, 52–60, doi:10.1109/TCOM.1967.1089532.

[44] Karunamuni, R., and T. Alberts (2005), On boundary correction in kernel density estimation, *Statistical Methodology*, *2*(3), 191 – 212, doi:DOI:10.1016/j.stamet.2005.04.001.

[45] Lakhina, A., M. Crovella, and C. Diot (2005), Mining anomalies using traffic feature distributions, in *In ACM SIGCOMM*, pp. 217–228.

[46] Lanckriet, G., N. Cristianini, P. Bartlett, and L. E. Ghaoui (2002), Learning the kernel matrix with semi-definite programming, *Journal of Machine Learning Research*, *5*, 2004.

[47] Laurent, B. (1996), Efficient estimation of integral functionals of a density, *The Annals of Statistics*, *24*(2), 659–681.

[48] Lee, J. (1997), *Riemannian manifolds: an introduction to curvature*, Springer.

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

[50] Levina, E., and P. Bickel (2004), Maximum likelihood estimation of intrinsic dimension, *Advances in Neural Information Processing Systems 17*, *48109*(C), 777–784.

[51] Liitiäinen, E., A. Lendasse, and F. Corona (2009), On the statistical estimation of rényi entropies, in *Proceedings of IEEE/MLSP 2009 International Workshop on Machine Learning for Signal Processing, Grenoble (France)*.

[52] Liu, F. T., K. M. Ting, and Z. Zhou (2008), Isolation forest, in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, IEEE Computer Society, Washington, DC, USA, doi:10.1109/ICDM.2008.17.

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

[54] M. Bernstein, J. C. l., V. de Silva, and J. B. Tanenbaum (2000), Graph approximations to geodesics on embedded manifolds, *Manuscript, Department of Statistics, Stanford University*.

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

[56] Miller, E. G., and J. W. Fisher III (2003), ICA using spacings estimates of entropy, *Proc. 4th Intl. Symp. on ICA and BSS*, pp. 1047–1052.

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

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

[59] Nguyen, X., M. J. Wainwright, and M. I. Jordan (2010), Estimating divergence functionals and the likelihood ratio by convex risk minimization, *Information Theory, IEEE Transactions on*, *56*(11), 5847 –5861, doi:10.1109/TIT.2010.2068870.

[60] Pál, D., B. Póczos, and C. Szepesvári (2010), Estimation of R\'enyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs, *ArXiv e-prints*.

[61] Park, C., J. Z. Huang, and Y. Ding (2010), A computable plug-in estimator of minimum volume sets for novelty detection, *Operations Research*, *58*(5), 1469–1480.

[62] Penrose, M. D. (1999), A strong law for the largest nearest-neighbour link between random points, *Journal of the London Mathematical Society*, *60*(3), 951–960, doi:10.1112/S0024610799008157.

[63] Pentland, A. P. (1984), Fractal-Based Description of Natural Scenes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 661–674, doi:10.1109/TPAMI.1984.4767591.

[64] Perone Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman (2004), False discovery control for random fields, *J. Amer. Statist. Assoc.*, *99*(468), 1002–1014, doi:10.1198/0162145000001655.

[65] Phothisonothai, M., and M. Nakagawa (2010), EEG-Based Classification of Motor Imagery Tasks Using Fractal Dimension and Neural Network for Brain-Computer Interface, *IEICE Transactions on Information and Systems*, *91*, 44–53, doi:10.1093/ietisy/e91-d.1.44.

[66] Ramaswamy, S., R. Rastogi, and K. Shim (2000), Efficient algorithms for mining outliers from large data sets, *SIGMOD Rec.*, *29*, 427–438, doi:http://doi.acm.org/10.1145/335191.335437.

[67] Ranneby, B. (1984), The maximum spacing method. an estimation method related to the maximum likelihood method., *Scandinavian Journal of Statistics*, *11*, 93–112.

[68] Rao, A., A. O. Hero, D. J. States, and J. D. Engel (2008), Using directed information for influence discovery in interconnected dynamical systems, in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 7074, doi:10.1117/12.801360.

[69] Raykar, V. C., and R. Duraiswami (2006), Fast optimal bandwidth selection for kernel density estimation, in *Proceedings of the sixth SIAM International Conference on Data Mining*, edited by J. Ghosh, D. Lambert, D. Skillicorn, and J. Srivastava, pp. 524–528.

[70] Raymond, X. S. (1991), *Elementary Introduction to the Theory of Pseudodifferential Operators*, CRC Press.

[71] Rocke, D. M., and D. L. Woodruff (1996), Identification of Outliers in Multivariate Data, *Journal of the American Statistical Association*, *91*(435), 1047–1061.

[72] Schapire, R. E. (1990), The strength of weak learnability, *Machine Learning*, *5*(2), 197–227–227, doi:10.1007/BF00116037.

[73] Schölkopf, B., R. Williamson, A. Smola, J. Shawe-Taylor, and J.Platt (2000), Support vector method for novelty detection.

[74] Scott, C., and R. Nowak (2006), Learning minimum volume sets, *J. Mach. Learn. Res.*, *7*, 665–704.

[75] Scott, C., and R. Nowak (2006), Learning minimum volume sets, *J. Machine Learning Res*, *7*, 665–704.

[76] Singh, A., C. Scott, and R. Nowak (2009), Adaptive Hausdorff estimation of density level sets, *Ann. Statist.*, *37*(5B), 2760–2782, doi:10.1214/08-AOS661.

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

[78] Solé, A., V. Caselles, G. Sapiro, and F. Arándiga (2004), Morse description and geometric encoding of digital elevation maps, *IEEE Trans. Image Process.*, *13*(9), 1245–1262, doi:10.1109/TIP.2004.832864.

[79] Sricharan, K., and A. O. Hero III (2012), Ensemble estimators for efficient estimation, *Technical Report, Communications and Signal Processing Laboratory, The University of Michigan*.

[80] Sricharan, K., R. Raich, and A. O. Hero (2010), Empirical estimation of entropy functionals with confidence, *ArXiv e-prints*.

[81] Steinwart, I., D. Hush, and C. Scovel (2005), A classification framework for anomaly detection, *J. Mach. Learn. Res.*, *6*, 211–232 (electronic).

[82] Stuetzle, W. (2003), Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample, *J. Classification*, *20*(1), 25–47, doi:10.1007/s00357-003-0004-6.

[83] Ting, K. M., G. Zhou, T. F. Liu, and J. S. C. Tan (2010), Mass estimation and its applications, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pp. 989–998, ACM, New York, NY, USA, doi:http://doi.acm.org/10.1145/1835804.1835929.

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

[85] Vasicek, O. (1976), A test for normality based on sample entropy., *Journal of the Royal Statistical Society. Series B (Methodological)*, *38*, 54–59.

[86] Vert, R., and J.-P. Vert (2006), Consistency and convergence rates of one-class SVMs and related algorithms, *J. Mach. Learn. Res.*, *7*, 817–854.

[87] Wang, Q., S. R. Kulkarni, and S. Verdú (2005), Divergence estimation of continuous distributions based on data-dependent partitions, *Information Theory, IEEE Transactions on*, *51*(9), 3064–3074.

[88] Yang, Y. H., M. J. Buckley, S. Dudoit, and T. P. Speed (2002), Comparison of methods for image analysis on cDNA microarray data, *J. Comput. Graph. Statist.*, *11*(1), 108–136, doi:10.1198/106186002317375640.

[89] Zhao, M., and V. Saligrama (2009), Anomaly detection with score functions based on nearest neighbor graphs, *Computing Research Repository*, *abs/0910.5461*.