Nonparametric Estimation of Distributional Functionals and Applications

by

Kevin R. Moon

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2016

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Associate Professor Rajesh Rao Nadakuditi
Associate Professor Long Nguyen
Associate Professor Clayton Scott

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Nonparametric Estimation of Distributional Functionals and Applications

by

Kevin R. Moon

Chair: Alfred O. Hero III

Distributional functionals are integrals of functionals of probability densities and include functionals such as information divergence, mutual information, and entropy. Distributional functionals have many applications in the fields of information theory, statistics, signal processing, and machine learning. Many existing nonparametric distributional functional estimators have either unknown convergence rates or are difficult to implement. In this thesis, we consider the problem of nonparametrically estimating functionals of distributions when only a finite population of independent and identically distributed samples are available from each of the unknown, smooth, d-dimensional distributions. We derive mean squared error (MSE) convergence rates for leave-one-out kernel density plug-in estimators and k-nearest neighbor estimators of these functionals. We then extend the theory of optimally weighted ensemble estimation to obtain estimators that achieve the parametric MSE convergence rate when the densities are sufficiently smooth. These estimators are simple to implement and do not require knowledge of the densities' support set, in contrast with many competing estimators. The asymptotic distribution of these estimators is also derived.

The utility of these estimators is demonstrated through their application to sunspot image data and neural data measured from epilepsy patients. Sunspot images are clustered by estimating the divergence between the underlying probability distributions of image pixel patches. The problem of overfitting is also addressed in both applications by performing dimensionality reduction via intrinsic dimension estimation and by benchmarking classification via Bayes error estimation.

CHAPTER I

Introduction

1.1 Background

The large growth of digital technology and devices in recent years has led to an explosion in the availability of signals and data. This increase in data corresponds to increased research opportunities and challenges in signal processing and machine learning. One such challenge is the increased risk of overfitting in machine learning problems due to increased dimensionality. Large dimensional data can provide a very high resolution description of the objects being analyzed. However, with this high resolution comes the "curse of dimensionality" which requires a rapidly increasing amount of data samples as dimension increases to analyze the data without overfitting [13, 86].

Two approaches to reducing the risk of overfitting are considered in this thesis. The first approach is to use estimated bounds on the Bayes error as a benchmark for classification. The second is to perform dimensionality reduction.

Another opportunity caused by the increase in data is that in a growing number of applications, an object may be represented by a collection of measurements [162, 182], e.g., patch features of an image represented as the collection of its 3×3 image patches [143]. In classical machine learning problems, an object is represented by a single feature vector. Some measure of difference or distance, such as the Euclidean distance, calculated between each pair of objects feature vectors might then be used as input into a machine learning algorithm such as classification, regression, or clustering.

In contrast, when an object is represented by a collection of measurements, the object is not well represented as a single vector. Simply combining all measurements into a single vector, e.g., by concatenation of the grey levels of all image patches, leads to a feature of overly high dimension whose dimension may vary if image size varies. Therefore, to compare these objects for a machine learning task, some measure of dissimilarity between collections of measurements is necessary [162].

We approach the problem of comparing collections of measurements by formulating the problem as comparing probability distributions. To measure the dissimilarity between collections of measurements, the vectors from a collection can be viewed as samples from some underlying distribution. The difference between the underlying distributions of two sample collections can be measured by a quantity known as the information divergence.

1.1.1 Nonparametric Estimation of Distributional Functionals

The common theme in these approaches is the estimation of information theoretic quantities and distributional functionals such as information divergence, mutual information, and entropy. These distributional functionals are integrals of functionals of the underlying densities of the data. However, the data densities are rarely known in practice.

A parametric approach can be used where the densities are fit to a parametric model such as a Gaussian distribution [97, 148]. The distributional functional is then calculated using the formula. This approach can be problematic for two reasons. First, for higher dimensions, evaluating the integral to obtain the distributional functional may require numerical integration which can be computationally intensive. Second, the parametric model may be a poor fit for the underlying distributions which would result in inaccurate estimates. This is especially likely for high dimensional data.

This thesis considers the problem of nonparametrically estimating functionals of distributions when only a finite population of independent and identically distributed (i.i.d.) samples are available from each of the unknown, smooth, d-dimensional distributions. Although the methods in this thesis can be applied to functionals of one or more distributions, we focus primarily on functionals of two distributions which we refer to as divergence functionals. Within the last few years, recent work has focused on defining nonparametric divergence functional estimators with known convergence rates [102, 111, 152, 180, 181]. However, these approaches are often either computationally difficult or require the use of an optimal kernel density estimator (KDE). These optimal KDEs require explicit knowledge of the support set of the densities and are difficult to construct when the support set contains boundaries. Furthermore, the asymptotic distributions of these divergence functional estimators is unknown for nearly all of them. Thus these estimators cannot be used to perform inference tasks on the divergence such as testing that two populations have identical distributions or constructing confidence intervals. See Section 1.2.1 for more details on these estimators.

In the context of this problem, we derive mean squared error (MSE) convergence rates for leave-one-out kernel density plug-in estimators of divergence functionals. We then extend the theory of optimally weighted ensemble entropy estimation developed in [187] to obtain two divergence functional estimators with a MSE convergence rate of $O\left(\frac{1}{T}\right)$, (the parametric rate) where T is the sample size, when the densities are sufficiently smooth. These estimators are simple to implement and do not require knowledge of the densities' support set. We then derive the asymptotic distribution of the weighted ensemble estimators which enables us to perform inference.

In addition to the kernel density plug-in estimators, we analyze the MSE convergence rates for k-nearest neighbor (nn) plug-in estimators of divergence functionals and apply the same optimally weighted ensemble estimation theory to derive k-nn estimators that achieve the parametric rate. The asymptotic distribution is similarly derived.

These nonparametric estimators of divergence functionals enable us to estimate bounds on the Bayes error for a classification problem and the divergence between the underlying distributions of two sample collections as input in machine learning problems. A similar entropy estimator can be used to estimate the intrinsic dimension of data which is useful for dimensionality reduction (see Section 1.1.3). We apply the estimators to sunspot image data and neural data in these contexts.

1.1.2 Bounds on the Bayes Error

In a classification problem, a common goal is to learn a classifier that minimizes the average probability of error for future samples. However, there exist many different classifiers of varying complexity and it is not known a priori which one will perform the best on a given data set. A common approach is to apply a large corpus of classifiers to the data and choose the classifier with the lowest test error. But this can be very computationally intensive, especially if some of the classifiers require the selection of tuning parameters. Additionally, many classifiers can potentially overfit the data, especially when the dimension is large, which will result in a poor generalization error [2, 86, 107]. These problems can be avoided by knowing the Bayes error.

The Bayes error is the lowest average probability of error that any classifier can achieve [86]. Thus it can serve as a benchmark for classification performance. For example, if we apply a simple linear classifier to a classification problem and obtain a test error that is significantly higher than the Bayes error, then we know that it is worth applying a more complicated classifier until the test error and the Bayes error are more closely aligned. On the other hand, if we apply a more complex classifier, such as a convolutional neural network, and we obtain a test error that is below the Bayes error, we know that we are overfitting and should either adjust tuning parameters or try a less complex classifier. Additionally, if we find that the Bayes error is very high for the given feature space, then this indicates that it may be better to use a different feature space.

The expression for the Bayes error is a distributional functional that depends on the underlying distributions of the data which are typically unknown in practice. Additionally, the functional of the distributions in the expression for the Bayes error is not differentiable everywhere which makes estimation more difficult. However, there are many bounds on the Bayes error that are related to divergences with smooth functionals [6, 15, 33, 85]. Therefore, we can estimate these bounds using a nonparametric estimator of distributional functions.

1.1.3 Dimensionality Reduction via Intrinsic Dimension Estimation

As mentioned previously, large dimensional data can lead to overfitting and poor performance in machine learning problems due to the curse of dimensionality. In practice data often lie on a lower dimensional manifold or subspace plus noise [40]. Dimensionality reduction techniques are therefore applied to mitigate the effects of the curse of dimensionality as well as denoise the data [12, 28, 86, 137, 203, 208].

Two key components of dimensionality reduction are choosing the size (i.e. dimension) and type (linear vs. nonlinear) of the lower dimensional manifold. These choices are often made with regards to heuristics and computational considerations instead of the natural geometry of the data. We estimate the intrinsic dimension as a measure of the natural geometry. The intrinsic dimension of the data is the dimension of the lower dimensional manifold on which the data lie. This can be used to choose the size of the lower dimensional manifold when performing dimensionality reduction. To determine the type of manifold, we estimate the intrinsic dimension using both a nonlinear and a linear method. The nonlinear method is based on a nonparametric k-nearest neighbor (nn) approach that is related to the entropy of the underlying data [31]. By comparing the estimates from the two approaches, we determine whether the lower dimensional subspace is linear or not.

1.1.4 Importance of Entropy and Divergence Measures in Other Applications

In addition to these problems, divergence estimation is useful for estimating the decay rates of error probabilities [41], testing the hypothesis that two sets of samples come from the same probability distribution [144], clustering [8, 45, 123], blind source separation [93, 136], image segmentation [82, 126, 195], and steganography [108]. For many more applications of divergence measures, see [11]. Although these applications are not explored in this thesis, our divergence functional estimators can be used for them as well.

As mentioned above, our methods of analysis can be easily extended to derive entropy functional estimators that achieve the parametric convergence rate as long as the density is sufficiently smooth. In addition to intrinsic dimension estimation [31, 40], these estimators can be used in applications such as texture classification and image registration [90], anomaly detection [185], goodness-of-fit testing [76], and many others.

While mutual information is a special case of divergence, applying our analysis methods to the problem of mutual information estimation requires a bit more care due to possible dependencies between samples. We extend the theory to derive nonparametric ensemble estimators of general mutual information measures that achieve the parametric rate. We consider two cases: 1) the data have purely continuous components; 2) the data have a mixture of continuous and discrete components. These estimators can then be used in applications such as determining channel capacity [41], feature selection [114, 159, 193, 196], fMRI data processing [32], independent subspace analysis [156], forest density estimation [127], clustering [123], neuron classification [175], and intrinsically motivated reinforcement learning [138, 172].

1.2 Related Work

1.2.1 Nonparametric Estimation of Divergence Functionals

Several nonparametric estimators for some functionals of two distributions including some divergences already exist. For example, *Póczos and Schneider* [161] established weak consistency of a bias-corrected k-nn estimator for Rényi-α and other divergences of similar form where k is fixed. *Wang et al.* [200] provided a k-nn based estimator for the Kullback-Leibler divergence. Mutual information and divergence estimators based on plug-in histogram schemes have been proven to be consistent [43, 118, 179, 199]. *Hero III et al.* [90] provided an estimator for Rényi-α divergence but assumed that one of the densities was known. However none of these works study the mean squared error convergence rates nor the asymptotic distribution of their estimators.

More recent work has focused on deriving convergence rates for divergence estimators. These estimators typically derive the rates in terms of a smoothness condition on the densities such as the Hölder condition which is a standard definition of smoothness:

Definition I.1 (Hölder Class)**.** Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space. For $r = (r_1, \dots, r_d)$, $r_i \in \mathbb{N}$, define $|r| = \sum_{i=1}^d r_i$ and $D^r = \frac{\partial^{|r|}}{\partial x_1^{r_1} \dots \partial x_d^{r_d}}$. The Hölder class $\Sigma(s, K)$ of functions on $L_2(\mathcal{X})$ consists of the functions f that satisfy

$$|D^r f(x) - D^r f(y)| \leq K \|x - y\|^{s - \lfloor s \rfloor},$$

for all x, $y \in \mathcal{X}$ and for all r s.t. $|r| \leq \lfloor s \rfloor$.

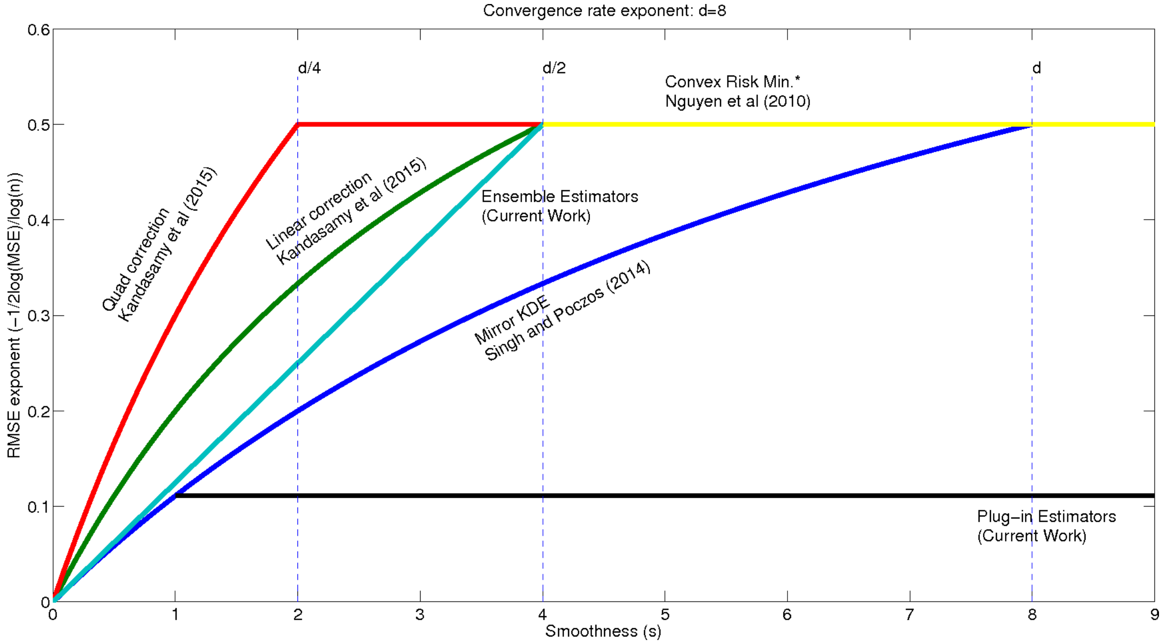

Figure 1.1: A comparison of the convergence rates of various divergence functional estimators (including this work) as a function of the smoothness of the densities when the dimension is $d = 8$. Smoothness is measured in terms of the Hölder condition given in Definition I.1. The convergence rate is given in terms of the negative root MSE exponent, i.e. $-\frac{1}{2}\log(\mathrm{MSE})/\log(N)$. Thus the maximum value that can be achieved (the parametric rate) is 1/2. Note that the estimators in *Krishnamurthy et al.* [111] achieve the same rates as those in *Kandasamy et al.* [102]. Also, for the estimator in *Nguyen et al.* [152], the smoothness condition is applied to the likelihood ratio instead of the densities.

From Definition I.1, it is clear that if a function f belongs to $\Sigma(s, K)$, then f is continuously differentiable up to order $\lfloor s \rfloor$. In this work, we propose estimators that achieve the parametric MSE convergence rate of $O(1/T)$ when $s \geq d$ and $s > \frac{d}{2}$, respectively. Figure 1.1 provides a comparison of the convergence rates for various divergence functional estimators as a function of the densities' smoothness s when $d = 8$. Specifically, our work is compared with the estimators in [102, 111, 152, 180, 181], all of which are discussed in the following.

Nguyen et al. [152] proposed a method for estimating f-divergences by estimating the likelihood ratio of the two densities by solving a convex optimization problem and then plugging it into the divergence formulas. For this method they prove that the minimax MSE convergence rate is parametric ($O\left(\frac{1}{T}\right)$ where T is the number of samples from each density) when the likelihood ratio is in the bounded Hölder class $\Sigma(s, K)$ with $s \geq d/2$. However, this estimator is restricted to true f-divergences instead of the broader class of divergence functionals. Additionally, solving the convex problem of [152] is similar in complexity to training the support vector machine (SVM) (between $O(T^2)$ and $O(T^3)$) which can be demanding when T is very large. In contrast, our method of optimally weighted ensemble estimation depends only on simple density plug-in estimates and an offline convex optimization problem. Thus the most computationally demanding step in our approach is the calculation of the density estimates which has complexity no greater than $O(T^2)$.

Singh and Póczos [180, 181] provided an estimator for Rényi-α divergences as well as general density functionals that uses a mirror image kernel density estimator. They prove a convergence rate of $O\left(\frac{1}{T}\right)$ when $s \geq d$ for each of the densities. However this method requires several computations at each boundary of the support of the densities which becomes difficult to implement as d gets large. Also, this method requires knowledge of the support of the densities which may not be possible for some problems. In contrast, while our assumptions require the density supports to be bounded, knowledge of the support is not required for implementation.

The linear and quadratic estimators given by *Krishnamurthy et al.* [111] estimate divergence functionals that include the form $\int f_1^\alpha(x) f_2^\beta(x) d\mu(x)$ for given α and β. These estimators achieve the parametric rate when $s \geq d/2$ and $s \geq d/4$ for the linear and quadratic estimators, respectively. However, the latter estimator is computationally infeasible in most cases and the former requires numerical integration for some divergence functionals, which can be computationally difficult. Additionally, while a suitable $\alpha - \beta$ indexed sequence of divergence functionals of this form can be made to converge to the KL divergence, this does not guarantee convergence of the corresponding sequence of divergence estimators in [111], whereas our estimator can be used to estimate the KL divergence. Other important f-divergence functionals are also excluded from this form including some that bound the Bayes error [6, 15, 139].

Kandasamy et al. [102] derived similar linear and quadratic estimators for more general divergence functionals by the use of influence functionals and both data-splitting and leave-one-out approaches. However, again the quadratic estimator is computationally infeasible and the linear estimator also requires numerical integration for some functionals. Furthermore, the estimators derived in [102, 111] require the use of an optimal KDE. If the support set of the densities is bounded (as is often assumed), an optimal KDE requires complicated techniques at the boundary of the support such as those used by [180, 181]. This is again in contrast with our work which requires no knowledge of the support.

Asymptotic normality has been established for certain appropriately normalized divergences between a specific density estimator and the true density [16, 17, 19]. However, this differs from our setting where we assume that both densities are unknown. Additionally, the asymptotic distributions of the estimators in [111, 152, 180, 181] are currently unknown. *Kandasamy et al.* [102] derived a central limit theorem for their data-splitting estimator but not their leave-one-out estimator.

Divergence functional estimation is also related to the problem of entropy functional estimation which has received a lot of attention. Some examples include [21, 73, 116] which used specialized kernel density estimators to achieve the parametric convergence rate when the density has smoothness parameter $s \geq d/4$. *Sricharan et al.* [187] derived an entropy functional estimator that uses a weighted average of an ensemble of simple estimators. While their approach requires the density to have smoothness parameter $s \geq d$ to achieve the parametric rate, their approach is simpler to implement compared to the other estimators [21, 73, 116]. Additionally, our work in this thesis can be applied to achieve a simple entropy estimator that only requires $s > d/2$ to achieve the parametric rate.

Many estimators for Shannon mutual information between continuous random variables have been developed. A popular k-nn-based estimator was proposed by*Kraskov et al.* [110] which is a modification of the entropy estimator derived by *Kozachenko and Leonenko* [109]. However, recent work has found that these estimators only achieve the parametric convergence rate when the dimension of each of the random variables is less than 3 [69]. Similarly, Pál et al's estimator of Rényi information [156] does not achieve the parametric rate. Other methods include estimators based on maximum likelihood estimation of the likelihood ratio [191] and minimal spanning trees [106].

Finally, *Gao et al.* [68] showed that k-nn or KDE based approaches underestimate the mutual information when the mutual information is large. As mutual information increases, the dependencies between random variables becomes more deterministic which results in less smooth densities. This is consistent with the work in [102, 111, 180, 181, 187] and this work which require the densities to be smooth to achieve the parametric rate.

1.2.2 Bayes Error Estimation

Accurately estimating the Bayes Error is not an easy task. It has been shown that without any assumptions on the distribution of the data, no convergence rate results can be obtained for an estimator of the Bayes error rate [5]. *Antos and Kontoyiannis* [4] showed similar results for additive functionals of a discrete distribution such as entropy and mutual information. Thus we can only obtain convergence rates for the Bayes error or its bounds for a subset of the set of all distributions on the data. Additionally, *Frigyik et al.* [61] found that the popular parametric approach of assuming the densities are Gaussian is not very robust and tends to underestimate the Bayes error. Thus to obtain our convergence rates, we use a nonparametric approach and assume that the data densities belong to $\Sigma(s, K)$ for $s \geq 2$.

Past work has attempted to estimate the Bayes error directly using nonparametric

k-nn or kernel density estimator approaches [48, 63, 65]. A Bayes error estimator can also be derived using the error rates of an ensemble of classifiers [194]. However the authors of these works did not derive the convergence rates for their estimators. *Fukunaga and Hummels* [64] derived the bias of the error of the finite sample 1-nn classifier with respect to the asymptotic error which converges to the Bayes error in probability [190]. They found that the bias converges very slowly to zero as a function of the number of samples when the dimension is high.

Multiple bounds on the Bayes error based on divergence functionals exist. Mutual information is related to the Bayes error [57, 88] and has been used widely as a proxy for the Bayes error in feature selection. Another bound based on the Henze-Penrose divergence can be consistently estimated by constructing a minimal spanning tree [15]. While this estimator performs well empirically, the convergence rates are currently unknown. Other bounds based on divergence functionals are discussed in Section 3.4.

1.2.3 Machine Learning on Distributional Features

Divergence measures have recently become more popular as measures of dissimilarity between objects modeled as probability distributions. *Dhillon et al.* [45] derived word clusters for dimensionality reduction in text classification based on the KL divergence between discrete word distributions. Parametric models have been used to embed distributions in a Hilbert space and then use kernel methods to solve a machine learning problem [99, 148].

More recently, nonparametric approaches have been used. In [182], collections of samples from distributions are compared using set kernels. *Muandet et al.* [149] extend the representer theorem to the space of probability distributions. *Póczos et al.* [162] estimated the Rényi and L_2 divergences to embed distributions, in image clustering and classification, and group anomaly detection. In particular, they find that the divergence-based approach to image classification outperforms other conventional approaches such as bag of words. Regression has also been applied to the cases where samples from probability distributions form the inputs [163] and possibly the outputs [155].

1.3 Thesis Contributions

1.3.1 Theoretical Work

In Chapter II, we present the analysis of leave-one-out KDE plug-in estimators of general divergence functionals. We derive expressions for the bias and the variance of these plug-in estimators without boundary correction when the support set of the densities is bounded. We generalize the theory of optimally weighted ensemble estimation derived in [187] to obtain two KDE divergence functional estimators that achieve the parametric MSE convergence rate when the densities have smoothness parameter $s \geq d$ and $s > d/2$ under different conditions on the functional. The estimators are computationally tractable as the weights are calculated via an offline convex optimization problem. We then derive the asymptotic distribution of the weighted ensemble estimators which enables us to construct confidence intervals and perform hypothesis testing.

A similar analysis of leave-one-out k-nn plug-in estimators of general divergence functionals is given in Chapter III. Expressions for the bias and variance of these k-nn plug-in estimators without boundary correction are derived. The generalized theory of optimally weighted ensemble estimation presented in Chapter II is applied to obtain two k-nn divergence functional estimators that achieve the parametric MSE convergence rate when the densities have smoothness parameter $s \geq d$ and $s > d/2$ under different conditions on the functional. The asymptotic distribution of these weighted ensemble estimators are also derived. These k-nn estimators are typically more computationally tractable than the KDE estimators in Chapter II as there exist many methods for computing the k-nearest neighbors that are computationally easier than calculating the KDE.

The analysis techniques used in Chapters II and III extend easily to the problem of estimating functionals of one (i.e. entropy functionals) or more distributions. Thus ensemble estimators for both KDE and k-nn plug-in estimators of entropy functionals (and functionals of 3 or more distributions) can be derived. However, extending these techniques to mutual information functionals requires a bit more care due to the possible dependencies between different samples. Under a similar setting, we extend the theory derived in Chapters II and III to provide nonparametric estimators general mutual information functionals under two cases: 1) the data have purely continuous components; 2) the data have a mixture of continuous and discrete components. To the best of our knowledge, our work is the first to derive MSE convergence rates for the latter case. The theory of optimally weighted ensemble estimation is applied to obtain estimators that achieve the parametric rate and the asymptotic distribution of these estimators is derived. This work is contained in Chapter IV.

1.3.2 Applications of Theory

The remaining chapters of this thesis are devoted to applications of our theoretical work described in Chapters II through IV. In Chapter V, we estimate the intrinsic dimension of sunspot image data using entropy-based estimators. We use the intrinsic dimension estimates to determine the size of a reduced dimension representation and to determine whether linear methods of dimensionality reduction are appropriate. The results of Chapter V are used in Chapter VI to reduce the dimension of the data and then cluster the sunspot images using divergence estimates as input to the clustering algorithm. Bounds on the Bayes error of a sunspot image classification problem are also estimated using our divergence functional estimators derived in the previous chapters.

We apply similar methods to high frequency oscillations (HFOs) measured from the brain in epilepsy patients. Typical analyses of HFOs have assumed that the data lie on a linear manifold that is global across time, channels, and patients. We estimated the intrinsic dimension of the data using entropy-based estimators to examine these assumptions and to aid in dimensionality reduction. We further estimate bounds on the Bayes error to quantify the distinction between two classes of HFOs (those occurring during seizures and those occurring due to other processes). This analysis provides the foundation for future clinical use of HFO features and guides the analysis for other discrete events such as individual action potentials or multi-unit activity.

1.3.3 Publications

Thereotical Work

1. K. Moon and A. Hero, "Ensemble estimation of multivariate f-divergence," in *IEEE Internatoinal Symposium on Information Theory* (ISIT), 2014 [140].

2. K. Moon and A. Hero, "Multivariate f-divergence estimation witn confidence," in *Advances in Neural Information Processing Systems* (NIPS), 2014 [141].

3. K. Moon, K. Sricharan, K. Greenewald, and A. Hero, "Improving convergence of divergence functional ensemble estimators," in *IEEE International Symposium on Information Theory* (ISIT), 2016 [146].

4. K. Moon, K. Sricharan, K. Greenewald, and A. Hero, "Nonparametric ensemble estimation of distributional functionals," submitted to *IEEE Transactions on Information Theory*, March 2016 [145].

5. K. Moon, K. Sricharan, and A. Hero, "Ensemble Estimation of Mutual Information," submitted to *Advances in Neural Information Processing Systems*

(NIPS), 2016.

6. K. Moon, K. Sricharan, and A. Hero, "Nearest neighbor ensemble estimation of distributional functionals," in preparation for submission to *IEEE Transactions on Information Theory.*

7. K. Moon, M. Noushad, S. Sekeh, and A. Hero, "Nonparametric mutual information measures," in preparation for submission to ICASSP.

Application to Sunspot Data

1. K. Moon, J. Li, V. Delouille, F. Watson, and A. Hero, "Image patch analysis and clustering of sunspots: A dimensionality reduction approach," in *IEEE International Conference on Image Processing* (ICIP), 2014 [142].

2. K. Moon, V. Delouille, A. Hero, "Meta learning of bounds on the Bayes classifier error," in *IEEE Signal Processing and SP Education Workshop*, 2015 [139].

3. K. Moon, J. Li, V. Delouille, R. De Visscher, F. Watson, and A. Hero, "Image patch analysis of sunspots and active regions. I. Intrinsic dimension and correlation analysis," *Journal of Space Weather and Space Climate*, 2016 [144].

4. K. Moon, V. Delouille, J. Li, R. De Visscher, F. Watson, and A. Hero, "Image patch analysis of sunspots and active regions. II. Clustering via matrix factorization," *Journal of Space Weather and Space Climate*, 2016 [143].

Application to HFO Data

1. S. Gliske, K. Moon, W. Stacey, and A. Hero, "The intrinsic value of HFO features as a biomarker of epileptic activity," in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), 2016 [75].

1.4 Conclusion

In conclusion, this thesis first develops a framework for simple and accurate estimation of distributional functionals such as entropy, divergence, and mutual information functionals. These estimators can be used to reduce the risk of overfitting data via intrinsic dimension estimation and Bayes error estimation. Additionally, these estimators can be used to extend machine learning techniques to distributional features. This thesis then demonstrates the use of these estimators in these applications by applying them to sunspot image data and HFO data.

CHAPTER II

Kernel Density Ensemble Estimation of Divergence Functionals

This chapter provides new theoretical results required to derive nonparametric estimators of distributional functionals using kernel density estimators (KDE). As mentioned in Chapter I, this thesis focuses primarily on functionals of two distributionals which are referred to as divergence functionals. Specifically, we consider the problem of estimating divergence functionals when only two finite populations of independent and identically distributed (i.i.d.) samples are available from some unknown, nonparametric, smooth, d-dimensional distributions. We derive mean squared error (MSE) convergence rates for kernel density plug-in divergence functional estimators. We then extend the theory of optimally weighted ensemble entropy estimation developed in [187] to obtain two divergence functional estimators with a MSE convergence rate of $O\left(\frac{1}{T}\right)$, where T is the sample size, when the densities are sufficiently smooth. We then derive the asymptotic distribution of the weighted ensemble estimators which enables us to perform hypothesis testing.

Notation

Bold face type is used for random variables and random vectors. The conditional expectation given a random variable $\mathbf{Z}$ is denoted $\mathbb{E}_{\mathbf{Z}}$. The variance of a random variable is denoted $\mathbb{V}$ and the bias of an estimator is denoted $\mathbb{B}$.

2.1 The Divergence Functional Weak Estimator

This chapter focuses on estimating functionals of the form

$$G(f_1, f_2) = \int g(f_1(x), f_2(x)) f_2(x) dx, \tag{2.1}$$

where $g(x, y)$ is a smooth functional, and f_1 and f_2 are smooth d-dimensional probability densities. If $g(f_1(x), f_2(x)) = g\left(\frac{f_1(x)}{f_2(x)}\right)$, g is convex, and $g(1) = 0$, then $G(f_1, f_2)$ defines the family of f-divergences. Some common divergences that belong to this family include the KL divergence ($g(t) = -\ln t$), the Rényi-α divergence ($g(t) = t^\alpha$), and the total variation distance ($g(t) = |t - 1|$). We consider a broader class of functionals than the f-divergences.

2.1.1 The Kernel Density Plug-in Estimator

We use a kernel density plug-in estimator of the divergence functional in (2.1). Assume that N_1 i.i.d. realizations $\{\mathbf{Y}_1, \dots, \mathbf{Y}_{N_1}\}$ are available from f_1 and N_2 i.i.d. realizations $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}\}$ are available from f_2. Let $h_i > 0$ be the kernel bandwidth for the density estimator of f_i. Let $K(\cdot)$ be a kernel function with $||K||_\infty < \infty$ where $\|K\|_\infty$ is the ℓ_∞ norm of the kernel K. The kernel density estimates (KDE) are defined as follows

$$\begin{aligned}
\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j) &= \frac{1}{N_1 h_1^d} \sum_{i=1}^{N_1} K\left(\frac{\mathbf{X}_j - \mathbf{Y}_i}{h_1}\right), \\
\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) &= \frac{1}{M_2 h_2^d} \sum_{\substack{i=1 \\ i \neq j}}^{N_2} K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h_2}\right),
\end{aligned}$$

where $M_2 = N_2 - 1$. The functional $G(f_1, f_2)$ is then approximated as

$$\tilde{\mathbf{G}}_{h_1,h_2} = \frac{1}{N_2} \sum_{i=1}^{N_2} g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i)\right). \tag{2.2}$$

2.1.2 Convergence Rates

Similar to [140, 141, 187], the principal assumptions we make on the densities f_1 and f_2 and the functional g are that: 1) f_1, f_2, and g are smooth; 2) f_1 and f_2 have common bounded support set $\mathcal{S}$; 3) f_1 and f_2 are strictly lower bounded on $\mathcal{S}$. We also assume 4) that the density support set is smooth with respect to the kernel $K(u)$. Our full assumptions are:

- ($\mathcal{A}.0$): Assume that the kernel K is symmetric, is a product kernel, and has bounded support in each dimension. Also assume that it has order ν which means that the jth moment of the kernel K_i defined as $\int t^j K_i(t)dt$ is zero for all $j = 1, \ldots, \nu - 1$ and $i = 1, \ldots, d$ where K_i is the kernel in the ith coordinate.

- ($\mathcal{A}.1$): Assume there exist constants $\epsilon_0, \epsilon_\infty$ such that $0 < \epsilon_0 \leq f_i(x) \leq \epsilon_\infty < \infty$, $\forall x \in S$.

- ($\mathcal{A}.2$): Assume that the densities $f_i \in \Sigma(s, K)$ in the interior of $\mathcal{S}$ with $s \geq 2$ (see Definition I.1).

- ($\mathcal{A}.3$): Assume that g has an infinite number of mixed derivatives.

- ($\mathcal{A}.4$): Assume that $\left|\frac{\partial^{k+l} g(x,y)}{\partial x^k \partial y^l}\right|$, $k, l = 0, 1, \ldots$ are strictly upper bounded for $\epsilon_0 \leq x, y \leq \epsilon_\infty$.

- ($\mathcal{A}.5$): Assume the following boundary smoothness condition: Let $p_x(u) : \mathbb{R}^d \to \mathbb{R}$ be a polynomial in u of order $q \leq r = \lfloor s \rfloor$ whose coefficients are a function of x and are $r-q$ times differentiable. Then assume that

$$\int_{x\in\mathcal{S}}\left(\int_{u:K(u)>0,\,x+uh\notin\mathcal{S}} K(u)p_x(u)du\right)^t dx = v_t(h),$$

where $v_t(h)$ admits the expansion

$$v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t}h^i + o\left(h^{r-q}\right),$$

for some constants $e_{i,q,t}$.

As stated in assumption $\mathcal{A}.0$, we focus on finite support kernels for simplicity in the proofs although it is likely that our results extend to some infinitely supported kernels as well. The smoothness assumptions on the densities in assumption $\mathcal{A}.2$ are weaker as compared to [140, 141, 187]. However, we assume stronger conditions on the smoothness of g to enable us to achieve good convergence rates without knowledge of the boundary of the support set. These assumptions are not overly restrictive as most divergence functionals of interest are infinitely differentiable (e.g. the KL divergence). Those that are not infinitely differentiable are typically not differentiable everywhere (e.g. the total variation distance) which violates the assumptions of current nonparametric estimators that achieve the parametric rate.

Densities for which assumptions $\mathcal{A}.1-\mathcal{A}.2$ hold include the truncated Gaussian distribution and the Beta distribution on the unit cube. Functions for which assumptions $\mathcal{A}.3-\mathcal{A}.4$ hold include $g(x,y) = -\ln\left(\frac{x}{y}\right)$ and $g(x,y) = \left(\frac{x}{y}\right)^\alpha$.

Assumption $\mathcal{A}.5$ requires the boundary of the density support set to be smooth wrt the kernel $K(u)$ in the sense that the expectation of the area outside of $\mathcal{S}$ wrt any random variable u with smooth distribution is a smooth function of the bandwidth h. It is not necessary for the boundary of $\mathcal{S}$ to have smooth contours with no edges or corners as this assumption is satisfied by the following cases:

Theorem II.1. *Assumption $\mathcal{A}.5$ is satisfied when $\mathcal{S} = [-1,1]^d$ and when K is the uniform rectangular kernel; that is $K(x) = 1$ for all $x : ||x||_1 \leq 1/2$. Assumption $\mathcal{A}.5$ is also satisfied when $\mathcal{S} = [-1,1]^d$ and when K is the uniform Euclidean kernel; that is $K(x) = 1$ for all $x : ||x||_2 \leq 1/2$.*

The proof is given in Appendix A. Given the simple nature of this density support set and kernels, it is likely that other kernels and supports will satisfy $\mathcal{A}.5$ as well.

The following theorem on the bias follows under assumptions $\mathcal{A}.0 - \mathcal{A}.5$:

Theorem II.2. *For general g, the bias of the plug-in estimator $\tilde{\mathbf{G}}_{h_1,h_2}$ is of the form*

$$\begin{aligned} \mathbb{B}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right] &= \sum_{j=1}^{r}\left(c_{4,1,j}h_1^j + c_{4,2,j}h_2^j\right) + \sum_{j=1}^{r}\sum_{i=1}^{r} c_{5,i,j}h_1^j h_2^i + O\left(h_1^s + h_2^s\right) \\ &\quad + c_{9,1}\frac{1}{N_1 h_1^d} + c_{9,2}\frac{1}{N_2 h_2^d} + o\left(\frac{1}{N_1 h_1^d} + \frac{1}{N_2 h_2^d}\right). \end{aligned} \tag{2.3}$$

Furthermore, if $g(x,y)$ has k,l-th order mixed derivatives $\frac{\partial^{k+l} g(x,y)}{\partial x^k \partial y^l}$ that depend on x,y only through $x^\alpha y^\beta$ for some $\alpha,\beta \in \mathbb{R}$, then for any positive integer $\lambda \geq 2$, the bias is of the form

$$\begin{aligned} \mathbb{B}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right] &= \sum_{j=1}^{r}\left(c_{4,1,j}h_1^j + c_{4,2,j}h_2^j\right) + \sum_{j=1}^{r}\sum_{i=1}^{r} c_{5,i,j}h_1^j h_2^i + O\left(h_1^s + h_2^s\right) \\ &\quad \sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\left(c_{9,1,j,m}\frac{h_1^m}{\left(N_1 h_1^d\right)^j} + c_{9,2,j,m}\frac{h_2^m}{\left(N_2 h_2^d\right)^j}\right) \\ &\quad + \sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\sum_{i=1}^{\lambda/2}\sum_{n=0}^{r} c_{9,j,i,m,n}\frac{h_1^m h_2^n}{\left(N_1 h_1^d\right)^j \left(N_2 h_2^d\right)^i} \\ &\quad + O\left(\frac{1}{\left(N_1 h_1^d\right)^{\frac{\lambda}{2}}} + \frac{1}{\left(N_2 h_2^d\right)^{\frac{\lambda}{2}}}\right). \end{aligned} \tag{2.4}$$

Divergence functionals that satisfy the mixed derivatives condition required for (2.4) include the KL divergence and the Rényi-α divergence. Obtaining similar terms for other divergence functionals requires us to separate the dependence on h_i of the derivatives of g evaluated at $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$. This is left for future work. See Appendix B.1 for details.

The following variance result requires much less strict assumptions:

Theorem II.3. *Assume that the functional g in (2.1) is Lipschitz continuous in both of its arguments with Lipschitz constant C_g. Then the variance of the plug-in estimator $\tilde{\mathbf{G}}_{h_1,h_2}$ is bounded by*

$$\mathbb{V}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right] \leq C_g^2||K||_\infty^2 \left(\frac{10}{N_2} + \frac{N_1}{N_2^2}\right).$$

From Theorems II.2 and II.3, it is clear that we require $h_i \to 0$ and $N_i h_i^d \to \infty$ for $\tilde{\mathbf{G}}_{h_1,h_2}$ to be unbiased while the variance of the plug-in estimator depends primarily on the number of samples. Note that the constants in front of the terms that depend on h_i and N_i may not be identical for different i, j, m, n in (2.3) and (2.4). However, these constants depend on the densities f_1 and f_2 and their derivatives which are often unknown. The rates given in Thm. II.2 and II.3 are similar to the rates derived for the entropy plug-in estimator in [187] if $h_i^d = k_i/N_i$. The differences lie in the constants in front of the rates and the dependence on the number of samples from two distributions instead of one. Additionally, as compared to (2.3), in (2.4) there are many more terms. These terms enable us to achieve the parametric MSE convergence rate when $s > d/2$ for an appropriate choice of bandwidths whereas the terms in (2.3) require $s \geq d$ to achieve the same rate.

The Lipschitz assumption on g is comparable to other nonparametric estimators of distributional functionals [102, 111, 145, 180, 181]. Specifically, assumption $\mathcal{A}.1$ ensures that functionals such as those for Shannon and Renyi divergences are Lipschitz on the space ϵ_0 to ϵ_∞.

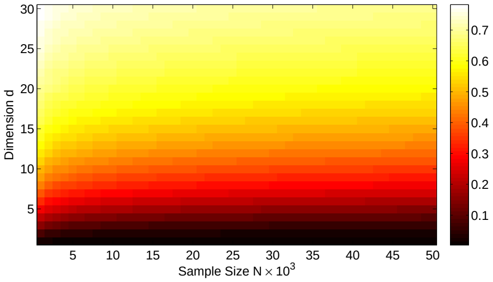

Figure 2.1: Heat map of predicted bias of divergence funtional plug-in estimator based on Theorem II.2 as a function of dimension and sample size when $h = N^{\frac{-1}{d+1}}$. Note the phase transition in the bias as dimension d increases for fixed sample size N: bias remains small only for relatively small values of d. The proposed weighted ensemble estimator removes this phase transition when the densities are sufficiently smooth.

2.1.3 Optimal MSE Rate

From Theorem II.2, the dominating terms in the bias are $\Theta\left(h_i\right)$ and $\Theta\left(\frac{1}{N_i h_i^d}\right)$. If no attempt is made to correct the bias, the optimal choice of h_i in terms of minimizing the MSE is

$$h_i^* = \Theta\left(N_i^{\frac{-1}{d+1}}\right).$$

This results in a dominant bias term of order $\Theta\left(N_i^{\frac{-1}{d+1}}\right)$. Note that this differs from the standard result for the optimal KDE bandwidth for minimum MSE density estimation which is $\Theta\left(N^{-1/(d+4)}\right)$ for a symmetric uniform kernel [83].

Figure 2.1 gives a heatmap showing the leading term $O\left(h\right)$ as a function of d and N when $h = N^{\frac{-1}{d+1}}$. The heatmap indicates that the bias of the plug-in estimator in (2.2) is small only for relatively small values of d.

2.1.4 Proof Sketches of Theorems II.2 and II.3

To prove the expressions for the bias, the bias is first decomposed into two parts by adding and subtracting $g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)$ within the expectation creating a "bias" term and a "variance" term. Applying a Taylor series expansion on the bias and variance terms results in expressions that depend on powers of $\mathbb{B}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right] :=$

$\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})-f_i(\mathbf{Z})$ and $\tilde{\mathbf{e}}_{i,h_i}(\mathbf{Z}) := \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})-\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$, respectively. Within the interior of the support, moment bounds can be derived from properties of the KDEs and a Taylor series expansion of the densities. Near the boundary of the support, the smoothness assumption on the boundary $\mathcal{A}.5$ is also required. Note that this approach differs from that in [187] which corrected the KDEs near the boundary of the support set and also used concentration inequalities for the KDEs. The full proof of Thm. II.2 is given in Appendix B.1.

The proof of the variance result takes a different approach. It uses the Efron-Stein inequality which bounds the variance by analyzing the expected squared difference between the plug-in estimator when one sample is allowed to differ. The full proof of Thm. II.3 is given in Appendix B.2.

2.2 Weighted Ensemble Estimation

As pointed out in Sec. 2.2.2, Thm. II.2 shows that when the dimension of the data is not small, the bias of the MSE-optimal plug-in estimator $\tilde{\mathbf{G}}_{h_1,h_2}$ decreases very slowly as a function of sample size, resulting in large MSE. However, by applying the theory of optimally weighted ensemble estimation, originally developed in [187] for entropy estimation, we can modify the minimum MSE estimator by taking a weighted sum of an ensemble of estimators where the weights are chosen to significantly reduce the bias.

2.2.1 The Weighted Ensemble Estimator

The bias expressions in Theorem II.2 are quite complicated due to their dependence on the sample size of two different distributions. We can simplify them significantly by assuming that $N_1 = N_2 = N$ and $h_1 = h_2 = h$. Define $\tilde{\mathbf{G}}_h := \tilde{\mathbf{G}}_{h,h}$.

Corollary II.4. *For general g, the bias of the plug-in estimator $\tilde{\mathbf{G}}_h$ is given by*

$$\mathbb{B}\left[\tilde{\mathbf{G}}_h\right] = \sum_{j=1}^{\lfloor s\rfloor} c_{10,j}h^j + c_{11}\frac{1}{Nh^d} + O\left(h^s + \frac{1}{Nh^d}\right).$$

If $g(x,y)$ has k,l-th order mixed derivatives $\frac{\partial^{k+l}g(x,y)}{\partial x^k \partial y^l}$ that depend on x,y only through $x^\alpha y^\beta$ for some $\alpha,\beta \in \mathbb{R}$, then for any positive integer $\lambda \geq 2$, the bias is

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_h\right] &= \sum_{j=1}^{\lfloor s\rfloor} c_{10,j}h^j + \sum_{q=1}^{\lambda/2}\sum_{j=0}^{\lfloor s\rfloor} c_{11,q,j}\frac{h^j}{\left(Nh^d\right)^q} \\ &\quad + O\left(h^s + \frac{1}{\left(Nh^d\right)^{\frac{\lambda}{2}}}\right).\end{aligned}$$

Note that the corollary still holds if N_1 and N_2 are linearly related, i.e., $N = N_1 = \Theta(N_2)$ and similarly if h_1 and h_2 are linearly related, i.e., $h = h_1 = \Theta(h_2)$. We form an ensemble of estimators by choosing different values of h. Choose $\mathcal{L} = \{l_1, \dots, l_L\}$ to be real positive numbers that index $h(l_i)$. Thus the parameter l indexes over different neighborhood sizes for the kernel density estimates. Define $w := \{w(l_1), \dots, w(l_L)\}$ and $\tilde{\mathbf{G}}_w := \sum_{l\in\mathcal{L}} w(l)\tilde{\mathbf{G}}_{h(l)}$. The key to reducing the MSE is to choose the weight vector w to reduce the lower order terms in the bias without substantially increasing the variance.

2.2.2 Finding the Optimal Weight

The theory of optimally weighted ensemble estimation is a general theory originally presented by Sricharan et al [187] that can be applied to many estimation problems as long as the bias and variance of the estimator can be expressed in a specific way. We generalize the conditions given in [187] that were required to apply the theory. Let $\mathcal{L} = \{l_1, \dots, l_L\}$ be a set of index values and let N be the number of samples available. For an indexed ensemble of estimators $\left\{\hat{\mathbf{E}}_l\right\}_{l\in\mathcal{L}}$ of a parameter E, the weighted ensemble estimator with weights $w = \{w(l_1), \dots, w(l_L)\}$ satisfying

$\sum_{l \in \mathcal{L}} w(l) = 1$ is defined as

$$\hat{\mathbf{E}}_w = \sum_{l \in \mathcal{L}} w\left(l\right) \hat{\mathbf{E}}_l.$$

$\hat{\mathbf{E}}_w$ is asyptotically unbiased if the estimators $\left\{\hat{\mathbf{E}}_l\right\}_{l \in \mathcal{L}}$ are asymptotically unbiased. Consider the following conditions on $\left\{\hat{\mathbf{E}}_l\right\}_{l \in \mathcal{L}}$:

- $\mathcal{C}.1$ The bias is expressible as

  $$\mathbb{B}\left[\hat{\mathbf{E}}_l\right] = \sum_{i \in J} c_i \psi_i(l) \phi_{i,d}(N) + O\left(\frac{1}{\sqrt{N}}\right),$$

  where c_i are constants depending on the underlying density, $J = \{i_1, \dots, i_I\}$ is a finite index set with $I < L$, and $\psi_i(l)$ are basis functions depending only on the parameter l and not on the sample size.

- $\mathcal{C}.2$ The variance is expressible as

  $$\mathbb{V}\left[\hat{\mathbf{E}}_l\right] = c_v \left(\frac{1}{N}\right) + o\left(\frac{1}{N}\right).$$

Theorem II.5. *Assume conditions $\mathcal{C}.1$ and $\mathcal{C}.2$ hold for an ensemble of estimators $\left\{\hat{\mathbf{E}}_l\right\}_{l \in \mathcal{L}}$. Then there exists a weight vector w_0 such that the MSE of the weighted ensemble estimator attains the parametric rate of convergence:*

$$\mathbb{E}\left[\left(\hat{\mathbf{E}}_{w_0} - E\right)^2\right] = O\left(\frac{1}{N}\right).$$

The weight vector w_0 is the solution to the following convex optimization problem:

$$\begin{array}{rl} \min_w & ||w||_2 \\ \textit{subject to} & \sum_{l \in \mathcal{L}} w(l) = 1, \\ & \gamma_w(i) = \sum_{l \in \mathcal{L}} w(l) \psi_i(l) = 0, \; i \in J. \end{array} \tag{2.5}$$

A more restrictive version of Theorem II.5 was originally presented in [187] with the stricter condition of $\phi_{i,d}(N) = N^{-1/(2d)}$. The proof of our generalized version (Theorem II.5) is sketched below.

Proof. From condition $\mathcal{C}.1$, the bias of the weighted estimator is

$$\mathbb{B}\left[\hat{\mathbf{E}}_w\right] = \sum_{i \in J} c_i \gamma_w(i) \phi_{i,d}(N) + O\left(\frac{\sqrt{L}||w||_2}{\sqrt{N}}\right).$$

The variance of the weighted estimator is bounded as

$$\mathbb{V}\left[\hat{\mathbf{E}}_w\right] \leq \frac{L||w||_2^2}{N}. \tag{2.6}$$

The optimization problem in (2.5) zeroes out the lower-order bias terms and limits the ℓ_2 norm of the weight vector w to limit the variance contribution. This results in an MSE rate of $O(1/N)$ when the dimension d is fixed and when L is fixed independently of the sample size N. Furthermore, a solution to (2.5) is guaranteed to exist as long as $L > I$ and the vectors $a_i = [\psi_i(l_1), \ldots, \psi_i(l_L)]$ are linearly independent. This completes our sketch of the proof of Thm. II.5. □

2.2.3 Optimally Weighted Distributional Functional (ODin) Estimators

To achieve the parametric rate $O\left(1/N\right)$ in MSE convergence it is not necessary that $\gamma_w(i) = 0$, $i \in J$. Solving the following convex optimization problem in place of the optimization problem in Theorem II.5 retains the $O(1/N)$ rate:

$$\begin{array}{rl} \min_w & \epsilon \\ \textit{subject to} & \sum_{l \in \mathcal{L}} w(l) = 1, \\ & \left|\gamma_w(i) N^{\frac{1}{2}} \phi_{i,d}(N)\right| \leq \epsilon, \ i \in J, \\ & \|w\|_2^2 \leq \eta, \end{array} \tag{2.7}$$

where the parameter η is chosen to achieve a trade-off between bias and variance. Instead of forcing $\gamma_w(i) = 0$, the relaxed optimization problem uses the weights to decrease the bias terms at the rate of $O\left(1/\sqrt{N}\right)$ yielding an MSE of $O(1/N)$.

We refer to the distributional functional estimators obtained using this theory as **O**ptimally Weighted **Di**stributional Fu**n**ctional (ODin) estimators. Sricharan et al [187] applied the stricter version of Theorem II.5 to obtain an entropy estimator with convergence rate $O(1/N)$. We also apply the same theory to obtain a divergence functional estimator with the same asymptotic rate. Let $h(l) = lN^{-1/(2d)}$. From Corollary II.4, we get $\psi_i(l) = l^i$, $i = 1, \dots, d$. Note that if $s \geq d$, then we are left with $O\left(\frac{1}{l^d\sqrt{N}}\right)$ in addition to the terms in the sum. To obtain a uniform bound on the bias with respect to w and $\mathcal{L}$, we also include the function $\psi_{d+1}(l) = l^{-d}$ in the optimization problem. The bias of the resulting base estimator satisfies condition $\mathcal{C}.1$ with $\phi_{i,d}(N) = N^{-i/(2d)}$ for $i = 1, \dots, d$ and $\phi_{d+1,d}(N) = N^{-1/2}$. The variance also satisfies condition $\mathcal{C}.2$. The optimal weight w_0 is found by using (2.7) to obtain a plug-in divergence functional estimator $\tilde{\mathbf{G}}_{w_0,1}$ with an MSE convergence rate of $O\left(\frac{1}{N}\right)$ as long as $s \geq d$. Otherwise, if $s < d$ we can only guarantee the MSE rate up to $O\left(\frac{1}{N^{s/d}}\right)$. We refer to this estimator as the ODin1 estimator.

Another weighted ensemble estimator can be defined that requires less strict assumptions on the smoothness of the densities. This is accomplished by letting $h(l)$ decrease at a faster rate. Let $\delta > 0$ and $h(l) = lN^{\frac{-1}{d+\delta}}$. From Theorem II.2, we have that if $g(x, y)$ has mixed derivatives of the form of $x^\alpha y^\beta$, then the bias has terms proportional to $l^{j-dq}N^{-\frac{j+\delta q}{d+\delta}}$ where $j, q \geq 0$ and $j + q > 0$. Theorem II.5 can be applied to the ensemble of estimators to derive an estimator that achieves the parametric convergence rate under these conditions. Let $\phi_{j,q,d}(N) = N^{-\frac{j+\delta q}{d+\delta}}$ and $\psi_{j,q}(l) = l^{j-dq}$.

Let

$$J = \left\{\{j,q\} : 0 < j + \delta q < (d+\delta)/2,\, q \in \{0,1,2,\ldots,\lambda/2\},\, j \in \{0,1,2,\ldots,\lfloor s \rfloor\}\right\}. \tag{2.8}$$

Then from (2.5), the bias of $\tilde{\mathbf{G}}_{h(l)}$ satisfies condition $\mathcal{C}.1$. If $L > |J| = I$, then Theorem II.5 can be applied to obtain the optimal weight vector. The estimator $\tilde{\mathbf{G}}_{w_0,2} = \sum_{l \in \mathcal{L}} w_0(l)\tilde{\mathbf{G}}_{h(l)}$ achieves the parametric convergence rate if $\lambda \geq d/\delta + 1$ and if $s \geq (d+\delta)/2$. Otherwise, if $s < (d+\delta)/2$ we can only guarantee the MSE rate up to $O\left(\frac{1}{N^{2s/(d+\delta)}}\right)$. $\tilde{\mathbf{G}}_{w_0,2}$ is referred to as the ODin2 estimator and is summarized in Algorithm 1 when $\delta = 1$.

Algorithm 1 Optimally weighted ensemble estimator of divergence functionals

```
Input: $\eta$, $L$ positive real numbers $\mathcal{L}$, samples $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_N\}$ from $f_1$, samples
    $\{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$ from $f_2$, dimension $d$, function $g$, kernel $K$
Output: The optimally weighted divergence estimator $\tilde{\mathbf{G}}_{w_0,2}$
Solve for $w_0$ using (2.7) with $\phi_{j,q,d}(N) = N^{-\frac{j+q}{d+1}}$ and basis functions $\psi_{j,q}(l) = l^{j-dq}$,
$l \in \bar{l}$, and $\{i,j\} \in J$ defined in (2.8)
for all $l \in \bar{l}$ do
    $h(l) \leftarrow lN^{\frac{-1}{d+1}}$
    for $i = 1$ to $N$ do
        $\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_i) \leftarrow \frac{1}{Nh(l)^d}\sum_{j=1}^{N} K\left(\frac{\mathbf{X}_i - \mathbf{Y}_j}{h(l)}\right)$,
        $\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_i) \leftarrow \frac{1}{(N-1)h(l)^d}\sum_{\substack{j=1 \\ j \neq i}}^{N} K\left(\frac{\mathbf{X}_i - \mathbf{X}_j}{h(l)}\right)$
    end for
    $\tilde{\mathbf{G}}_{h(l)} \leftarrow \frac{1}{N}\sum_{i=1}^{N} g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_i)\right)$
end for
$\tilde{\mathbf{G}}_{w_0,2} \leftarrow \sum_{l \in \mathcal{L}} w_0(l)\tilde{\mathbf{G}}_{h(l)}$
```

2.2.4 Comparison of ODin1 and ODin2 Estimators

For the ODin1 estimator $\tilde{\mathbf{G}}_{w_0,1}$, $h \propto N^{\frac{-1}{2d}}$ and the parametric convergence rate is guaranteed when $s \geq d$. This can be achieved with $L \geq d$ parameters and applies to any functional g in (2.1) that is infinitely differentiable.

In contrast, for the ODin2 estimator $\tilde{\mathbf{G}}_{w_0,2}$, $h \propto N^{\frac{-1}{d+\delta}}$ if $g(x,y)$ has mixed derivatives of the form of $x^\alpha y^\beta$ and the parametric convergence rate is guaranteed when $s \geq \frac{d+\delta}{2}$. Thus the parametric rate can be achieved with $\tilde{\mathbf{G}}_{w_0,2}$ under less strict assumptions on the smoothness of the densities than those required for $\tilde{\mathbf{G}}_{w_0,1}$. Since $\delta > 0$ can be arbitrary, it is theoretically possible to construct an estimator that achieves the parametric rate as long as $s > d/2$. This is consistent with the rate achieved by the more complex estimators proposed in [111].

These rate improvements come at a cost in the number of parameters L required to implement the weighted ensemble estimator. If $s \geq \frac{d+\delta}{2}$ then the size of J for ODin2 is on the order of $\frac{d^2}{8\delta}$. This may lead to increased variance of the ensemble estimator as indicated by (2.6) and highlights the practical limits in the choice of δ. Also, so far $\tilde{\mathbf{G}}_{w_0,2}$ can only be applied to functionals $g(x,y)$ with mixed derivatives of the form of $x^\alpha y^\beta$. Future work is required to extend this estimator to other functionals of interest.

2.2.5 Central Limit Theorem

The following theorem shows that the appropriately normalized ensemble estimator $\tilde{\mathbf{G}}_w$ converges in distribution to a normal random variable. This enables us to perform hypothesis testing on the divergence functional. The proof is based on the Efron-Stein inequality and an application of Slutsky's Theorem (Appendix B.3).

Theorem II.6. *Assume that the functional g is Lipschitz in both arguments with Lipschitz constant C_g. Further assume that $h = o(1)$, $N \to \infty$, and $Nh^d \to \infty$. Then for fixed L, the asymptotic distribution of the weighted ensemble estimator $\tilde{\mathbf{G}}_w$ is*

$$Pr\left(\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right]\right)/\sqrt{\mathbb{V}\left[\tilde{\mathbf{G}}_w\right]} \leq t\right) \to Pr(\mathbf{S} \leq t),$$

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

2.2.6 Uniform Convergence Rates

In this section, we show that the optimally weighted ensemble estimators achieve the parametric MSE convergence rate uniformly. Denote the subset of $\Sigma(s, K)$ with densities bounded between ϵ_0 and ϵ_∞ as $\Sigma(s, K, \epsilon_0, \epsilon_\infty)$.

Theorem II.7. *Let $\tilde{\mathbf{G}}_{w_0}$ be an optimally weighted ensemble estimator of the functional*

$$G(p,q) = \int g\left(p(x), q(x)\right) q(x) dx,$$

where p and q are d-dimensional probability densities. That is, $\tilde{\mathbf{G}}_{w_0}$ corresponds to either the ODin1 or ODin2 estimator described in Section 2.2.3 with w_0 calculated using (2.7) and s sufficiently large according to the estimator; i.e. $s \geq d$ for ODin1 and $s \geq (d+\delta)/2$ for ODin2. Additionally, let $r = \lfloor d \rfloor$ (ODin1) or $r = \lfloor (d+\delta)/2 \rfloor$ and $s > r$. Then

$$\sup_{p,q \in \Sigma(s,K,\epsilon,\epsilon_\infty)} \mathbb{E}\left[\left(\tilde{\mathbf{G}}_{w_0} - G(p,q)\right)^2\right] \leq \frac{C}{N}, \tag{2.9}$$

where C is a constant.

The proof decomposes the MSE into the variance plus the square of the bias. The variance is bounded easily by using Theorem II.3. To bound the bias, we show that the constants in the bias terms are continuous with respect to the densities p and q under an appropriate norm. We then show that $\Sigma(s, K, \epsilon, \epsilon_\infty)$ is compact with respect to this norm and then apply the Extreme Value Theorem. Details are given in Appendix B.4.

2.3 Numerical Validation

Throughout this section, we choose $\delta = 1$ for all experiments involving the ODin2 estimator.

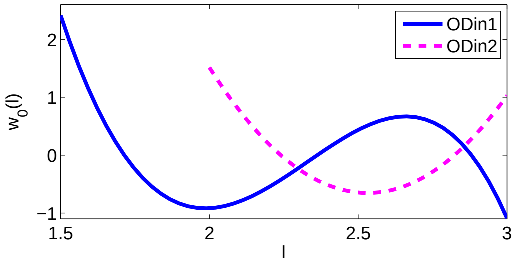

Figure 2.2: Examples of the optimal weights for $g(x,y) = \left(\frac{x}{y}\right)^{\alpha}$, $d = 4$, $N = 3100$, $L = 50$, and l is uniformly spaced between 1.5 (ODin1) or 2 (ODin2) and 3. The lowest values of l are given the highest weight. Thus the minimum value of bandwidth parameters $\mathcal{L}$ should be sufficiently large to render an adequate estimate.

2.3.1 Tuning Parameter Selection

The optimization problem in (2.7) has parameters η, L, and $\mathcal{L}$. The parameter η provides an upper bound on the norm of the weight vector, which gives an upper bound on the constant in the variance of the ensemble estimator. If all the constants in (2.3) or (2.4) and an exact expression for the variance of the ensemble estimator were known, then η could be chosen to minimize the MSE. Since the constants are unknown, by applying (2.7), the resulting MSE of the ensemble estimator is $O\left(\epsilon^2/N\right)+O\left(L\eta^2/N\right)$, where each term in the sum comes from the bias and variance, respectively. Since there is a tradeoff between η and ϵ, in principle setting $\eta = \epsilon/\sqrt{L}$ would minimize these terms. In practice, we find that the variance of the ensemble estimator is less than the upper bound of $L\eta^2/N$ and setting $\eta = \epsilon/\sqrt{L}$ is therefore overly restrictive. Setting $\eta = \epsilon$ instead works well in practice.

For fixed L, the set of kernel widths $\mathcal{L}$ can in theory be chosen by minimizing ϵ in (2.7) over $\mathcal{L}$ in addition to w. However, this results in a nonconvex optimization problem since w does not lie in the non-negative orthant. A parameter search may not be practical as ϵ generally decreases as the size and spread of $\mathcal{L}$ increases. This decrease in ϵ does not always correspond to a decrease in MSE as high and low values of $h(l)$ can lead to inaccurate density estimates. Denote the value of the minimum value of l so that $\tilde{\mathbf{f}}_{i,h(l_{min})}(\mathbf{X}_j) > 0 \ \forall i = 1, 2$ as l_{min} and the diameter of the support $\mathcal{S}$ as D. To ensure the density estimates are bounded away from zero, we require that $\min(\mathcal{L}) \geq l_{min}$. The weights in w_0 are generally largest for the smallest values of $\mathcal{L}$ (see Fig. 2.2) so $\min(\mathcal{L})$ should also be sufficiently larger than l_{min} to render an adequate estimate. Similarly, $\max(\mathcal{L})$ should be sufficiently smaller than D as high bandwidth values lead to high bias. The remaining $\mathcal{L}$ values are chosen to be equally spaced between $\min(\mathcal{L})$ and $\max(\mathcal{L})$.

As L increases, the similarity of bandwidth values $h(l)$ and basis functions $\psi_{i,d}(l)$ increases, resulting in a negligible decrease in the bias. Hence L should be chosen large enough to decrease the bias but small enough so that the $h(l)$ values are sufficiently distinct (typically $30 \leq L \leq 60$).

2.3.2 Convergence Rates Validation: Rényi-α Divergence

To validate our theory, we estimated the Rényi-α divergence integral between two truncated multivariate Gaussian distributions with varying dimension and sample sizes. The densities have means $\bar{\mu}_1 = 0.7 * \bar{1}_d$, $\bar{\mu}_2 = 0.3 * \bar{1}_d$ and covariance matrices $0.1 * I_d$ where $\bar{1}_d$ is a d-dimensional vector of ones, and I_d is a $d \times d$ identity matrix. We used $\alpha = 0.5$ and restricted the Gaussians to the unit cube.

The left plots in Fig. 2.3 show the MSE (200 trials) of the standard plug-in estimator implemented with a uniform kernel, the two proposed optimally weighted estimators ODin1 and ODin2, and a linear combination of ODin1 and ODin2, $\tilde{\mathbf{G}}_\rho = (1-\rho)\tilde{\mathbf{G}}_{w_0,1} + \rho\tilde{\mathbf{G}}_{w_0,2}$, for various dimensions and sample sizes. The set of kernel widths $\mathcal{L}$, L, and ρ are tuned to minimize the MSE. The bandwidth used for the standard plug-in estimator was selected from the set $\mathcal{L}$ that resulted from the ODin2 optimization; specifically the member of the set that empirically minimized the MSE of the plug-in estimator. Note that for $d = 4$, the standard plug-in estimator performs comparably with the optimally weighted estimators. However, for $d = 7, 10$,

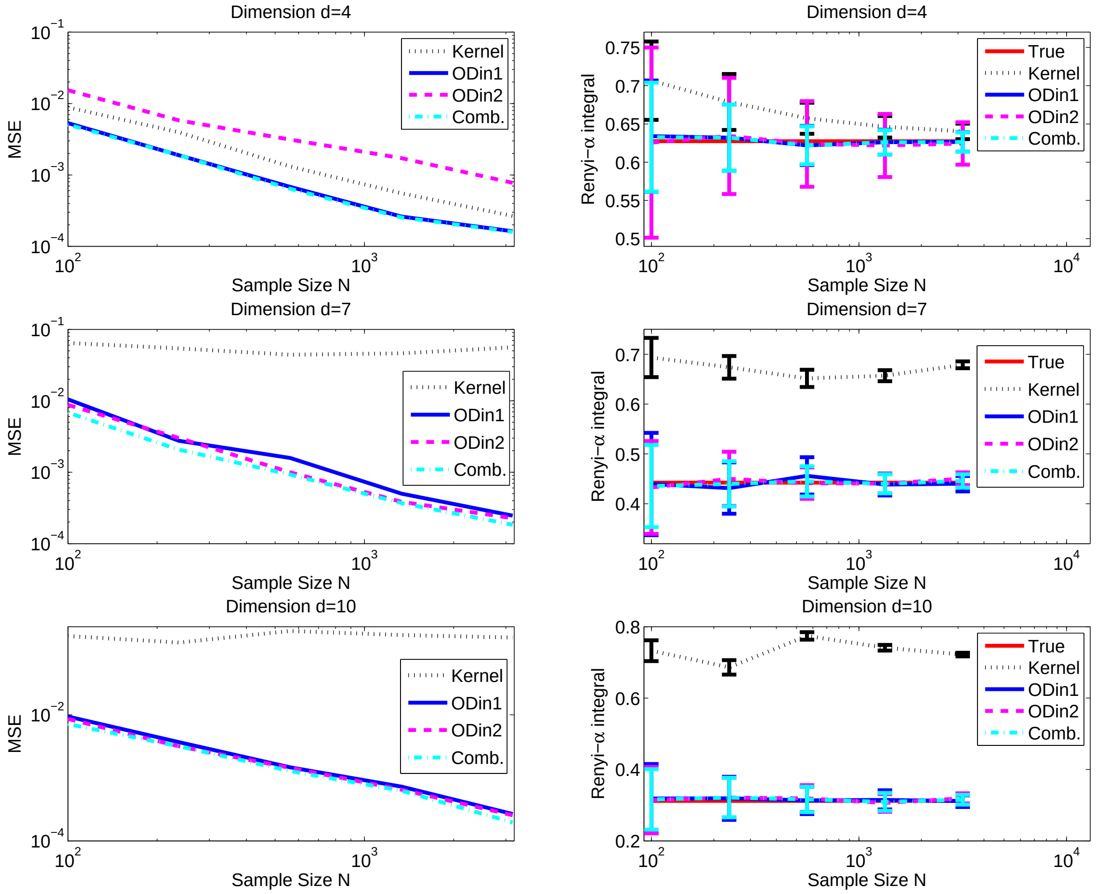

Figure 2.3: (Left) Log-log plot of MSE of the uniform kernel plug-in ("Kernel"), the two proposed optimally weighted estimators (ODin1 and ODin2), and the optimal linear combination of ODin1 and ODin2 for various dimensions and sample sizes. (Right) Plot of the average value of the same estimators with standard error bars compared to the true values being estimated. The proposed weighted ensemble estimators generally match the theoretical rate (see Table 2.1) and perform much better than the plug-in estimator for high dimensions.

Estimator	$d = 4$	$d = 7$	$d = 10$
ODin1	1.04	1.07	1.01
ODin2	0.83	1.08	1.00
Comb.	1.03	1.04	1.02

Table 2.1: Negative log-log slope of the MSE as a function of sample size for various dimensions and estimators

Dim.	$N = 100$	$N = 240$	$N = 560$	$N = 1330$	$N = 3200$
$d = 4$	0.15	0	0.1	0.05	0.05
$d = 7$	0.6	0.45	0.75	0.75	0.55
$d = 10$	0.55	1	0.5	0.65	0.5

Table 2.2: Values of the weight ρ for the estimator $\tilde{\mathbf{G}}_\rho = (1-\rho)\tilde{\mathbf{G}}_{w_0,1} + \rho\tilde{\mathbf{G}}_{w_0,2}$ that minimize MSE

the plug-in estimator performs considerably worse. This reflects the strength of ensemble estimators: the weighted sum of a set of poor estimators can result in a very good estimator. Note also that for most cases, the ensemble estimators' MSE rates match the theoretical rate based on the estimated log-log slope given in Table 2.1.

ODin1 tends to do better than ODin2 when the dimension is lower ($d = 4$) while the opposite occurs for the higher dimensions. Further evidence for this is given in the right figures in Fig. 2.3 that show the corresponding average estimates with standard error bars compared to the true values. ODin1 has smaller variance than ODin2 when $d = 4$ and slightly larger variance when $d = 10$. This seems to account for the differences in MSE between ODin1 and ODin2. The values for the weight ρ are given in Table 2.2 which indicate a preference for ODin1 when $d = 4$ and a preference for ODin2 for higher dimensions. Paired t-tests on the MSE (125 trials) of the two methods indicate that the MSE differences are statistically significant (see Table 2.3).

Dim.	ODin1>ODin2	ODin1<ODin2	ODin1=ODin2
4	1	0	1.8×10^{-58}
7	8.7×10^{-52}	1	1.8×10^{-51}
10	0	1	1.0×10^{-52}

Table 2.3: p-values of paired t-tests of ODin1 vs. ODin2 MSE ($N = 1300$). Null hypotheses for the p-values reported in each column are given on the top row.

	ODin1		ODin2	
Set	min($\mathcal{L}$)	max($\mathcal{L}$)	min($\mathcal{L}$)	max($\mathcal{L}$)
1	1.5	3	2	3
2	1.75	3	2.25	3
3	2	3	2.5	3
4	2.25	3	2.75	3
5	2.5	3	2.75	3.25

Table 2.4: Values of min($\mathcal{L}$) and max($\mathcal{L}$) for different experiments.

2.3.3 Tuning Parameter Robustness

The results in Section 2.3.2 were obtained by selecting the tuning parameters $\mathcal{L}$ and L for each pair of dimension and samples to minimize the MSE. Here we demonstrate the robustness of the estimators to variations in the tuning parameters.

In all experiments, we estimated the Rényi-α divergence integral between the same distributions described in Section 2.3.2 (truncated Gaussians with same covariance and different mean) and chose $L = 50$. In the first set of experiments, we set $\eta = \epsilon$ and chose the set of kernel bandwidths $\mathcal{L}$ to be linearly spaced between min($\mathcal{L}$) and max($\mathcal{L}$). Table 2.4 provides the values chosen for min($\mathcal{L}$) and max($\mathcal{L}$).

Figure 2.4 shows the results for these experiments when $d = 5$. As the number of samples increase, choosing a larger range of values for $\mathcal{L}$ (Sets 1 and 2) generally gives better performance for both ODin1 and ODin2 in terms of MSE than choosing a smaller range for $\mathcal{L}$ (e.g. Sets 4 and 5). This suggests that for large sample sizes, the estimators will perform well if a reasonably large range for $\mathcal{L}$ is chosen. In contrast, choosing a smaller range of values for $\mathcal{L}$ when the sample size is small results in smaller bias and variance compared to the larger ranges. Thus for small sample sizes, it may be useful to tighten the range of kernel bandwidths.

Comparing the results for ODin1 and ODin2 indicates that ODin2 is more robust to the choice of $\mathcal{L}$ as the difference in MSE under the different settings is smaller for ODin2 than ODin1. This is due primarily to the relatively smaller bias of ODin2 as the variances of the two estimators under each setting are comparable (see the bottom plots in Fig. 2.4). Similar results hold when the dimension is increased to $d = 7$ (see Fig. 2.5). For larger sample sizes, a large range for $\mathcal{L}$ gives better results than a smaller range. However, for smaller sample sizes, the larger range for $\mathcal{L}$ does not perform as well as other configurations. Additionally, ODin2 again appears to be more robust to the choice of $\mathcal{L}$ as the difference in MSE at larger sample sizes is smaller for ODin2. Additionally, the Set 5 configuration of ODin1 does not even appear to be converging to the true value yet when $N = 10000$.

For the second set of experiments, we fixed $\mathcal{L}$ to be linearly spaced values between 2 and 3. We then varied the values of η from 0.5 to 10. Figure 2.6 provides heatmaps of the MSE of the two ensemble estimators under this configuration with $d = 5, 7$. For $d = 5$, choosing $\eta = 0.5$ gives the lowest MSE when $N \geq 10^{3.5}$ for ODin1 and for all sample sizes for ODin2. In fact, when $d = 5$, ODin2 with $\eta = 0.5$ outperforms all other configurations at all sample sizes, including those shown in Fig. 2.4. Increasing d to 7 changes this somewhat as choosing $\eta = 0.5$ results in the lowest MSE when $N \geq 1000$ for ODin2 and for no sample sizes for ODin1. However, generally lower values of η ($\eta < 2$) result in the lowest MSE for ODin2 when $N < 1000$ and for ODin1 when $N \geq 10^{3.5}$ ($\eta \leq 3$). Both ODin1 and ODin2 are fairly robust to the choice of η when $d = 5$ as the MSE is relatively constant at each sample size for most η values. However, ODin2 has generally lower MSE values for $N \geq 10^{2.5}$ (see Table 2.5). When $d = 7$, ODin2 is more robust than ODin1 for larger samples ($N \geq 1000$).

Overall, based on our experiments, ODin1 has lower MSE on average for smaller sample sizes while ODin2 is generally more robust to the tuning parameters for larger

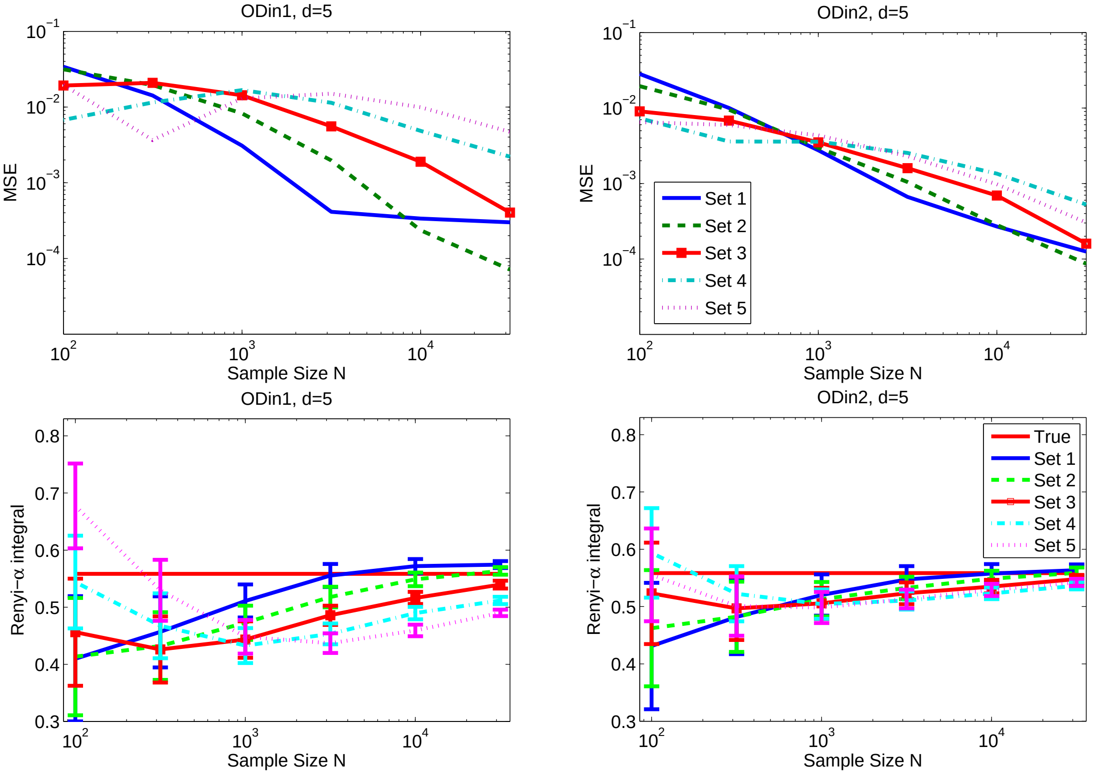

Figure 2.4: (Top) Log-log plot of MSE of the two proposed optimally weighted estimators (ODin1 and ODin2) as a function of sample size using different values for the range of kernel bandwidths $\mathcal{L}$ (see Table 2.4) when $d = 5$. (Bottom) Plot of the average value of the same estimators with standard error bars compared to the true value being estimated. For larger sample sizes, a larger range in $\mathcal{L}$ results in smaller MSE (see Sets 1 and 2), while a smaller range in $\mathcal{L}$ is more accurate at smaller sample sizes. ODin2 is generally more robust to the choice of $\mathcal{L}$.

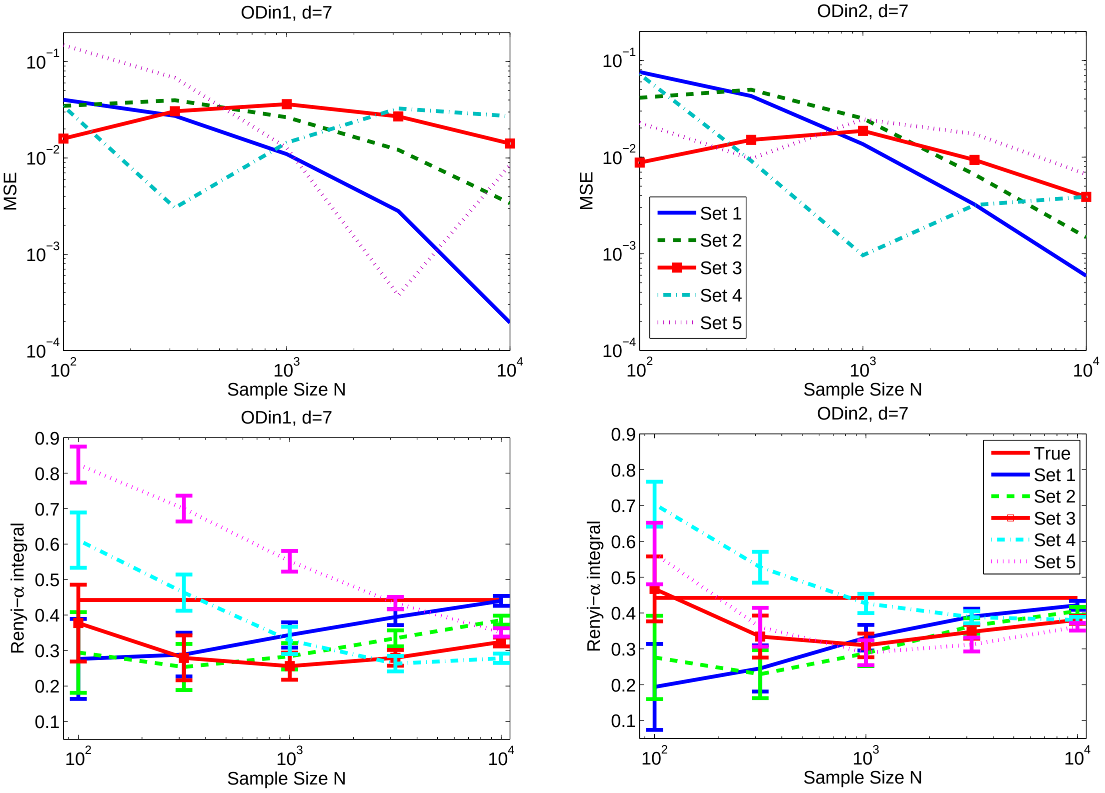

Figure 2.5: (Top) Log-log plot of MSE of the two proposed optimally weighted estimators (ODin1 and ODin2) as a function of sample size using different values of the parameter $\mathcal{L}$ (see Table 2.4) when $d = 7$. (Bottom) Plot of the average value of the same estimators with standard error bars compared to the true value being estimated. Again, a larger range for $\mathcal{L}$ at large sample sizes results in smaller MSE. Note from the bottom plots that the non-monotonicity of the MSE for Sets 4 and 5 is due to the fact that the estimators happen to be less biased for some mid-range values of N ; i.e. the asymptotics have not yet taken effect. As can be seen from the bottom plots, the bias is again decreasing for most of these estimators.

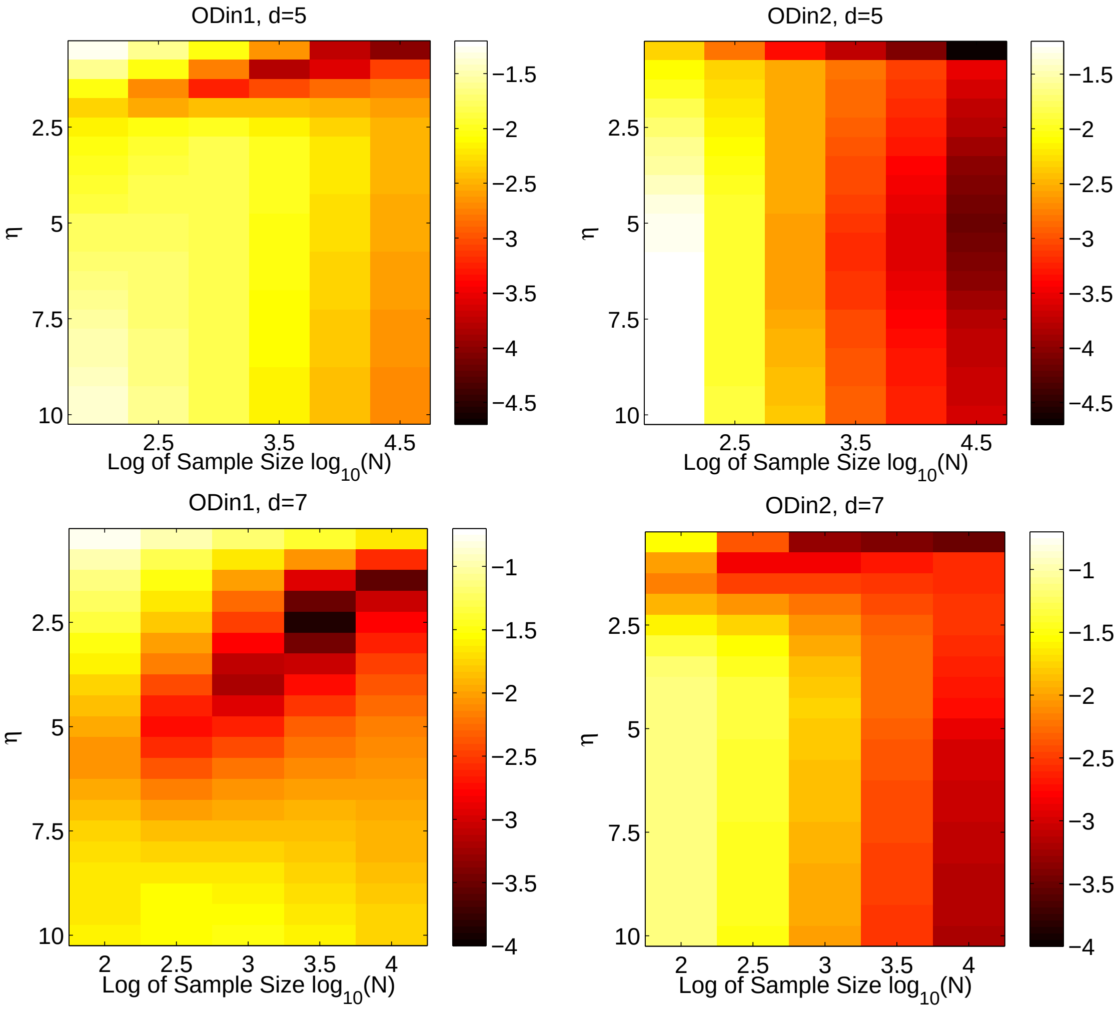

Figure 2.6: Heatmaps of the ensemble estimators' MSE ($\log_{10}$ scale) as a function of sample size and the tuning parameter η. Lower values of η tend to give the smallest MSE, especially for ODin2. Both estimators are fairly robust to the choice of η as the MSE is relatively constant at each sample size for most η values.

Sample Size N	10^2	$10^{2.5}$	10^3	$10^{3.5}$	10^4	$10^{4.5}$
Mean MSE, ODin1, $d = 5$	0.0225	0.0159	0.0118	0.0071	0.0040	0.0022
Mean MSE, ODin2, $d = 5$	0.0358	0.0092	0.0029	0.0011	0.0005	0.0002
Mean MSE, ODin1, $d = 7$	0.0394	0.0233	0.0155	0.0115	0.0091	N/A
Mean MSE, ODin2, $d = 7$	0.0557	0.0292	0.0120	0.0046	0.0018	N/A

Table 2.5: Average MSE over all values of η used in Figure 2.6 for a fixed sample size.

sample sizes. Additionally, choosing a low value for η with ODin2 may result in better performance. Thus unless the sample size is small, ODin2 should be preferred over ODin1.

2.3.4 Central Limit Theorem Validation: KL Divergence

To verify the central limit theorem of both ensemble estimators, we estimated the KL divergence between two truncated Gaussian densities again restricted to the unit cube. We conducted two experiments where 1) the densities are different with means $\bar{\mu}_1 = 0.7 * \bar{1}_d$, $\bar{\mu}_2 = 0.3 * \bar{1}_d$ and covariance matrices $\sigma_i * I_d$, $\sigma_1 = 0.1$, $\sigma_2 = 0.3$; and where 2) the densities are the same with means $0.3 * \bar{1}_d$ and covariance matrices $0.3 * I_d$. For both experiments, we chose $d = 6$ and $N = 1000$.

Figure 2.7 shows Q-Q plots of the normalized optimally weighted ensemble estimators ODin1 (left) and ODin2 (right) of the KL divergence when the two densities are the same (top) and when they are different (bottom). The linear relationship between the quantiles of the normalized estimators and the standard normal distribution validates Theorem II.6 for both estimators under the two cases.

2.4 Conclusion

We derived convergence rates for a kernel density plug-in estimator of divergence functionals. We generalized the theory of optimally weighted ensemble estimation and derived an estimator that achieves the parametric rate when the densities belong to the Hölder smoothness class with smoothness parameter greater than $d/2$. The

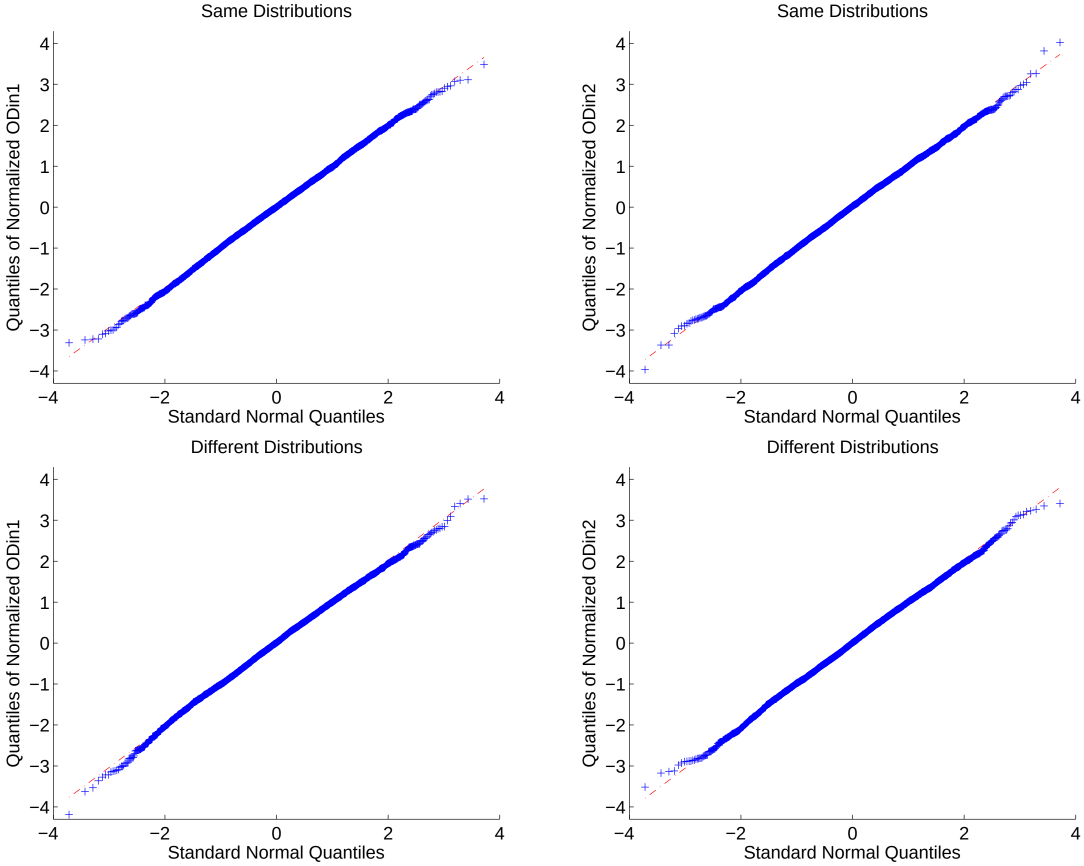

Figure 2.7: Q-Q plots comparing quantiles from the normalized weighted ensemble estimators ODin1 (left) and ODin2 (right) of the KL divergence (vertical axis) to the quantiles from the standard normal distribution (horizontal axis) when the two distributions are the same (top) and when they are different (bottom). The red line shows a reference line passing through the first and third quantiles. The linearity of the plot points validates the central limit theorem (Theorem II.6) for all four cases.

estimators we derive apply to general bounded density support sets and do not require knowledge of the support which is a distinct advantage over other estimators. We also derived the asymptotic distribution of the estimator, provided some guidelines for tuning parameter selection, and validated the convergence rates for the case of empirical estimation of the Rényi-α divergence. We then performed experiments to examine the estimators' robustness to the choice of tuning parameters and validated the central limit theorem for KL divergence estimation.

The generalized theory of optimally weighted ensemble estimators derived in this chapter can be used to derive k-nn based estimators of entropy and divergence functionals that have similar performance. This is presented in Chapter III. Similarly, we extend our methods of analysis to derive optimally weighted ensemble estimators of mutual information measures based on KDE plug-in methods. This is presented in Chapter IV.

CHAPTER III

k-Nearest Neighbors Ensemble Estimation of Divergence Functionals

The theory derived in Chapter II for KDE plug-in estimators of divergence functionals does not easily extend to...This chapter focuses on nonparametric k-nearest neighbor (nn) plug-in estimators of divergence functionals. We consider the same problem as in Chapter II of estimating divergence functionals when only two finite populations of i.i.d. samples are available from unknown, nonparametric, smooth, d-dimensional distributions. In this chapter, the MSE convergence rates for k-nn plug-in divergence functional estimators is derived. We use the same general theory of optimally weighted ensemble estimation developed in Chapter II to again obtain two divergence functional estimators with a MSE convergence rate of $O(1/T)$ when the densities are sufficiently smooth. We also derive the asymptotic distribution of the weighted ensemble estimators.

3.1 The Divergence Functional Plug-in Estimator

As in Chapter II, we focus on estimating functionals of two distributions of the form

$$G\left(f_1, f_2\right) = \int g\left(f_1(x), f_2(x)\right) f_2(x) dx, \tag{3.1}$$

where f_1 and f_2 are smooth d-dimensional probability densities and $g(t_1, t_2)$ is a smooth functional.

3.1.1 The k-nn Plug-in Estimator

We use a k-nn density plug-in estimator of the divergence functional in (3.1). Assume that N_1 i.i.d. samples $\{\mathbf{Y}_1, \dots, \mathbf{Y}_{N_1}\}$ are available from f_1 and N_2 i.i.d. samples $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}\}$ are available from f_2. Let $M_1 = N_1$, $M_2 = N_2 - 1$, and $k_i \leq M_i$. Denote the distance of the k_1th nearest neighbor of the sample $\mathbf{Y}_i$ in $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}\}$ as $\rho_{1,k_1}(i)$. Similarly, denote the distance of the k_2th nearest neighbor of the sample $\mathbf{X}_i$ in $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}\} \setminus \{\mathbf{X}_i\}$ as $\rho_{2,k_2}(i)$. The standard k-nn density estimator is [128]

$$\hat{\mathbf{f}}_{i,k_i}(\mathbf{X}_j) = \frac{k_i}{M_i c_d \rho_{i,k_i}^d(j)},$$

where c_d is the volume of a d-dimensional unit ball. The functional $G(f_1, f_2)$ is estimated as

$$\hat{\mathbf{G}}_{k_1,k_2} = \frac{1}{N_2} \sum_{i=1}^{N_2} g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_i)\right).$$

3.1.2 Convergence Rates

We make similar assumptions on the densities and the functional g as in Chapter II. For completeness, they are

- $(\mathcal{B}.1)$: Assume there exist constants $\epsilon_0, \epsilon_\infty$ such that $0 < \epsilon_0 \leq f_i(x) \leq \epsilon_\infty < \infty$, $\forall x \in \mathcal{S}$.

- $(\mathcal{B}.2)$: Assume that the densities $f_i \in \Sigma(s, K)$ in the interior of $\mathcal{S}$ with $s \geq 2$ and $r = \lfloor s \rfloor$.

- $(\mathcal{B}.3)$: Assume that g has an infinite number of mixed derivatives.

- $(\mathcal{B}.4)$: Assume that $\left|\frac{\partial^{k+l} g(x,y)}{\partial x^k \partial y^l}\right|$, $k, l = 0, 1, \dots$ are strictly upper bounded for $\epsilon_0 \leq x, y \leq \epsilon_\infty$.

- $(\mathcal{B}.5)$: Assume that the support $\mathcal{S} = [0,1]^d$.

The following theorem on the bias of the plug-in estimator follows under assumptions $\mathcal{B}.1 - \mathcal{B}.5$

Theorem III.1. *For general g, the bias of the plug-in estimator $\hat{\mathbf{G}}_{k_1,k_2}$ is of the form*

$$\begin{aligned}\mathbb{B}\left[\hat{\mathbf{G}}_{k_1,k_2}\right] &= \sum_{j=1}^{r}\left(\left(c_{17,1,j} + \frac{c_{17,1,j,0}}{\sqrt{k_1}}\right)\left(\frac{k_1}{N_1}\right)^{\frac{j}{d}} + \left(c_{17,2,j} + \frac{c_{17,2,j,0}}{\sqrt{k_2}}\right)\left(\frac{k_2}{N_2}\right)^{\frac{j}{d}}\right) \\ &+\sum_{j=0}^{r}\sum_{\substack{i=0\\i+j\neq 0}}^{r} c_{18,i,j}\left(\frac{k_1}{N_1}\right)^{\frac{i}{d}}\left(\frac{k_2}{N_2}\right)^{\frac{j}{d}} \\ &+O\left(\frac{1}{\sqrt{k_1k_2}} + \frac{1}{k_1} + \frac{1}{k_2} + \max\left(\frac{k_1}{N_1}, \frac{k_2}{N_2}\right)^{\frac{\min(s,d)}{d}}\right). \qquad (3.2)\end{aligned}$$

Furthermore, if $g(x,y)$ has m, l-th order mixed derivatives $\frac{\partial^{m+l}g(x,y)}{\partial x^m \partial y^l}$ that depend on x, y only through $x^\alpha y^\beta$ for some $\alpha, \beta \in \mathbb{R}$, then for any positive integer $\nu \geq 0$, the bias is of the form

$$\begin{aligned}\mathbb{B}\left[\hat{\mathbf{G}}_{k_1,k_2}\right] &= \sum_{j=0}^{\lfloor s\rfloor}\sum_{\substack{i=0\\i+j\neq 0}}^{\lfloor s\rfloor} c_{18,i,j}\left(\frac{k_1}{N_1}\right)^{\frac{i}{d}}\left(\frac{k_2}{N_2}\right)^{\frac{j}{d}} + O\left(\max\left(\frac{k_1}{N_1}, \frac{k_2}{N_2}\right)^{\frac{\min(s,d)}{d}} + \frac{1}{\min(k_1,k_2)^{\frac{2+\nu}{2}}}\right) \\ &+\sum_{m=0}^{\nu}\sum_{\substack{j=0\\j+m\neq 0}}^{r}\left(\frac{c_{20,1,j,m}}{k_1^{\frac{1+m}{2}}}\left(\frac{k_1}{N_1}\right)^{\frac{j}{d}} + \frac{c_{20,2,j,m}}{k_2^{\frac{1+m}{2}}}\left(\frac{k_2}{N_2}\right)^{\frac{j}{d}}\right) \\ &+\sum_{j=1}^{r}\left(c_{17,1,j}\left(\frac{k_1}{N_1}\right)^{\frac{j}{d}} + c_{17,2,j}\left(\frac{k_2}{N_2}\right)^{\frac{j}{d}}\right) \\ &+\sum_{m=0}^{\nu}\sum_{\substack{j=0\\m+j\neq 0}}^{\lfloor s\rfloor}\sum_{n=0}^{\nu}\sum_{\substack{i=0\\n+i\neq 0}}^{\lfloor s\rfloor}\frac{c_{18,i,j,m,n}}{k_1^{\frac{1+m}{2}}k_2^{\frac{1+n}{2}}}\left(\frac{k_1}{N_1}\right)^{\frac{i}{d}}\left(\frac{k_2}{N_2}\right)^{\frac{j}{d}}. \qquad (3.3)\end{aligned}$$

The following variance result requires much less strict assumptions:

Theorem III.2. *If the functional g is Lipschitz continuous in both of its arguments* *with Lipschitz constant C_g, then the variance of $\hat{\mathbf{G}}_{k_1,k_2}$ is*

$$\mathbb{V}\left[\hat{\mathbf{G}}_{k_1,k_2}\right] = O\left(\frac{1}{N_2} + \frac{N_1}{N_2^2}\right). \tag{3.4}$$

From Theorems III.1 and III.2, it is clear that we require $k_i \to \infty$ and $k_i/N_i \to 0$ for $\hat{\mathbf{G}}_{k_1,k_2}$ to be unbiased. For the variance to decrease to zero, we require $N_2 \to \infty$. As in Chapter II, the additional terms in (3.3) enable us to achieve the parametric MSE convergence rate when $s > d/2$ for an appropriate choice of k values whereas the terms in (3.2) require $s \geq d$ to achieve the same rate. Moreover, the additional terms in (3.3) enable us to achieve the parametric rate for smaller values of k which is more computationally efficient. For a discussion on the Lipschitz condition in Theorem III.2 and the extra condition required for (3.3), see Section 2.1.2.

3.1.3 Optimal MSE Rate

From Theorem III.1, the dominating terms in the bias are $\Theta\left(\left(\frac{k_i}{N_i}\right)^{\frac{1}{d}}\right)$ and $\Theta\left(\frac{1}{k_i}\right)$. If no bias correction is made, the optimal choice of k_i that minimizes the MSE is

$$k_i^* = \Theta\left(N_i^{\frac{1}{d+1}}\right).$$

This results in a dominant bias term of order $\Theta\left(N_i^{\frac{-1}{d+1}}\right)$, which is large whenever d is not small.

3.1.4 Proof Sketches of Theorems III.1 and III.2

The proof of the bias result uses a conditioning argument on the k-nn distances by viewing the k-nn estimator as a kernel density estimator with uniform kernel and random bandwidth. This allows us to leverage some of the KDE plug-in estimator proof techniques. For fixed bandwidth (i.e. k-nn distance), we then consider separately the cases where the k-nn ball is contained within the support and when it intersects the boundary of the support. See Appendix C.1 for the full proof.

The proof of the variance result uses the Efron-Stein inequality as in the proof of Theorem II.3. However, the proof of the k-nn result is more complicated due to the dependencies between different k-nn neighborhoods. Thus we analyze the possible effects on the k-nn graph when one sample is allowed to differ in order to use the Efron-Stein inequality. See Appendix C.2 for the full proof of Theorem III.2.

3.2 Weighted Ensemble Estimation

As for the KDE plug-in estimator in Chapter II, the k-nn plug-in estimator $\hat{\mathbf{G}}_{k_1,k_2}$ in Section 3.1 has slowly decreasing bias when the dimension of the data is not small. By applying the theory of optimally weighted ensemble estimation derived in Section 2.2, we can take a weighted sum of an ensemble of estimators where the weights are chosen to reduce the bias.

3.2.1 Optimally Weighted k-nn Estimators

We simplify the bias expressions in Theorem III.1 by assuming that $N_1 = N_2 = N$ and $k_1 = k_2 = k$. Define $\hat{\mathbf{G}}_k := \hat{\mathbf{G}}_{k,k}$.

Corollary III.3. *For general g, the bias of the plug-in estimator $\hat{\mathbf{G}}_k$ is given by*

$$\mathbb{B}\left[\hat{\mathbf{G}}_k\right] \;=\; \sum_{j=1}^{r}\left(c_{21,1,j}+\frac{c_{21,2,j}}{\sqrt{k}}\right)\left(\frac{k}{N}\right)^{\frac{j}{d}} + O\left(\frac{1}{k}+\left(\frac{k}{N}\right)^{\frac{\min(s,d)}{d}}\right).$$

If $g(x,y)$ has m, l-th order mixed derivatives $\frac{\partial^{m+l}g(x,y)}{\partial x^m \partial y^l}$ that depend on x, y only through $x^\alpha y^\beta$ for some $\alpha,\beta\in\mathbb{R}$, then for any positive integer $\nu \geq 2$, the bias is of the form

$$\mathbb{B}\left[\hat{\mathbf{G}}_k\right] = \sum_{j=1}^{r} c_{22,j}\left(\frac{k}{N}\right)^{\frac{j}{d}} + \sum_{m=0}^{\lambda}\sum_{\substack{j=0\\ j+m\neq 0}}^{r}\frac{c_{22,j,m}}{k^{\frac{1+m}{2}}}\left(\frac{k}{N}\right)^{\frac{j}{d}} + O\left(\frac{1}{k^{\frac{\nu}{2}}}+\left(\frac{k}{N}\right)^{\frac{\min(s,d)}{d}}\right)$$

The corollary still holds if N_1 and N_2 are linearly reated and if k_1 and k_2 are linearly related. An ensemble of estimators is formed by choosing different neighborhood sizes by choosing different values of k. Choose $\mathcal{L} = \{l_1, \ldots, l_L\}$ to be real positive numbers that index $h(l_i)$. As in Chapter II, define $w := \{w(l_1), \ldots, w(l_L)\}$ and $\hat{\mathbf{G}}_w := \sum_{l \in \mathcal{L}} w(l) \hat{\mathbf{G}}_{k(l)}$. The weights can be used to decrease the bias as before.

For general g, let $k(l) = l\sqrt{N}$. From Corollary III.3, we have $\psi_i(l) = l^{i/d}$ for $i = 1, \ldots, d$. If $s \geq d$, then we have a $O\left(\frac{1}{l\sqrt{N}}\right)$. We also include the function $\psi_{d+1}(l) = l^{-1}$. The bias of the resulting base estimator satisfies condition $\mathcal{C}.1$ with $\phi_{i,d}(N) = N^{-i/(2d)}$ for $i = 1, \ldots, d$ and $\phi_{i,d+1}(N) = N^{-1/2}$. The variance also satisfies condition $\mathcal{C}.2$. The optimal weight w_0 is found using (2.7) to obtain a plug-in divergence functional estimator $\hat{\mathbf{G}}_{w_0,1}$ with an MSE convergence rate of $O\left(\frac{1}{N}\right)$ as long as $s \geq d$. Otherwise, if $s < d$ we can only guarantee the MSE rate up to $O\left(\frac{1}{N^{s/d}}\right)$. We refer to this estimator as the ODin1 k-nn estimator.

As for the KDE case, we can define another weighted ensemble estimator that achieves the parametric rate under less strict assumptions on the smoothness of the densities if the functional g satisfies the assumption required for (3.3). Let $\delta > 0$ and $k(l) = lN^{\delta}$. From Corollary III.3, the bias has terms proportional to $l^{j-\frac{q}{2}} N^{-\frac{(1-\delta)j}{d} - \frac{q\delta}{2}}$ where $j, q \geq 0$ and $j + \frac{q}{2} > \frac{1}{2}$. Let $\phi_{j,q,d}(N) = N^{-\frac{(1-\delta)j}{d} - \frac{q\delta}{2}}$ and $\psi_{j,q}(l) = l^{j-\frac{q}{2}}$. Let

$$\begin{aligned} J &= \Bigg\{ \{j, q\} : 0 < \frac{(1-\delta)j}{d} + \frac{q\delta}{2} < \frac{1}{2},\, q \in \{0, 1, 2, \ldots, \nu\}, j \in \{0, 1, 2, \ldots, r\}, \\ &\qquad j + \frac{q}{2} > \frac{1}{2} \Bigg\}. \end{aligned}$$

Then the bias of the resulting base estimator satisfies condition $\mathcal{C}.1$ and the variance satisfies condition $\mathcal{C}.2$. If $L > |J|$, then the optimal weight can be found using (2.7). The resulting weighted ensemble estimator $\hat{\mathbf{G}}_{w_0,2}$ achieves the parametric convergence rate if $\nu \geq 1/\delta$ and if $s \geq \frac{d}{2(1-\delta)}$. Otherwise, if $s < d/(2(1-\delta))$ we can only guarantee the MSE rate up to $O\left(\frac{1}{N^{\frac{2(1-\delta)s}{d}}}\right)$. We refer to this estimator as the ODin2 k-nn estimator.

Comparing the ODin1 and ODin2 k-nn estimators yields similar results to that in Section 2.2.4. The parametric rate can be achieved with $\hat{\mathbf{G}}_{w_0,2}$ under less strict assumptions on the smoothness of the densities than those required for $\hat{\mathbf{G}}_{w_0,1}$. Since $\delta > 0$ can be arbitrary, it is theoretically possible to construct an estimator that achieves the parametric rate as long as $s > d/2$. However, $\hat{\mathbf{G}}_{w_0,2}$ requires more parameters to implement the weighted ensemble estimator than $\hat{\mathbf{G}}_{w_0,1}$ which may have an effect on the variance.

3.2.2 Central Limit Theorem

The following theorem shows that the appropriately normalized ensemble estimator $\hat{\mathbf{G}}_w$ converges in distribution to a normal random variable, which enables us to perform hypothesis testing on the divergence functional. The proof is different from the proof of Theorem II.6 in that we use a lemma modified from [186] that gives sufficient conditions on an interchangeable process for a central limit theorem. The details are given in Appendix C.3.

Theorem III.4. *Assume that the mixed derivatives of g of order 2 are bounded and $k(l) \to \infty$ as $N \to \infty$ for each $l \in \mathcal{L}$. Then for fixed L, and if $\mathbf{S}$ is a standard normal random variable,*

$$\Pr\left(\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right]\right)/\sqrt{\mathbb{V}\left[\tilde{\mathbf{G}}_w\right]} \leq t\right) \to \Pr\left(\mathbf{S} \leq t\right).$$

3.3 Numerical Validation

We validate our theory on the MSE convergence rates by estimating the Rényi-α divergence integral between two truncated multivariate Gaussian distributions with varying dimension and sample sizes. The densities have means $\bar{\mu}_1 = 0.7 * \bar{1}_d$, $\bar{\mu}_2 = 0.3 * \bar{1}_d$ and covariance matrices $0.1 * I_d$ where $\bar{1}_d$ is a d-dimensional vector of ones,

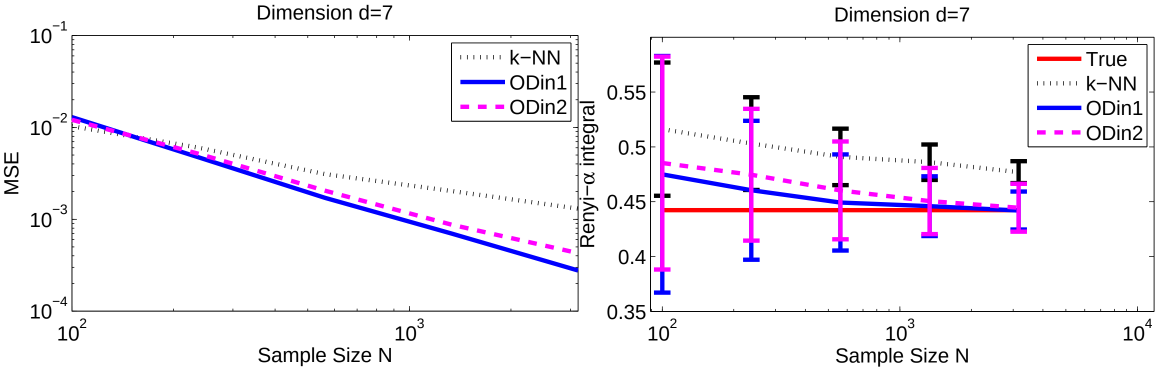

Figure 3.1: (Left) Log-log plot of MSE of the k-nn plug-in estimator ("k-NN") and the two proposed optimally weighted estimators (ODin1 and ODin2) for $d = 7$. (Right) Plot of the average value of the same estimators with standard error bars compared to the true value being estimated. The proposed weighted ensemble estimators outperform the plug-in estimator.

and I_d is a $d \times d$ identity matrix. We used $\alpha = 0.5$ and restricted the Gaussians to the unit cube.

The left plot in Fig. 3.1 shows the MSE (200 trials) of the standard plug-in k-nn estimator where $k = \sqrt{N}$ and the two proposed optimally weighted estimators ODin1 and ODin2. We show the case where $d = 7$ and the sample size varies. For the ODin1 estimator, we chose $\bar{\ell}$ to be linearly spaced between 0.3 and 3 with $L = 50$. For the ODin2 estimator, we chose the minimum value of $\bar{\ell}$ to be 1.4 and then chose the next 24 values for k (i.e. $L = 25$). Note that in comparison to Fig. 2.3, the plug-in k-nn estimator does better than the plug-in uniform kernel plug-in estimator. This is likely due to the adaptive nature of the k-nn estimator. Additionally, both ODin1 and ODin2 outperform both plug-in estimators which validates our theory.

3.4 Estimation of Bounds on the Bayes Error

3.4.1 The Bayes Error

Consider the problem of classifying a feature vector x into one of two classes C_1 or C_2. Denote the *a priori* class probabilities as $q_1 = \Pr(C_1) > 0$ and $q_2 = \Pr(C_2) =$

$1-q_1 > 0$. The conditional densities of x given that x belongs to C_1 or C_2 are denoted by $f_1(x)$ and $f_2(x)$, respectively, and the Bayes classifier assigns x to C_1 if and only if $q_1 f_1(x) > q_2 f_2(x)$. If $p(x) = q_1 f_1(x) + q_2 f_2(x)$, the average error rate of this classifier, known as the BER, is

$$\begin{aligned} P_e^* &= \int \min\left(\Pr\left(C_1|x\right), \Pr\left(C_2|x\right)\right) p(x)dx \\ &= \int \min\left(q_1 f_1(x), q_2 f_2(x)\right) dx. \end{aligned} \tag{3.5}$$

The BER is the minimum classification error rate that can be achieved by any classifier on x's feature space [86].

3.4.2 f-Divergence Bounds

While the expression for the BER is a divergence functional, the min function is not differentiable everywhere. Thus the theory derived in this and the previous chapter does not apply. However, multiple upper and lower bounds on the BER related to smooth divergence functionals exist. A classical bound is the Chernoff bound [33]. It is derived from the fact that for $a,\ b > 0$, $\min(a,b) \leq a^\alpha b^{1-\alpha}\ \forall \alpha \in (0,1)$. Replacing the minimum function in (3.5) with this bound gives

$$P_e^* \leq q_1^\alpha q_2^{1-\alpha} c_\alpha(f_1, f_2), \tag{3.6}$$

where $c_\alpha(f_1, f_2) = \int f_1^\alpha(x) f_2^{1-\alpha}(x)dx$ is the Chernoff α-coefficient. The Chernoff coefficient is found by minimizing the right hand side of (3.6) with respect to α:

$$c^*(f_1, f_2) = \min_{\alpha \in (0,1)} \int f_1^\alpha(x) f_2^{1-\alpha}(x)dx. \tag{3.7}$$

Combining this with (3.6) gives an upper bound on the BER.

In general, the Chernoff bound is not very tight. A tighter bound was presented in [15]. Consider the following quantity:

$$\begin{aligned} \tilde{D}_{q_1}(f_1, f_2) &= 1 - 4q_1q_2 \int \frac{f_1(x)f_2(x)}{q_1f_1(x) + q_2f_2(x)}dx && (3.8) \\ &= \int \frac{\left(q_1f_1(x) - q_2f_2(x)\right)^2}{q_1f_1(x) + q_2f_2(x)}dx. && (3.9) \end{aligned}$$

It was shown in [15] that the BER P_e^* is bounded above and below as follows:

$$\frac{1}{2} - \frac{1}{2}\sqrt{\tilde{D}_{q_1}(f_1, f_2)} \leq P_e^* \leq \frac{1}{2} - \frac{1}{2}\tilde{D}_{q_1}(f_1, f_2).$$

Arbitrarily tight upper and lower bounds to the BER were given in [6]. We consider only the lower bound here. Define

$$g_\alpha(f_1, f_2) = \ln\left(\frac{1 + e^{-\alpha}}{\exp\left(\frac{-\alpha q_1 f_1(x)}{p(x)}\right) + \exp\left(\frac{-\alpha q_2 f_2(x)}{p(x)}\right)}\right),$$

where $p(x) = q_1f_1(x) + q_2f_2(x)$ as before and $\alpha > 0$. Then the BER is bounded below as

$$P_e^* \geq \frac{1}{\alpha}\int g_\alpha(f_1, f_2)p(x)dx =: G_\alpha(f_1, f_2). \tag{3.10}$$

The functionals in (3.6) and (3.8)-(3.10) all contain the form of

$$D_\phi(f_1, f_2) = \int \phi\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x)dx. \tag{3.11}$$

To see this, note that for the Chernoff α coefficient, $\phi(t) = t^\alpha$. For the $\tilde{D}_{q_1}$ based bounds, the functions are more complicated with $\phi(t) = \frac{4q_1q_2t}{q_2+q_1t}$ and $\phi(t) = \frac{(q_1t-q_2)^2}{q_1t+q_2}$ for (3.8) and (3.9), respectively. The functions are even more complex for (3.10).

However, if $t = \frac{f_1(x)}{f_2(x)}$, then

$$\begin{aligned} \exp\left(\frac{-\alpha q_1 f_1(x)}{p(x)}\right) &= \exp\left(\frac{-\alpha q_1}{q_1 + q_2 t^{-1}}\right), \\ \exp\left(\frac{-\alpha q_2 f_2(x)}{p(x)}\right) &= \exp\left(\frac{-\alpha q_2}{q_2 + q_1 t}\right). \end{aligned}$$

Substituting these expressions into $G_\alpha(f_1, f_2)$ gives the required form. Thus we can use the optimally weighted ensemble divergence estimator from this chapter to estimate all of these bounds on the Bayes error. To estimate $c^*(f_1, f_2)$, we estimate $c_\alpha(f_1, f_2)$ for multiple values of α (e.g. $0.01, 0.02, \dots, 0.99$) and choose the minimum.

3.4.3 Simulations

We use the ODin1 k-nn estimator. In addition, we use an alternate estimator for $\tilde{D}_{q_1}$ based on an extension of the Friedman-Rafsky (FR) multivariate two sample test statistic for comparison [60]. This estimator is derived from the MST of the combined data set $\mathcal{X}_T \cup \mathcal{Y}_M$ and does not require direct estimation of the densities f_1 and f_2 [14, 15]. However, the convergence rate and asymptotic distribution of this estimator are currently unknown.

To compare the estimation performance of the various bounds on the BER, we consider 200 trials of two samples from two Gaussian distributions with unit variance and varying mean. In practice, we use a leave one out approach for the weighted k-nn estimator and so the number of samples from both distributions is equal to T. In the first experiment, we fix the dimension $d = 5$ and vary the number of samples from each distribution. Figure 3.2 shows the cases where $T = 5000$ and 50. We choose $\alpha = 500$ for G_α. In the large sample regime, the bounds vary smoothly as the separation between the means of the distributions increases. The two methods for estimating $\tilde{D}_{q_1}$ have nearly identical results when Eq. 3.8 is used for the weighted k-nn method. If Eq. 3.9 is used, then the estimated bounds (not shown) are inaccurate. This

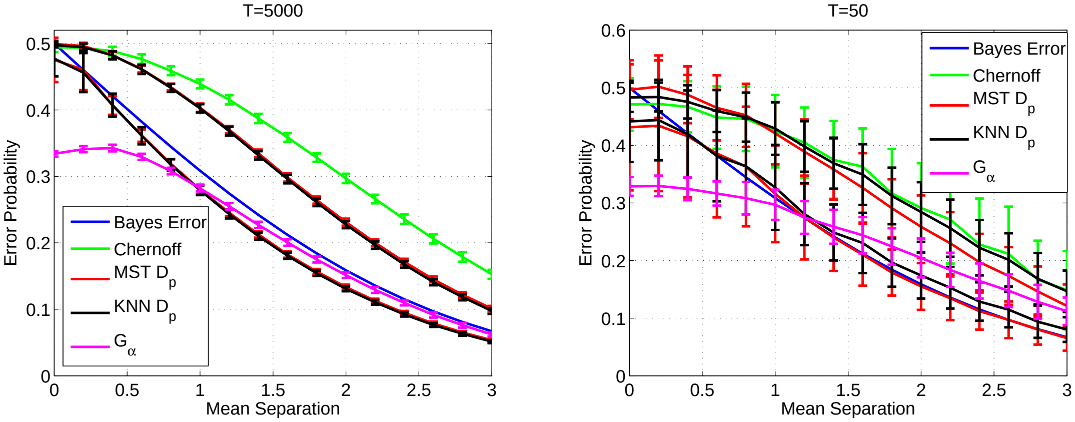

Figure 3.2: Estimated bounds on the Bayes error rate for two unit variance Gaussians with dimension $d = 5$, varying sample sizes ($T = 5000$, 50), and varying means over 200 trials. Error bars correspond to a single standard deviation. The $\tilde{D}_{q_1}$ based lower bounds are close to the actual Bayes error for both the large and small sample regimes but are much more variant with a smaller sample size. The arbitrarily tight lower bound (G_α with $\alpha = 500$) is very close to the Bayes error when $T = 5000$ and when the Bayes error is low.

underscores the importance of using an appropriate representation of the function ϕ when using plug-in based estimation methods as numerical errors may lead to varying results.

In the low sample regime, the estimates have much higher variance and are more biased as the lower bounds often cross the Bayes error. However, the $\tilde{D}_{q_1}$ based lower bounds are still fairly close to the true BER and are thus valuable for assessing the potential performance of a given feature space. Increasing the sample size to as little as 150 greatly improves the performance (not shown).

In the second experiment, we fixed the number of samples at $T = 1000$ and varied the dimension. The results for $d = 1$ and 10 are given in Fig. 3.3. In the higher dimension, the $\tilde{D}_{q_1}$ lower bounds are closer to the BER which results in these estimates crossing over the BER more often. The variance in all of the estimates is also higher when $d = 10$.

Several trends are apparent in both Figs. 3.2 and 3.3. One is that the variance of the $\tilde{D}_{q_1}$ lower bounds decreases as the BER decreases. In general, the MST-based

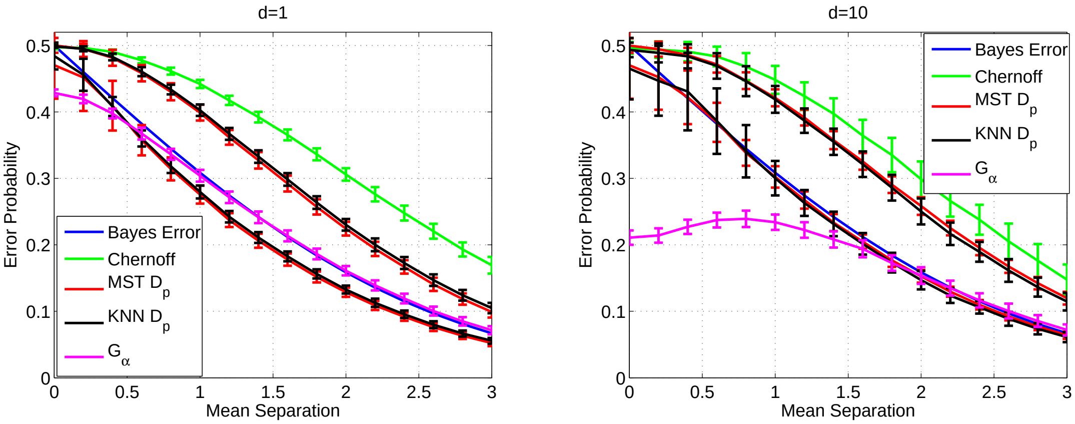

Figure 3.3: Estimated bounds on the Bayes error rate for two unit variance Gaussians with varying dimension ($d = 1,\ 10$) and a fixed sample size of $T = 1000$ over 200 trials. The estimated $\tilde{D}_{q_1}$ based bounds are more biased and variant when the dimension is higher.

estimator is more variant than the k-nn estimator except when the dimension or number of samples is high (e.g. $d = 10$ or $T = 5000$). This is not a substantial problem as an accurate estimate of the BER is less useful at higher values. This is because if the BER is around 0.4, then the feature space being considered does not improve the classification much beyond random guessing. Thus time and energy may be better spent on finding a new feature space for the problem instead of attempting to achieve the BER on the given feature space.

Another observation is that for $d > 1$, the G_α based lower bound is not tight for higher BER when using $\alpha = 500$. Increasing α does not substantially improve the tightness at these values due to numerical precision errors. However, it may be possible to manipulate the expression for g_α so that this is not an issue.

Overall, these results suggest that estimating the $\tilde{D}_{q_1}$ lower bound provides a value that is fairly close to the true BER. The weighted k-nn estimator appears to be less variant than the MST based estimator except when the dimension or number of samples is sufficiently high. Thus we recommend using the $\tilde{D}_{q_1}$ bounds to estimate the location of the BER. If this gives a range for the BER that is low (approximately less than 0.2) and there are enough samples, then G_α may be estimated for a more

	Setosa-Versicolor	Setosa-Virginica	Versicolor-Virginica
Estimated Confidence Interval	$(0, 0.0013)$	$(0, 0.0002)$	$(0, 0.0726)$
QDA Misclassification Rate	0	0	0.04

Table 3.1: Estimated 95% confidence intervals for the bound on the pairwise Bayes error and the misclassification rate of a QDA classifier with 5-fold cross validation applied to the Iris dataset. The right endpoint of the confidence intervals is nearly zero when comparing the Setosa class to the other two classes while the right endpoint is much higher when comparing the Versicolor and Virginica classes. This is consistent with the QDA performance and the fact that the Setosa class is linearly separable from the other two classes.

precise estimate of the BER. Similar results are obtained for truncated Gaussians.

3.4.4 Application to the Iris Dataset

We use the optimally weighted ensemble estimator to obtain confidence intervals on the Chernoff bound of the Iris data set from the UCI machine learning repository [7, 59]. We estimated a bound on the pairwise Bayes error between the three classes (Setosa, Versicolor, and Virginica) and used bootstrapping to calculate confidence intervals. We compared the bounds to the performance of a quadratic discriminant analysis classifier (QDA) with 5-fold cross validation. The pairwise estimated 95% confidence intervals and the misclassification rates of the QDA are given in Table 3.1. Note that the right endpoint of the confidence interval is less than 1/50 when comparing the Setosa class to either of the other two classes. This is consistent with the performance of the QDA and the fact that the Setosa class is linearly separable from the other two classes. In contrast, the right endpoint of the confidence interval is higher when comparing the Versicolor and Virginica classes which are not linearly separable. This is also consistent with the QDA performance. Thus the estimated bounds provide a measure of the relative difficulty of distinguishing between the classes, even though the small number of samples for each class (50) limits the accuracy of the estimated bounds.

3.5 Conclusion

In this chapter, we derived convergence rates for a k-nearest neighbor plug-in estimator of divergence functionals. We applied the generalized theory of optimally weighted ensemble estimation derived previously to derive an estimator that achieves the parametric rate when the densities belong to the Hölder smoothness class with smoothness parameter greater than $d/2$. The estimators we derive apply when the densities have support $[0,1]^d$ although they do not require knowledge of the support. We also derived the asymptotic distribution of the estimator and showed how certain divergence functionals can be used to estimate bounds on the Bayes error for a classification problem.

CHAPTER IV

Extension to Mutual Information Estimation

This chapter extends the work on estimating divergence functionals in Chapters II and III to the problem of estimating general mutual information measures. We focus on two cases: 1) the data have purely continuous components; 2) the data have a mixture of discrete and continuous components. Section 4.1 focuses on the first case while Section 4.2 focuses on the second.

4.1 Mutual Information Estimation with KDEs: Continuous Random Variables

We obtain mutual information estimators by modifying the general divergence functional estimators developed in Chapters II and III, which focused on estimating functionals of two distributions of the form:

$$G\left(f_1, f_2\right) = \int g\left(f_1(x), f_2(x)\right) f_2(x) dx, \tag{4.1}$$

where f_1 and f_2 are smooth d-dimensional probability densities and $g(t_1, t_2)$ is a smooth functional. To derive a general class of mutual informations from (4.1), let $f_X(x)$, $f_Y(y)$, and $f_{XY}(x,y)$ be d_X, d_Y, and $d_X + d_Y = d$-dimensional densities. Let $g(t_1, t_2) = g\left(\frac{t_1}{t_2}\right)$. Then $G(\mathbf{X};\mathbf{Y}) := G\left(f_X \cdot f_Y, f_{XY}\right)$ defines a family of mutual informations:

$$G(\mathbf{X};\mathbf{Y}) = \int g\left(\frac{f_X(x)f_Y(y)}{f_{XY}(x,y)}\right) f_{XY}(x,y)dxdy. \tag{4.2}$$

Throughout this section, we assume that all of the components of $\mathbf{X}$ and $\mathbf{Y}$ lie in some continuous space.

4.1.1 The KDE Plug-in Estimator

When both $\mathbf{X}$ and $\mathbf{Y}$ are continuous random variables or vectors with marginal densities f_X and f_Y, the mutual information functional $G(\mathbf{X};\mathbf{Y})$ can be estimated using KDEs. Let $\mathcal{S}_X$ be the support of f_X and $\mathcal{S}_Y$ the support of f_Y. Assume that N i.i.d. samples $\{\mathbf{Z}_1,\dots,\mathbf{Z}_N\}$ are available from the joint density f_{XY} with $\mathbf{Z}_i = (\mathbf{X}_i,\mathbf{Y}_i)^T$. Let $M = N-1$ and let h_X, h_Y be kernel bandwidths. Let $K_X(\cdot)$ and $K_Y(\cdot)$ be kernel functions with $||K_X||_\infty$, $||K_Y||_\infty < \infty$ where $||K||_\infty = \sup_x |K(x)|$. The KDEs for f_X and f_Y are

$$\begin{aligned}
\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_j) &= \frac{1}{Mh_X^{d_X}}\sum_{\substack{i=1\\ i\neq j}}^{N} K_X\left(\frac{\mathbf{X}_j-\mathbf{X}_i}{h_X}\right), && (4.3)\\
\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_j) &= \frac{1}{Mh_Y^{d_Y}}\sum_{\substack{i=1\\ i\neq j}}^{N} K_Y\left(\frac{\mathbf{Y}_j-\mathbf{Y}_i}{h_Y}\right). && (4.4)
\end{aligned}$$

To estimate the joint distribution f_{XY}, we use the product kernel:

$$\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X}_j,\mathbf{Y}_j) = \frac{1}{Mh_X^{d_X}h_Y^{d_Y}}\sum_{\substack{i=1\\ i\neq j}}^{N} K_X\left(\frac{\mathbf{X}_j-\mathbf{X}_i}{h_X}\right) K_Y\left(\frac{\mathbf{Y}_j-\mathbf{Y}_i}{h_Y}\right), \tag{4.5}$$

where $h_Z = (h_X, h_Y)$. The functional $G(\mathbf{X};\mathbf{Y})$ is then estimated as

$$\tilde{\mathbf{G}}_{h_X,h_Y} = \frac{1}{N}\sum_{i=1}^{N} g\left(\frac{\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_i)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_i)}{\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X}_i,\mathbf{Y}_i)}\right). \tag{4.6}$$

4.1.2 Convergence Rates of the Plug-in Estimator

To derive the convergence rate of $\tilde{\mathbf{G}}_{h_X,h_Y}$, we make similar assumptions about the densities and the functional g as in Chapter II. The full assumptions are:

- $(\mathcal{D}.0)$: Assume that the kernels K_X and K_Y are symmetric product kernels and have bounded support in each dimension.

- $(\mathcal{D}.1)$: Assume there exist constants $\epsilon_0, \epsilon_\infty$ such that $0 < \epsilon_0 \leq f_X(x) \leq \epsilon_\infty < \infty$, $\forall x \in \mathcal{S}_X$, $\epsilon_0 \leq f_X(x) \leq \epsilon_\infty$, $\forall y \in \mathcal{S}_Y$, and $\epsilon_0 \leq f_{XY}(x,y) \leq \epsilon_\infty$, $\forall (x,y) \in \mathcal{S}_X \times \mathcal{S}_Y$.

- $(\mathcal{D}.2)$: Assume that each of the densities belong to $\Sigma(s,K)$ in the interior of their support sets with $s \geq 2$.

- $(\mathcal{D}.3)$: Assume that $g\left(t_1/t_2\right)$ has an infinite number of mixed derivatives wrt t_1 and t_2.

- $(\mathcal{D}.4)$: Assume that $\left|\frac{\partial^{k+l} g(t_1,t_2)}{\partial t_1^k \partial t_2^l}\right|$, $k,l = 0,1,\ldots$ are strictly upper bounded for $\epsilon_0 \leq t_1, t_2 \leq \epsilon_\infty$.

- $(\mathcal{D}.5)$: Assume the following boundary smoothness condition: Let K be either K_X or K_Y, $\mathcal{S}$ either $\mathcal{S}_X$ or $\mathcal{S}_Y$, h either h_X or h_Y. Let $p_x(u) : \mathbb{R}^d \rightarrow \mathbb{R}$ be a polynomial in u of order $q \leq r = \lfloor s \rfloor$ whose coefficients are a function of x and are $r-q$ times differentiable. Then assume that

$$\int_{x \in \mathcal{S}} \left(\int_{u:K(u)>0,\, x+uh \notin \mathcal{S}} K(u) p_x(u) du \right)^t dx = v_t(h),$$

  where $v_t(h)$ admits the expansion

$$v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o\left(h^{r-q}\right),$$

for some constants $e_{i,q,t}$.

For a discussion on these assumptions, see Section 2.1.2.

Theorem IV.1. *Under assumptions $\mathcal{D}.0-\mathcal{D}.5$ and for general g, the bias of $\tilde{\mathbf{G}}_{h_X,h_Y}$ is*

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_Y}\right] &= \sum_{\substack{j=0\\ i+j\neq 0}}^{r}\sum_{i=0}^{r} c_{10,i,j}h_X^i h_Y^j + \frac{c_{11,X}}{Nh_X^{d_X}} + \frac{c_{11,Y}}{Nh_Y^{d_Y}} \\ &\quad +O\left(h_X^s + h_Y^s + \frac{1}{Nh_X^{d_X}} + \frac{1}{Nh_Y^{d_Y}}\right). \qquad (4.7)\end{aligned}$$

If $g(t_1,t_2)$ has j,l-th order mixed derivatives $\frac{\partial^{j+l}}{\partial t_1^j \partial t_2^l}$ that depend on t_1 and t_2 only through $t_1^\alpha t_2^\beta$ for some $\alpha,\beta\in\mathbb{R}$, then for any positive integer $\lambda\geq 2$, the bias is

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_Y}\right] &= \sum_{\substack{j=0\\ i+j\neq 0}}^{r}\sum_{i=0}^{r} c_{10,i,j}h_X^i h_Y^j + \sum_{j=1}^{\lambda/2}\sum_{i=1}^{\lambda/2}\sum_{m=0}^{r}\sum_{n=0}^{r} c_{12,j,i,m,n}\frac{h_X^m h_Y^n}{\left(Nh_X^{d_X}\right)^j\left(Nh_Y^{d_Y}\right)^i} \\ &\quad +\sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\sum_{n=0}^{r}\left(c_{11,m,n,j,X}\frac{h_X^m h_Y^n}{\left(Nh_X^{d_X}\right)^j} + c_{11,m,n,j,Y}\frac{h_X^m h_Y^n}{\left(Nh_Y^{d_Y}\right)^j}\right) \\ &\quad +O\left(h_X^s + h_Y^s + \frac{1}{\left(Nh_X^{d_X}\right)^{\lambda/2}} + \frac{1}{\left(Nh_Y^{d_Y}\right)^{\lambda/2}}\right) \qquad (4.8)\end{aligned}$$

The expression in (4.8) enables us to achieve the parametric convergence rate under less restrictive smoothness assumptions on the densities ($s > d/2$ for (4.8) compared to $s \geq d$ for (4.7)). The extra condition required on the mixed derivatives of g to obtain the expression in (4.8) is satisfied, for example, for Shannon and Renyi informations.

Theorem IV.2. *If the functional g is Lipschitz continuous in both of its arguments* *with Lipschitz constant C_g, then the variance of $\tilde{\mathbf{G}}_{h_X,h_Y}$ is*

$$\mathbb{V}\left[\tilde{\mathbf{G}}_{h_X,h_Y}\right] \leq \frac{22C_g^2||K_X \cdot K_Y||_\infty^2}{N}.$$

The proofs of Theorems IV.1 and IV.2 are similar to the proofs of the bias and variance results for the divergence functional estimators in Chapter II. The primary differences deal with the product of the marginal KDEs. See Appendix D for proof sketches.

Theorems IV.1 and IV.2 indicate that for the MSE to go to zero, we require $h_X, h_Y \to 0$ and $Nh_X^{d_X}$, $Nh_Y^{d_Y} \to \infty$. The Lipschitz assumption on g is comparable to other nonparametric estimators of distributional functionals [102, 111, 145, 180, 181]. Specifically, assumption $\mathcal{A}.1$ ensures that functionals such as those for Shannon and Renyi informations are Lipschitz on the space ϵ_0 to ϵ_∞.

4.1.3 Ensemble Estimation of Mutual Information

An ensemble of estimators can be formed by choosing different bandwidth values as in Section 2.2. Choose $\mathcal{L}_X = \{l_X(1), \ldots, l_X(L_X)\}$ and $\mathcal{L}_Y = \{l_Y(1), \ldots, l_Y(L_Y)\}$ to be real positive numbers that index $h_X(l_X(i))$ and $h_Y(l_Y(i))$ over different neighborhood sizes for the KDEs. Define w to be a weight matrix s.t. $w_{ij} = w(l_X(i), l_Y(j))$. Then the weighted ensemble estimator is $\tilde{\mathbf{G}}_w = \sum_{(l,l')\in\mathcal{L}_X\times\mathcal{L}_Y} w(l,l')\tilde{\mathbf{G}}_{h_X(l),h_Y(l')}$.

We use the general theory of optimally weighted ensemble estimation in Theorem II.5 to improve the MSE convergence rate of the mutual information plug-in estimator. For general g, (4.7) indicates that we need $h_X \propto N^{-1/(2d_X)}$ for the $O(1/(Nh_X^{d_X}))$ terms to be $O(1/\sqrt{N})$. Similarly, we require $h_Y \propto N^{-1/(2d_Y)}$. For the ensemble of estimators, we thus choose $h_X(l_X(i)) = l_X(i)N^{-1/(2d_X)}$ and $h_Y(l_Y(j)) = l_Y(j)N^{-1/(2d_Y)}$. From Theorems IV.1 and IV.2, conditions $\mathcal{C}.1$ and $\mathcal{C}.2$ (see Section 2.2.2) are satisfied with $\psi_{i,j}(l_X(m), l_Y(n)) = l_X^i(m)l_Y^j(n)$ and $\phi_{i,j}(N) = N^{-i/(2d_X)-j/(2d_Y)}$

for $0 \leq i \leq d_X$ and $0 \leq j \leq d_Y$ s.t. $0 < i/d_X + j/d_Y \leq 1$. The optimal weight w_0 is calculated using (2.7). The resulting estimator $\tilde{\mathbf{G}}_{w_0,1}$ achieves the parametric MSE rate when $s \geq \max(d_X, d_Y)$. We refer to this estimator as the ODin1 estimator of mutual information.

If the mixed derivatives of the functional g satisfy the extra condition required for (4.8), then we can define an estimator that achieves the parametric MSE rate under less strict smoothness assumptions. Choose $h_X(l_X(i)) = l_X(i)N^{-1/(d_X+\delta)}$ and $h_Y(l_Y(j)) = l_Y(j)N^{-1/(d_Y+\delta)}$. Then conditions $\mathcal{C}.1$ and $\mathcal{C}.2$ are satisfied with $\psi_{i,j,p,q}(l_X(m), l_Y(n)) = l_X^{i-pd_X}(m)l_Y^{j-qd_Y}(n)$ and $\phi_{i,j,p,q}(N) = N^{-(i-pd_X)/(d_X+\delta)-(j-qd_Y)/(d_Y+\delta)-p-q}$ for $i+j+p+q > 0$ and $(i - pd_X)/(d_X + \delta) + (j - qd_Y)/(d_Y + \delta) + p + q \leq 1/2$. The optimal weight w_0 is again calculated using (2.7) and the resulting estimator $\tilde{\mathbf{G}}_{w_0,2}$ achieves the parametric MSE convergence rate when $s \geq (\max(d_X, d_Y) + \delta)/2$. We refer to this estimator as the ODin2 estimator of mutual information.

As for divergence functionals (see Section 2.2.3), the ODin2 estimator has better statistical properties as the parametric convergence rate is guaranteed under less restrictive smoothness assumptions on the densities. On the other hand, the number of parameters required for the optimization problem in (2.7) is larger for the ODin2 estimator than the ODin1 estimator. In theory, this could lead to larger variance. Algorithm 2 summarizes the estimator $\tilde{\mathbf{G}}_{w_0,1}$ for the case when $l_X(i) = l_Y(i)$ (e.g. when the scales for both spaces are similar).

We finish this section with a central limit theorem:

Theorem IV.3. *Assume that the functional g is Lipschitz in both arguments with Lipschitz constant C_g and that h_X, $h_Y = o(1)$, $N \to \infty$, and $Nh_X^{d_X}$, $Nh_Y^{d_Y} \to \infty$. Then for fixed L_X and L_Y, and if $\mathbf{S}$ is a standard normal random variable,*

$$\Pr\left(\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right]\right)/\sqrt{\mathbb{V}\left[\tilde{\mathbf{G}}_w\right]} \leq t\right) \to \Pr\left(\mathbf{S} \leq t\right).$$

Algorithm 2 Optimally weighted ensemble mutual information estimator $\tilde{\mathbf{G}}_{w_0,1}$

```
Input: L positive real numbers $\mathcal{L}$, samples $\{\mathbf{Z}_1, \dots, \mathbf{Z}_N\}$ from $f_{XY}$, dimensions $d_X$ and $d_Y$, function $g$, kernels $K_X$ and $K_Y$
Output: The optimally weighted divergence estimator $\tilde{\mathbf{G}}_{w_0,1}$
Solve for $w_0$ using (2.7) with basis functions $\psi_{i,j}(l) = l^{i+j}$, $\phi_{i,j}(N) = N^{-i/(2d_X)-j/(2d_Y)}$, $l \in \mathcal{L}$, $0 \leq i \leq d_X$, and $0 \leq j \leq d_Y$ s.t. $0 < i/d_X + j/d_Y \leq 1$
for all $l \in \mathcal{L}$ do
    $h_X(l) = lN^{-1/(2d_X)}$, $h_Y(l) = lN^{-1/(2d_Y)}$
    for $i = 1$ to $N$ do
        Calculate $\tilde{\mathbf{f}}_{X,h(l)}(\mathbf{X}_i)$, $\tilde{\mathbf{f}}_{Y,h(l)}(\mathbf{Y}_i)$, and $\tilde{\mathbf{f}}_{Z,h(l)}(\mathbf{X}_i, \mathbf{Y}_i)$ from (4.3), (4.4), and (4.5)
    end for
    $\tilde{\mathbf{G}}_{h_X(l),h_Y(l)} \leftarrow \frac{1}{N}\sum_{i=1}^{N} g\left(\frac{\tilde{\mathbf{f}}_{X,h(l)}(\mathbf{X}_i)\tilde{\mathbf{f}}_{Y,h(l)}(\mathbf{Y}_i)}{\tilde{\mathbf{f}}_{Z,h(l)}(\mathbf{X}_i,\mathbf{Y}_i)}\right)$
end for
$\tilde{\mathbf{G}}_{w_0,1} \leftarrow \sum_{l\in\mathcal{L}} w_0(l)\tilde{\mathbf{G}}_{h_X(l),h_Y(l)}$
```

The proof is based on an application of Slutsky's Theorem preceded by an application of the Efron-Stein inequality that is very similar to the proof of the central limit theorem for the divergence functional ensemble estimators in Chapter II. The extension of the central limit theorem to the mutual information estimator is analagous to the extension required in the proof of the variance result in Theorem IV.2. Due to this similarity, we omit the proof of Theorem IV.3.

4.2 Mutual Information Estimation: Mixed Random Variables

Another important case in mutual information estimation is when $\mathbf{X}$ has only continuous components and $\mathbf{Y}$ has only discrete components. For example, if $\mathbf{Y}$ is a predictor variable (e.g. classification labels), then the mutual information between $\mathbf{X}$ and $\mathbf{Y}$ indicates the value of $\mathbf{X}$ as a predictor of $\mathbf{Y}$. Although $\mathbf{Y}$ is discrete, $f_{XY} = f_Z$ is also a density. Let $\mathcal{S}_X$ be the support of the density f_X and $\mathcal{S}_Y$ be the support of the probability mass function f_Y. The generalized mutual information can be written as

$$\begin{aligned} G\left(\mathbf{X};\mathbf{Y}\right) &= \sum_{y\in\mathcal{S}_Y}\int g\left(\frac{f_X(x)f_Y(y)}{f_{XY}(x,y)}\right)f_{XY}(x,y)dx \\ &= \sum_{y\in\mathcal{S}_Y}f_Y(y)\int g\left(\frac{f_X(x)}{f_{X|Y}(x|y)}\right)f_{X|Y}(x|y)dx. \qquad (4.9)\end{aligned}$$

Let $\mathbf{N}_y=\sum_{i=1}^N 1_{\{\mathbf{Y}_i=y\}}$ where $y\in\mathcal{S}_Y$. Let $\tilde{\mathbf{f}}_{X,h_X}$ be as in (4.3) and define $\mathcal{X}_y=\{\mathbf{X}_i\in\{\mathbf{X}_1,\dots,\mathbf{X}_N\}\,|\mathbf{Y}_i=y\}$. Then if $\mathbf{X}_i\in\mathcal{X}_y$, the KDE of $f_{X|Y}(x|y)$ is

$$\tilde{\mathbf{f}}_{X|y,h_{X|y}}(\mathbf{X}_i) \;=\; \frac{1}{\left(\mathbf{N}_y-1\right)h_{X|y}^{d_X}}\sum_{\substack{\mathbf{X}_j\in\mathcal{X}_y\\ i\neq j}}K_X\left(\frac{\mathbf{X}_i-\mathbf{X}_j}{h_{X|y}^{d_X}}\right).$$

We define the plug-in estimator $\tilde{\mathbf{G}}_{h_X,h_{X|Y}}$ of (4.9) to be

$$\tilde{\mathbf{G}}_{h_X,h_{X|Y}}=\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y}{N}\tilde{\mathbf{G}}_{h_X,h_{X|y}}, \qquad (4.10)$$

where

$$\tilde{\mathbf{G}}_{h_X,h_{X|y}}=\frac{1}{\mathbf{N}_y}\sum_{\mathbf{X}\in\mathcal{X}_y}g\left(\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})/\tilde{\mathbf{f}}_{X|y,h_{X|y}}(\mathbf{X})\right).$$

4.2.1 Convergence Rates

Theorem IV.4. *(Bias) Assume that assumptions $\mathcal{D}.0-\mathcal{D}.5$ apply to the functional g, the kernel K_X, and the densities f_X and $f_{X|Y}$. Assume that $\mathbf{h}_{X|y}=l\mathbf{N}_y^{-\beta}$ with $0<\beta<\frac{1}{d_X}$ and l a positive number. Then the bias of $\tilde{\mathbf{G}}_{h_X,h_{X|Y}}$ is*

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\right] &= \sum_{\substack{j=0\\ i+j\neq 0}}^{r}\sum_{i=0}^{r}c_{13,i,j}h_X^i l^j N^{-j\beta}+\frac{c_{14,X}}{Nh_X^{d_X}}+\frac{c_{14,y}}{l^{d_X}N^{1-\beta d_X}} \\ &\quad +O\left(h_X^s+N^{-s\beta}+\frac{1}{Nh_X^{d_X}}+\frac{1}{N^{1-\beta d_X}}+\frac{1}{N}\right). \qquad (4.11)\end{aligned}$$

Furthermore, if $g(t_1, t_2)$ has j, l-th order mixed derivatives $\frac{\partial^{j+l}}{\partial t_1^j \partial t_2^l}$ that depend on t_1 and t_2 only through $t_1^\alpha t_2^\beta$ for some $\alpha, \beta \in \mathbb{R}$, then for any positive integer $\lambda \geq 2$, the bias is

$$\begin{aligned}
\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X, h_{X|Y}}\right] &= \sum_{\substack{j=0 \\ i+j\neq 0}}^{r} \sum_{i=0}^{r} c_{13,i,j} h_X^i l^j N^{-j\beta} + \sum_{j=1}^{\lambda/2} \sum_{i=1}^{\lambda/2} \sum_{m=0}^{r} \sum_{n=0}^{r} c_{14,j,i,m,n} \frac{h_X^m l^n N^{-n\beta}}{\left(N h_X^{d_X}\right)^j \left(l^{d_X} N^{1-\beta d_X}\right)^i} \\
&\quad + \sum_{j=1}^{\lambda/2} \sum_{m=0}^{r} \sum_{n=0}^{r} \left(c_{14,m,n,j,X} \frac{h_X^m l^n N^{-n\beta}}{\left(N h_X^{d_X}\right)^j} + c_{14,m,n,j,Y} \frac{h_X^m l^n N^{-n\beta}}{\left(l^{d_X} N^{1-\beta d_X}\right)^j} \right) \\
&\quad + O\left(h_X^s + N^{-s\beta} + \frac{1}{\left(N h_X^{d_X}\right)^{\lambda/2}} + \frac{1}{\left(N^{1-\beta d_X}\right)^{\lambda/2}} + \frac{1}{N} \right). \qquad (4.12)
\end{aligned}$$

Proof. For brevity, we only sketch the main ideas of the proof here. See Appendix D for more details. The conditional bias of $\tilde{\mathbf{G}}_{h_X, h_{X|y}}$ given $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ can be obtained from Theorem IV.1. Then given that $\mathbf{h}_{X|y} \propto \mathbf{N}_y^{-\beta}$, (4.10) gives terms of the form of $\mathbf{N}_y^{1-\gamma}$ with $\gamma > 0$. $\mathbf{N}_y$ is a binomial random variable with parameter $f_Y(y)$, N trials, and mean $N f_Y(y)$. Thus we need to compute the fractional moments of a binomial random variable. By the generalized binomial theorem, we have that

$$\begin{aligned}
\mathbf{N}_y^\alpha &= \left(\mathbf{N}_y - N f_Y(y) + N f_Y(y)\right)^\alpha \\
&= \sum_{i=0}^{\infty} \begin{pmatrix} \alpha \\ i \end{pmatrix} \left(N f_Y(y)\right)^{\alpha - i} \left(\mathbf{N}_y - N f_Y(y)\right)^i, \\
\implies \mathbb{E}\left[\mathbf{N}_y^\alpha\right] &= \sum_{i=0}^{\infty} \begin{pmatrix} \alpha \\ i \end{pmatrix} \left(N f_Y(y)\right)^{\alpha - i} \mathbb{E}\left[\left(\mathbf{N}_y - N f_Y(y)\right)^i\right]. \qquad (4.13)
\end{aligned}$$

From Riordan [170], the i-th central moment of $\mathbf{N}_y$ has the form of

$$\mathbb{E}\left[\left(\mathbf{N}_Y - N f_Y(y)\right)^i\right] = \sum_{n=0}^{\lfloor i/2 \rfloor} c_{n,i}(f_Y(y)) N^n.$$

Thus $\mathbb{E}\left[\mathbf{N}_y^{1-\gamma}\right]$ has terms proportional to $N^{1-\gamma-i+n} \leq N^{1-\gamma-\lfloor i/2 \rfloor}$ for $i = 0, 1, \ldots$ since $n \leq \lfloor i/2 \rfloor$. Then since there is an N in the denominator of (4.10), this leaves terms of the form of $N^{-\gamma}$ when $i = 0, 1$ and N^{-1} for $i \geq 2$. This completes the proof for the bias. □

Theorem IV.5. *If the functional g is Lipschitz continuous in both of its arguments with Lipschitz constant C_g and $\mathcal{S}_Y$ is finite, then the variance of $\tilde{\mathbf{G}}_{h_X, h_{X|Y}}$ is $O(1/N)$.*

Proof. We again sketch the main ideas for brevity. By the law of total variance, we have

$$\mathbb{V}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\right] = \mathbb{E}\left[\mathbb{V}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\middle|\mathbf{Y}_1,\ldots,\mathbf{Y}_N\right]\right] + \mathbb{V}\left[\mathbb{E}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\middle|\mathbf{Y}_1,\ldots,\mathbf{Y}_N\right]\right]. \tag{4.14}$$

From Theorem IV.2, we know that $\mathbb{V}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\middle|\mathbf{Y}_1,\ldots,\mathbf{Y}_N\right] = O\left(\sum_{y\in\mathcal{S}_Y}\mathbf{N}_y/N^2\right)$. Taking the expectation then yields $O(1/N)$.

For the second term, we know from the proof of Theorem IV.4 that $\mathbb{E}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\middle|\mathbf{Y}_1,\ldots,\mathbf{Y}_N\right]$ yields a sum of terms of the form of $\mathbf{N}_y^\gamma/N$ for $0 < \gamma \leq 1$. Taking the variance of the sum of these terms yields a sum of terms of the form $\mathbb{V}\left[\mathbf{N}_y^\gamma\right]/N^2$. Note that the covariance terms can be bounded by the Cauchy Schwarz inequality to yield similar terms. The variance term can be bounded by taking a Taylor series expansion of the functions $\mathbf{N}_y^\gamma$ and $\mathbf{N}_y^{2\gamma}$ at the point $Nf_Y(y)$ which yields an expression that depends on the central moments of $\mathbf{N}_y$. The variance can then be calculated from these equations to obtain $\mathbb{V}\left[\mathbf{N}_y^\gamma\right] = O(N)$ which completes the proof. See Appendix D for more details. □

4.2.2 Ensemble Estimation

Given Theorems IV.4 and IV.5, we can obtain ensemble estimators for $G(\mathbf{X};\mathbf{Y})$ that achieve the parametric rate when either $s \geq d_X$ (general g) or $s \geq (d_X+\delta)/2$ (g has mixed derivatives of the form specified in Theorem IV.4). For the former, choose

$h_X \propto N^{-1/(2d_X)}$ and $\mathbf{h}_{X|y} \propto \mathbf{N}_y^{-1/(2d_X)}$ and for the latter, choose $h_X \propto N^{-1/(d_X+\delta)}$ and $\mathbf{h}_{X|y} \propto \mathbf{N}_y^{-1/(d_X+\delta)}$. For the most general case, choose $\mathcal{L}_X = \{l_X(1), \ldots, l_X(L_X)\}$ and $\mathcal{L}_{X|y} = \left\{l_{X|y}(1), \ldots, l_{X|y}(L_{X|y})\right\}$ to be real positive numbers that index $h_X(l_X(i))$ and $\mathbf{h}_{X|y}(l_{X|y})$ for all $y \in \mathcal{S}_Y$. For simplicity of exposition, we assume that $\mathcal{L}_{X|y} = \mathcal{L}_{X|Y}$ for each $y \in \mathcal{S}_Y$. The weighted ensemble estimator is then

$$\tilde{\mathbf{G}}_w = \sum_{(l,l')\in\mathcal{L}_X\times\mathcal{L}_{X|Y}} w(l,l') \sum_{y\in\mathcal{S}_Y} \frac{\mathbf{N}_y}{N} \tilde{\mathbf{G}}_{h_X(l),h_{X|y}(l')}.$$

To construct the estimator $\tilde{\mathbf{G}}_{w_0,1}$, set $h_X(l_X(i)) = l_X(i)N^{-1/(2d_X)}$ and $\mathbf{h}_{X|y}(l_{X|Y}(i)) = l_{X|Y}(i)\mathbf{N}_y^{-1/(2d_X)}$. w_0 can then be obtained by solving (2.7) with $\psi_{i,j}(l_X, l_{X|Y}) = l_X^i l_{X|Y}^j$ and $\phi_{i,j}(N) = N^{-(i+j)/(2d_X)}$ for $i, j \in \{0, 1, \ldots, d_X\}$ and $i + j \neq 0$. The resulting estimator $\tilde{\mathbf{G}}_{w_0,1}$ has a MSE rate of $O(1/N)$ as long as $s \geq d$.

A similar procedure as above gives $\tilde{\mathbf{G}}_{w_0,2}$ when $h_X \propto N^{-1/(d_X+\delta)}$ and $\mathbf{h}_{X|y} \propto \mathbf{N}_y^{-1/(d_X+\delta)}$. Furthermore, if the space $\mathcal{S}_Y$ is finite, then the ensemble estimators obey a central limit theorem.

4.3 Experimental Validation

In this section, we validate our theory by estimating the Rényi-α mutual information integral (i.e. $g(x) = x^\alpha$ in (4.9); see *Principe* [165]) where $\mathbf{X}$ is a mixture of truncated Gaussian random variables and $\mathbf{Y}$ is a categorical random variable with three possible outcomes and respective probabilities $\Pr(\mathbf{Y} = 0) = \Pr(\mathbf{Y} = 1) = 2/5$ and $\Pr(\mathbf{Y} = 2) = 1/5$. The conditional covariance matrices are all $0.1 * I_d$ and the conditional means are, respectively, $\bar{\mu}_0 = 0.25 * \bar{1}_d$, $\bar{\mu}_0 = 0.75 * \bar{1}_d$, and $\bar{\mu}_0 = 0.5 * \bar{1}_d$, where I_d is the $d \times d$ identity matrix and $\bar{1}_d$ is a d-dimensional vector of ones. We chose $\alpha = 0.5$ and $d = 6$.

Figure 4.1 shows the MSE (125 trials) of the plug-in KDE estimator of mutual information using a uniform kernel and the two optimally weighted ensemble esti-

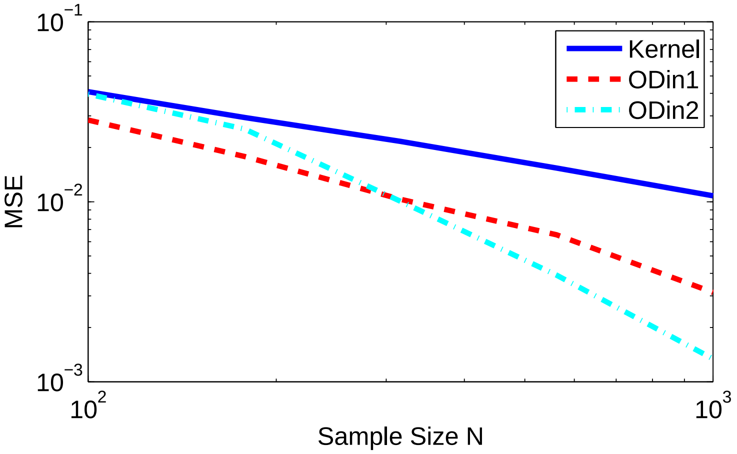

Figure 4.1: MSE log-log plot as a function of sample size for the uniform kernel plug-in mutual information estimator ("Kernel") and the two proposed optimally weighted ensemble estimators $\tilde{\mathbf{G}}_{w_0,1}$ ("ODin1") and $\tilde{\mathbf{G}}_{w_0,2}$ ("ODin2") for the distributions described in the text. The ensemble estimators generally outperform the kernel plug-in estimator, especially for larger sample sizes.

mators $\tilde{\mathbf{G}}_{w_0,1}$ and $\tilde{\mathbf{G}}_{w_0,2}$ for various sample sizes. For $\tilde{\mathbf{G}}_{w_0,2}$, we chose $\delta = 1$. Both ensemble estimators generally outperform the standard plug-in estimator, especially for larger sample sizes. $\tilde{\mathbf{G}}_{w_0,2}$ performs the best at higher sample sizes while $\tilde{\mathbf{G}}_{w_0,1}$ performs the best at lower sample sizes.

4.4 Conclusion

We derived the MSE convergence rates for plug-in KDE-based estimators of mutual information measures between $\mathbf{X}$ and $\mathbf{Y}$ when they have only continuous components. Using these rates, we defined ensemble estimators that achieve an MSE rate of $O(1/N)$ when the densities are sufficiently smooth and showed that a central limit theorem also holds. We then extended this theory to the case where $\mathbf{Y}$ is discrete and $\mathbf{X}$ is continuous. To the best of our knowledge, this is the first nonparametric mutual information estimator that achieves the MSE convergence rate of $O(1/N)$ in this setting.

CHAPTER V

Application to Sunspot Images: Dimensionality Reduction

The remainder of this thesis focuses on applications of nonparametric distributional functional estimation. This chapter and the next focus on applications to sunspot and active region images.

5.1 Background

Active regions (AR) in the solar atmosphere have intense and intricate magnetic fields that emerge from subsurface layers to form loops which extend into the corona. When active regions undergo external forcing such as flux emergence and rearrangement, the system may destabilize. The stored magnetic energy is then suddenly released as accelerated particles (electrons, protons, ions) and an increase in radiation called a *flare* is observed across the entire electromagnetic spectrum [160].

The morphology of sunspots is correlated with flare occurrence and has therefore received a lot of attention. The Mount Wilson classification scheme [81] groups sunspots into four main classes based on the magnetic structure, that is, on the relative locations and sizes of concentrations of opposite polarity magnetic flux. The sunspots with simplest morphology belong to the unipolar α and the bipolar β groups.

Class	Classification Rule	Number of AR	Number of Patches
α	A single dominant spot	50	13,358
β	A pair of dominant spots of opposite polarity	192	75,463
$\beta\gamma$	A β sunspot where a single north-south polarity inversion line cannot divide the two polarities	130	95,631
$\beta\gamma\delta$	A $\beta\gamma$ sunspot where umbrae of opposite polarity are together in a single penumbra	52	66,195

Table 5.1: Mount Wilson classification rules, number of each AR, and total number of joint patches or pixels per Mt. Wilson class used in this chapter when using the STARA masks.

More complex morphologies are described as $\beta\gamma$ when a bipolar sunspot is such that a single north-south polarity inversion line cannot divide the two polarities. When a $\beta\gamma$ sunspot group contains in addition a δ spot, that is, umbrae of different polarities inside a single penumbra, it is labeled as a $\beta\gamma\delta$ group. The presence of a δ configuration, where large values of opposite polarity exist close together, was identified as a warning of the build up of magnetic energy stress with an increased probability of a large flare [132, 173]. See Table 5.1 for a summary of the Mount Wilson classes.

McIntosh [135] proposes another classification scheme containing 60 classes, thus describing the magnetic structure in greater details. The McIntosh classification is the basis for several flare forecasting methods which estimate the flare occurrence rate from historical records of flares and active region classes [26], possibly combining such information with observed waiting time distribution between flares [24, 66].

The McIntosh and Mount Wilson classifications are in general carried out manually, and this results in inconsistencies that stem from human observation bias as well as non-reproducible catalogs. To overcome these caveats, some supervised machine learning methods have been proposed to automatically classify sunspot groups according to these schemes. [188] extract various measurements from continuum and magnetogram images, and then feed these into a machine learning classifier which reproduces the Mount Wilson classification. [35] employ neural networks and supervised classification techniques to reproduce the McIntosh scheme and use those results in a flare forecasting system [36]. While these approaches reduce the human bias, they do not use the information present in sunspot images in an optimal way and make the study of AR dynamic behavior impractical.

Several attempts were made to find quantitative descriptors of an active region's complexity. [133] showed that fractal dimension of an active region alone cannot distinguish between the various Mount Wilson classes. The generalization to multifractal spectrum, where each scale has its own fractal dimension, allowed to study in greater details the evolution of active region in view of distinguishing between quiet and flare-productive active regions. Box counting method [1, 38, 70] as well as more accurate methods based on continuous wavelet transform [39, 105] were employed. Continuous wavelet transforms and energy spectrum were also used with a similar purpose in [91, 134].

Wavelet basis functions act as a microscope to describe local discontinuities and gradients in an image, and [98] used two multiresolution analyses to compute at various length scales the gradients of the magnetic field along lines separating opposite polarities. Using a data set of about 10 000 magnetogram images, they showed that, at all length scales, those gradients increase going from α to β, $\beta\gamma$, and $\beta\gamma\delta$ classes.

However, a wavelet analysis is known to generate artifacts due to the particular shape of the specific wavelet functions. Signal representations based on a set of redundant functions called a *dictionary*, were therefore introduced [130]. [54] proposed the use of a small sized dictionary to find a sparse representation of *patches*. Specifically, a patch is a $m \times m$-pixel neighborhood, and a patch analysis of a n-pixel image will process the $m^2 \times n$ data matrix that collects the overlapping patches. See Figure 6.2 for a representation.

As an example of image patch analysis, [54] considered the problem of denoising an image corrupted by additive Gaussian noise. They computed a sparse representation of patches over a dictionary, thus effectively denoising the patches. The dictionary

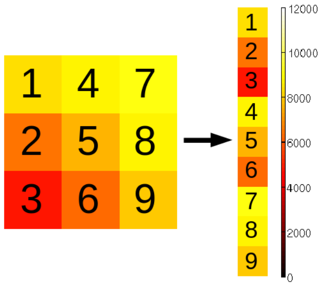

Figure 5.1: An example patch from the edge of a sunspot in a continuum image and its column representation.

itself may either be fixed *a priori* or *learned* from the corrupted patches. An estimate of the noise-free image is then obtained by averaging the denoised overlapping patches. [54] showed that dictionary learning methods based on patch analysis are more flexible and provide superior results in the context of image denoising.

In this chapter, we carry out a patch analysis of a set of sunspots and active region magnetogram images that span the four main Mount Wilson classes. We estimate the intrinsic dimension of the local patches, and show how it relates to the Mount Wilson classification. We also study patterns of local correlation using partial correlation and canonical correlation analysis, which reveal some characteristics of simple and more complex active regions. Such analysis also serves as a preparation to an unsupervised clustering of active region using patch-based matrix factorization which will be presented in Chapter VI.

Section 5.2 describes our data set. Unlike previous works, our approach combines information from two modalities: photospheric continuum images and magnetograms, both obtained by the *Michelson Doppler Imager* (MDI) on board the *Solar and Heliospheric Observatory* (SOHO) spacecraft. We consider 424 active regions spanning the four main Mount Wilson classes. We use SMART masks [92] to delineate the boundaries of magnetic active regions, and the STARA algorithm [201] which provides masks for umbrae and penumbrae from the continuum images. These two masks enable us to differentiate between pixels belonging to the actual sunspots and pixels featuring the region surrounding the sunspots.

In Section 5.3, the intrinsic dimension of the image patches extracted from the two modalities is estimated using both linear and non-linear methods. The linear method relies on Principal Component Analysis (PCA) [100], while the non-linear method relies on a k-Nearest Neighbor graph approach [31, 40]. The latter method also estimates the local intrinsic dimension, which has several advantages over a global estimate. We show that the intrinsic dimension is related to the complexity of the sunspot groups.

Section 5.4 identifies the spatial and modal interactions of the patches at different scales by estimating the partial correlation and by using canonical correlation analysis (CCA) [150, 153]. This gives insight about relationships that may exist between active region complexity and the correlation patterns.

This chapter expands and refines some of the work in [142]. Whereas [142] used fixed size square pixel regions centered on the sunspot group as input to the analyses, in this chapter SMART detection masks are used. A larger set of images is considered in all methods which enables us to analyze the relationships of intrinsic dimension and correlation with AR complexity. We also explore the partial correlation of patches which was not included in [142].

5.2 Data

The data used in this study are taken from the *Michelson Doppler Imager* (MDI) instrument [174] on board the SOHO Spacecraft.

Within the time range of 1996-2010, we select a set of 424 ARs as follows. Using the information from the Solar Region Summary reports compiled by the Space Weather Prediction Center of NOAA `http://www.swpc.noaa.gov/ftpdir/forecasts/SRS/`, we consider ARs located within 30° of the solar meridian. We looked at a maximum of two-hundred instances per Mount Wilson types $\alpha, \beta, \beta\gamma$,

	α	β	$\beta\gamma$	$\beta\gamma\delta$	Simple	Complex	Total
Number of AR	50	192	130	52	242	182	424

Table 5.2: Number of each AR per Mt. Wilson class. Simple ARs include α and β groups while complex ARs are $\beta\gamma$ and $\beta\gamma\delta$ groups.

and $\beta\gamma\delta$. Out of this first selection, we removed AR with a longitudinal extent smaller than four, and finally we checked if MDI continuum and magnetogram data were available. This provides us with a number of ARs in each Mount Wilson class as given by Table 5.2. In our analysis, we also divide the ARs into two groups: simple ARs (α and β) and complex ARs ($\beta\gamma$ and $\beta\gamma\delta$).

AR are observed using two modalities: photospheric continuum images and magnetogram. SOHO-MDI provides almost continuous observations of the Sun in the white-light continuum, in the vicinity of the Ni I 676.78 nm photospheric absorption line. These photospheric intensity images are primarily used for sunspot observations. MDI data are available in several processed "levels". We use level-1.8 images, and rotate them with North up. SOHO provides two to four MDI photospheric intensity images per day with continuous coverage since 1995. Using the same instrument level, 1.8 line-of-sight (LOS) MDI magnetograms are recorded with a nominal cadence of 96 minutes. The magnetograms show the magnetic fields of the solar photosphere, with negative (represented as black) and positive (as white) areas indicating opposite LOS magnetic-field orientations.

As stated in Section 7.1, SMART masks [92] are used to determine the boundaries of magnetic active regions from MDI magnetograms. Those masks are applied also on continuum images to determine the surrounding part of the sunspot that is affected by magnetic fragments as seen in magnetogram images. Similarly, the STARA algorithm [201] provides masks for sunspots (umbrae and penumbrae) from MDI continuum and those masks are applied on magnetogram images to determine the AR cores corresponding to the sunspots. Combining these two types of masks provides thus two sets of pixels within each AR: those belonging to the *sunspots* themselves as found by STARA and those belonging to the *magnetic fragments* (or background) within an AR as found by the difference set between the SMART and STARA masks.

As in [142] we use image patch features to account for spatial dependencies using square patches of pixels. Thus if a SMART mask of an image has n pixels and we use a $m \times m$ patch, the corresponding continuum data matrix $\mathbf{X}$ is $m^2 \times n$ where the ith column contains the pixels in the patch centered at the ith pixel. The magnetogram data matrix $\mathbf{Y}$ is formed in the same way and the full data matrix is $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$ with size $2m^2 \times n$. We let $\mathbf{z}_i$ denote the ith column of $\mathbf{Z}$. The images from both modalities are also normalized prior to analyzing them.

In image patch analysis, the size of the patch should be no larger than the smallest feature that is to be captured. Otherwise, the relevant feature may be suppressed. Additionally, large patches lead to high-dimensional estimates which suffer in accuracy from "the curse of dimensionality," which refers to the fact that the number of observations must increase at least linearly in the number of parameters for accurate estimates to be possible in statistical inference [29]. Since some sunspot and active region features can be quite small and to limit the effects of high dimensionality on our analysis, we primarily use 3×3 patches in each modality although larger patches are used in Section 5.4 when analyzing spatial correlations in the images.

5.3 Intrinsic Dimension Estimation

The goal of this section is to determine the number of intrinsic parameters or degrees of freedom required to describe the spatial and modal dependencies using image patches. We consider 3×3 patches within both the continuum and magnetogram images giving an extrinsic dimension of 18. The intrinsic dimension will determine how redundant these 18 dimensions are. In addition, intrinsic dimension provides an indicator of complexity which we compare against the Mount Wilson classification, similarly to what [133] and [98] did using fractal dimension and gradient strength along polarity separating lines, respectively. More details on the concept of intrinsic dimension on manifolds are included in Appendix E.

It is also important to know whether *linear* analyses can be accurately applied to the data or whether *non-linear* techniques are required. Linear methods have been applied successfully to solar images before such as in [49]. However, it is not guaranteed that natural images are best represented using linear methods as there are cases where non-linear models have superior performance [47]. Thus this is important to investigate both for further analysis of the data in Chapter VI and for the correlation analysis in Section 5.4. If the data lie on a nonlinear subspace and we perform a linear analysis of the data (e.g. partial correlation, canonical correlation analysis, or principal component analysis), then the results will be only a linear approximation of the true relationships and dependencies of the data. Nonlinear methods of analysis would be necessary to obtain higher accuracy in this case. To answer this question, we estimate the local intrinsic dimension using a method appropriate for linear subspaces and a method appropriate for any (linear or non-linear) smooth subspace and then compare the results.

5.3.1 PCA: A Linear Estimator

Principal Component Analysis (PCA) [100] finds a set of linearly uncorrelated vectors (principal components) that can be used to represent the data. PCA has been used previously for various purposes in solar-physics and space-weather literature, e.g. to study the background and sunspot magnetic fields [30, 117, 209], for analysis of solar wind data [94], or to reduce dimensionality [50].

In PCA, the principal components are the eigenvectors of the covariance matrix

Σ:

$$\Sigma = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix},$$

where $\mathbf{x}$ and $\mathbf{y}$ are random vectors of dimension 9, $\mathbf{x}$ being a patch from the continuum image, and $\mathbf{y}$ the corresponding patch from the magnetogram image. The eigenvalues indicate the amount of variance accounted for by the corresponding principal component. A linear estimate of intrinsic dimension is the number of principal components that are required to explain a certain percentage of the variance.

By nature, PCA is a global operation and so it provides a global estimate of the intrinsic dimension. We can obtain more local estimates by performing PCA separately on the areas within the sunspots and on the magnetic fragments. These areas are separated using the STARA and SMART masks.

5.3.2 *k*-NN: A General Estimator

The general method we use is a k-Nearest-Neighbor (k-NN) graph approach with neighborhood smoothing [31, 40]. The intuition behind the method is that we grow the k-NN graph from a point $\mathbf{z}_i$ by adding an edge from $\mathbf{z}_i$ to $\mathbf{z}_j$ if $\mathbf{z}_j$ is within the k nearest neighbors of $\mathbf{z}_i$. The growth rate of the total edge length of the graph is related to the intrinsic dimension of the data in a way that enables us to estimate it.

One advantage of the k-NN method, in contrast to global methods such as [122], is it provides an estimate of the *local* intrinsic dimension by limiting the growth of the graph to a smaller neighborhood. This provides an estimate of intrinsic dimension at each pixel location in the image which allows us to more easily visualize the intrinsic dimension estimates. Additionally, when the number of samples within a region of interest is small (such as within a small sunspot), this local method provides more accurate estimates of intrinsic dimension than applying a global method (such as PCA) since the inclusion of the neighboring pixels results in a higher number

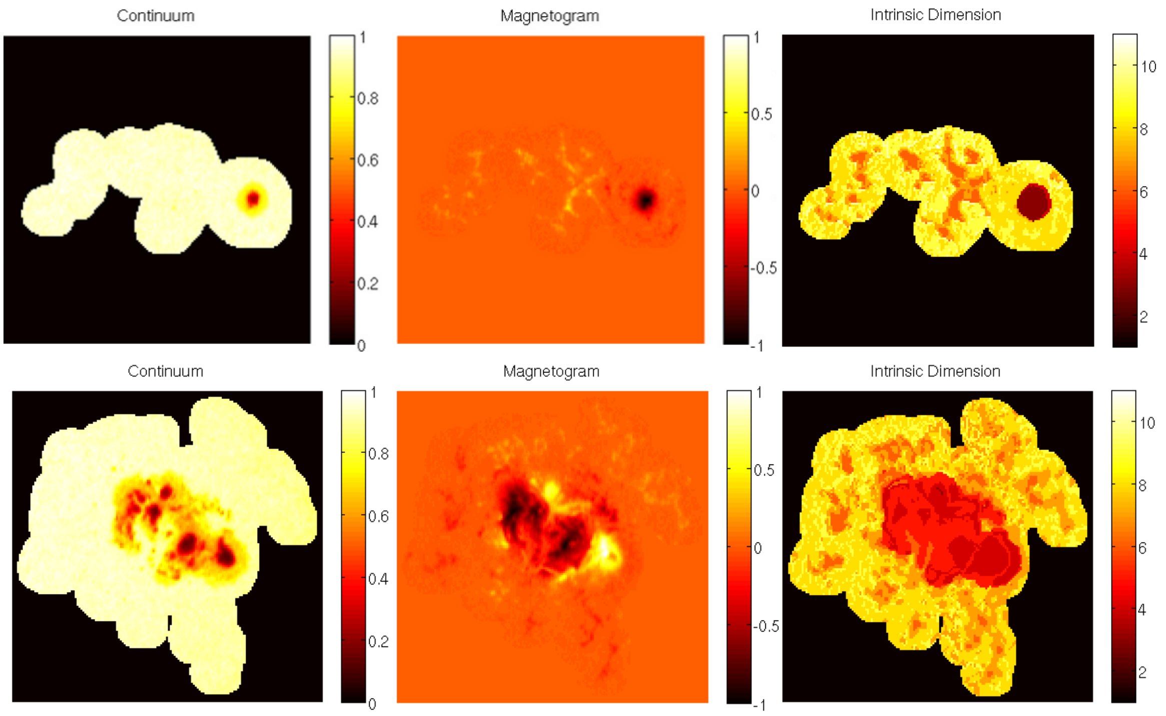

Figure 5.2: Examples of the estimated local intrinsic dimension using the k-NN method for an α group (top) and a $\beta\gamma\delta$ group (bottom). Regions with more spatial structure have lower intrinsic dimension.

of samples. Technical explanation of the k-NN method and more details on local intrinsic dimension are given in Appendix E.

5.3.3 General Results

We estimate the intrinsic dimension of the image patches within the sunspots and magnetic fragments for all 424 ARs using both the k-NN approach and PCA, where the extrinsic dimension of the joint patches is 18. Figure 5.2 shows two examples of the estimated local intrinsic dimension using the k-NN method and the corresponding continuum and magnetogram images. One set of images corresponds to an α group while the other set is a $\beta\gamma\delta$ group. In these examples, areas with more spatial structure, such as within the sunspots, have lower intrinsic dimension. Fewer parameters are required to accurately represent structured data than noise and so the intrinsic dimension is lower.

	α	β	$\beta\gamma$	$\beta\gamma\delta$	All
Sunspots k-NN, pooled	3.9 ± 1.2	4.4 ± 1.0	4.4 ± 0.9	4.5 ± 0.7	4.3 ± 1.0
Sunspots k-NN, means	4.0 ± 1.0	4.8 ± 1.2	4.4 ± 0.6	4.4 ± 0.4	4.5 ± 1.0
Sunspots PCA 97%	3.7 ± 0.8	4.4 ± 0.9	4.3 ± 0.6	4.2 ± 0.5	4.3 ± 0.8
Sunspots PCA 98%	4.5 ± 0.9	5.2 ± 1.0	5.1 ± 0.8	5.0 ± 0.5	5.0 ± 0.9
Fragments k-NN, pooled	8.0 ± 1.1	8.0 ± 1.2	7.7 ± 1.2	7.6 ± 1.2	8.0 ± 1.2
Fragments k-NN, means	8.0 ± 0.4	7.9 ± 0.5	7.6 ± 0.5	7.6 ± 0.5	7.8 ± 0.5
Fragments PCA 97%	7.7 ± 1.6	7.1 ± 1.2	6.2 ± 1.2	6.2 ± 1.3	6.8 ± 1.4
Fragments PCA 98%	9.1 ± 1.6	8.5 ± 1.3	7.5 ± 1.2	7.4 ± 1.3	8.1 ± 1.4

Table 5.3: Estimated intrinsic dimension results for different groups of ARs in the form of mean±standard deviation. The complex ARs have higher intrinsic dimension within the sunspots than the simple ARs but lower intrinsic dimension within the magnetic fragments.

Table 5.3 provides the mean and standard deviation of the intrinsic dimension estimates within the sunspots and magnetic fragments. These statistics are also provided for ARs within the main Mount Wilson classes (α, β, $\beta\gamma$, and $\beta\gamma\delta$). We provide PCA results for the cases where we estimate the intrinsic dimension as the number of components required to explain 97% and 98% of the variance, respectively. For the k-NN method, we provide the results in two ways. For one, we take the mean of local intrinsic dimensions within each image (separating the 'sunspot' from the 'magnetic fragments') and then calculate the mean and standard deviation of these *means*. The statistics in this category correspond to the mean and standard deviation of the average intrinsic dimension of each image and are more directly comparable to the PCA results. However these results may be affected slightly by small sunspot groups. For the other approach, we *pool* all of the local estimates (again separating sunspots from magnetic fragments) and then calculate the mean and standard deviation. These results correspond to the mean and standard deviation of the pixels within each region and category and are less affected by small sunspot groups.

From Table 5.3, it is clear that the intrinsic dimension is lower within the sunspots than in the magnetic fragments for all methods. This is expected as there is more spatial structure within the images inside the sunspots than in the magnetic fragments, especially in the continuum image.

The average PCA estimate with a 97% threshold and the average mean k-NN estimate give similar results inside the sunspot while the average 98% PCA estimate is closest to the average mean k-NN estimate within the magnetic fragments. If linear methods were not sufficient to represent the spatial and modal dependencies, we would expect the PCA results to be much higher than the k-NN results when using comparable thresholds as more linear than nonlinear components would be required to accurately represent the data. However, this close agreement between the general and linear results suggests that linear methods are sufficient and that linear dictionary methods would be appropriate for these data.

5.3.4 Patterns Within the Mount Wilson Groups

For both the PCA and k-NN methods, the average estimated intrinsic dimension is lower within the sunspots in α groups than in the more complex groups such as $\beta\gamma\delta$. This is consistent with Figure 5.2 and may be related to the lower complexity of α groups. These exhibit more spatially coherent images, which can be described using a lower number of basis elements, and hence have a lower intrinsic dimension.

Within the magnetic fragments, the opposite trend occurs where the less complex groups have higher intrinsic dimension. This suggests that the magnetic fragments are fewer, weaker, and less structured outside of the α and β groups compared to the more complex regions, leading to a more noise-like background in their magnetic fragments. This hypothesis is supported by the normalized histograms of the mean k-NN estimates of intrinsic dimension and the normalized histograms of the pooled k-NN estimates in Figures 5.3 and 5.4. The histograms of mean intrinsic dimension show that within the magnetic fragments, α groups generally have higher mean intrinsic dimension than $\beta\gamma\delta$ groups. In fact, no α groups have a mean intrinsic dimension

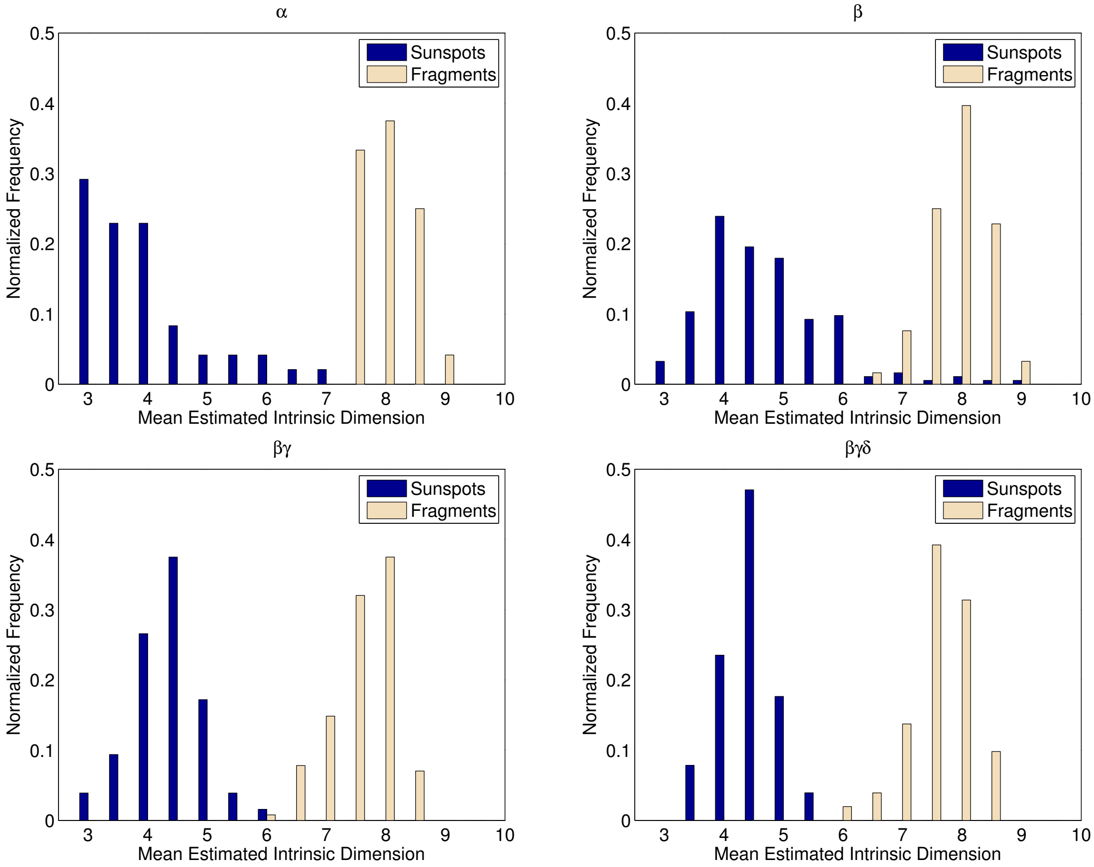

Figure 5.3: Normalized histograms of mean estimated intrinsic dimension of α, β, $\beta\gamma$, and $\beta\gamma\delta$ groups using the k-NN method. The distributions of intrinsic dimension differ by complexity with simpler AR groups having higher (resp. lower) intrinsic dimension within the sunspot (resp. magnetic fragments).

less than 7.5 within the magnetic fragments. However, the normalized histograms of the individual patch estimates show a significant number of patches with intrinsic dimension less than 7.5 within the fragments. This suggests that for each α group, the majority of the patches have higher intrinsic dimension in the magnetic fragments and are thus more noise-like. In contrast, there are some $\beta\gamma\delta$ groups where the mean intrinsic dimension of the magnetic fragments is lower (less than 7.5) and so these magnetic fragments are dominated by patches with more structure.

Table 5.3 also shows that the standard deviation of the estimates within the sunspots decreases as the complexity increases as measured by the Mount Wilson classification scheme. The histograms in Figures 5.3 and 5.4 can be used to determine

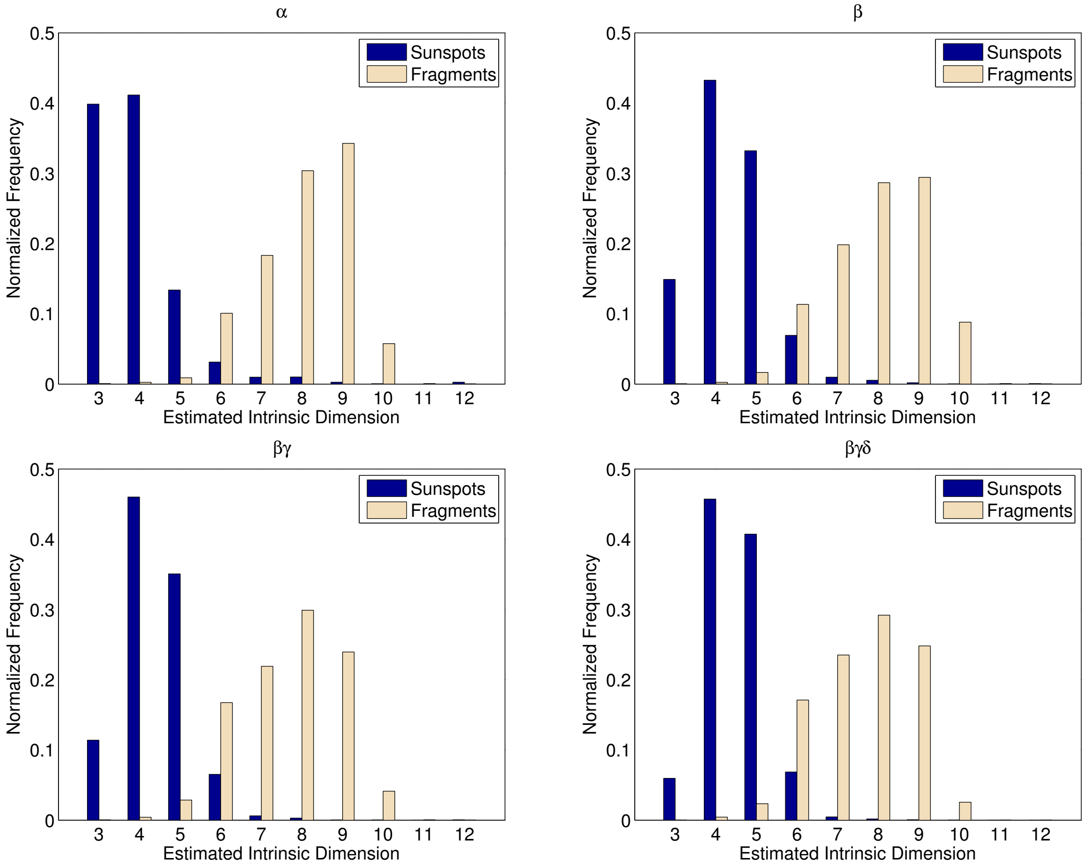

Figure 5.4: Normalized histograms of pooled local estimates of intrinsic dimension of α, β, $\beta\gamma$, and $\beta\gamma\delta$ groups using the k-NN method. The distributions of intrinsic dimension differ by complexity with simpler AR groups having higher (resp. lower) intrinsic dimension within the sunspot (resp. magnetic fragments).

the cause. From the histograms, it is clear that within the sunspots the intrinsic dimension of α groups does not have a Gaussian distribution. In this case, most of the estimates are between 3 and 5. However, there are a significant number of outliers with intrinsic dimension greater than 5. The presence of these outliers contributes to the high standard deviation. This is in contrast to the intrinsic dimension of $\beta\gamma$ and $\beta\gamma\delta$ groups inside the sunspot which have fewer outliers and thus smaller standard deviations.

The outliers in the α groups correspond to small sunspots. The number of pixels within the α sunspots with average intrinsic dimension ≥ 6 range between 10 and 53 with a median of 16. In these cases, the spatial structure of the sunspots may be more similar to the magnetic fragments than the spatial structure of larger sunspots. Thus the intrinsic dimension is higher in the small sunspots.

A similar phenomenon occurs within the β groups. Note that in Table 5.3, the average and standard deviation of the mean intrinsic dimension of the β groups within the sunspots is higher than for all other groups. This is also caused by a few outliers that have high average intrinsic dimension due to the small size of the sunspots. When individual local intrinsic dimension estimates of the patches from these small sunspots are pooled with the estimates from all other β patches, the average intrinsic dimension is more aligned with that of the other Mount Wilson types. Additionally, ignoring the biggest outliers in the mean intrinsic dimension (defined as having mean intrinsic dimension > 6.25) gives an average mean intrinsic dimension of 4.6 for the β groups which is more aligned with the other groups.

The distribution of intrinsic dimension within the magnetic fragments also differs by complexity based on Figures 5.3 and 5.4. The complex ARs have more patches and images with lower intrinsic dimension than the simple sunspots which is consistent with Table 5.3.

In summary, based on the estimated intrinsic dimension of the image patches, relatively few parameters are required to accurately represent the data. We have found that the distribution of local intrinsic dimension varies based on the complexity of the sunspot group with the more complex sunspots having higher (resp. lower) intrinsic dimension within the sunspot (resp. magnetic fragments). Additionally, the standard deviation of the intrinsic dimension is higher within the sunspot in the simpler sunspots than the complex ARs. This is due to the presence of small sunspots among the simpler ARs that tend to have less spatial structure and thus a higher intrinsic dimension than typical sunspots. We have also shown that linear methods should be sufficient to accurately analyze the data.

5.4 Spatial and Modal Correlations

The results in the previous section indicate that linear methods are likely sufficient to represent the spatial and modal dependencies within a sunspot. We therefore analyze the linear correlation over patches using partial correlation and canonical correlation analysis (CCA).

The partial correlation is proportional to the inverse of the correlation matrix and analyzes the pixel-to-pixel correlation when the influence of all other pixels has been removed. It provides insight into how large a patch should be used to sufficiently capture the spatial and modal correlations in future analysis.

CCA on the other hand is determined by finding the most correlated linear combinations of pixels from each image, solved as a generalized eigenvalue problem, which is useful for determining the degree of mutual correlation between two modalities. If the two modalities are independent, there is no benefit in processing them together, while if the two modalities are strongly dependent, processing only one of the modalities is sufficient since the other modality would not contain any additional information.

5.4.1 Partial Correlation: Methodology

The partial correlation measures the correlation between two random variables while conditioning on the remaining random variables. The intuition behind partial correlation can be best explained with the linear regression concept. Suppose you want to compute the partial correlation between two variables X_1 and X_2 given a set of variables $\mathcal{X}$. First, compute the linear regression using variables in $\mathcal{X}$ to explain X_1 and obtain the associated residuals r_{X_1}. Proceed similarly for X_2 and get residuals r_{X_2}. The partial correlation between X_1 and X_2 is then equal to the (usual) correlation between r_{X_1} and r_{X_2}, for which the effect of variables $\mathcal{X}$ have been removed.

In our context, let $\mathbf{x}$ be a patch from the continuum image, and $|\mathbf{y}|$ be the magnitude (entry-wise absolute value) of the corresponding patch from the magnetogram. The partial correlation matrix $\mathbf{P} = \begin{pmatrix} \mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{x}|\mathbf{y}|} \\ \mathbf{P}_{|\mathbf{y}|\mathbf{x}} & \mathbf{P}_{|\mathbf{y}||\mathbf{y}|} \end{pmatrix}$ and its off-diagonal elements can be derived from the inverse correlation matrix. We use the magnitude of the magnetogram data since both positive and negative polarities affect the continuum image in similar ways.

5.4.2 Partial Correlation: Results

Figure 5.5 gives the estimated partial correlation matrices when using 3×3 and 5×5 patches. The patches are extracted from all of the active regions and divided using the STARA and SMART masks into sunspots and magnetic fragments as before. The partial correlation of 3×3 patches is quite strong within both modalities. Based on a false alarm rate of 0.05, the theoretical thresholds for significance for the partial correlation [89] of the 3×3 patches are approximately 0.0070 and 0.0014 for within the sunspots and magnetic fragments, respectively. For the 5×5 patches, the thresholds are 0.0080 and 0.0016, respectively. Given these thresholds, the partial correlation is statistically significant for nearly all values within the modalities ($\mathbf{P}_{\mathbf{xx}}$ and $\mathbf{P}_{|\mathbf{y}||\mathbf{y}|}$)

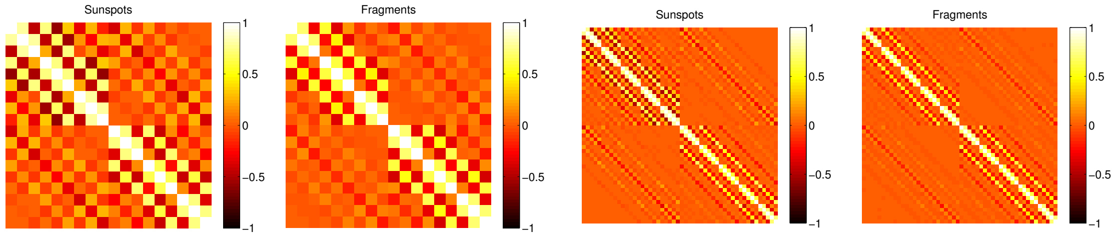

Figure 5.5: Estimated partial correlation matrices of patch data from within the sunspots and the magnetic fragments using 3×3 (left) and 5×5 (right) patches. The theoretical thresholds [89] for significance to attain a 0.05 false alarm rate are 0.0070 and 0.0014 for within the sunspots and magnetic fragments, respectively when using a 3×3 patch. For the 5×5 patch, the thresholds are 0.0080 and 0.0016, respectively. Statistically insignificant values are set to zero.

using the 3×3 patches.

The cross-partial correlation when using the signed magnetic field (i.e. $\mathbf{P}_{\mathbf{xy}}$ and $\mathbf{P}_{\mathbf{yx}}$) is very near zero in both regions (not shown). However, when we take the absolute value of the magnetogram patches, then the magnitude of the cross-partial correlation ($\mathbf{P}_{\mathbf{x}|\mathbf{y}|}$ and $\mathbf{P}_{|\mathbf{y}|\mathbf{x}}$) is much higher in both regions suggesting that the correlation between the modalities is significant. The partial correlation is also stronger in magnitude in all cases within the sunspots than within the magnetic fragments.

The partial correlation matrices are very structured. In both sunspots and magnetic fragments, the pentadiagonal-like structure within the modalities suggests that the image is generally stationary with approximately a third order nearest neighbor Markov structure in the pixels. Such structure is clearly seen in the matrices for 5×5 patches. The cross-partial correlation also has a pentadiagonal-like structure although the correlation is not as strong as within the modalities.

To better see the spatial correlations, in Figure 5.6 we plot the partial correlation patches taken from the columns of the sunspot partial correlation matrix corresponding to the center pixels when using 5×5 patches. Figure 5.6 shows clearly the greater partial correlation within the continuum. It also highlights that correlation is slightly higher in magnitude in the vertical direction than the horizontal direction. Nearly all

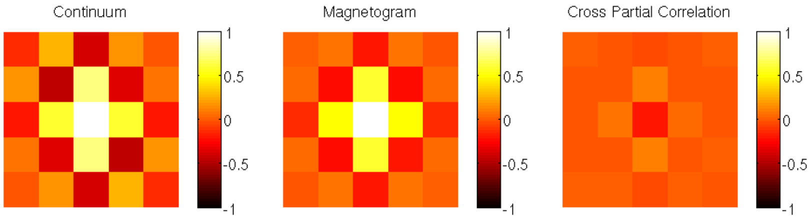

Figure 5.6: Partial correlation patches extracted from the columns in the sunspot partial correlation matrix corresponding to the center pixels. The partial correlation is stronger within the continuum.

sunspots in this study are located within $(-30^\circ, +30^\circ)$ from both the central meridian and the equator, and so projection effect are unlikely to cause this difference. The difference in correlation may be a feature of the sunspots themselves, but this may be difficult to determine since the difference in partial correlation is small.

Some slight differences exist in the partial correlation matrices restricted to certain Mount Wilson classes. As an example, Figure 5.7 contains the partial correlation matrices within the sunspots after restricting the data to α and β groups as well as the difference between the absolute value of the two matrices. The α partial correlation matrix is higher in magnitude within the modalities than the β matrix but lower between the modalities. Within modalities, the strongest differences (a maximum of 0.056 and 0.067 within the continuum and magnetogram, respectively) are in the entries that correspond to pixels that are farther away from each other. In contrast, within the cross-partial correlation, the strongest differences (a maximum of 0.072) between the two AR types are in the entries that correspond to pixels that are close to each other. A similar pattern holds when comparing the α matrix to the matrices of the more complex groups.

Overall, the partial correlation matrices indicate that no larger than a 5×5 patch is necessary to capture the local spatial dependencies. A 5×5 patch of pixels corresponds roughly to the size of a mesogranule [168, 169]. This suggests that within the magnetic fragments, it is likely that the granules and mesogranules within the

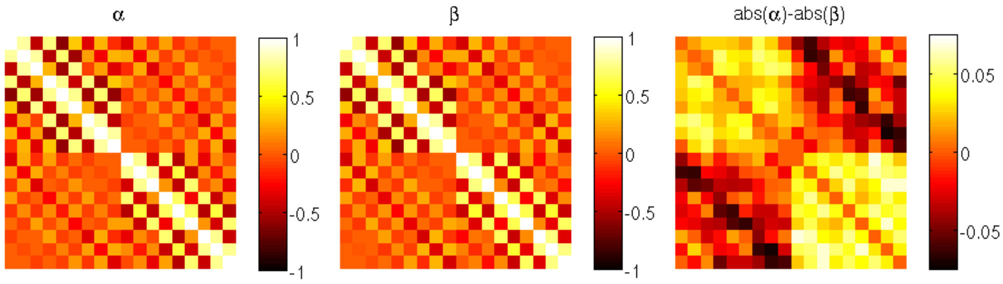

Figure 5.7: Partial correlation matrices within the sunspots using the data from α (left) and β ARs (center). Statistically insignificant values are again set to zero. (Right) difference between the absolute value of the α and β matrices. The α sunspots are more (resp. less) strongly correlated within (resp. between) the modalities than the β groups.

photosphere contribute to the local spatial dependencies. Within the sunspots, a 5×5 patch corresponds to the size of the characteristic length of the largest penumbral filaments [192] which suggests that on average the local spatial dependencies are minimal beyond this scale. This analysis, however, does not rule out long-range spatial dependencies, which are more difficult to assess due to the large dimensionality. Future work will focus on this.

In the remainder of our analysis, we choose a 3×3 patch for the reasons mentioned in Section 5.2: to ensure that we capture the features of small sunspots and to limit the effects high dimension on the accuracy of the analysis. Given these concerns, we see that 3×3 patches capture most of the spatial correlation. This is evident from Figure 5.5 where the partial correlation between pixels on opposite corners of a 3×3 patch is near zero and other pixels that are similarly far away from each other have low partial correlation. Thus a 3×3 patch strikes a good balance between scale, extrinsic dimension, and capturing the spatial correlation.

5.4.3 Canonical Correlation Analysis: Methodology

To further investigate the correlation between the modalities, we use canonical correlation analysis (CCA) on the continuum patch $\mathbf{x}$ and the magnitude of the magnetogram patch $\mathbf{y}$. CCA finds patterns and correlations between two multivariate data sets [150, 153] and was used previously in the context of space weather e.g. for the combined analysis of solar wind and geomagnetic index data sets [27].

In our application, CCA provides linear combinations of continuum patches $\mathbf{x}$ that are most correlated with linear combinations of magnetogram patches $\mathbf{y}$. In other words, all correlations between the continuum and magnetogram patches are channeled through the canonical variables. Formally, CCA finds vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ for $i = 1, \ldots, m^2$ such that the correlation $\rho_i = \mathrm{corr}(\mathbf{a}_i^T \mathbf{x}, \mathbf{b}_i^T |\mathbf{y}|)$ is maximized and the pair of random variables $u_i = \mathbf{a}_i^T \mathbf{x}$ and $v_i = \mathbf{b}_i^T |\mathbf{y}|$ are uncorrelated with all other pairs u_j and v_j, $j \neq i$. The variables u_i and v_i are called the ith pair of canonical variables while the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ are the canonical vectors. The solution $\mathbf{a}_i$ is the ith eigenvector of the matrix $\Sigma_{\mathbf{xx}}^{-1} \Sigma_{\mathbf{x}|\mathbf{y}|} \Sigma_{|\mathbf{y}||\mathbf{y}|}^{-1} \Sigma_{|\mathbf{y}|\mathbf{x}}$ which are taken from the covariance matrix. The vector $\mathbf{b}_i$ is found similarly [84].

5.4.4 Canonical Correlation Analysis: Results

Here we focus on 3 × 3 patches and apply CCA to all 424 image pairs using the magnitude of the magnetogram patches. Figure 5.8 (left and center) shows histograms of the estimated values of ρ_1. Within the sunspots, there are many groups with a near perfect correlation between the modalities and none of the groups have an estimated value below 0.41. The right plot in Figure 5.8 plots the estimated values of ρ_1 vs. the number of samples used within the sunspots. Based on this plot, there are many ARs with high correlation and few patch samples suggesting that the correlation may be spurious. However, all of the estimated values are statistically significant as defined by the threshold given by [89] using a false alarm rate of 0.05 (shown as the magenta line in Figure 5.8).

The histogram of ρ_1 within the magnetic fragments (Fig. 5.8, center) is quite different from the sunspot histogram (Fig. 5.8, left). ρ_1 is generally lower within the

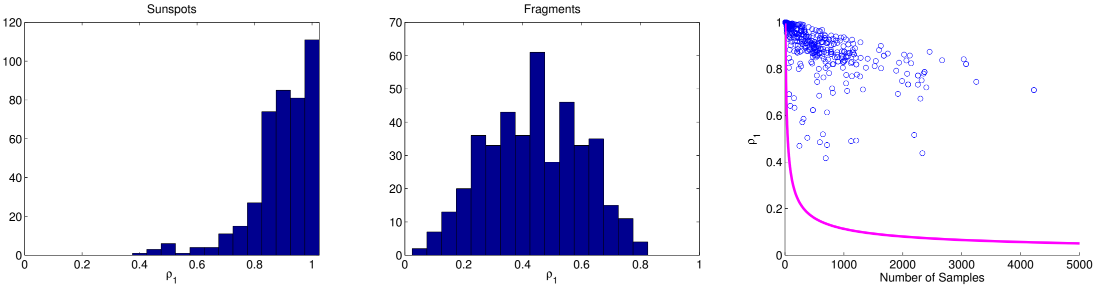

Figure 5.8: Histograms of estimated ρ_1 using CCA for within the sunspots (left) and the magnetic fragments (center) using 3×3 patches. Right: Scatter plot of ρ_1 values and the number of samples available for within the sunspots. All points are above the magenta line which gives the threshold for statistical significance at a false alarm rate of 0.05 [89].

magnetic fragments than within the sunspots which is consistent with the results in Figure 5.5. All of the ρ_1 values are statistically significant.

The distributions of ρ_1 differ slightly when comparing simple sunspot groups (α and β) with complex groups ($\beta\gamma$ and $\beta\gamma\delta$). Figure 5.9 shows that complex groups generally have lower correlation between the modalities within the sunspots than the simpler groups. The estimated Hellinger distance (see Appendix **??**) between the distributions using the divergence estimator in [140, 141] is 0.22. Based on the central limit theorem of the estimator [141], this value is statistically significant with a p-value of 1.6×10^{-12}. At least some of this difference is likely due to the smaller size of the simpler groups (and thus smaller sample size). However, it is unlikely to fully explain the difference given that there are many simple sunspot groups with high correlation and sufficient sample size.

Within the magnetic fragments, there are many more simple regions than complex regions with $\rho_1 < 0.4$ (see the histogram in Figure 5.9, right). This could be related to the same phenomena that causes the intrinsic dimension to be higher within the magnetic fragments of simple sunspot groups observed in Section 5.3.3. However, the estimated Hellinger distance between these distributions is 0.016. Using the same statistical test, this estimate is not statistically significant with a p-value of 0.31.

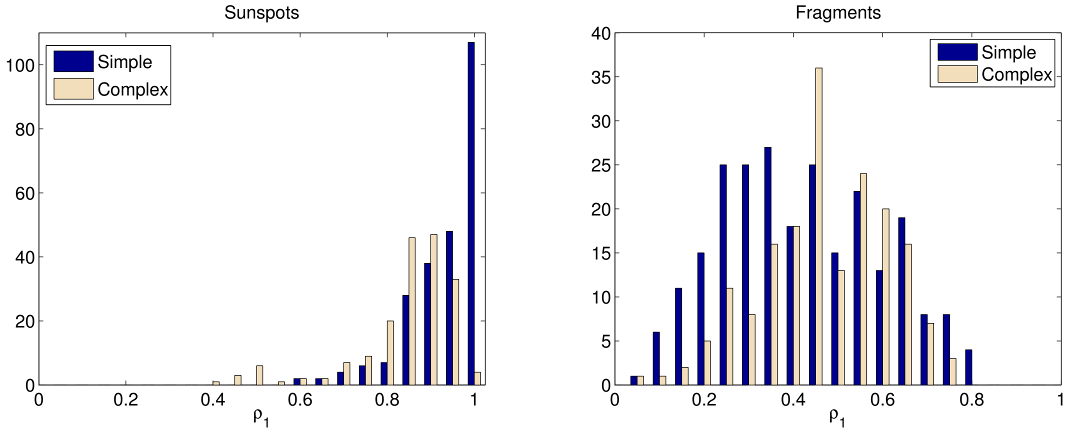

Figure 5.9: ρ_1 histograms of complex ($\beta\gamma$ and $\beta\gamma\delta$) and simple (α and β) regions within the sunspots (left) and the magnetic fragments (right). The simple ARs are generally more correlated within the sunspots but less correlated within the magnetic fragments. The difference between the sunspot distributions, as measured by the Hellinger distance, is statistically significant.

Thus the distributions are not statistically different from each other.

To analyze the spatial patterns that produce the highest correlation between modalities, we apply CCA to the entire data set. Figure 5.10 plots ρ_i for $i = 1, \dots, 9$ for within the sunspots and within the magnetic fragments. The ρ_i are all statistically significant. Notice that the ρ_i are higher within the sunspots than the magnetic fragments which is consistent with the results in Figures 5.5 and 5.8.

Figure 5.10 shows the canonical patches $\mathbf{a}_i$ and $\mathbf{b}_i$ for $i = 1, \dots, 6$ when using all the data from within the sunspots. These are the spatial patterns within the two modalities that are most correlated with each other. The canonical patches have a "saddle-like" appearance where the gradient is positive in some directions and negative in others. For example, in $\mathbf{a}_4$, the pixels to the left and right of the center are very negative but the pixels in the corners are all very positive. Note that these vectors correspond to centered values with respect to the mean patches.

Comparing the $\mathbf{a}_i$s to the $\mathbf{b}_i$s shows that the $\mathbf{b}_i$s are approximately equal to the negative of the $\mathbf{a}_i$s. This makes sense as sunspots within the continuum images correspond to a decrease in value relative to the background while ARs within the

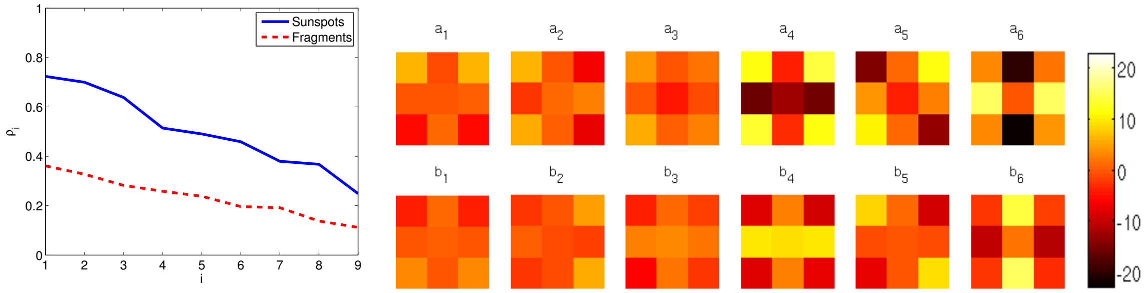

Figure 5.10: (Left) Plot of the estimated ρ_i using CCA on the entire data set for $i = 1, \ldots, 9$. All values are statistically significant [89]. (Right) Canonical patches $\mathbf{a}_i$ (top) and $\mathbf{b}_i$ (bottom) for $i = 1, \ldots, 6$ when using the entire data set from within the sunspots. The $\mathbf{b}_i$s are approximately equal to the negative of the $\mathbf{a}_i$s.

magnitude of the magnetogram images correspond to an increase in value relative to the background.

We also performed CCA separately on the data from the Mount Wilson classes. Figure 5.11 plots the ρ_i values for each class and the first canonical patches $\mathbf{a}_1$ and $\mathbf{b}_1$. For ρ_1 and ρ_2, the values for each class decrease in order of complexity (α, β, $\beta\gamma$, $\beta\gamma\delta$). This is consistent with our comparison of the partial correlation matrices in Figure 5.7 where the partial correlation was generally higher (in magnitude) for the α groups than the others. This is also consistent with the intrinsic dimension analysis in Section 5.3 where the intrinsic dimension generally increases with complexity. This is because if the correlation between and within modalities is higher, then fewer parameters are required to accurately describe the data which results in a lower intrinsic dimension.

The canonical patches $\mathbf{a}_1$ and $\mathbf{b}_1$ have similar patterns across the different classes although the patches for the $\beta\gamma$ class are flipped compared to the others. The magnitude of the values in the $\beta\gamma\delta$ patches are also smaller than the those of the other patches.

Overall, the results of this section suggest that the two modalities are correlated in both the sunspots and the magnetic fragments and are therefore not independent.

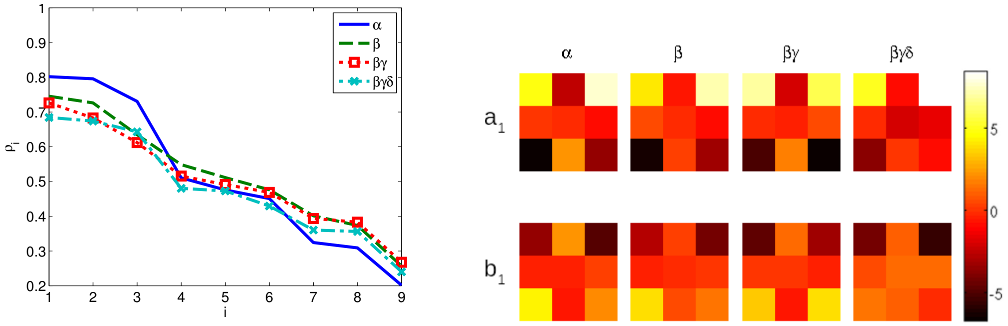

Figure 5.11: (Left) Plot of the estimated ρ_i using CCA on data segregated by Mount Wilson classes for $i = 1, \ldots, 9$ within the sunspots α groups start out with the highest correlation. (Right) Canonical patches $\mathbf{a}_1$ (top) and $\mathbf{b}_1$ (bottom) for the Mount Wilson classes within the sunspots. Again, the $\mathbf{b}_1$s are approximately equal to the negative of the $\mathbf{a}_1$s as in Figure 5.10 but the patches differ slightly from class to class.

The correlation is stronger within the sunspots compared to the magnetic fragments and stronger within the sunspots in simple ARs compared to complex ARs. However, the correlation is not perfect and so there may be an advantage to including both modalities in the classification of sunspots and flare prediction.

5.5 Conclusion

Existing AR categorical classification systems such as the Mount Wilson and McIntosh schemes describe geometrical arrangements of the magnetic field at the *largest* length scale. In this chapter, we have focused on the properties of the ARs at *fine* length scale. We showed that when we analyze the global statistics or attributes of these local properties, we find differences between the simple and complex ARs as defined using the large scale characteristics. Thus by this approach, we are analyzing both the large and fine scale properties of the images. Such results may be due to the multi-scale properties of the magnetic fields, as evidenced previously in [98].

This chapter also highlighted specific behaviors of the core of active regions (that corresponds to the sunspot masks in continuum) and magnetic fragments (the surrounding part of AR), as well as the difference between these two regions as a function of the Mount Wilson classification. We found that within the sunspots, the spatial and modal correlations are stronger than within the magnetic fragments. Additionally, simpler ARs were found to have higher correlation between the modalities within the sunspots than the complex ARs.

This chapter paves the way for further analysis based on matrix factorization. The results of Section 5.3 justifies the use of linear methods of analysis. Knowledge of the intrinsic dimension allows us to choose the number of factors. The spatial and modal correlation analysis in Section 5.4 also justifies the choice of a patch size of 3×3 and confirms that both modalities (continuum and magnetogram) should be used in the analysis. This analysis is provided in Chapter VI.

CHAPTER VI

Application to Sunspot Images: Clustering via Divergence Estimation and Bayes Error Estimation

6.1 Introduction

6.1.1 Context

While the Mount Wilson classification scheme has been effective in relating a sunspot's large scale magnetic configuration with its ability to produce flares, the categorical nature of the Mount Wilson classification prevents the differentiation between two sunspots with the same classification and makes the study of an AR's evolution cumbersome. Moreover, the Mount Wilson classification is generally carried out manually which results in human bias. Several papers [35, 36, 188] have used supervised techniques to reproduce the Mount Wilson and other schemes which has resulted in a reduction in human bias.

To go beyond categorical classification in the flare prediction problem, the last decade has seen many efforts in describing the photospheric magnetic configuration in more details. Typically, a set of scalar properties is derived from line of sight (LOS) or vector magnetogram and analyzed in a supervised classification context to derive which combination of properties is predictive of increased flaring activity [3, 10, 25, 56, 71, 77, 96, 120, 121, 176, 183, 207]. Examples of scalar properties include: sunspot area, total unsigned magnetic flux, flux imbalance, neutral line length, maximum gradients along the neutral line, or other proxies for magnetic connectivity within ARs. These scalar properties are features that can be used as input in flare prediction. However, there is no guarantee that these selected features exploit the information present in the data in an optimal way for the flare prediction problem.

6.1.2 Contribution

We introduce a new data-driven method to cluster ARs using information contained in magnetogram and continuum. Instead of focusing on the best set of properties that summarizes the information contained in those images, we study the natural geometry present in the data via a reduced-dimension representation of such images. The reduced-dimension is implemented via matrix factorization of an image patch representation as explained in Section 6.1.3. We show how this geometry can be used for classifying ARs in an unsupervised way, that is, without including AR labels as input to the analysis. We consider the same dataset as in Chapter V.

Our method can be adapted to any definition of the support of an AR, or Region of Interest (ROI), and such ROI must be given a priori. We consider three types of ROIs:

1. Umbrae and penumbrae masks obtained with the Sunspot Tracking and Recognition Algorithm (STARA) [201] from continuum images. These sunspot masks encompass the regions of highest variation observed in both continuum and magnetogram images, and hence are used primarily to illustrate our method. Figure 6.1 provides some examples of AR images overlaid with their respective STARA masks.

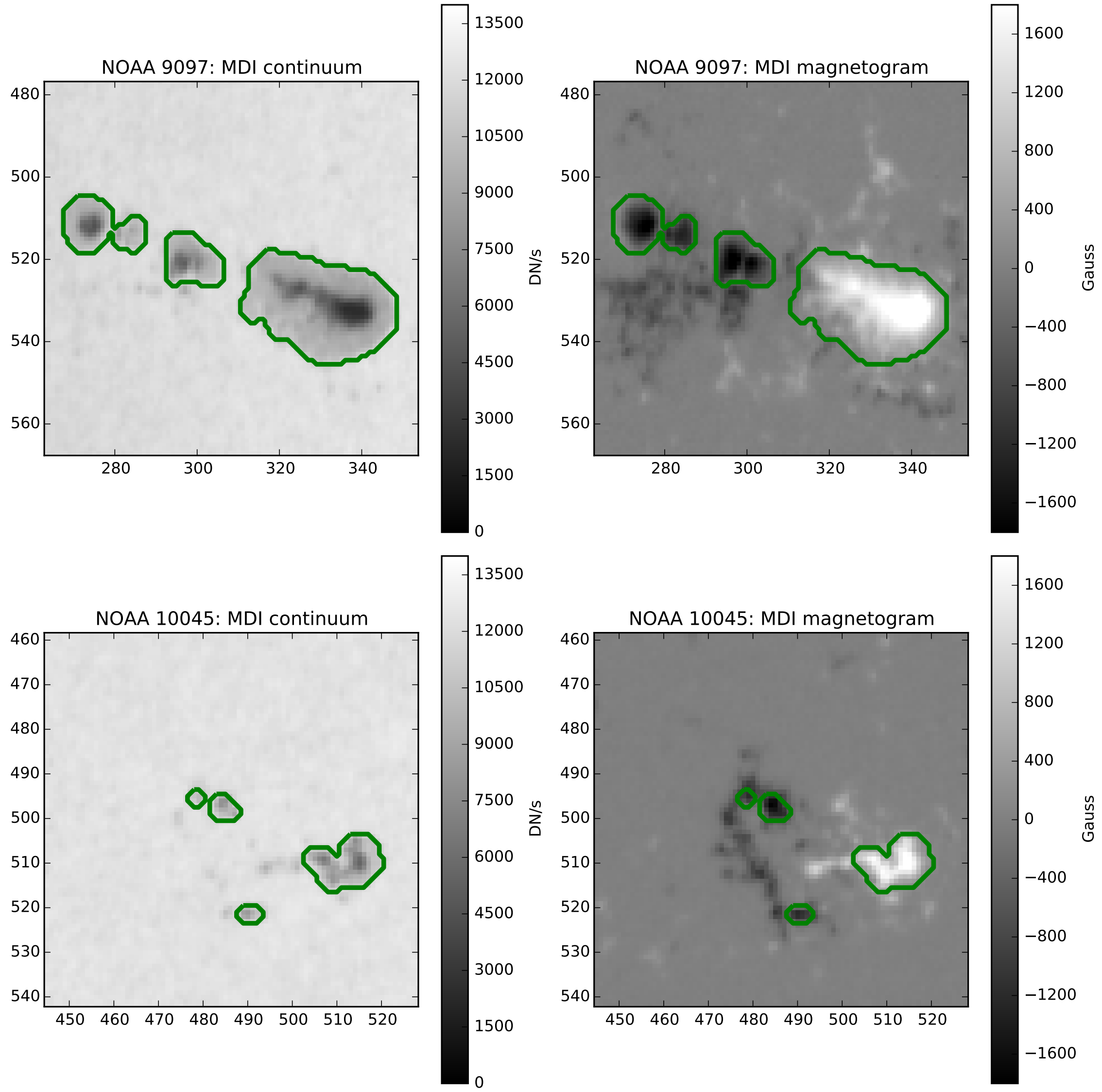

Figure 6.1: MDI continuum and magnetogram from NOAA 9097 on July 23, 2000 (top) and from NOAA 10045 on July 25, 2002 (bottom) overlaid with the corresponding STARA masks in green.

2. The neutral line region, defined as the set of pixels situated no more than 10 pixels (20 arcsec) away from the neutral line, and located within the Solar Monitor Active Region Tracker (SMART) masks [92], which defines magnetic AR boundaries.

3. The set of pixels that are used as support for the computation of the R-value defined in [176]. The R-value measures a weighted absolute magnetic flux, where the weights are positive only around the neutral line.

Our patch-based matrix factorization method investigates the fine scale structures encoded by localized gradients of various directions and amplitudes, or locally smooth areas for example. In contrast, the Mount Wilson classification encodes the relative locations and sizes of concentrations of opposite polarity magnetic flux on a large scale. Although both classification schemes rely on completely different methods, using the first ROI defined above, we find some similarities (see Section 6.5). Moreover the Mount Wilson classification can guide us in the interpretation of the results and clusters obtained.

The shape of the neutral line separating the two main polarities in an AR is a key element in the Mount Wilson classification scheme, and the magnetic field gradients observed along the neutral line are important information in the quest for solar activity prediction [56, 176]. We therefore analyze the effect of including the neutral line region in Section 6.5.

Results based on the third ROI are compared directly to the R-value. The various comparisons enable us to evaluate the potential of our method for flare prediction.

6.1.3 Reduced dimension via matrix factorization

Our data-driven method is based on a reduced-dimension representation of an AR ROI via matrix factorization of image patches. Matrix factorization is a widely used tool to reveal patterns in high dimensional datasets. Applications outside of solar physics are numerous and range e.g. from multimedia activity correlation, neuroscience, gene expression [12], to hyperspectral imaging [137].

The idea is to express a $k-$multivariate observation $\mathbf{z}_1$ as a linear combination of a reduced number of $r < k$ components $\mathbf{a}_j$, each weighted by some (possibly random) coefficients $h_{j,1}$:

$$\mathbf{z}_1 = \sum_{j=1}^{r} \mathbf{a}_j h_{j,1} \; + \mathbf{n}_1, \tag{6.1}$$

where $\mathbf{n}_1$ represents residual noise. With $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]$, the equivalent matrix factorization representation is written as

$$\mathbf{Z} = \mathbf{AH} + \mathbf{N} \tag{6.2}$$

where $\mathbf{Z}$ is a $k \times n$ data matrix containing n observations of k different variables, $\mathbf{A}$ is the $k \times r$ matrix containing the ‘dictionary elements’ (called ’factor loadings’ in some applications) and $\mathbf{H}$ is the $r \times n$ matrix of coefficients (or ‘factor scores’). The $k \times n$ matrix $\mathbf{N}$ contains residuals from the matrix factorization model fitting. Finding $\mathbf{A}$ and $\mathbf{H}$ from the knowledge of $\mathbf{Z}$ alone is a severely ill-posed problem, hence prior knowledge is needed to constrain the solution to be unique.

Principal Component Analysis (PCA) [100] is probably the most widely used dimensionality reduction technique. It seeks principal directions that capture the highest variance in the data under the constraints that these directions are mutually orthogonal, thereby defining a subspace of the initial space that exhibit information rather than noise. The PCA solution can be written as a matrix factorization thanks to the Singular Value Decomposition (SVD) [147], and so we use SVD in the clustering method presented here.

The Nonnegative Matrix Factorization (NMF) [119] is also considered in this chapter. Instead of imposing orthogonality, it constrains elements of matrices $\mathbf{A}$ and $\mathbf{H}$ to be nonnegative. We further impose that each column of $\mathbf{H}$ has elements that sum up to one, thereby effectively using a formulation identical to one used in hyperspectral unmixing [20].

Unmixing techniques exploit the high redundancy observed in similar bandpasses. They aim at separating the various contributions and at estimating a smaller set of less dependent source images. Matrix factorization, known as 'blind source separation' in this context, has many applications, ranging from biomedical imaging, chemometrics, to remote sensing [37], and recently to the extraction of salient morphological features from multi-wavelength extreme ultraviolet solar images [49].

In this chapter, we wish to factorize a $k \times n$ data matrix $\mathbf{Z}$ containing n observations of k different variables as in (6.2) where the dictionary matrix $\mathbf{A}$ spans a subspace of the initial space, with $r < k$. We consider the cases where $\mathbf{Z}$ is formed from a single image as well as from multiple images.

When a single image is used, the data matrix $\mathbf{Z}$ is built from a n pixel image by taking overlapping $m \times m$-pixel neighborhoods called *patches*. Figure 6.2 (left) presents such a patch and its column representation. The k rows of the $i-$th column of $\mathbf{Z}$ are thus given by the m^2 pixel values in the neighborhood of pixel i. The right plot in Figure 6.2 provides the number of patches in each pair of AR images when using the STARA masks. When multiple images are used, such as when analyzing collectively all images from a given Mount Wilson class, the patches are combined into a single data matrix. Table 5.1 gives the total number of patches from each Mount Wilson class when using the STARA masks.

A factorization of a data matrix containing image patches is illustrated in Figure 6.3. In this figure, let $\mathbf{z}_1$ be the first column of $\mathbf{Z}$, containing the intensity values for the first patch. These intensity values are decomposed as a sum of r elements as in Equation (6.1) where $\mathbf{a}_j$ is the $j-$th column of $\mathbf{A}$ and $h_{j,1}$ is the $(j, 1)-$th element of $\mathbf{H}$. In this representation, the vectors $\mathbf{a}_j, j = 1, \ldots, r$ are the elementary building blocks common to all patches, whereas the $h_{j,1}$ are the coefficients specific to the first

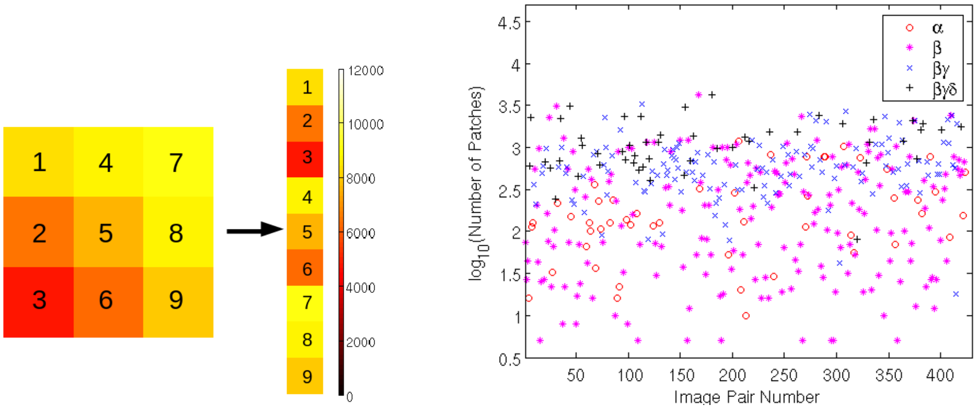

Figure 6.2: (Left) An example of a 3 × 3 pixel neighborhood or *patch* extracted from the edge of a sunspot in a continuum image and its column representation. (Right) The number of patches extracted from each pair of AR images when using the STARA masks.

patch.

To compare ARs and cluster them based on this reduced dimension representation, some form of distance is required. To measure the distance between two ARs, we apply some metrics to the corresponding matrices $\mathbf{A}$ or $\mathbf{H}$ obtained from the factorizations of the two ARs. These distances are further introduced into a clustering algorithm that groups ARs based on the similarity of their patch geometry.

6.1.4 Outline

Section 6.2 describes two matrix factorization methods: the singular value decomposition (SVD) and nonnegative matrix factorization (NMF). While more sophisticated methods exist that may lead to improved performance, we focus on SVD and NMF to demonstrate the utility of an analysis of a reduced dimension representation of image patches for this problem. Future work will include further refinement in the choice of matrix factorization techniques. To compare the results from this factorization we need a metric, and so we use the Hellinger distance for this purpose. To obtain some insight on how these factorizations separate the data, we make some general comparisons in Section 6.3. In particular, with the defined metric, we compute the

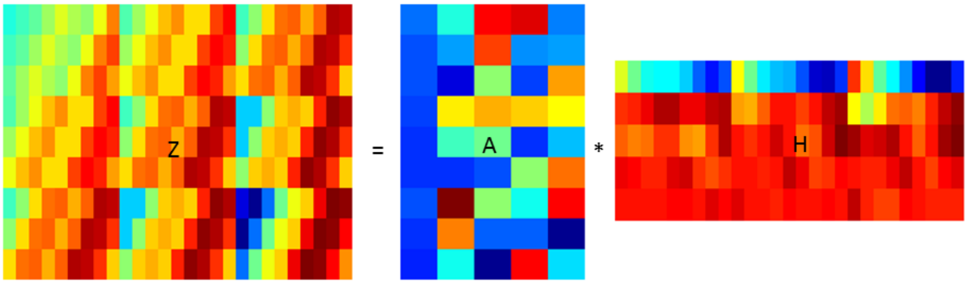

Figure 6.3: An example of linear dimensionality reduction where the data matrix of AR image patches **Z** is factored as a product of a dictionary **A** of representative elements and the corresponding coefficients in the matrix **H**. The **A** matrix consists of the basic building blocks for the data matrix **Z** and **H** contains the corresponding coefficients.

pairwise distances between Mount Wilson classes to identify which classes are most similar or dissimilar according to the matrix factorization results.

Section 6.4 describes the clustering procedures that take the metrics' output as input. The method called 'Evidence Accumulating Clustering with Dual rooted Prim tree Cuts' (EAC-DC) was introduced by [67] and is used to cluster the ARs. By combining the two matrix factorization methods, a total of two procedures are used to analyze the data. Besides analyzing the whole sunspot data, we also look at information contained in patches situated along the neutral lines. The results of the clustering analyses are provided in Section 6.5.

6.2 Matrix Factorization

The intrinsic dimension analysis in Chapter V showed that linear methods (e.g. matrix factorization) are sufficient to represent the data, and hence we focus on those. Matrix factorization methods aim at finding a set of basis vectors or dictionary elements such that each data point (in our case, pair of pixel patches) can be accurately expressed as a linear combination of the dictionary elements. Mathematically, if we use $m \times m$ patches then this can be expressed as $\mathbf{Z} \approx \mathbf{AH}$, where $\mathbf{Z}$ is the $2m^2 \times n$

data matrix with n data points being considered, $\mathbf{A}$ is the $2m^2 \times r$ dictionary with the columns corresponding to the dictionary elements, and $\mathbf{H}$ is the $r \times n$ matrix of coefficients. The goal is to find matrices $\mathbf{A}$ and $\mathbf{H}$ whose product nearly approximates $\mathbf{Z}$. The degree of approximation is typically measured by the squared error $||\mathbf{Z} - \mathbf{AH}||_F^2$, where $|| \cdot ||_F$ denotes the Frobenius norm [204]. Additional assumptions on the structure of the matrices $\mathbf{A}$ and $\mathbf{H}$ can be applied in matrix factorization depending on the application. Examples include assumptions of orthonormality of the columns of the dictionary $\mathbf{A}$, sparsity of the coefficient matrix $\mathbf{H}$ [166], and nonnegativity on $\mathbf{A}$ and $\mathbf{H}$ [125].

We consider two popular matrix factorization methods: the singular value decomposition (SVD) and nonnegative matrix factorization (NMF).

6.2.1 Factorization using SVD

To perform matrix factorization using SVD, we take the singular value decomposition of the data matrix $\mathbf{Z} = \mathbf{U\Sigma V}^T$ where $\mathbf{U}$ is the matrix of the left singular vectors, $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values, and $\mathbf{V}$ is a matrix of the right singular vectors. If the size of the dictionary r is fixed and is less than $2m^2$, then the matrix of rank r that is closest to $\mathbf{Z}$ in terms of the Frobenius norm is the matrix product $\mathbf{U}_r \Sigma_r \mathbf{V}_r^T$, where $\mathbf{U}_r$ and $\mathbf{V}_r$ are matrices containing only the first r singular vectors and Σ_r contains only the first r singular values [147]. Thus for SVD, the dictionary and coefficient matrices are $\mathbf{A} = \mathbf{U}_r$ and $\mathbf{H} = \Sigma_r \mathbf{V}_r^T$, respectively. Note that SVD enforces orthonormality on the columns of $\mathbf{U}_r$.

The intrinsic dimension estimated in Chapter V determines the number of parameters required to accurately represent the data. It is used to provide an initial estimate for the size of the dictionaries r. For SVD, we choose r to be one standard deviation above the mean intrinsic dimension estimate, that is, $r \simeq 5$ or 6. The choice of r is then further refined by a comparison of dictionaries in Section 6.3. See

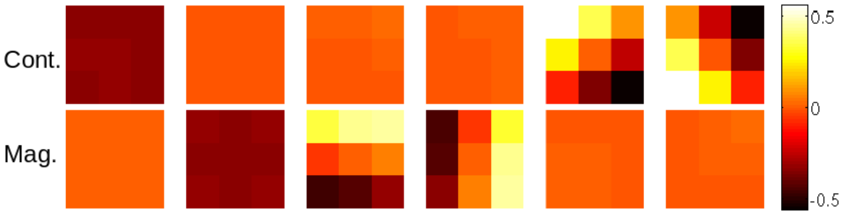

Figure 6.4: Learned dictionary elements using SVD. Dictionary elements are constrained to be orthonormal. The patches consist of uniform patches and gradients in varied directions. The magnetogram patches are essentially zero when the continuum components are nonzero and vice versa. The dictionary size r is first chosen based on the intrinsic dimension estimates in Chapter V and then refined by comparing dictionaries of various sizes in Section 6.3. Section 6.4.2 contains more details on choosing r.

Section 6.4.2 for more on selecting the dictionary size.

Figure 6.4 shows the learned dictionaries using SVD on the entire data set of 424 image pairs of ARs. Interestingly, the SVD seems to consider the continuum and magnetogram separately as the magnetogram elements are essentially zero when the continuum elements are not and vice versa. This is likely caused by the orthonormality constraint. The dictionary patches largely consist of a mix of uniform patches and patches with gradients in varied directions. The second dictionary element is associated with the average magnetic field value of a patch.

6.2.2 Factorization using NMF

Non-negative matrix factorization (NMF) [119] solves the problem of minimizing $||\mathbf{Z}-\mathbf{AH}||_F^2$ while constraining $\mathbf{A}$ and $\mathbf{H}$ to have nonnegative values. Thus NMF is a good choice for matrix factorization when the data is nonnegative. For our problem, the continuum data is nonnegative while the magnetogram data is not. Therefore we use a modified version of NMF using projected gradient where we only constrain the parts of $\mathbf{A}$ corresponding to the continuum to be nonnegative. An effect of using this modified version of NMF is that since the coefficient matrix $\mathbf{H}$ is still constrained to be nonnegative, we require separate dictionary elements that are either positive or

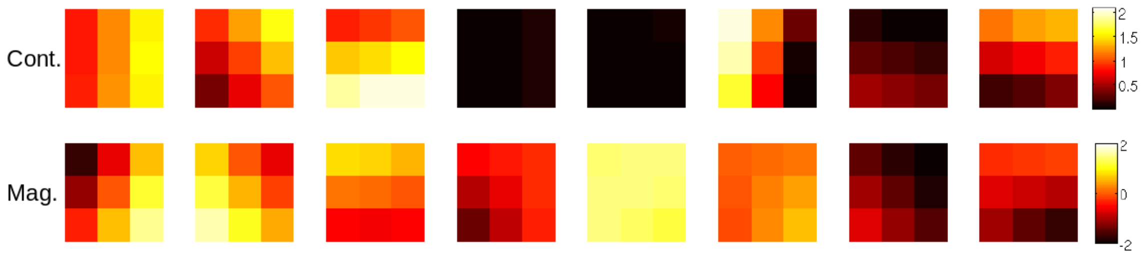

Figure 6.5: Learned dictionary elements using NMF where the continuum dictionary elements are constrained to be nonnegative. All the dictionary patches consist of uniform patches or gradients in varied directions. The order of the elements is not significant. The dictionary size r is chosen to be approximately 1.5 times larger than the SVD dictionary size, which is chosen based on the intrinsic dimension estimates in [**?**] and then refined using the results in Section 6.3.

negative in the magnetogram component. Thus we use approximately 1.5 times more dictionary elements for NMF than SVD.

Since we apply NMF to the full data matrix $\mathbf{Z}$, this enforces a coupling between the two modalities by forcing the use of the same coefficient matrix to reconstruct the matrices $\mathbf{X}$ and $\mathbf{Y}$. This is similar to coupled NMF which has been used in applications such as hyperspectral and multispectral data fusion [206].

Figure 6.5 shows the learned dictionary elements using NMF on the entire dataset. For NMF, the modalities are not treated separately as in the SVD results. But as for SVD, the patches largely consist of a mix of uniform patches and patches with gradients in varied directions.

6.2.3 SVD vs. NMF

There are advantages to both SVD and NMF matrix factorization methods which are summarized in Table 6.1. SVD produces the optimal rank r approximation of $\mathbf{Z}$, is fast and unique, and results in orthogonal elements [115]. NMF has the advantages of nonnegativity, sparsity, and interpretability. The interpretability comes from the additive parts-based representation inherent in NMF [115]. In contrast, the SVD results are not sparse which can make interpretation more difficult. However, NMF is

SVD Advantages	NMF Advantages
Optimal rank r approximation	Results are nonnegative
Fast to compute	Results are sparse
Unique	Sparsity and nonnegativity lead to improved interpretability

Table 6.1: Summary of the advantages of SVD and NMF matrix factorization methods. The advantages of one method complement the disadvantages of the other [115]. For example, the NMF optimization problem is nonconvex with local minima resulting in solutions that depend on the initialization of the algorithm.

not as robust as SVD as the NMF algorithm is a numerical approximation to a nonconvex optimization problem having local minima. Thus the solution provided by the NMF algorithm depends on the initialization. More details on matrix factorization using NMF and SVD are included in Appendix E.2.

6.2.4 Methods for Comparing Matrix Factorization Results

To compare the results from matrix factorization, we primarily seek a difference between the coefficients from $\mathbf{H}$. To aid us in choosing a dictionary size r, we also require a measure of difference between dictionaries $\mathbf{A}$. We use the Hellinger distance and Grassmannian projection metric to measure the respective differences.

In [142], the Frobenius norm was used to compare the dictionaries. However, this fails to take into account the fact that two dictionaries may have the same elements but in a different order. In this case, the Frobenius norm of the difference between two dictionaries may be high even though the dictionaries span the same subspace. A better way to measure the difference would be to compare the subspaces spanned by the dictionaries. The Grassmannian $\mathbf{Gr}(r, V)$ is a space which parameterizes all linear subspaces with dimension r of a vector space V. As an example, the Grassmannian $\mathbf{Gr}\left(2, \mathbb{R}^n\right)$ is the space of planes through the origin of the standard Euclidean vector space in $\mathbb{R}^n$. In our case, we are concerned with the Grassmannian $\mathbf{Gr}\left(r, \mathbb{R}^{18}\right)$, where r is the size of the dictionary. The space spanned by a given dictionary $\mathbf{A}$ is then a single point in $\mathbf{Gr}\left(r, \mathbb{R}^{18}\right)$. Several metrics have been defined on this space including the Grassmannian projection metric [52, 189]. It can be defined as

$$d_G(\mathbf{A}, \mathbf{A}') = \|\mathbf{P}_{\mathbf{A}} - \mathbf{P}_{\mathbf{A}'}\|,$$

where $\mathbf{P}_{\mathbf{A}} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$ is the projection matrix of $\mathbf{A}$ and $\|\cdot\|$ is the ℓ_2 norm. This metric is invariant to the order of the dictionary elements and compares the subspaces spanned by the dictionaries. This metric has a maximum value of 1.

To compare the coefficient matrices, we assume that the 18-dimensional pixel patches within an AR are samples from an 18-dimensional probability distribution, and that each AR has a corresponding (potentially unique) probability distribution of pixel patches. We project these samples onto a lower dimensional space by matrix factorization. In other words, we learn a dictionary $\mathbf{A}$ and the coefficient matrix $\mathbf{H}$ for the entire dataset $\mathbf{Z}$, and then separate the coefficients in $\mathbf{H}$ according to the K different ARs (or groups of ARs) considered: $\mathbf{Z} = \mathbf{A}\begin{pmatrix} \mathbf{H}_1 & \mathbf{H}_2 & \dots & \mathbf{H}_K \end{pmatrix}$. The columns of $\mathbf{H}_i$ are a collection of projected, low-dimensionality, samples from the ith AR (or group), and we let f_i denote the corresponding probability density function. Given two such collections, we can estimate the difference between their probability distributions by estimating the *information divergence*. Many kinds of divergences exist such as the popular Kullback-Leibler divergence [113]. We use the Hellinger distance which is defined as [18, 42, 87]

$$H(f_i, f_j) = 1 - \int \sqrt{f_i(x) f_j(x)}dx,$$

where f_i and f_j are the two probability densities being compared. The Hellinger distance has the advantage over other divergences of being a metric which is not true of divergences in general. To estimate the Hellinger distance, we use the nonparametric estimator derived in Chapter III that is based on the k-nearest neighbor density estimators for the densities f_i and f_j.

6.3 Comparisons of General Matrix Factorization Results

We apply the metrics mentioned in Section 6.2.4 and compare the local features as extracted by matrix factorization per Mount Wilson class. One motivation for these comparisons is to investigate differences between the Mount Wilson classes based on the Hellinger distance. Another motivation is to further refine our choice of dictionary size r in preparation for clustering the ARs. When comparing the dictionary coefficients using the Hellinger distance, we want a single, representative dictionary that is able to accurately reconstruct all of the images. Then the ARs will be differentiated based on their respective distributions of dictionary coefficients instead of the accuracy of their reconstructions. The coefficient distributions can then be compared to interpret the clustering results as is done in Section 6.5.1.

Recall that our goal is to use unsupervised methods to separate the data based on the natural geometry. Our goal is not to replicate the Mount Wilson results. Instead we use the Mount Wilson labels in this section as a vehicle for interpreting the results.

6.3.1 Grassmannian Metric Comparisons

We first learn dictionaries for each of the Mount Wilson types by applying matrix factorization to a subset of the patches corresponding to sunspot groups of the respective type. We then use the Grassmannian metric to compare the dictionaries. For example, if we want to compare the α and β groups, we collect a subset of patches from all ARs designated as α groups into a single data matrix $\mathbf{Z}_\alpha$. We then factor this matrix with the chosen method to obtain $\mathbf{Z}_\alpha = \mathbf{A}_\alpha \mathbf{H}_\alpha$. Similarly, we obtain $\mathbf{Z}_\beta = \mathbf{A}_\beta \mathbf{H}_\beta$ and then calculate $d_G(\mathbf{A}_\alpha, \mathbf{A}_\beta)$.

The reason we use only a subset of patches is that each AR type has a different number of total patches (see Table 5.1) which may introduce bias in the comparisons. One source of potential bias in this case is due to the potentially increased patch variability in groups with more patches, which would result in increased difficulty in characterizing certain homogeneities of the patch features. This is mitigated somewhat by the fact that the local intrinsic dimension is typically less than 6. However, it is possible that there may be different local subspaces with the same dimension. A second source of potential bias is in the different levels of variance of the estimates due to difference in patch numbers. To circumvent these potential biases, we use the same number of patches in each group for each comparison. For the inter-class comparison, we randomly take 13,358 patches (the number of patches in the smallest class) from each class to learn the dictionary, and then calculate the Grassmannian metric. For the intra-class comparison, we take two disjoint subsets of 6,679 patches (half the number of patches in the smallest class) from each class to learn the dictionaries. This process is repeated 100 times and the resulting mean and standard deviation are reported.

Table 6.2 shows the corresponding average Grassmannian distance metrics when using SVD and NMF and for different sizes of dictionaries r. For SVD, the results are very sensitive to r. Choosing $r = 5$ results in large differences between the different dictionaries but for $r = 6$, the dictionaries are very similar. This suggests that for SVD, 6 principal components are sufficient to accurately represent the subspace upon which the sunspot patches lie. This is consistent with the results in Chapter **??** where the intrinsic dimension is found to be less than 6 for most patches.

Interestingly, for the $r = 5$ SVD results, the $\beta\gamma$ group is the most dissimilar to the other groups while being relatively similar to itself. In contrast, the β group is fairly dissimilar to itself and relatively similar to the α and $\beta\gamma\delta$ groups.

The NMF results are less sensitive to r. The average difference between the dictionaries and its standard deviation is larger when $r = 9$ compared to when $r = 8$. However, for a given r, all of the mean differences are within a standard deviation of each other. Thus on aggregate, the NMF dictionaries learned from large collections of patches from multiple images differ from each other to the same degree regardless

SVD, Pooled Grassmannian, $r = 5$

	α	β	$\beta\gamma$	$\beta\gamma\delta$
α	0.15 ± 0.10	0.26 ± 0.18	0.89 ± 0.06	0.34 ± 0.18
β		0.50 ± 0.29	0.89 ± 0.14	0.43 ± 0.27
$\beta\gamma$			0.24 ± 0.16	0.7 ± 0.2
$\beta\gamma\delta$				0.45 ± 0.28

SVD, Pooled Grassmannian, $r = 6$

	α	β	$\beta\gamma$	$\beta\gamma\delta$
α	0.02 ± 0.004	0.03 ± 0.003	0.02 ± 0.004	0.04 ± 0.005
β		0.02 ± 0.005	0.02 ± 0.004	0.03 ± 0.006
$\beta\gamma$			0.03 ± 0.006	0.03 ± 0.006
$\beta\gamma\delta$				0.03 ± 0.007

NMF, Pooled Grassmannian, $r = 8$

	α	β	$\beta\gamma$	$\beta\gamma\delta$
α	0.40 ± 0.13	0.40 ± 0.10	0.33 ± 0.09	0.37 ± 0.10
β		0.29 ± 0.13	0.35 ± 0.09	0.37 ± 0.11
$\beta\gamma$			0.37 ± 0.12	0.34 ± 0.10
$\beta\gamma\delta$				0.41 ± 0.11

NMF, Pooled Grassmannian, $r = 9$

	α	β	$\beta\gamma$	$\beta\gamma\delta$
α	0.62 ± 0.25	0.41 ± 0.15	0.45 ± 0.19	0.40 ± 0.13
β		0.54 ± 0.23	0.49 ± 0.19	0.44 ± 0.19
$\beta\gamma$			0.53 ± 0.23	0.44 ± 0.20
$\beta\gamma\delta$				0.49 ± 0.20

Table 6.2: Difference between dictionaries learned from the collection of sunspot patches corresponding to the different Mount Wilson types as measured by the Grassmannian metric d_G, e.g. $d_G(\mathbf{A}_\alpha, \mathbf{A}_\beta)$. Dictionaries are learned using random subsets of the data and the results are reported in the form of mean±standard deviation using 100 trials. Different sizes of dictionaries r are used. The SVD results are sensitive to r.

of the Mount Wilson type.

6.3.2 Hellinger Distance Comparisons

For the Hellinger distance, we learn a dictionary $\mathbf{A}$ and the coefficient matrix $\mathbf{H}$ for the entire data set $\mathbf{Z}$. We then separate the coefficients in $\mathbf{H}$ according to the Mount Wilson type and compare the coefficients' distributions using the Hellinger distance. For example, suppose that the data matrix is arranged as $\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_\alpha & \mathbf{Z}_\beta & \mathbf{Z}_{\beta\gamma} & \mathbf{Z}_{\beta\gamma\delta} \end{pmatrix}$. This is factored as $\mathbf{Z} = \mathbf{A}\begin{pmatrix} \mathbf{H}_\alpha & \mathbf{H}_\beta & \mathbf{H}_{\beta\gamma} & \mathbf{H}_{\beta\gamma\delta} \end{pmatrix}$. To compare the α and β groups, we assume that the columns in $\mathbf{H}_\alpha$ are samples from the distribution f_α and similarly $\mathbf{H}_\beta$ contains samples from the distribution f_β. We then estimate the Hellinger distance $H(f_\alpha, f_\beta)$ using the divergence estimator from Chapter II.

When the Hellinger distance is used to compare the collections of dictionary coefficients within the sunspots, the groups are very similar, especially when using SVD (Table 6.3). This indicates that when the coefficients of all ARs from one class are grouped together, the distribution looks similar to the distribution of the other classes. However, there are some small differences. First the intraclass distances are often much smaller than the interclass distances which indicates that there is some relative difference between most classes. Second, for both matrix factorization methods, the $\beta\gamma\delta$ groups are the most dissimilar. This could be due to the presence of a δ spot configuration, where umbrae of opposite polarities are within a single penumbra. Such a configuration may require specific linear combinations of the dictionary elements as compared to the other classes. The presence and absence of these linear combinations in two Mount Wilson types would result in a higher Hellinger distance between them.

Again, for clustering, we compute the pairwise Hellinger distance between each AR's collection of coefficients. This is done by forming the data matrix from the 424

SVD, Pooled Hellinger

	α	β	$\beta\gamma$	$\beta\gamma\delta$
α	0.0006 ± 0.004	0	0	0.03
β		0.0005 ± 0.002	0.01	0.08
$\beta\gamma$			0.0003 ± 0.002	0.05
$\beta\gamma\delta$				0.0004 ± 0.002

NMF, Pooled Hellinger

	α	β	$\beta\gamma$	$\beta\gamma\delta$
α	0 ± 0	0.08	0.05	0.10
β		0.00007 ± 0.0004	0.03	0.12
$\beta\gamma$			0.000002 ± 0.00003	0.11
$\beta\gamma\delta$				0.00001 ± 0.0002

Table 6.3: Difference between the collection of dictionary coefficients pooled from the different Mount Wilson classes as measured by the Hellinger distance. Intraclass distances are reported in the form of mean±standard deviation and are calcuated by randomly splitting the data and then calculating the distance over 100 trials. The size of the dictionaries is $r = 6$ and 8 for SVD and NMF, respectively. The $\beta\gamma\delta$ group is most dissimilar to the others.

ARs as $\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 & \dots & \mathbf{Z}_{424} \end{pmatrix}$ and factoring it as $\mathbf{Z} = \mathbf{A} \begin{pmatrix} \mathbf{H}_1 & \mathbf{H}_2 & \dots & \mathbf{H}_{424} \end{pmatrix}$. The columns of $\mathbf{H}_i$ are samples from a distribution f_i and the distributions f_i and f_j are compared by estimating $H(f_i, f_j)$. The corresponding dictionaries for the two methods are shown in Figures 6.4 and 6.5.

Table 6.4 gives the average pairwise Hellinger distance between the ARs. The average distances differ more with the NMF based coefficients resulting in larger dissimilarity. The average distance is smallest when comparing the β groups to all others and largest when comparing the $\beta\gamma$ groups to the rest. The standard deviation is also larger when comparing α and β groups. This may be partially related to the variability in estimation due to smaller sample sizes as the α and β groups contain more of the smaller ARs (see Figure 6.2). Overall, the average distances show that there are clear differences between the ARs within the sunspots using this metric.

SVD, Average Hellinger

	α	β	$\beta\gamma$	$\beta\gamma\delta$
α	0.83 ± 0.21	0.80 ± 0.20	0.82 ± 0.16	0.80 ± 0.14
β		0.75 ± 0.22	0.78 ± 0.18	0.77 ± 0.17
$\beta\gamma$			0.83 ± 0.15	0.81 ± 0.14
$\beta\gamma\delta$				0.78 ± 0.13

NMF, Average Hellinger

	α	β	$\beta\gamma$	$\beta\gamma\delta$
α	0.85 ± 0.21	0.81 ± 0.20	0.84 ± 0.17	0.82 ± 0.14
β		0.76 ± 0.23	0.80 ± 0.19	0.80 ± 0.17
$\beta\gamma$			0.85 ± 0.15	0.85 ± 0.14
$\beta\gamma\delta$				0.83 ± 0.12

Table 6.4: Average pairwise difference between dictionary coefficients from each AR from different Mount Wilson types as measured by the Hellinger distance. Results are reported in the form of mean±standard deviation. The size of the dictionaries is $r = 6$ and 8 for SVD and NMF, respectively. The $\beta\gamma$ ARs are most dissimilar to each other and the other classes while the β ARs are most similar.

6.4 Clustering of Active Regions: Methods

6.4.1 Clustering Algorithm

The clustering algorithm we use is the Evidence Accumulating Clustering with Dual rooted Prim tree Cuts (EAC-DC) method in [67] which scales well for clustering in high dimensions. EAC-DC clusters the data by defining a metric based on the growth of two minimal spanning trees (MST) grown sequentially from a pair of points. To grow the MSTs, a base dissimilarity measure is required as input such as the Hellinger distance described in Section 6.2.4. From the new metric defined using the MSTs, a similarity measure between inputs is created. It is fed into a spectral clustering algorithm that groups together inputs which are most similar. The similarity measure based on the MSTs adapts to the geometry of the data, and this results in a clustering method that is robust and competitive with other algorithms [67]. See Appendix E.3 for more details.

Measure	Type	Properties
Grassmannian metric	Dissimilarity	Compares dictionaries by comparing the subspace spanned by the dictionary elements
Hellinger distance	Dissimilarity	Compares the underlying distributions of dictionary coefficients
EAC-DC based measure	Similarity	Based on sequentially grown MSTs of the data; requires a base dissimilarity measure as input

Table 6.5: Summary of the dissimilarity and similarity measures used.

6.4.2 Clustering Input: Dictionary Sizes

As input to the clustering algorithm, we use the matrix factorization results as described in Section 6.2.4. We learn a single dictionary from the entire dataset. We then project the data onto a lower dimensional space, i.e. we learn the coefficient matrices $\mathbf{H}_i$. These matrices are the inputs in this method and the base dissimilarity measure is the Hellinger distance estimated using each AR's respective coefficients. Table 6.5 provides a summary of the various dissimilarity and similarity measures that we use.

As mentioned in Section 6.2, the estimated intrinsic dimension from Chapter V is used to provide an initial estimate for the size of the dictionaries r. The choice of r is further refined by the dictionary comparison results from Section 6.3 . For SVD, we choose r to be one standard deviation above the mean intrinsic dimension estimate which is approximately 5 or 6. When comparing the dictionary coefficients, we want the single dictionary to accurately represent the images. The dictionary should not be too large as this may add spurious dictionary elements due to the noise. The results in Table 6.2 suggest that for SVD, the dictionaries are essentially identical for $r = 6$. This means that 6 dictionary elements are sufficient to accurately reconstruct most of the images. This is consistent with the intrinsic dimension estimates in Chapter V. Thus we choose $r = 6$ when using the Hellinger distance for the SVD dictionary coefficients.

Since our mixed version of NMF requires approximately 1.5 times the number of dictionary elements as SVD (see Section 6.2.1), we choose $r = 8$ within the sunspots. Since the differences between classes were similar for $r = 8$ and $r = 9$, choosing $r = 8$ strikes a balance between accurate representation of the data and limiting the effects of noise.

6.4.3 Clustering Input: Patches within Sunspots and Along the Neutral Line

Our main focus up to this point in this chapter has been on data matrices $\mathbf{Z}$ containing the patches within the STARA masks, that is, within sunspots. The clustering based on these patches is discussed in Sections 6.5.1-6.5.2.

It is well-known, however, that the shape of the neutral line separating the main polarities plays an important role in the Mount Wilson classification. For this reason, we conduct two experiments involving data from along the neutral line.

The results of the first experiment are in Section 6.5.3 where we apply matrix factorization on a data matrix containing only the patches situated along the neutral line using the same ARs as in Sections 6.5.1-6.5.2. To compute the location of a neutral line in this experiment, we assume it is situated in the middle between regions of opposite polarity, and proceed as follows. First, we determine regions of high magnetic flux of each polarity using an absolute threshold at 50 Gauss. Second, we compute for each pixel the distance to the closest high flux region in each polarity using the Fast Marching method [178]. Once the two distance fields (one for each polarity) are calculated, the neutral line can be obtained by finding the pixels that lie on or are close to the zero level set of the difference of these two distance fields. In this chapter, we choose a maximum distance of 10 pixels to determine the neutral line region.

We extract the patches that lie in the neutral line region and within the SMART mask associated to the AR. Call the resulting data matrix $\mathbf{Z}_N$. We then apply SVD or

# of clusters	Mount Wilson Comparison
2	Simple (α and β), complex ($\beta\gamma$ and $\beta\gamma\delta$)
3	α, β, and complex
4	Mount Wilson (α, β, $\beta\gamma$, and $\beta\gamma\delta$)

Table 6.6: The labels used to compare with the clustering results when analyzing the effects of including the neutral line.

NMF matrix factorization as before and calculate the pairwise distance between each AR neutral line using the Hellinger distance on the results. Define the resulting 424×424 dissimilarity matrix as $\mathbf{D}_N$ for whichever factorization method we are currently using. Similarly, define $\mathbf{Z}_S$ and $\mathbf{D}_S$ as the respective data matrix and dissimilarity matrix of the data from within the sunspots using the same configuration. The base dissimilarity measure $\mathbf{D}$ inputted in the clustering algorithm is now a *weighted average* of the distances computed within the neutral line regions and within the sunspots: $\mathbf{D} = w\mathbf{D}_N + (1-w)\mathbf{D}_S$ where $0 \leq w \leq 1$. Using a variety of weights, we then compare the clustering output to different labeling schemes based on the Mount Wilson labels as shown in Table 6.6.

For the second experiment, we perform clustering on a ROI that selects pixels along a strong field polarity reversal line. The high gradients near strong field polarity reversal lines in LOS magnetograms are a proxy for the presence of near-photospheric electrical currents, and thus might be indicative of a non-potential configuration [176]. To compute this ROI, the magnetograms are first reprojected using an equal-area, sinusoidal re-projection that uses Singular-Value Padding [44]. The latter is known to be more accurate than image interpolation on transformed coordinates. To conserve flux, the magnetograms are also area-normalized.

In the reprojected magnetograms, we delimit an AR using sunspot information from the Debrecen catalog [79] to obtain the location of all the pixels that belong to the spots related to an AR (called ”Debrecen spots” hereafter). The ROI consists of a binary array constructed as follows:

1. The pixels that belong to the Debrecen spots are assigned the scalar value 1 and all others are assigned the scalar value -1.

2. From this two-valued array, we retrieve the distances from the zero level-set using a fast marching method [178] implemented in the Python SciKits' module "scikit-fmm".

3. We mask out the pixels within a distance of 80 pixels from the zero level-set (in the equal-area-reprojected coordinate system).

4. The ROI is delimited by the convex hull of the resulting mask, that is, the smallest convex polygon that surrounds all 1-valued pixels. The resulting mask is a binary array with the pixels inside the convex hull set to True.

Within the ROI, the R-value is calculated similarly to the method described in [176]. We dilate the image using a dilation factor of 3 pixels, and extract the flux using overlapping gaussian masks with $\sigma = 2\,px$. Integrating the non-zero values outputs the R-value, i.e, the total flux in the vicinity of the polarity-inversion line.

We then perform an image patch analysis using image patches from this region. We do this by using either SVD or NMF to do dimensionality reduction on the patches, and then estimate the Hellinger distance between ARs using the reduced dimension representation. We exclude α groups from the analysis as they do not have a strong field polarity reversal line. This leaves 420 images to be clustered. The clustering assignments are then compared to the calculated R value via the correlation coefficient. The results are presented in Section 6.5.4.

6.5 Clustering of Active Regions: Results

Given the choices of matrix factorization techniques (NMF and SVD) we have two different clustering results on the data. Section 6.5.1 focuses on the clusterings using data from within the sunspots, and Section 6.5.2 provides some recommendations for which metrics and matrix factorization techniques to use to study different ARs. The neutral line clustering results are then given in Section 6.5.3 followed by the R value based experiment in Section 6.5.4.

6.5.1 Clustering within the Sunspot

We now present the clustering results when using the Hellinger distance as the base dissimilarity. The corresponding dictionary elements to the coefficients are represented in Figure 6.4 (for the SVD factorization) and in Figure 6.5 (for the NMF).

The EAC-DC algorithm does not automatically choose the number of clusters. We use the mean silhouette heuristic to determine the most natural number of clusters as a function of the data [171]. The silhouette is a measure of how well a data point belongs to its assigned cluster. The heuristic chooses the number of clusters that results in the maximum mean silhouette. In both clustering configurations, the number of clusters that maximize the mean silhouette is 2 so we focus on the two clustering case for all clustering schemes throughout.

To compare the clustering correspondence, we use the adjusted Rand index (ARI). The ARI is a measure of similarity between two clusterings (or a clustering and labels) and takes values between -1 and 1. A 1 indicates perfect agreement between the clusterings and a 0 corresponds to the agreement from a random assignment [167]. Thus a positive ARI indicates that the clustering correspondence is better than a random clustering while a negative ARI indicates it is worse. The ARI between the NMF and SVD clusterings is 0.27 which indicates some overlap.

Visualizing the clusters in lower dimensions is done with multidimensional scaling (MDS) as in [142]. Let $\mathbf{S}$ be the 424×424 symmetric matrix that contains the AR pair similarities as created by EAC-DC algorithm. MDS projects the similarity matrix $\mathbf{S}$ onto the eigenvectors of the normalized Laplacian of $\mathbf{S}$ [112]. Let $\mathbf{c}_i \in \mathbb{R}^{424}$ be the projection onto the ith eigenvector of $\mathbf{S}$ using NMF. The first eigenvector represents the direction of highest variability in the matrix $\mathbf{S}$, and hence a high value of the $k-$th element of $\mathbf{c}_1$ indicates that the $k-$th AR is dissimilar to other ARs.

Figure 6.6 displays the scatter plot of $\mathbf{c}_1$ vs. $\mathbf{c}_2$ (top) and $\mathbf{c}_1$ vs. $\mathbf{c}_3$ (bottom) using NMF. Comparing them with the Mount Wilson classification we see a concentration of simple ARs in the region with highest $\mathbf{c}_1$ values (most dissimilar ARs), and a concentration of complex ARs in the region with lowest $\mathbf{c}_1$ (more similar ARs). We can show this more precisely by computing the mean similarity of the ith AR to all other ARs as the mean of the ith row (or column) of $\mathbf{S}$. The value from $\mathbf{c}_1$ is then inversely related to this mean similarity as seen in Figure 6.7.

The similarity defined under this clustering scheme gathers in Cluster 2 ‘similar’ AR that are for a large part of the type $\beta\gamma$ and $\beta\gamma\delta$, whereas Cluster 1 contains AR that are more ‘dissimilar’ to each other, with a large part of α or β active regions. The other clustering configuration has a similar relationship between the first MDS coefficient and the mean similarity.

Table 6.7 makes this clearer by showing the mean similarity measure within each cluster and between the two clusters, which is calculated in the following manner. Suppose that the similarity matrix is organized in block form where the upper left block corresponds to Cluster 1 and the lower right block corresponds to Cluster 2. The mean similarity of Cluster 1 is calculated by taking the mean of all the values in the upper left block of this reorganized similarity matrix. The mean similarity of Cluster 2 is found similarly from the lower right block and the mean similarity between the clusters is found from either the lower left or upper right blocks. These means show that under the NMF clustering scheme, ARs in Cluster 2 are very similar to each other while ARs in Cluster 1 are not very similar to each other on average. In fact, the ARs in Cluster 1 are more similar to the ARs in Cluster 2 on average than to each other. The other clustering configuration has a similar relationship between

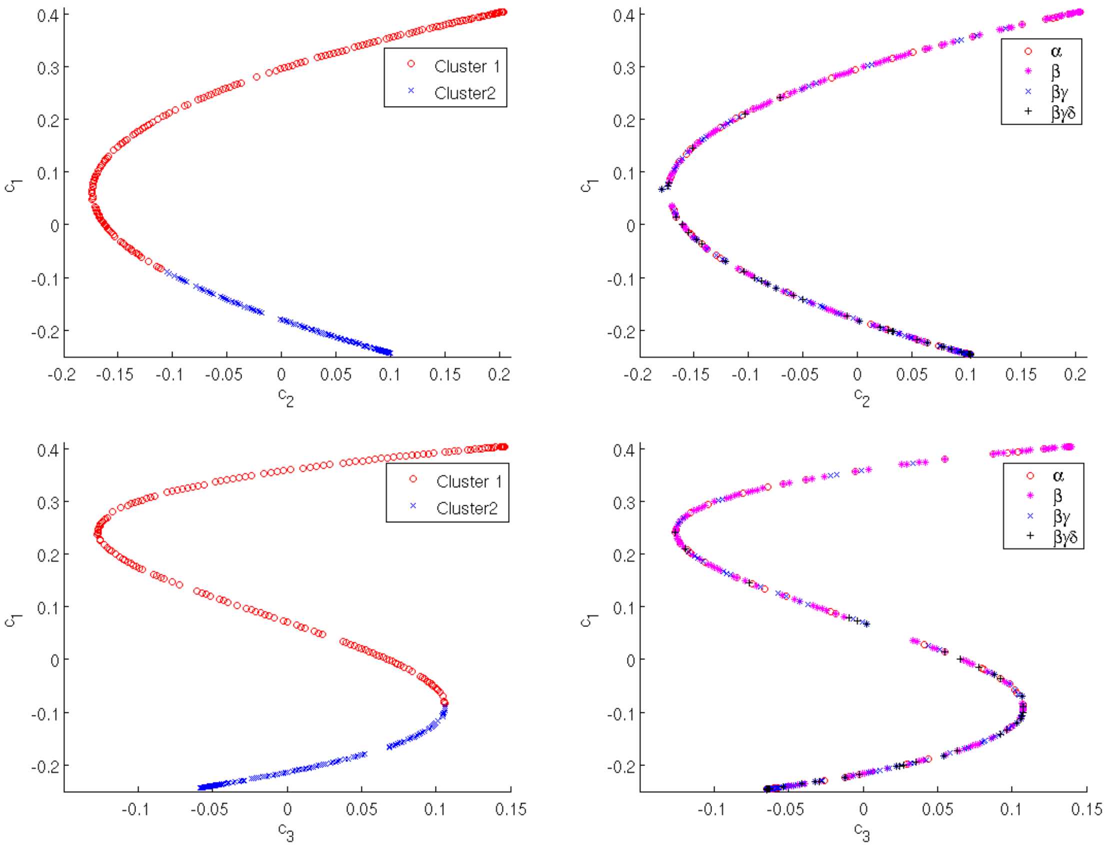

Figure 6.6: Scatter plot of MDS variables $\mathbf{c}_1$ vs. $\mathbf{c}_2$ (top) and $\mathbf{c}_1$ vs. $\mathbf{c}_3$ (bottom) where $\mathbf{c}_i \in \mathbb{R}^4 24$ is the projection of the similarity matrix onto the ith eigenvector of the normalized Laplacian of the similarity matrix when using the NMF coefficients. Each point corresponds to one AR and they are labeled according to the clustering (left) and the Mount Wilson labels (right). In this space, the clusters data appear to be separable and there are concentrations of complex ARs in the region with lowest $\mathbf{c}_1$ values. Other patterns are described in the text.

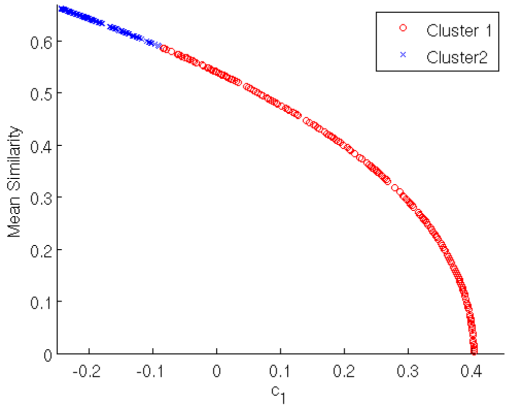

Figure 6.7: Mean similarity of an AR with all other ARs as a function of its MDS variable c_1 using NMF. Cluster 2 is associated with those ARs that are most similar to all other ARs while Cluster 1 contains those that are least similar to all others.

	Mean Similarity		
	1 vs. 1	1 vs. 2	2 vs. 2
SVD, Hellinger	0.29	0.42	0.87
NMF, Hellinger	0.30	0.42	0.88

Table 6.7: Mean similarity of ARs to other ARs either in the same cluster (1 vs. 1 or 2 vs. 2) or in the other cluster (1 vs. 2) under the different schemes. Cluster 1 contains ARs that are very dissimilar to each other while Cluster 2 contains ARs that are very similar to each other.

cluster assignment and mean similarity.

This relationship between AR complexity and clustering assignment is further noticeable in Figure 6.8 which gives a histogram of the Mount Wilson classes divided by clustering assignment. This figure shows clear patterns between the clusterings and Mount Wilson type distribution, where the clustering separates somewhat the complex sunspots from the simple sunspots. This suggests that these configurations are clustering based on some measure of AR complexity.

The Hellinger-based clusterings are correlated with sunspot size for some of the Mount Wilson classes, see Table 6.8. Based on the mean and median number of pixels, the Hellinger distance on the NMF coefficients tends to gather in Cluster 2 the smallest AR from classes α, β, and $\beta\gamma$. Similarly, the Hellinger distance on the

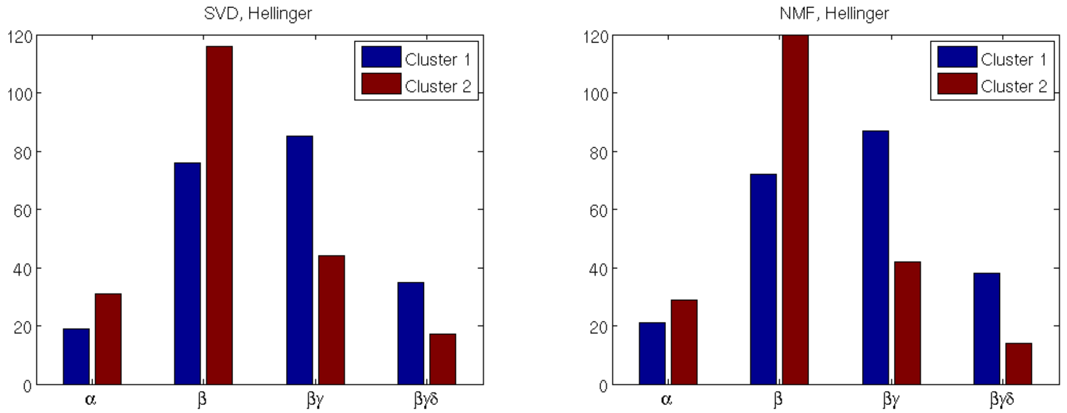

Figure 6.8: Histograms of the Mount Wilson classes divided by clustering assignment using the Hellinger distance. Cluster 1 contains more of the complex ARs while Cluster 2 contains more of the simple ARs.

	Number of Pixels									
	α		β		$\beta\gamma$		$\beta\gamma\delta$		All	
Cluster	1	2	1	2	1	2	1	2	1	2
Mean, SVD Hellinger	278	260	582	270	823	577	1234	1354	756	422
Mean, NMF Hellinger	384	183	788	156	847	515	1418	880	882	283
Median, SVD Hellinger	148	128	477	70	612	472	1012	1172	588	174
Median, NMF Hellinger	265	121	580	56	631	393	1157	665	677	105

Table 6.8: Mean and median number of pixels of the ARs in each cluster under the Hellinger clustering schemes. Cluster 1 contains the larger sunspots for all groups when using NMF and for some of the groups when using SVD.

SVD coefficients separates the β and $\beta\gamma$ AR by size with Cluster 1 containing the largest and Cluster 2 containing the smallest AR.

Since the Hellinger distance calculates differences between ARs based on their respective distribution of dictionary coefficients, we can examine the coefficient distribution to obtain insight on what features the clustering algorithm is exploiting. For simplicity, we examine the marginal histograms of the coefficients pooled from ARs of a given cluster. When looking at the SVD coefficients, we see that their marginal distributions are similar across clusters, except for the coefficients that correspond to the second dictionary element of Figure 6.4. Recall that this second dictionary element is associated with the average magnetic field value of a patch. If the corresponding coefficient is close to zero, it means the average magnetic field in the patch is also close to zero.

Figure 6.9 shows histograms of the coefficients of the second dictionary element. The histograms correspond to patches from all ARs separated by cluster assignment. The histograms show that Cluster 1 has a high concentration of patches with near zero average magnetic field. In contrast, the larger peaks for Cluster 2 are centered around $+1$ and -1. This suggests that the clustering assignments are influenced somewhat by the amount of patches in an AR that have near zero average magnetic field values. As we are considering only the core (sunspot) part of the AR, having 3×3 patch with a near zero average magnetic field entails that the corresponding patch is likely to be located along the neutral line separating strong magnetic fields of opposite polarity. Thus the local distribution of magnetic field values is related to cluster assignments when using the SVD coefficients. This is consistent with Figure 6.8 where cluster 1 contains more of the complex ARs ($\beta\gamma$ and $\beta\gamma\delta$) and fewer simple ARs (α and β) than cluster 2 as measured by the Mt. Wilson scheme.

Checking the individual ARs and their coefficient distributions in each cluster, we see indeed that Cluster 1 does contain more ARs with patches having near zero

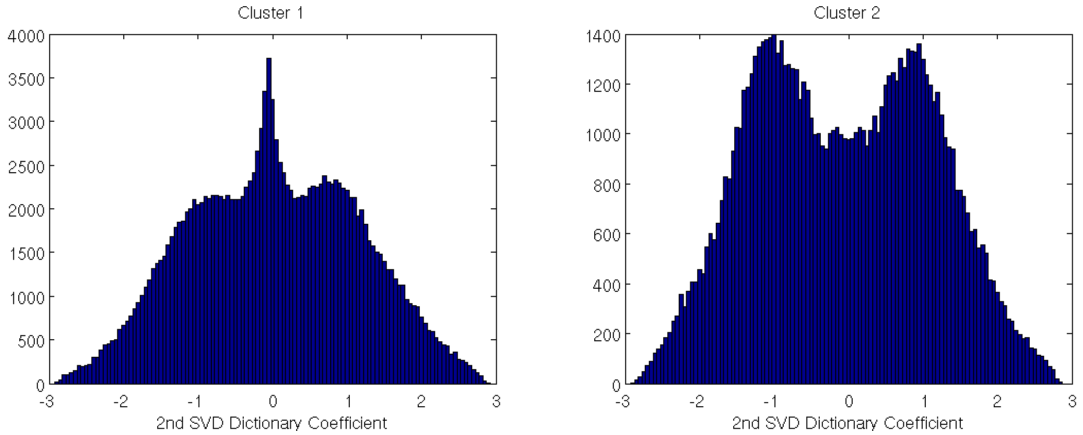

Figure 6.9: Histograms of the marginal distributions of the coefficients corresponding to the mean magnetic field value (dictionary element 2 in Figure 6.4) for Cluster 1 (left) and Cluster 2 (right) using the SVD coefficients. Cluster 1 ARs contain more patches with near neutral magnetic field values.

average magnetic field. This tends to include more of the complex ARs in Cluster 1 since they are more likely to have a neutral line close to the regions of strong magnetic fields that will therefore be included in the STARA masks.

It should be noted however that the correspondence is not perfect. There are some ARs in Cluster 2 where the regions of opposing polarity are close to each other and some ARs in Cluster 1 where the regions of opposing polarity are far apart. Thus the distribution of average magnetic field values is only one factor in the natural geometry of the ARs defined by the Hellinger distance. As mentioned previously, the size of the AR is another factor, especially for β groups. This is consistent with Figure 6.8 where cluster 1 does contain some of the simple ARs which are less likely to have strong magnetic fields around the neutral line.

Investigating the joint histograms of the NMF dictionary elements corresponding to positive and negative magnetic field values reveals that the NMF Hellinger clustering results are also influenced by the local magnetic field distribution.

All these observations indicate that the natural geometry exploited by both clustering configurations is related to some form of complexity of the ARs.

6.5.2 Discussion of Sunspot Results

We note that the Cluster 2 ARs containing the smallest sunspots are most similar to each other while the Cluster 1 ARs are more dissimilar (see Tables 6.7 and 6.8). This indicates that the Hellinger distance approaches are best for distinguishing between different types of larger or complex ARs.

When NMF is applied on datasets where all values in the dictionary and coefficient matrices are constrained to be nonnegative, its results are generally more interpretable than SVD. In our application however, the magnetogram components can be negative. Hence the NMF results are not particularly sparse and lose some benefits of nonnegativity since the positive and negative magnetogram components can cancel each other. This results in some loss of interpretability. Additionally, the SVD results seem to be more interpretable due to separate treatment of the continuum and magnetogram components. However, there is still some value in the NMF approach as we see that the clustering on NMF coefficients are better at separating the ARs by size than the SVD approach. Additionally, the NMF approaches tend to agree more strongly with the Mount Wilson labels than their SVD counterparts as is seen in Section 6.5.3 below. Future work could include using alternate forms of NMF such as in [46] where sparsity and interpretability is preserved even when the dictionary is no longer constrained to be nonnegative. Another variation on coupled NMF that may be applicable is soft NMF [177] where the requirement that the two modalities share the same regression coefficients is relaxed somewhat. Finally, future work could perform factorization using a composite objective function comprised of two terms corresponding to the two modalities that are scaled according to their noise characteristics.

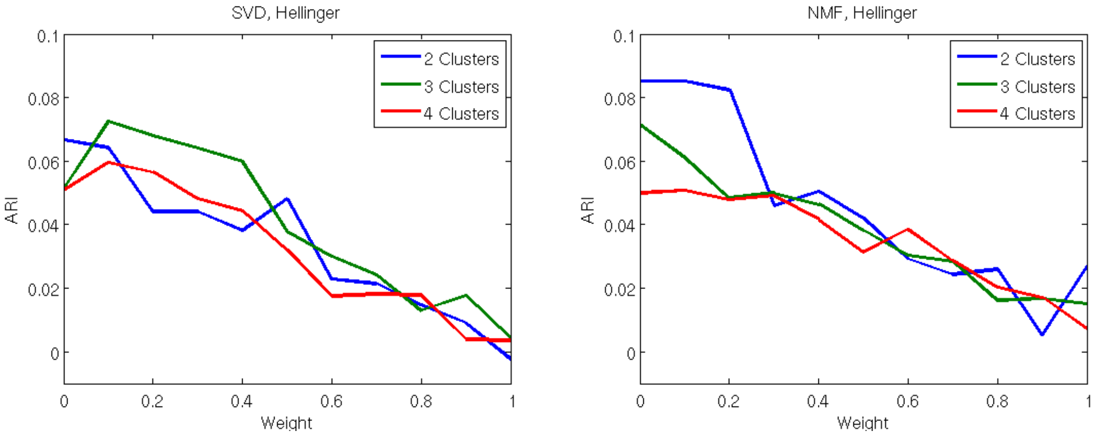

Figure 6.10: Plot of the adjusted Rand index (ARI) using and Hellinger distances within the neutral line and sunspots as a function of the weight. A weight of 0 corresponds to clustering with only the sunspots while a weight of 1 clusters with only the neutral line. The different lines correspond to different numbers of clusters and the corresponding labels from Table 6.6. Higher ARI indicates greater correspondence.

6.5.3 Clustering with Neutral Line Data

As described in Section 6.4.3, we analyze the effects of including data from the neutral line in the clustering. We proceed by taking a weighted average of the dissimilarities calculated from the sunspots and from the neutral line data matrices. Using the ARI, we compare the results to labels based on the Mount Wilson classification scheme, see Table 6.6 for the label definition. We use a grid of weights, starting from a weight of 0 for a clustering using only patches within sunspots up to a weight of 1 for a clustering that takes into account only the neutral line data.

Figure 6.10 plots the ARI for the four different schemes as a function of the weight. In nearly all cases, the ARI is above zero which indicates that the clustering does slightly better than a random assignment. In general, the correspondence of the clustering results with the Mount Wilson based labels decreases as the weight approaches 1. This means that the natural clusterings associated with only the neutral line data do not correspond as well with the Mount Wilson based labels. However in several cases, including some information from the neutral line at lower weights appears to increase the correspondence, e.g., the ARI increases for the 3 and 4 cluster cases for the SVD coefficients. This suggests that the neutral line and the sunspots contain information about AR complexity that may be different.

Note that clustering separates the ARs based on the natural geometry in the spaces we are considering. Thus we can influence the clustering by choosing the space. For example, if we restrict our analysis to include only coefficients corresponding to specific dictionary patches, then this will influence the clustering.

The gradients of magnetic field values across the neutral line is a key quantity used in several indicators of potential eruptive activity [176]. We therefore repeated the neutral line experiment where we focused only on the gradients within the magnetogram as follows. When we applied SVD to the data matrix **Z** extracted from the neutral line, the resulting dictionary matrix **A** was very similar to that shown in Figure 6.4. Note that elements 3 and 4 correspond to the gradient patterns within the magnetogram data. Therefore, after learning **A** and **H** from **Z**, we kept only the coefficients corresponding to dictionary elements 3 and 4, i.e. the 3rd and 4th rows of **H**. We then estimated the Hellinger distance between the ARs' underlying distributions of these two coefficients. This restricted the neutral line analysis to include only the coefficients corresponding to magnetogram gradients. For the data within the sunspots, we included all coefficients as before.

Figure 6.11 shows the ARI as a function of the weight for this experiment. For all cases, the ARI stays fairly constant until the weight increases to 0.9, after which it drops dramatically. We can compare this to the results in Figure 6.10 (bottom left) to determine if using only the neutral line gradient coefficients results in increased correspondence with the Mount Wilson labels relative to using all of the neutral line coefficients. From this comparison, the ARI is higher when using only the gradient components for weights greater than 0.1 and less than 1. Thus the correspondence with the Mount Wilson labels and the clustering is higher when we only include the magnetogram gradient coefficients. Since the Mount Wilson scheme is related to the

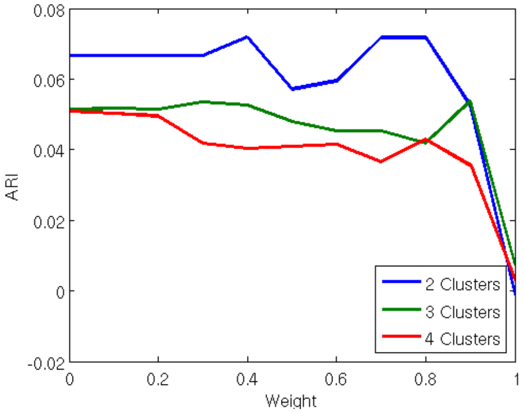

Figure 6.11: Plot of the ARI using the Hellinger distance within the neutral line on only the coefficients corresponding to the SVD dictionary elements associated with magnetogram gradients. The corresponding dictionary elements are similar to the 3rd and 4th elements in Figure 6.4. Focusing on the gradients results in a higher ARI for higher weights than when all coefficients are used as seen in Figure 6.10.

complexity of the neutral line, this higher correspondence suggests that focusing on the gradients in the neutral line results in a natural geometry that is more closely aligned with the complexity of the neutral line than simply using all of the coefficients. Applying supervised techniques would lead to improved correspondence.

The clear patterns in the ARI indicate that the relationship between the weight and the ARI is unlikely to be due entirely to noise. Thus including data from the neutral line with the data from the sunspots would add value in an unsupervised setting and would likely lead to improved performance in a supervised setting.

We can investigate this further by estimating bounds on the Bayes error where β groups are labeled as 'simple' and $\beta\gamma$ and $\beta\gamma\delta$ groups are labeled as 'complex.' α groups are excluded to keep the number of simple and complex ARs roughly the same (192 and 182, respectively). We estimate both the lower and upper bounds formed from $\tilde{D}_{q_1}$ using the MST estimator described in Chapter III. We use this estimator as the k-nn density estimator and the KDE are not easily defined in the space of probability distributions.

Figure 6.12 shows the estimated bounds when using SVD. When the Hellinger

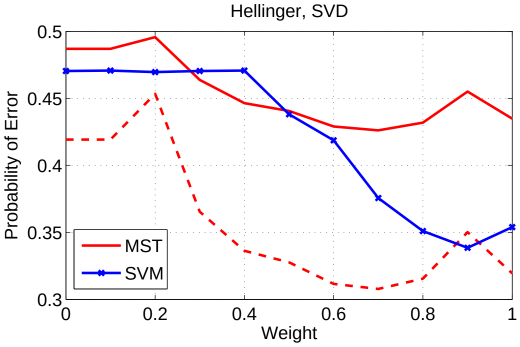

Figure 6.12: $\tilde{D}_{q_1}$-based upper (plain line) and lower (dashed line) bounds on the Bayes error when classifying sunspot groups as simple or complex for a variety of weights compared to the error from an SVM classifier using SVD dictionaries. A weight of $r = 0$ corresponds to using only the data from within the sunspots while $r = 1$ corresponds to using only the neutral line data. The area around the neutral line gives better results.

distance is used on the dictionary coefficients, the estimated bounds and SVM error rate are generally lower when the weight r favors the neutral line data. This is consistent with our previous results.

The NMF results are not shown, but similar trends are observed. The estimated bounds and the SVM error rate generally decrease as the weight increases, suggesting that the neutral line is better suited for this classification problem than the data from within the sunspots when using NMF dictionaries. However, the estimated bounds and error rates are generally still high (> 0.25).

In general, these results indicate that if the goal is to accurately classify ARs into complex or simple ARs based on the Mount Wilson definition, then additional or different features are required. The dictionary features may still be relevant for other learning tasks such as predicting and detecting solar eruptive events.

6.5.4 Clustering of Regions Exhibiting Strong Field Polarity Reversal Lines

We now analyze ARs exhibiting strong field polarity reversal lines by comparing the natural clustering of these ARs to the calculated R value as described in Section 6.4.3. When we apply dimensionality reduction on this data using SVD, the resulting dictionary is very similar to Figure 6.4 with the first two patches consisting of uniform nonzero patches, the third and fourth patches consisting of gradients in the magnetogram, and the fifth and sixth patches consisting of gradients in the continuum.

As before, the mean silhouette width indicates that the appropriate number of clusters is 2. When we cluster the ARs using the SVD coefficients corresponding to all six patches, we obtain a correlation between cluster assignment and R of 0.09 (see Table 6.9). This isn't particularly high which suggests that the natural geometry based on the distribution of all six coefficients does not correlate well with R. However, since the clustering is separating the ARs based on the natural geometry in the spaces we are considering, we can influence the clustering by choosing the space. In other words, if we restrict our analysis to only coefficients corresponding to specific dictionary patches, then this will influence the clustering.

Restricting the clustering analysis to the SVD coefficients corresponding only to the magnetogram components (i.e. elements 2, 3, and 4 in Figure 6.4) results in a correlation of 0.30 between cluster assignment and R value. If we only consider the gradient components (elements 3 and 4), then the correlation is 0.34.

The relationship between cluster assignment and R may not be linear as the correlation between the clustering assignment using only the gradient components and $\log R$ is 0.45. Comparing the magnetogram only components based clustering with $\log R$ similarly increases the correlation coefficient. Given that clustering is an unsupervised method and that we are only clustering into two groups, this correlation is quite high. This suggests that the natural geometry of the image patch analysis increasingly corresponds with R as we restrict the analysis to magnetogram gradients. Supervised methods, such as regression, should lead to an even greater correspondence.

	SVD			NMF
	All	Mag. only	Grad. only	All
R	0.09	0.30	0.34	0.15
$\log R$	0.02	0.37	0.45	0.08

Table 6.9: Magnitude of the correlation coefficient of the clustering assignment with either R or $\log R$ when using all of the coefficients, only the coefficients corresponding to the magnetogram component (SVD elements 2-4 in Figure 4), or only magnetogram gradient coefficients (SVD elements 3-4 in Figure 4). For NMF, all of the dictionary elements are associated with the magnetogram and many of them have gradient components so we only perform clustering with all of the coefficients.

For NMF, when we include all of the coefficients, the correlation between R and clustering assignment is 0.15. While this is small, we are again comparing the labels of an unsupervised approach to a continuum of values. Thus we can expect that the performance would be better in a supervised setting. If we compare the clusering to $\log R$, the correlation decreases to 0.08. It is difficult to restrict the NMF dictionary to only continuum and magnetogram parts and gradients as most of the components contain a gradient component in the magnetogram. Therefore we only cluster the ARs using all NMF coefficients.

6.6 Conclusion

In this chapter, we introduced a reduced-dimension representation of an AR that allows a data-driven unsupervised classification of ARs based on their local geometry. The ROI that surrounds and includes the AR represents its most salient part and must be provided by the user. We used STARA masks in conjunction with masks situated around the neutral line, and compared our results with the Mount Wilson classification in order to ease interpretation of the unsupervised scheme.

The Mount Wilson scheme focuses on the largest length scale when describing the geometrical arrangements of the magnetic field, whereas our method focuses on classifying ARs using information from fine length scale. We have shown that when

Class. Scheme	Cluster 1	Cluster 2
SVD Hellinger	largest $\beta, \beta\gamma$ sunspots; majority of $\beta\gamma\delta$; high concentration of patches with average magnetic field value $\simeq 0$; large Hellinger distance between ARs	smallest $\beta, \beta\gamma$ sunspots; high concentration of patches with average magnetic field value close to $+1$ or -1; small Hellinger distance between ARs
NMF Hellinger	largest $\alpha, \beta, \beta\gamma$ sunspots; majority of $\beta\gamma\delta$; large Hellinger distance between ARs	smallest $\alpha, \beta, \beta\gamma$ sunspots; small Hellinger distance between ARs

Table 6.10: Summary of features distinguishing the clusters under the various classification schemes tested.

we analyze and cluster the ARs based on the global statistics of the local properties, there are similarities to the classification based on the large scale characteristics. For example, when clustering using the Hellinger distance, one cluster contained most of the complex ARs. Other large scale properties such as the size of the AR also influenced the clustering results. Table 6.10 summarizes the properties that are found to influence the clustering under the two schemes.

In this comparison with the Mount Wilson scheme, we found that the STARA masks were sometimes too restrictive which led to a mismatch between the Mount Wilson label and the extracted data. For example, there were several cases where an AR was labeled as a β class but the STARA mask only extracted magnetic field values of one polarity. We showed that the neutral line contains additional information about the complexity of the AR. For this reason, we expect that including information beyond the STARA masks will lead to improved matching with the Mount Wilson labels.

To investigate the possibility for our method to distinguish between potential and non-potential fields, we considered a ROI made of pixels situated along high-gradient, strong field polarity reversal lines. This is the same ROI as that used in the computation of the R value, which has proved useful in flare prediction in a supervised context. We found that our clustering was correlated with the R value, that is, the clustering based on the reduced dimension representation separates ARs corresponding to low R from the ones with large R.

CHAPTER VII

Application to HFO Data: Dimensionality Reduction and Bayes Error Estimation

7.1 Introduction

About one third of epilepsy patients fail to obtain seizure control with available pharmaceuticals. One of the few options for these refractory patients is resective surgery—removing the portion of the brain thought to be causing the seizures. This region is denoted the seizure onset zone (SOZ). In some cases, determining the SOZ involves a highly invasive surgery to place electrodes on the brain's surface, followed by one to two weeks of recording and monitoring. A second invasive surgery is performed if the SOZ can be identified and safely resected. A schematic relating the implanted electrodes with the recorded data is shown in Fig 7.1.

A proposed biomarker to improve the localization of the SOZ are high frequency oscillations (HFOs) [34, 157]. HFOs are high frequency (about 80–300 Hz), short (< 50 ms), rare events occurring in intracranial EEG recorded at sampling rates of several kHz. Example HFO detections and a recorded seizure are shown in Fig 7.2.

Much of the published research on HFOs uses human identified HFOs in short (10 to 20 minute) recordings [80, 104]. These results have shown that a high HFO occurrence rate is correlated with the SOZ. However, recent work is moving towards auto-

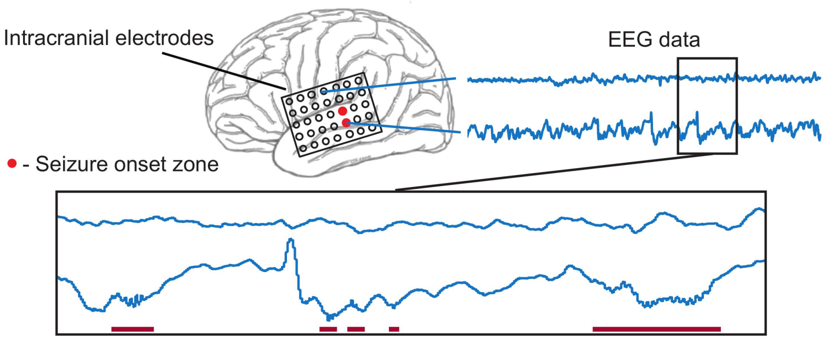

Figure 7.1: Diagram relating the recorded data with the implanted electrodes. A 5×7 grid of electrodes placed over a region of cortex. Each channel produces a separate time series of data, with some channels being identified by clinicians as seizure onset zone (SOZ). HFO detections are marked by solid magenta lines under the EEG trace.

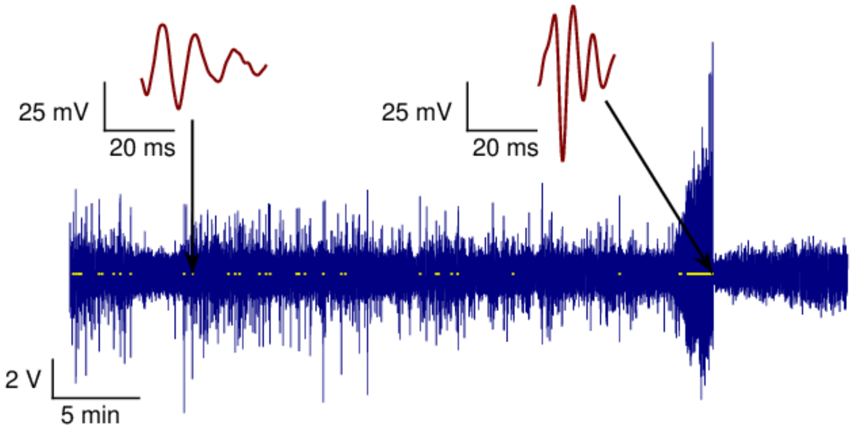

Figure 7.2: Example HFO detections within 45 min of one channel of intracranial EEG data. A seizure occurs at about 35 minutes. HFO detections (72 interictal and preictal, 32 ictal) are shown as small yellow dots. Two HFOs are also shown using a *much* smaller scale.

mated identification and analysis of HFOs in long term, high resolution data, which requires advanced computational and statistical techniques [74]. Quality recordings may span 7-14 days, with over 100,000 HFO detections in several terabytes of data. Thus, the next advances are expected to come through big data analysis of HFO features, utilizing millions of recorded HFOs across as many patients as possible [202].

Relatively few research groups have analyzed HFO features in detail. The most advanced analysis computed six features of about 300,000 HFOs in nine patients and two controls, and utilized a global PCA across all channels followed by k-means clustering [22, 23, 158]. The authors implicitly assumed that the distribution of these HFO features lies on a linear manifold in feature space, and that the manifold is consistent across time and space (i.e., recording electrode).

The other most advanced analysis compared HFOs produced in the motor cortex via movement versus HFOs occurring in the SOZ, utilizing three features and a support vector machine (SVM) classifier [131]. Some differences were noted, but a more general analysis that addresses the degree to which HFOs produced by pathological activity or networks (denoted pathological HFOs or pHFOs) and HFOs produced by normal, physiological activity (denoted normal HFOs or nHFOs) are observably different has not been performed.

The goals of this chapter are to test the implicit assumptions previously used in HFO feature analysis. Specifically, the goals are to 1) assess the type of manifold on which HFO features lie, and 2) assess how discernible pHFOs are from nHFOs, based on their feature-space distributions.

The general outline is as follows. To assess the linearity of the HFO-feature manifold, a non-linear, local estimate of intrinsic dimension [31, 40] is compared with a linear global estimate (PCA) applied to local subsets. This approach is similar to that in Chapter V. The corresponding reduced dimension subspaces are individually compared across time and space using a modified Grassmann distance, building off the work in [205]. We additionally use a greedy Fisher LDA algorithm (similar to the greedy LDA in [198]) to identify a basis that maximally separates pHFOs from nHFOs. Unsupervised clustering of the subspaces is then compared with channel groups based on clinical markings of the SOZ and physical groupings of the electrodes. Lastly, we assess the discernibility of nHFOs and pHFOs by estimating bounds on the Bayes Error using the Henze-Penrose divergence (HPD) [15, 139] (see Chapter III).

7.2 Patient population and data

EEG data from adult patients with refractory epilepsy who underwent intracranial EEG monitoring were selected from the IEEG Portal [197] and from the University of Michigan. All patient data was included which met the following criteria: sampling rate of at least 5 kHz, recording time greater than one hour, data recorded with traditional intracranial electrodes, and available meta-data regarding seizure times and the resected volume or SOZ. This yielded 17 patients, (nine IEEG portal, eight U. of M.). All data were acquired with approval of local institutional review boards (IRB), and all patients consented to share their de-identified data. Of these 17 patients, 13 had recorded seizures, nine were known to have resection and obtained seizure-freedom (ILAE class I), with eight patients in common between these two categories. Because these surgeries are relatively rare, this patient population size is moderately large for analysis of intracranial EEG data.

HFOs were detected using the qHFO algorithm [74], resulting in over 1.6 million HFOs in nearly 100,000 channel-hours of 5 kHz data. Each HFO was band pass filtered between 80 to 500 Hz using an elliptical filter and 33 features were computed, including duration, peak power, mean of the Teager-Kaiser energy [101], and various spectral properties. Ictal is defined as during seizures, with seizures assumed to be five minutes long if the length was not specified in available meta-data. Interictal is defined as at least 30 minutes from the start or end of a seizure, based on [158].

7.3 Dimensionality reduction

7.3.1 Consistency of local, non-linear intrinsic dimension

To assess both the linearity and local versus global nature of the HFO feature manifold, the non-linear intrinsic dimension was computed via the k-nearest neighbor (NN) based estimator in [31] and described in Appendix E. This nonlinear estimator provides a *local* estimate of intrinsic dimension which enables us to identify local variations in data manifolds. The result is an estimate of the intrinsic dimension for each given HFO, which are then averaged to obtain the mean intrinsic dimension for a given partition of the data. The consistency of the manifold is measured by comparing the distribution of intrinsic dimension across time, space and patients. The specific comparisons are: a) interictal versus ictal times, per channel, b) time variation within interictal periods, per channel, and c) comparisons between channels, integrated over time.

A variety of methods could be employed to compare the intrinsic dimension for two disjoint sets of HFOs. However, the final dimension selected will be an integer. Thus, small differences in the intrinsic dimension between two sets, no matter how statistically significant, are not meaningfully different if the mean value for each set round to the same integer.

To compare two sets of intrinsic dimension, we define a distance measure, θ_I, between collections of integers. Let the two sets of integers be A and B, and let n_i be the fraction of elements in A equal to i, and m_i be the fraction of elements in B equal to i. The probability that two elements in set A are equal is $\boldsymbol{n}^T\boldsymbol{n}$, and the probability that an element of A is equal to an element of B is $\boldsymbol{n}^T\boldsymbol{m}$. A measure of distance between A and B is how likely an element of A is equal to an element of B, normalized by the likelihoods of elements being equal another element in the same

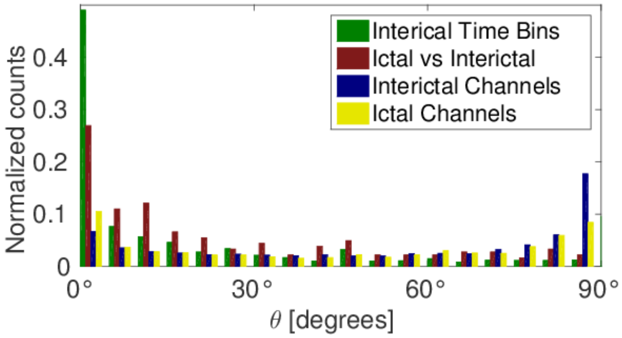

Figure 7.3: Comparisons of the consistency of the intrinsic dimension (θ_I from Eq. 7.1) for four comparisons, as described in the text. Note, 0° implies no difference and 90° implies maximal difference between the two collections of intrinsic dimension being compared.

group:

$$\theta_I = \arccos\left(\frac{\boldsymbol{n}^T\boldsymbol{m}}{\sqrt{\boldsymbol{n}^T\boldsymbol{n}}\sqrt{\boldsymbol{m}^T\boldsymbol{m}}}\right), \tag{7.1}$$

The value θ_I is the angle between $\boldsymbol{n}$ and $\boldsymbol{m}$ and provides an easy interpretation as to the consistency of two different collections of local intrinsic dimension. This quantity is also know as the angular distance or angular dissimilarity, and is the inverse cosine of the Ochiai-Barkman coefficient [9, 154].

We compute the θ_I-distance for three different comparisons of HFOs: 1) a comparison of each pair of 30 minute time windows on a given channel for a given patient, 2) a comparison of ictal versus interictal periods, again on a given channel for a given patient, and 3) a comparison of different channels in a given patient during interictal periods. Interictal-ictal comparisons where one set of events is less than 50 HFOs are ignored. Histograms of the distribution of θ_I for each type of comparison are shown in Fig. 7.3. This figure involves over 5 million interictal time bin comparisons, 163 ictal versus interictal comparisons, over 36 thousand interictal channel-channel comparisons and almost 10 thousand ictal channel-channel comparisons.

The intrinsic dimension is quite consistent across different time segments during interictal times (strong peak near 0°), but has some variance between ictal and interictal times (still peaked near 0° but the peak is wider). The dimension is less consistent across channels, with comparisons during ictal times showing small peaks at both 0° and 90°, and interictal having the largest peak at 90°, showing maximal difference. Thus we see that the intrinsic dimension varies significantly across channels, especially for interictal HFOs.

7.3.2 Comparison with Global Linear Intrinsic Dimension

Next we compare the k-NN intrinsic dimension estimate of Section 7.3.1 with a global, linear method to assess how linear and/or local the feature manifold is. The most common global, linear method of intrinsic dimension estimation and dimensionality reduction is principle component analysis (PCA), which we perform by first centering the data and then using singular value decomposition (SVD) as in Chapter V. PCA is performed multiple times for the same divisions of the HFO data as done for Fig. 7.3. Subsets of HFOs with less than 50 events are ignored, as these are deemed insufficient to estimate the PCA vectors in 33D. This results in PCA vectors being computed for 606 of the 1318 channels (12 patients) for interictal HFOs, 171 channels for ictal HFOs (8 patients), and 163 channels comparing ictal versus interictal (8 patients). In patients that had multiple recording sessions, channels are counted once per each session.

To compare the k-NN (non-linear) and PCA intrinsic dimension per channel, we 1) select the number of principle components equal to the non-linear intrinsic dimension estimate, and then 2) report the fraction of the variance accounted for by that number of principle components. This is repeated for all 606 channels (interictal) and 171 channels (ictal). When using either ictal or interictal HFOs, the median fraction of variance was 99.8%, with 99%-tile of the channels being above 89.5% (interictal) and 97.1% (ictal). It is likely that that noise in the data could account for up to 10% of the variance, and thus we conclude that the feature manifolds are approximately linear over the locality of a given channel and ictal state.

7.3.3 Comparison between Local Manifolds

In addition to the earlier comparison of the subspace dimensionality across channels, the next step is to directly compare the subspaces selected by the dimensionality reduction. We use a generalization of distance in the Grassmann space, which allows comparison of affine subspaces with unequal dimension [205]. The method augments the principle angles (defined in [95]) with enough additional angles (all equal to $\pi/2$) to increase the number of angles to the dimensionality of the larger space. An additional "direction vector" is also added to account for the affine offset.

We apply this generalization [205] to a new modification of the chordal distance, defined as

$$\theta_C = \arcsin\left(\left(\frac{1}{k}\sum_{i=1}^{k}\sin^2\theta_i\right)^{1/2}\right), \tag{7.2}$$

for k principle angles $\{\theta_i\}_{i=1}^{k}$. The two modifications are 1) dividing by k, which allows the distance to be independent on the dimensionality of the spaces being compared, and 2) converting the distance measure back to an angle, which is more intuitive.

We then compare the subspaces obtained by PCA dimensionality reduction, for the same divisions of the data as used for Fig. 7.3. However, we now ignore any subsets of less than 50 HFOs, as these are deemed unreliable for computing the PCA in 33D. Note this figure still involves over 200 thousand time bin comparisons, the same 163 ictal versus interictal comparisons, and nearly 3,500 of each type of channel-channel comparisons.

Results for these comparisons are shown in Fig. 7.4. We observe that the PCA subspaces are quite consistent across different time bins, with the distributions for PCA subspaces all peaking less than 15° and not extending much past 20°.

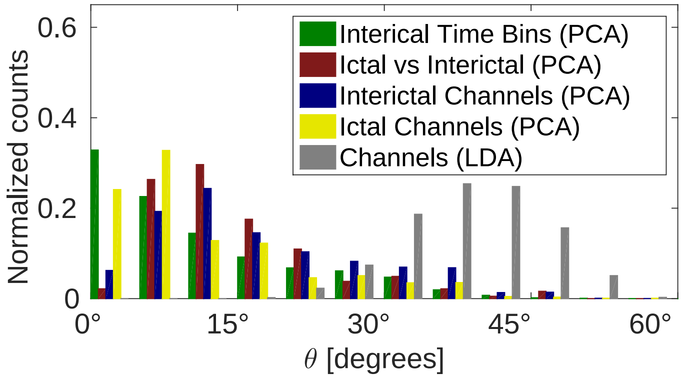

Figure 7.4: Comparison of the subspaces from the PCA and greedy Fisher LDA dimensionality reduction, using θ_C (Eq. 7.2), for the same types of comparisons as Fig. 7.3. Again, 0° implies maximally similar and 90° implies maximally different.

7.4 Pathological versus Normal HFOs

7.4.1 Dimensionality Reduction: Greedy Fisher LDA

While the subspaces obtained via PCA represent the variance well, they are not necessarily the optimal directions for separating the feature distributions of pHFOs and nHFOs. An alternate dimensionality reduction method is used: greedy Fisher's LDA. In this method, Fisher's LDA is applied, resulting in a single basis direction. The projection of the data in this direction is then subtracted from the data, and the process is repeated.

Recall that Fisher's LDA uses the sum of the covariance of each group, rather than the covariance of the pooled groups [58]. Letting the mean and covariance of the two groups be denoted μ_A, μ_B and Σ_A, Σ_B, respectively, the specific direction is given by

$$w \propto (\Sigma_A + \Sigma_B)^{-1} (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B) . \tag{7.3}$$

Note, the rank of the sum of the covariances will be reduced by one in each step. Thus, to invert the matrix, the eigenvalue method is used, with the inverse of the eigenvalues corresponding to the removed projections being set to zero.

We compute this basis, comparing ictal versus interictal times, for the 13 patients with recorded seizures. We select the number of basis vectors equal to the mean intrinsic dimension. We again compare the basis vectors using Eq. 7.2, with the results shown in Fig. 7.4. The LDA subspaces vary much more across channels than the PCA subspaces, with the θ_C distribution for LDA being almost fully localized between 30° and 50°. Note that 45° implies that the subspaces overlap by half. Thus, the Fisher LDA subspaces show that some differences in ictal versus interictal HFO features are consistent between channels, while other differences are not conserved.

7.4.2 Bayes Error Estimates of pHFOs versus nHFOs

Next we quantify how distinct the feature distributions of pHFOs and nHFOs are, per channel. This quantification serves as a guide for future work. We utilize the Henze-Penrose divergence (HPD) [15, 139] to compute bounds on the Bayes Error. Note that small upper bounds imply highly separable classes, whereas upper bounds near 0.5 imply inseparable classes. We estimate the HPD bound using the nonparametric ensemble estimator described in Chapter III which achieves the parametric convergence rate.

The left panel of Fig. 7.5 shows the Bayes Error bound estimates for the 163 channels (eight patients) with at least 50 HFOs in each of the ictal and interictal states. Channels are separated between SOZ (27 channels) and non-SOZ (124 channels) for patients with ILEA Class I (the best) surgery outcome, and an “other” category (12 channels), including channels from patients with either worse surgery outcomes, no surgery, or missing meta-data.

Patients with Both SOZ and non-SOZ channels in ILAE class I patients (best surgery outcome) have the Bayes Error rate bound to relatively small values. However, channels in other patients have a more diffuse distribution of upper bounds, with the most probable upper bound about 0.25. It is expected that ictal HFOs are almost entirely pHFOs (on any channel), that interictal HFOs on non-SOZ channels are predominately nHFOs, and that interictal HFOs on SOZ channels are a mixture of both pHFOs and nHFOs. Thus, it is expected that the Bayes Error bounds would be lower for non-SOZ channels. Overall, these bounds suggest that there is sufficient separation between pHFO and nHFO feature distributions to allow classification of pHFOs and nHFOs in most channels.

The right panel of Fig. 7.5 displays the lower bound versus a linear error estimate. The linear estimate was computed with 10-fold cross validation in the Fisher LDA space by using a "box" classification boundary, with the threshold in each dimension being the value for which the receiver operator curve has largest transverse distance from the diagonal. Linear regression was also performed, resulting in an offset of 0.06 (0.04–0.08 at 95% C.L.) and a slope of 1.05 (0.82–1.28 at 95% C.L.). Thus, in aggregate this "box" classifier is already relatively close to the bound on the Bayes Error, though many individual channels are still quite far from the bound.

7.5 Clustering Channels based on subspaces

Given the observed variations in greedy Fisher LDA subspaces across channels, we seek to compare the natural clustering of these subspaces with known groupings of channels. The most relevant groupings of channels are the groups based on the physical configuration of the recording electrodes (several grids or strips of electrodes are implanted for each patient), as well as the clinically determined SOZ and resected volume for ILAE Class I patients.

The unsupervised clustering is obtained by first converting the matrix of modified chordal distances (7.2) between channels per each patient to a metric using EAC-DC algorithm (see Appendix E.3). We select either two groups (as the SOZ and resected volume clustering is binary) or the number of groups equal to the number of strips and grids. The unsupervised clustering results are compared with these labels using the adjusted Rand index (ARI) [167]. Unfortunately, the requirement that there be at

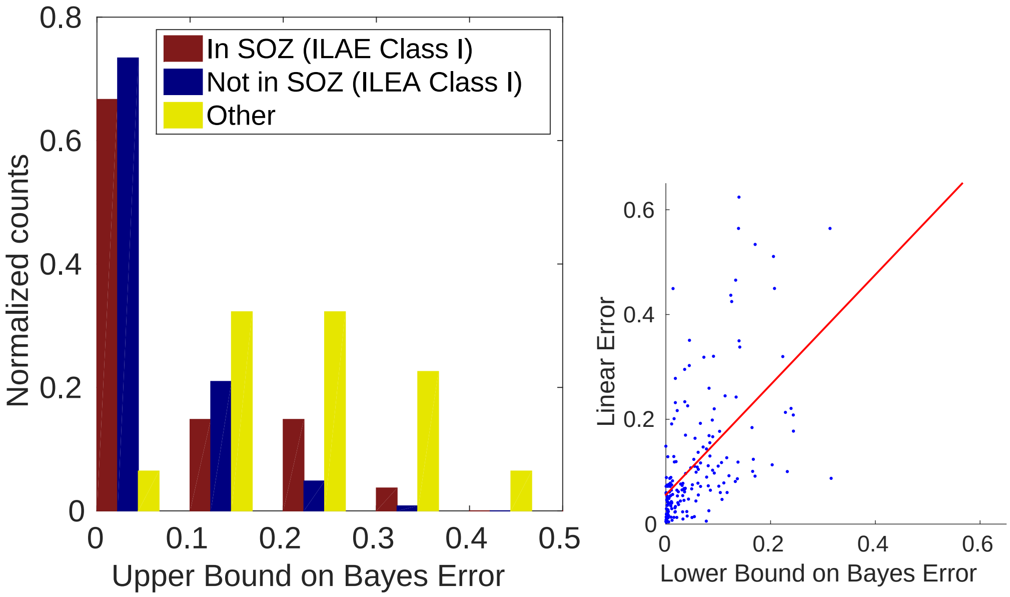

Figure 7.5: Distribution of upper Bayes Error bound based on Henze-Penrose divergences (HPD) between ictal and interictal HFOs per channel. Left panel: upper bounds stratified by channel type. Bounds near zero imply high distinction between pHFOs and nHFOs, and bounds near 0.5 imply no distinction. Right panel: comparison between lower bound estimate and a linear error estimate (see the text), including the linear regression fit.

least 50 ictal HFOs reduces the number of channels per patient significantly, resulting in only a few (4–5) patients having enough channels to cluster. However, the ARI never exceeded 0.02 for any of these comparisons in any patients, suggesting that the primary distinction between channels is not the pathology or the grid/strip placing, but other effects.

7.6 Conclusion

Overall, we observe that the HFO features tend to cluster on linear manifolds. Both the subspaces of these manifolds and the local intrinsic dimension tend to be consistent within interictal periods, but may change between interictal and ictal periods. We especially note significant differences in both the intrinsic dimension and feature manifolds between different channels within the same patient. Thus, dimensionality reduction and feature analysis must account for variations between channels. The dominant cause of this inter-channel variation does not appear to be tissue pathology or grid/strip groups in the recording. We also observe that pHFOs and nHFOs are indeed distinct on a large number of channels, suggesting a strong potential for classifying individual HFOs.

This analysis also demonstrates methods applicable to other discrete events, including using the θ_I statistic for comparing local, intrinsic dimension for collections of events, an affine Grassmann distance θ_C for comparing consistency of subspaces, and estimating bounds on the Bayes error to assess the feasibility of low-error classification.

CHAPTER VIII

Conclusion and Future Work

In this thesis, I have presented useful methods for exploiting opportunities and countering challenges that arise in some big data problems. These methods are united in their reliance on accurate estimation of distributional functionals. In this chapter, I conclude the work included in this thesis and propose directions for future work.

8.1 Nonparametric Estimation of Distributional Functionals

In Chapter II, we presented two KDE-based ensemble estimators of functionals of two distributions or divergence functionals that achieve the parametric convergence rate. These estimators use basic kernel density estimators as the base estimators and choose the weights based on the convergence rates of the base estimators. These estimators are simpler to implement than many of the competing estimators. Variance and central limit proofs are given that are simpler than previous work and require less strict assumptions.

In Chapter III, we similarly presented two k-nn based ensemble estimators for divergence functionals that also achieve the parametric rate. Due to the properties of k-nn, these estimators can be computationally easier than the KDE-based estimators in Chapter II while enjoying many of the same advantages.

In Chapter IV, we adapt the KDE and k-nn divergence functional ensemble estimators to estimate mutual information measures between the random vectors $\mathbf{X}$ and $\mathbf{Y}$. We show that the parametric rate can be achieved both when the data only have continuous components, and when $\mathbf{X}$ has only continuous components and $\mathbf{Y}$ has only discrete components.

Some future work remains. In some cases, we are interested in the mutual information between $\mathbf{X}$ and $\mathbf{Y}$ when one or both of them contains a mix of continuous and discrete components. The convergence rates of plug-in estimators (such as KDE or k-nn estimators) are currently unknown for this case. Other future work could include extending the k-nn plug-in divergence results to mutual information estimation and to more general density support sets and estimating functionals that are less smooth (e.g. the total variation distance).

Another important area of future work involves extending the theoretical work in this thesis to time series data. There are many important applications that include time series including sunspot images and HFO data. Therefore, it is important to extend the theory to these cases. This would require a relaxation of the i.i.d. assumption of the data which complicates the analysis. It is likely that a simple plug-in estimator such as the kernel density estimator will have poor convergence rates. Thus to improve the convergence rate via an ensemble estimator, we will first need an expression for the convergence rate of the base estimator which will require more advanced analysis techniques. This will open up other opportunities for future work in this direction.

8.2 Sunspot Images

In Chapter V, we performed an image patch analysis of sunspot continuum and magnetogram images. We estimated the local intrinsic dimension of the data via a k-nn based entropy estimator and analyzed the correlation at different scales and between the modalities. This paved the way for further analysis based on matrix factorization in Chapter VI. Knowledge of the intrinsic dimension allowed us to choose the dictionary size. Moreover the results of Section 5.3 showed that linear methods are sufficient. The spatial and modal correlation analysis in Section 5.4 justified a choice of a patch size of 3×3 and confirmed that both modalities (continuum and magnetogram) should be used in matrix factorization.

Chapter VI focused on clustering the sunspot images using divergence as the base dissimilarity measure. The divergence was estimated using the nonparametric ensemble estimator described in Chapter III. We found that the resulting clusters are correlated with physical phenomena such as the scale of the ARs. The clusters were also correlated with measures associated with flares such as the R value and the complexity of the AR as measured by the Mount Wilson classification scheme.

The local intrinsic dimension based on the k-NN approach combines both continuum and magnetogram observations and provides some measure of local regularity for those images. Further differences between the Mount Wilson classes may be found by comparing the histograms or distributions of local intrinsic dimension of each individual AR instead of only comparing the means or pooled estimates as we did in this thesis. There are several options to perform such comparisons. Each histogram could be treated as a vector, or we could consider the underlying probability density function within the framework of functional analysis. Supervised (using Mount Wilson classes) or unsupervised classification could be performed. Another option would be to view the set of histograms belonging to a specific class as samples from a distribution of vectors (or a distribution of probability density functions). Different classes could then be compared using divergence measures.

In future work, we plan to study the efficiency of supervised techniques applied to the reduced dimension representation. Supervised classification can always do at least as well as unsupervised learning in the task of reproducing class labels (e.g. Mount Wilson label). Thus if the goal is for example to reproduce the Mount Wilson classes, or to detect nonpotentiality using global statistics of local properties, then supervised methods would lead to increased correspondence relative to our unsupervised results.

In case of flare prediction, the labels would be some indicator of flare activity such as the strength of the largest flare that occurred within a specified time period after the image was taken. Supervised techniques such as classification or regression could be applied depending on the nature of the label (i.e. categorical vs. continuum). For classification problems, estimation of the Bayes error via divergence estimators can be helpful.

These methods of comparing AR images can also be adapted to a time series of image pairs. For example, image pairs from a given point in time may be compared to the image pairs from an earlier period to measure how much the ARs have changed. The evolution of an AR may also be studied by defining class labels based on the results from one of the clustering schemes in this thesis. From the clustering results, a classifier may be trained that is then used to assign an AR to one of these clusters at each time step. The evolution of the AR's cluster assignment can then be examined. To properly do this, the future theoretical work involving time series discussed in Section 8.1 will be helpful.

8.3 HFO Data

In Chapter VII, we tested some of the implicit assumptions previously used in HFO feature analysis. We observed that the HFO features tend to cluster on linear manifolds that tend to be consistent within interictal periods. However, we found that the manifolds may not be consistent between interictal and ictal periods and between different channels within the same patient. Thus, dimensionality reduction and feature analysis must account for these variations. We also observe that pHFOs and nHFOs are indeed distinct on a large number of channels, suggesting a strong potential for classifying individual HFOs.

Future work will extend this analysis to a larger patient population, and classify HFOs and/or recording channels based on HFO features. By including more patients, it may be possible to extend this work to better detect the seizure onset zone within the brain. The machine learning on distributions framework will apply in this case. Given that HFOs occur in a time series, the future theoretical work discussed in Section 8.1 will also be helpful here.

APPENDICES

APPENDIX A

Boundary Conditions

This appendix contains the proof for the boundary condition assumption $\mathcal{A}.5$ for rectangular kernels and for spherical kernels (Theorem II.1).

A.1 Rectangular Kernels

Consider a uniform rectangular kernel $K(x)$ that satisfies $K(x) = 1$ for all x such that $||x||_1 \leq 1/2$. Also consider the family of probability densities f with rectangular support $\mathcal{S} = [-1,1]^d$. We will prove Theorem II.1 which is that that $\mathcal{S}$ satisfies the following smoothness condition ($\mathcal{A}.5$): for any polynomial $p_x(u) : \mathbb{R}^d \to \mathbb{R}$ of order $q \leq r = \lfloor s \rfloor$ with coefficients that are $r-q$ times differentiable wrt x,

$$\int_{x\in\mathcal{S}} \left(\int_{u:||u||_1 \leq \frac{1}{2},\, x+uh \notin \mathcal{S}} p_x(u)du \right)^t dx = v_t(h), \tag{A.1}$$

where $v_t(h)$ has the expansion

$$v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o\left(h^{r-q}\right).$$

Note that the inner integral forces the x's under consideration to be boundary points via the constraint $x + uh \notin \mathcal{S}$.

A.1.1 Single Coordinate Boundary Point

We begin by focusing on points x that are boundary points by virtue of a single coordinate x_i such that $x_i + u_i h \notin \mathcal{S}$. Without loss of generality, assume that $x_i + u_i h > 1$. The inner integral in (A.1) can then be evaluated first wrt all coordinates other than i. Since all of these coordinates lie within the support, the inner integral over these coordinates will amount to integration of the polynomial $p_x(u)$ over a symmetric $d-1$ dimensional rectangular region $|u_j| \leq \frac{1}{2}$ for all $j \neq i$. This yields a function $\sum_{m=1}^{q} \tilde{p}_m(x) u_i^m$ where the coefficients $\tilde{p}_m(x)$ are each $r-q$ times differentiable wrt x.

With respect to the u_i coordinate, the inner integral will have limits from $\frac{1-x_i}{h}$ to $\frac{1}{2}$ for some $1 > x_i > 1 - \frac{h}{2}$. Consider the $\tilde{p}_q(x) u_i^q$ monomial term. The inner integral wrt this term yields

$$\sum_{m=1}^{q} \tilde{p}_m(x) \int_{\frac{1-x_i}{h}}^{\frac{1}{2}} u_i^m du_i = \sum_{m=1}^{q} \tilde{p}_m(x) \frac{1}{m+1} \left(\frac{1}{2^{m+1}} - \left(\frac{1-x_i}{h} \right)^{m+1} \right). \tag{A.2}$$

Raising the right hand side of (A.6) to the power of t results in an expression of the form

$$\sum_{j=0}^{qt} \check{p}_j(x) \left(\frac{1-x_i}{h} \right)^j, \tag{A.3}$$

where the coefficients $\check{p}_j(x)$ are $r-q$ times differentiable wrt x. Integrating (A.3) over all the coordinates in x other than x_i results in an expression of the form

$$\sum_{j=0}^{qt} \bar{p}_j(x_i) \left(\frac{1-x_i}{h} \right)^j, \tag{A.4}$$

where again the coefficients$\bar{p}_j(x_i)$ are $r-q$ times differentiable wrt x_i. Note that since the other cooordinates of x other than x_i are far away from the boundary, the coefficients $\bar{p}_j(x_i)$ are independent of h. To evaluate the integral of (A.4), consider the $r-q$ term Taylor series expansion of $\bar{p}_j(x_i)$ around $x_i = 1$. This will yield terms of the form

$$\begin{aligned}\int\limits_{1-h/2}^{1} \frac{(1-x_i)^{j+k}}{h^k} dx_i &= -\frac{(1-x_i)^{j+k+1}}{h^k(j+k+1)}\Bigg|_{x_i=1-h/2}^{x_i=1} \\ &= \frac{h^{j+1}}{(j+k+1)2^{j+k+1}},\end{aligned}$$

for $0 \leq j \leq r-q$, and $0 \leq k \leq qt$. Combining terms results in the expansion $v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t}h^i + o\left(h^{r-q}\right)$.

A.1.2 Multiple Coordinate Boundary Point

The case where multiple coordinates of the point x are near the boundary is a straightforward extension of the single boundary point case so we only sketch the main ideas here. As an example, consider the case where 2 of the coordinates are near the boundary. Assume for notational ease that they are x_1 and x_2 and that $x_1 + u_1h > 1$ and $x_2 + u_2h > 1$. The inner integral in (A.1) can again be evaluated first wrt all coordinates other than 1 and 2. This yields a function $\sum_{m,j=1}^{q} \tilde{p}_{m,j}(x)u_1^m u_2^j$ where the coefficients $\tilde{p}_{m,j}(x)$ are each $r-q$ times differentiable wrt x. Integrating this wrt x_1 and x_2 and then raising the result to the power of t yields a double sum similar to (A.3). Integrating this over all the coordinates in x other than x_1 and x_2 gives a double sum similar to (A.4). Then a Taylor series expansion of the coefficients and integration over x_1 and x_2 yields the result.

A.2 Spherical (Euclidean) Kernels

Consider a uniform circular kernel $K(x)$ with $K(x) = 1$ for all x s.t. $||x||_2 \leq 1$. We also consider the family of probability densities with rectangular support $\mathcal{S} = [-1, 1]^d$. In this section, we show that the boundary condition is satisfied for this kernel and support. The smoothness condition reduces to the following: for any polynomial $p_x(u) : \mathbb{R}^d \to \mathbb{R}$ of degree $q \leq r = \lfloor s \rfloor$ with coefficients that are $r - q$ times differentiable wrt x,

$$\int\limits_{x \in \mathcal{S}} \left(\int\limits_{u:||u||_2 \leq 1, x+uh \notin \mathcal{S}} p_x(u) du \right)^t dx = v_t(h), \tag{A.5}$$

where $v_t(h)$ has the expansion

$$v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o(h^{r-q}).$$

Note that the inner integral forces the x terms to be boundary points through the constraint $x + uh \notin \mathcal{S}$. Note also that this proof is more difficult than for the uniform rectangular kernel since in that case, the kernel aligns better with the boundary.

A.2.1 Single Coordinate Boundary Point.

As before, we begin by focusing on points x that are boundary points due to a single coordinate x_i s.t. $x_i + u_i h \notin \mathcal{S}$. Without loss of generality, assume that $x_i + u_i h > 1$. We focus first on the inner integral in (A.5). We will use the following lemma:

Lemma A.1. *Let $D_d(\rho)$ be a d-sphere with radius d and let $\sum_{i=1}^d n_i = q$. Then*

$$\int\limits_{D_d(r)} u_1^{n_1} u_2^{n_2} \dots u_d^{n_d} du_1 \dots du_d = C\rho^{d+q},$$

where C is a constant that depends on the n_is and d.

Proof. We convert to d-dimensional spherical coordinates to handle the integration. Let r be the distance of a point u from the origin. We nave $d-1$ angular coordinates ϕ_i where ϕ_{d-1} ranges from 0 to 2π and all other ϕ_i range from 0 to π. The conversion from the spherical coordinates to Cartesian coordinates is then

$$\begin{aligned}
u_1 &= r\cos\left(\phi_1\right) \\
u_2 &= r\sin\left(\phi_1\right)\cos\left(\phi_2\right) \\
u_3 &= r\sin\left(\phi_1\right)\sin\left(\phi_2\right)\cos\left(\phi_3\right) \\
&\vdots \\
u_{d-1} &= r\sin\left(\phi_1\right)\cdots\sin\left(\phi_{d-2}\right)\cos\left(\phi_{d-1}\right) \\
u_d &= r\sin\left(\phi_1\right)\cdots\sin\left(\phi_{d-2}\right)\sin\left(\phi_{d-1}\right).
\end{aligned}$$

The spherical volume element is then

$$r^{d-1}\sin^{d-2}\left(\phi_1\right)\sin^{d-3}\left(\phi_1\right)\cdots\sin\left(\phi_{d-1}\right)dr\,d\phi_1\,d\phi_2\cdots d\phi_{d-1}.$$

Combining these results gives

$$\begin{aligned}
&\int\limits_{D_d(r)} u_1^{n_1}u_2^{n_2}\ldots u_d^{n_d}du_1\ldots du_d \\
&= \int\limits_0^\rho\int\limits_o^{2\pi}\int\limits_0^\pi\cdots\int\limits_0^\pi r^{q+d-1}\Big[\sin^{q-n_1+d-2}\left(\phi_1\right)\sin^{q-n_1-n_d+d-3}\left(\phi_2\right)\cdots \\
&\quad \sin^{n_d+n_{d-1}+1}\left(\phi_{d-2}\right)\sin^{n_d}\left(\phi_{d-1}\right)\Big]\left[\cos^{n_1}\left(\phi_1\right)\cdots\cos^{n_d}\left(\phi_{d-1}\right)\right]d\phi_1\cdots d\phi_{d-1}dr \\
&= C\rho^{q+d}.
\end{aligned}$$

□

The region of integration for the inner integral in (A.5) corresponds to a hyperspherical cap with radius 1 and height of $\frac{1-x_i}{h}$. The inner integral can be calculated using an approach similar to that used in [124] to calculate the volume of a hyperspherical cap. It is obtained by integrating the polynomial $p_x(u)$ over a $d-1$-sphere with radius $\sin\theta$ and height element $d\cos\theta$. This is done using Lemma A.1. We then integrate over θ which has a range of 0 to $\phi = \cos^{-1}\left(\frac{1-x_i}{h}\right)$. Thus we have

$$\begin{aligned}\int\limits_{u:||u||_2\leq 1, x+uh\notin\mathcal{S}} p_x(u)du &= \sum_{m=0}^{q} \tilde{p}_m(x) \int\limits_0^{\phi} \sin^{m+d-1}(\theta)\sin\theta u_d^m d\theta \\ &= \sum_{m=0}^{q} \tilde{p}_m(x) \int\limits_0^{\phi} \sin^{m+d}(\theta)\cos^m\theta d\theta. \qquad (A.6)\end{aligned}$$

From standard integral tables, we get that for $n \geq 2$ and $m \geq 0$

$$\int\limits_0^{\phi} \sin^n\theta\cos^m\theta d\theta = -\frac{\sin^{n-1}\phi\cos^{m+1}\phi}{n+m} + \frac{n-1}{n+m}\int\limits_0^{\phi}\sin^{n-2}\theta\cos^m\theta d\theta. \qquad (A.7)$$

If $n = 1$, then we get

$$\int\limits_0^{\phi} \sin\theta\cos^m\theta d\theta = \frac{1}{m+1} - \frac{\cos^{m+1}\phi}{m+1}.$$

Since $\phi = \cos^{-1}\left(\frac{1-x_i}{h}\right)$, we have

$$\begin{aligned}\cos\phi &= \frac{1-x_i}{h}, \\ \sin\phi &= \sqrt{1-\left(\frac{1-x_i}{h}\right)^2}.\end{aligned}$$

Therefore, if n is odd, we obtain

$$\int_0^\phi \sin^n\theta\cos^m\theta d\theta = \sum_{\ell=0}^{(n-1)/2} c_\ell \left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{2\ell} \left(\frac{1-x_i}{h}\right)^{m+1} + c, \qquad \text{(A.8)}$$

where the constants depend on m and n.

If n is even and $m > 0$, then the final term in the recursion in (A.7) reduces to

$$\int_0^\phi \cos^m\theta d\theta = \frac{\cos^{m-1}\phi\sin\phi}{m} + \frac{m-1}{m}\int_0^\phi \cos^{m-2}\theta d\theta.$$

If $m = 2$, then

$$\begin{aligned} \int_0^\phi \cos^2\theta d\theta &= \frac{\phi}{2} + \frac{1}{4}\sin(2\phi) \\ &= \frac{\phi}{2} + \frac{1}{2}\sin\phi\cos\phi. \end{aligned}$$

Therefore, if n and m are both even, then this gives

$$\begin{aligned} \int_0^\phi \sin^n\theta\cos^m\theta d\theta &= \sum_{\ell=0}^{(n-2)/2} c_\ell^{'} \left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{2\ell+1} \left(\frac{1-x_i}{h}\right)^{m+1} + c^{'}\cos^{-1}\left(\frac{1-x_i}{h}\right) \\ &\quad + \sum_{\ell=0}^{(m-2)/2} c_\ell^{''} \left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right) \left(\frac{1-x_i}{h}\right)^{2\ell+1}. \qquad \text{(A.9)} \end{aligned}$$

On the other hand, if n is even and m is odd, we get

$$\begin{aligned} \int_0^\phi \sin^n\theta\cos^m\theta d\theta &= \sum_{\ell=0}^{(n-2)/2} c_\ell^{'''} \left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{2\ell+1} \left(\frac{1-x_i}{h}\right)^{m+1} \\ &\quad + \sum_{\ell=0}^{(m-1)/2} c_\ell^{''''} \left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right) \left(\frac{1-x_i}{h}\right)^{2\ell}. \qquad \text{(A.10)} \end{aligned}$$

If d is odd, then combining (A.8) and (A.10) with (A.6) gives

$$\int\limits_{u:||u||_2\leq 1,x+uh\notin\mathcal{S}} p_x(u)du \;=\; \sum_{m=0}^{q}\sum_{\ell=0}^{d+q} p_{m,\ell}(x)\left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{\ell}\left(\frac{1-x_i}{h}\right)^{m} \text{(A,11)}$$

where the coefficients $p_{m,\ell}(x)$ are $r-q$ times differentiable wrt x. Similarly, if d is even, then

$$\begin{aligned}\int\limits_{u:||u||_2\leq 1,x+uh\notin\mathcal{S}} p_x(u)du \;=\; & \sum_{m=0}^{q}\sum_{\ell=0}^{d+q} p'_{m,\ell}(x)\left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{\ell}\left(\frac{1-x_i}{h}\right)^{m} \\ & +p'(x)\cos^{-1}\left(\frac{1-x_i}{h}\right),\end{aligned} \qquad \text{(A.12)}$$

where again the coefficients $p'_{m,\ell}(x)$ and $p'(x)$ are $r-q$ times differentiable wrt x. Raising (A.11) and (A.12) to the power of t gives respective expressions of the form

$$\sum_{m=0}^{qt}\sum_{\ell=0}^{(d+q)t} \check{p}_{m,\ell}(x)\left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{\ell}\left(\frac{1-x_i}{h}\right)^{m}, \qquad \text{(A.13)}$$

$$\sum_{m=0}^{qt}\sum_{\ell=0}^{(d+q)t}\sum_{n=0}^{t} \check{p}_{m,\ell,n}(x)\left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{\ell}\left(\frac{1-x_i}{h}\right)^{m}\left(\cos^{-1}\left(\frac{1-x_i}{h}\right)\right)^{n}, \qquad \text{(A.14)}$$

where the coefficients $\check{p}_{m,\ell}(x)$ and $\check{p}_{m,\ell,n}(x)$ are all $r-q$ times differentiable wrt x. Integrating (A.13) and (A.14) over all the coordinates in x except for x_i affects only the $\check{p}_{m,\ell}(x)$ and $\check{p}_{m,\ell,n}(x)$ coefficients, resulting in respective expressions of the form

$$\sum_{m=0}^{qt}\sum_{\ell=0}^{(d+q)t} \bar{p}_{m,\ell}(x_i)\left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{\ell}\left(\frac{1-x_i}{h}\right)^{m}, \qquad \text{(A.15)}$$

$$\sum_{m=0}^{qt} \sum_{\ell=0}^{(d+q)t} \sum_{n=0}^{t} \bar{p}_{m,\ell,n}(x_i) \left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{\ell} \left(\frac{1-x_i}{h}\right)^{m} \left(\cos^{-1}\left(\frac{1-x_i}{h}\right)\right)^{n}. \tag{A.16}$$

The coefficients $\bar{p}_{m,\ell}(x_i)$ and $\bar{p}_{m,\ell,n}(x_i)$ are $r-q$ times differentiable wrt x_i. Since the other coordinates of x other than x_i are far away from the boundary, the coefficients are independent of h. For the integral wrt x_i of (A.15), taking a Taylor series expansion of $\bar{p}_{m,\ell}(x_i)$ around $x_i = 1$ yields terms of the form

$$\begin{aligned} \int_{1-h}^{1} \left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{\ell} \left(\frac{1-x_i}{h}\right)^{m+j} h^j dx_i &= h^{j+1} \int_{0}^{1} (1-y_i)^{\frac{\ell}{2}} y_i^{\frac{m+j-1}{2}} dy_i \\ &= h^{j+1} B\left(\frac{\ell+2}{2}, \frac{m+j+1}{2}\right), \end{aligned}$$

where $0 \leq j \leq r-q$, $0 \leq \ell \leq (d+q)t$, $0 \leq m \leq qt$, and $B(x,y)$ is the beta function. Note that the first step uses the substitution of $y_i = \left(\frac{1-x_i}{h}\right)^2$.

If d is even (i.e. (A.16)), a simple closed-form expression is not easy to obtain due to the $\cos^{-1}\left(\frac{1-x_i}{h}\right)$ terms. However, by similarly applying a Taylor series expansion to $\bar{p}_{m,\ell,n}(x_i)$ and substituting $y_i = \frac{1-x_i}{h}$ gives terms of the form of

$$\begin{aligned} &\int_{1-h}^{1} \left(\sqrt{1-\left(\frac{1-x_i}{h}\right)^2}\right)^{\ell} \left(\frac{1-x_i}{h}\right)^{m+j} \left(\cos^{-1}\left(\frac{1-x_i}{h}\right)\right)^{n} h^j dx_i \\ &= h^{j+1} \int_{0}^{1} \left(1-y_i^2\right)^{\frac{\ell}{2}} y_i^{m+j} \left(\cos^{-1} y_i\right)^n dy_i \\ &= h^{j+1} c_{\ell,m,j,n}, \end{aligned}$$

for $0 \leq j \leq r-q$, $0 \leq \ell \leq (d+q)t$, $0 \leq m \leq qt$, and $0 \leq n \leq t$. Combining terms results in the expansion $v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o(h^{r-q})$.

A.2.2 Multiple Coordinate Boundary Point

The case where multiple coordinates of the point x are near the boundary is a fairly straightforward extension of the single boundary point case. Consider the case where 2 of the coordinates are near the boundary, e.g., x_1 and x_2 with $x_1+u_1h > 1$ and $x_2+u_2h > 1$. The region of integration for the inner integral can be decomposed into two parts: a hyperspherical cap wrt x_1 and the remaining area (denoted, respectively, as A_1 and A_2). The remaining area A_2 can be decomposed further into two other areas: a hyperspherical cap wrt x_2 (denoted B_1) and a height chosen s.t. B_1 just intersects A_1 on their boundaries. Integrating over the remainder of A_2 is achieved by integrating along x_2 over $d-1$-dimensional hyperspherical caps from the boundary of B_1 to the boundary of A_2. Thus integrating over these regions yields an expression similar to (A.6). Following a similar procedure will then yield the result.

APPENDIX B

Proofs for KDE Plug-in Estimators

This appendix contains the proofs for the KDE approaches.

B.1 Proof of Theorem II.2 (Bias)

In this appendix, we prove the bias results in Thm. IV.1. The bias of the base kernel density plug-in estimator $\tilde{\mathbf{G}}_{h_1,h_2}$ can be expressed as

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right] &= \mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\right]\\ &= \mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\right]\\ &\quad +\mathbb{E}\left[g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\right],\end{aligned}\tag{B.1}$$

where $\mathbf{Z}$ is drawn from f_2. The first term is the "variance" term while the second is the "bias" term. We bound these terms using Taylor series expansions under the assumption that g is infinitely differentiable. The Taylor series expansion of the variance term in (D.1) will depend on variance-like terms of the KDEs while the Taylor series expansion of the bias term in (D.1) will depend on the bias of the KDEs.

The Taylor series expansion of $g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)$ around $f_1(\mathbf{Z})$ and $f_2(\mathbf{Z})$

is

$$g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\left(\left.\frac{\partial^{i+j}g(x,y)}{\partial x^i\partial y^j}\right|_{\substack{x=f_1(\mathbf{Z})\\ y=f_2(\mathbf{Z})}}\right)\frac{\mathbb{B}_{\mathbf{Z}}^{i}\left[\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})\right]\mathbb{B}_{\mathbf{Z}}^{j}\left[\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right]}{i!j!} \tag{B.2}$$

where $\mathbb{B}_{\mathbf{Z}}^{j}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]=\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})-f_i(\mathbf{Z})\right)^j$ is the bias of $\tilde{\mathbf{f}}_{i,h_i}$ at the point $\mathbf{Z}$ raised to the power of j. This expansion can be used to control the second term (the bias term) in (D.1). To accomplish this, we require an expression for $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})-f_i(\mathbf{Z})=\mathbb{B}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]$.

To obtain an expression for $\mathbb{B}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]$, we consider separately the cases when $\mathbf{Z}$ is in the interior of the support $\mathcal{S}$ or when $\mathbf{Z}$ is near the boundary of the support. A point $X\in\mathcal{S}$ is defined to be in the interior of $\mathcal{S}$ if for all $Y\notin\mathcal{S}$, $K\left(\frac{X-Y}{h_i}\right)=0$. A point $X\in\mathcal{S}$ is near the boundary of the support if it is not in the interior. Denote the region in the interior and near the boundary wrt h_i as $\mathcal{S}_{I_i}$ and $\mathcal{S}_{B_i}$, respectively. We will need the following.

Lemma B.1. *Let $\mathbf{Z}$ be a realization of the density f_2 independent of $\tilde{\mathbf{f}}_{i,h_i}$ for $i=1,2$. Assume that the densities f_1 and f_2 belong to $\Sigma(s,L)$. Then for $\mathbf{Z}\in\mathcal{S}_{I_i}$,*

$$\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]=f_i(\mathbf{Z})+\sum_{j=\nu/2}^{\lfloor s/2\rfloor}c_{i,j}(\mathbf{Z})h_i^{2j}+O\left(h_i^s\right). \tag{B.3}$$

Proof. Obtaining the lower order terms in (B.3) is a common result in kernel density estimation. However, since we also require the higher order terms, we present the proof here. Additionally, some of the results in this proof will be useful later. From the linearity of the KDE, we have that if $\mathbf{X}$ is drawn from f_i and is independent of

$\mathbf{Z}$, then

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) &= \mathbb{E}_{\mathbf{Z}}\left[\frac{1}{h_i^d}K\left(\frac{\mathbf{X}-\mathbf{Z}}{h_i}\right)\right] \\ &= \int \frac{1}{h_i^d}K\left(\frac{x-\mathbf{Z}}{h_i}\right)f_i(x)dx \\ &= \int K\left(t\right)f_i(th_i+\mathbf{Z})dt, \qquad \text{(B.4)}\end{aligned}$$

where the last step follows from the substitution $t=\frac{x-\mathbf{Z}}{h_i}$. Since the density f_i belongs to $\Sigma(s,K)$, using multi-index notation we can expand it as

$$f_i(th_i+\mathbf{Z}) = f_i(\mathbf{Z}) + \sum_{0<|\alpha|\leq\lfloor s\rfloor}\frac{D^\alpha f_i(\mathbf{Z})}{\alpha!}(th_i)^\alpha + O\left(\|th_i\|^s\right), \qquad \text{(B.5)}$$

where $\alpha! = \alpha_1!\alpha_2!\dots\alpha_d!$ and $Y^\alpha = Y_1^{\alpha_1}Y_2^{\alpha_2}\dots Y_d^{\alpha_d}$. Combining (B.4) and (B.5) gives

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) &= f_i(\mathbf{Z}) + \sum_{0<|\alpha|\leq\lfloor s\rfloor}\frac{D^\alpha f_i(\mathbf{Z})}{\alpha!}h_i^{|\alpha|}\int t^\alpha K(t)dt + O(h_i^s) \\ &= f_i(\mathbf{Z}) + \sum_{j=\nu/2}^{\lfloor s/2\rfloor} c_{i,j}(\mathbf{Z})h_i^{2j} + O(h_i^s),\end{aligned}$$

where the last step follows from the fact that K is symmetric and of order ν. □

To obtain a similar result for the case when $\mathbf{Z}$ is near the boundary of $\mathcal{S}$, we use assumption $\mathcal{A}.5$.

Lemma B.2. *Let $\gamma(x,y)$ be an arbitrary function satisfying $\sup_{x,y}|\gamma(x,y)|<\infty$. Let $\mathcal{S}$ satisfy the boundary smoothness conditions of Assumption $\mathcal{A}.5$. Assume that the densities f_1 and f_2 belong to $\Sigma(s,L)$ and let $\mathbf{Z}$ be a realization of the density f_2 independent of $\tilde{\mathbf{f}}_{i,h_i}$ for $i=1,2$. Let $h^{'}=\min\left(h_1,h_2\right)$. Then*

$$\mathbb{E}\left[1_{\left\{\mathbf{Z}\in\mathcal{S}_{B_i}\right\}}\gamma\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\mathbb{B}_{\mathbf{Z}}^t\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]\right] = \sum_{j=1}^{r}c_{4,i,j,t}h_i^j + o\left(h_i^r\right) \qquad \text{(B.6)}$$

$$\mathbb{E}\left[1_{\left\{\mathbf{Z}\in\mathcal{S}_{B_1}\cap\mathcal{S}_{B_2}\right\}}\gamma\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\mathbb{B}_{\mathbf{Z}}^t\left[\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})\right]\mathbb{B}_{\mathbf{Z}}^q\left[\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right]\right] = \sum_{j=0}^{r-1}\sum_{i=0}^{r-1}c_{4,j,i,q,t}h_1^jh_2^ih^{'} + o\left(\left(h^{'}\right)^r\right) \quad \text{(B.7)}$$

Proof. For fixed X near the boundary of $\mathcal{S}$, we have

$$\begin{aligned}
\mathbb{E}\left[\tilde{\mathbf{f}}_{i,h_i}(X)\right]-f_i(X) &= \frac{1}{h_i^d}\int_{Y:Y\in\mathcal{S}}K\left(\frac{X-Y}{h_i}\right)f_i(Y)dY-f_i(X)\\
&= \left[\frac{1}{h_i^d}\int_{Y:K\left(\frac{X-Y}{h_i}\right)>0}K\left(\frac{X-Y}{h_i}\right)f_i(Y)dY-f_i(X)\right]\\
&\quad -\left[\frac{1}{h_i^d}\int_{Y:Y\notin\mathcal{S}}K\left(\frac{X-Y}{h_i}\right)f_i(Y)dY\right]\\
&= T_{1,i}(X)-T_{2,i}(X).
\end{aligned}$$

Note that in $T_{1,i}(X)$, we are extending the integral beyond the support of the density f_i. However, by using the same Taylor series expansion method as in the proof of Lemma B.1, we always evaluate f_i and its derivatives at the point X which is within the support of f_i. Thus it does not matter how we define an extension of f_i since the Taylor series will remain the same. Thus $T_{1,i}(X)$ results in an identical expression to that obtained from (B.3).

For the $T_{2,i}(X)$ term, we expand it as follows using multi-index notation as

$$\begin{aligned}
T_{2,i}(X) &= \frac{1}{h_i^d}\int_{Y:Y\notin\mathcal{S}}K\left(\frac{X-Y}{h_i}\right)f_i(Y)dY\\
&= \int_{u:h_iu+X\notin\mathcal{S},K(u)>0}K(u)\,f_i(X+h_iu)du\\
&= \sum_{|\alpha|\le r}\frac{h_i^{|\alpha|}}{\alpha!}\int_{u:h_iu+X\notin\mathcal{S},K(u)>0}K(u)\,D^\alpha f_i(X)u^\alpha du+o\left(h_i^r\right).
\end{aligned}$$

Recognizing that the $|\alpha|$th derivative of f_i is $r-|\alpha|$ times differentiable, we can apply assumption $\mathcal{A}.5$ to obtain the expectation of $T_{2,i}(X)$ wrt X:

$$\begin{aligned}
\mathbb{E}\left[T_{2,i}(\mathbf{X})\right] &= \frac{1}{h_i^d}\int_X\int_{Y:Y\notin\mathcal{S}} K\left(\frac{X-Y}{h_i}\right)f_i(Y)dYf_2(X)dx \\
&= \sum_{|\alpha|\le r}\frac{h_i^{|\alpha|}}{\alpha!}\int_X\int_{u:h_iu+X\notin\mathcal{S},K(u)>0} K\left(u\right)D^\alpha f_i(X)u^\alpha duf_2(X)dX+o\left(h_i^r\right) \\
&= \sum_{|\alpha|\le r}\frac{h_i^{|\alpha|}}{\alpha!}\left[\sum_{1\le|\beta|\le r-|\alpha|} e_{\beta,r-|\alpha|}h_i^{|\beta|}+o\left(h_i^{r-|\alpha|}\right)\right]+o\left(h_i^r\right) \\
&= \sum_{j=1}^r e_jh_i^j+o\left(h_i^r\right).
\end{aligned}$$

Similarly, we find that

$$\begin{aligned}
\mathbb{E}\left[\left(T_{2,i}(\mathbf{X})\right)^t\right] &= \frac{1}{h_i^{dt}}\int_X\left(\int_{Y:Y\notin\mathcal{S}} K\left(\frac{X-Y}{h_i}\right)f_i(Y)dY\right)^t f_2(X)dx \\
&= \int_X\left(\sum_{|\alpha|\le r}\frac{h_i^{|\alpha|}}{\alpha!}\int_{u:h_iu+X\notin\mathcal{S},K(u)>0} K\left(u\right)D^\alpha f_i(X)u^\alpha du\right)^t f_2(X)dX \\
&= \sum_{j=1}^r e_{j,t}h_i^j+o\left(h_i^r\right).
\end{aligned}$$

Combining these results gives

$$\begin{aligned}
&\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}_B\}}\gamma\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\left(\mathbb{E}_\mathbf{Z}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]-f_i(\mathbf{Z})\right)^t\right] \\
&= \mathbb{E}\left[\gamma\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\left(T_{1,i}(\mathbf{Z})-T_{2,i}(\mathbf{Z})\right)^t\right] \\
&= \mathbb{E}\left[\gamma\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\sum_{j=0}^t\binom{t}{j}\left(T_{1,i}(\mathbf{Z})\right)^j\left(-T_{2,i}(\mathbf{Z})\right)^{t-j}\right] \\
&= \sum_{j=1}^r c_{4,i,j,t}h_i^j+o\left(h_i^r\right),
\end{aligned}$$

where the constants are functionals of the kernel, γ, and the densities.

The expression in (B.7) can be proved in a similar manner. □

Applying Lemmas B.1 and B.2 to (B.2) gives

$$\begin{aligned}
&\mathbb{E}\left[g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\right] \\
&= \sum_{j=1}^{r}\left(c_{4,1,j}h_1^j+c_{4,2,j}h_2^j\right)+\sum_{j=0}^{r-1}\sum_{i=0}^{r-1}c_{5,i,j}h_1^jh_2^ih^{'}+o\left(h_1^r+h_2^r\right).
\end{aligned}$$

For the variance term (the first term) in (D.1), the truncated Taylor series expansion of $g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)$ around $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})$ and $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})$ gives

$$\begin{aligned}
g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) &= \sum_{i=0}^{\lambda}\sum_{j=0}^{\lambda}\left(\left.\frac{\partial^{i+j}g(x,y)}{\partial x^i\partial y^j}\right|_{\substack{x=\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})\\ y=\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})}}\right)\frac{\tilde{\mathbf{e}}_{1,h_1}^i(\mathbf{Z})\tilde{\mathbf{e}}_{2,h_2}^j(\mathbf{Z})}{i!j!} \\
&\quad +o\left(\tilde{\mathbf{e}}_{1,h_1}^{\lambda}(\mathbf{Z})+\tilde{\mathbf{e}}_{2,h_2}^{\lambda}(\mathbf{Z})\right) \qquad \text{(B.8)}
\end{aligned}$$

where $\tilde{\mathbf{e}}_{i,h_i}(\mathbf{Z}):=\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})-\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$. To control the variance term in (D.1), we thus require expressions for $\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{i,h_i}^j(\mathbf{Z})\right]$.

Lemma B.3. *Let* $\mathbf{Z}$ *be a realization of the density* f_2 *that is in the interior of the support and is independent of* $\tilde{\mathbf{f}}_{i,h_i}$ *for* $i=1,2$. *Let* $n(q)$ *be the set of integer divisors of* q *including* 1 *but excluding* q. *Then,*

$$\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{i,h_i}^q(\mathbf{Z})\right] = \begin{cases}\sum_{j\in n(q)}\frac{1}{\left(N_2h_2^d\right)^{q-j}}\sum_{m=0}^{\lfloor s/2\rfloor}c_{6,i,q,j,m}(\mathbf{Z})h_i^{2m}+O\left(\frac{1}{N_i}\right), & q\geq 2\\ 0, & q=1,\end{cases} \qquad \text{(B.9)}$$

$$\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{1,h_1}^q(\mathbf{Z})\tilde{\mathbf{e}}_{2,h_2}^l(\mathbf{Z})\right] = \begin{cases}\left(\sum_{i\in n(q)}\frac{1}{\left(N_1h_1^d\right)^{q-i}}\sum_{m=0}^{\lfloor s/2\rfloor}c_{6,1,q,i,m}(\mathbf{Z})h_1^{2m}\right)\times & q,\ l\geq 2\\ \left(\sum_{j\in n(l)}\frac{1}{\left(N_2h_2^d\right)^{l-j}}\sum_{t=0}^{\lfloor s/2\rfloor}c_{6,2,l,j,t}(\mathbf{Z})h_2^{2t}\right) & \\ +O\left(\frac{1}{N_1}+\frac{1}{N_2}\right), & \\ 0, & q=1\ or\ l=1\end{cases} \qquad \text{(B.10)}$$

where $c_{6,i,q,j,m}$ *is a functional of* f_1 *and* f_2.

Proof. Define the random variable $\mathbf{V}_i(\mathbf{Z}) = K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{h_2}\right) - \mathbb{E}_{\mathbf{Z}}K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{h_2}\right)$. This gives

$$\begin{aligned}\tilde{\mathbf{e}}_{2,h_2}(\mathbf{Z}) &= \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \\ &= \frac{1}{N_2h_2^d}\sum_{i=1}^{N_2}\mathbf{V}_i(\mathbf{Z}).\end{aligned}$$

Clearly, $\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i(\mathbf{Z}) = 0$. From (B.4), we have for integer $j \geq 1$

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[K^j\left(\frac{\mathbf{X}_i-\mathbf{Z}}{h_2}\right)\right] &= \int K^j\left(t\right)f_2(th_2+\mathbf{Z})dt \\ &= h_2^d\sum_{m=0}^{\lfloor s/2\rfloor}c_{3,2,j,m}(\mathbf{Z})h_2^{2m},\end{aligned}$$

where the constants $c_{3,2,j,m}$ depend on the density f_2, its derivatives, and the moments of the kernel K^j. Note that since K is symmetric, the odd moments of K^j are zero for $\mathbf{Z}$ in the interior of the support. However, all even moments may now be nonzero since K^j may now be nonnegative. By the binomial theorem,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\mathbf{V}_i^j(\mathbf{Z})\right] &= \sum_{k=0}^{j}\binom{j}{k}\mathbb{E}_{\mathbf{Z}}\left[K^k\left(\frac{\mathbf{X}_i-\mathbf{Z}}{h_2}\right)\right]\mathbb{E}_{\mathbf{Z}}\left[K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{h_2}\right)\right]^{j-k} \\ &= \sum_{k=0}^{j}\binom{j}{k}h_2^dO\left(h_2^{d(j-k)}\right)\sum_{m=0}^{\lfloor s/2\rfloor}c_{3,2,k,m}(\mathbf{Z})h_2^{2m} \\ &= h_2^d\sum_{m=0}^{\lfloor s/2\rfloor}c_{3,2,j,m}(\mathbf{Z})h_2^{2m} + O\left(h^{2d}\right).\end{aligned}$$

We can use these expressions to simplify $\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^q(\mathbf{Z})\right]$. As an example, let $q = 2$. Then since the $\mathbf{X}_i s$ are independent,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^2(\mathbf{Z})\right] &= \frac{1}{N_2h_2^{2d}}\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i^2(\mathbf{Z}) \\ &= \frac{1}{N_2h_2^d}\sum_{m=0}^{\lfloor s/2\rfloor}c_{3,2,2,m}(\mathbf{Z})h_2^{2m} + O\left(\frac{1}{N_2}\right).\end{aligned}$$

Similarly, we find that

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^{3}(\mathbf{Z})\right] &= \frac{1}{N_2^2h_2^{3d}}\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i^3(\mathbf{Z}) \\ &= \frac{1}{\left(N_2h_2^d\right)^2}\sum_{m=0}^{\lfloor s/2\rfloor} c_{3,2,3,m}(\mathbf{Z})h_2^{2m} + o\left(\frac{1}{N_2}\right).\end{aligned}$$

For $q = 4$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^{4}(\mathbf{Z})\right] &= \frac{1}{N_2^3h_2^{4d}}\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i^4(\mathbf{Z}) + \frac{N_2-1}{N_2^3h_2^{4d}}\left(\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i^2(\mathbf{Z})\right)^2 \\ &= \frac{1}{\left(N_2h_2^d\right)^3}\sum_{m=0}^{\lfloor s/2\rfloor} c_{3,2,4,m}(\mathbf{Z})h_2^{2m} + \frac{1}{\left(N_2h_2^d\right)^2}\sum_{m=0}^{\lfloor s/2\rfloor} c_{6,2,2,m}(\mathbf{Z})h_2^{2m} + o\left(\frac{1}{N_2}\right).\end{aligned}$$

The pattern is then for $q \geq 2$,

$$\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^{q}(\mathbf{Z})\right] = \sum_{i\in n(q)} \frac{1}{\left(N_2h_2^d\right)^{q-i}}\sum_{m=0}^{\lfloor s/2\rfloor} c_{6,2,q,i,m}(\mathbf{Z})h_2^{2m} + O\left(\frac{1}{N_2}\right).$$

For any integer q, the largest possible factor is $q/2$. Thus for given q, the smallest possible exponent on the $N_2h_2^d$ term is $q/2$. This increases as q increases. A similar expression holds for $\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{1,h_1}^{q}(\mathbf{Z})\right]$ except the $\mathbf{X}_i$s are replaced with $\mathbf{Y}_i$, f_2 is replaced with f_1, and N_2 and h_2 are replaced with N_1 and h_1, respectively, all resulting in different constants. Then since $\tilde{\mathbf{e}}_{1,h_1}(\mathbf{Z})$ and $\tilde{\mathbf{e}}_{2,h_2}(\mathbf{Z})$ are conditionally independent given $\mathbf{Z}$,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{1,h_1}^{q}(\mathbf{Z})\tilde{\mathbf{e}}_{2,h_2}^{l}(\mathbf{Z})\right] &= O\left(\frac{1}{N_1}+\frac{1}{N_2}\right) + \left(\sum_{i\in n(q)} \frac{1}{\left(N_1h_1^d\right)^{q-i}}\sum_{m=0}^{\lfloor s/2\rfloor} c_{6,1,q,i,m}(\mathbf{Z})h_1^{2m}\right) \\ &\quad \times \left(\sum_{j\in n(l)} \frac{1}{\left(N_2h_2^d\right)^{l-j}}\sum_{t=0}^{\lfloor s/2\rfloor} c_{6,2,l,j,t}(\mathbf{Z})h_2^{2t}\right).\end{aligned}$$

□

Applying Lemma B.3 to (B.8) when taking the conditional expectation given $\mathbf{Z}$ in the interior gives an expression of the form

$$\sum_{j=1}^{\lambda/2}\sum_{m=0}^{\lfloor s/2\rfloor}\left(c_{7,1,j,m}\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\frac{h_1^{2m}}{\left(N_1h_1^d\right)^j}\right.$$
$$\left.+\,c_{7,2,j,m}\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\frac{h_2^{2m}}{\left(N_2h_2^d\right)^j}\right)+O\left(\frac{1}{\left(N_1h_1^d\right)^{\frac{\lambda}{2}}}+\frac{1}{\left(N_2h_2^d\right)^{\frac{\lambda}{2}}}\right)$$
$$+\sum_{j=1}^{\lambda/2}\sum_{m=0}^{\lfloor s/2\rfloor}\sum_{i=1}^{\lambda/2}\sum_{n=0}^{\lfloor s/2\rfloor}c_{7,j,i,m,n}\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\frac{h_1^{2m}h_2^{2n}}{\left(N_1h_1^d\right)^j\left(N_2h_2^d\right)^i}. \tag{B.11}$$

Note that the functionals $c_{7,i,j,m}$ and $c_{7,j,i,m,n}$ depend on the derivatives of g and $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$ which depends on h_i. To apply ensemble estimation, we need to separate the dependence on h_i from the constants. If we use ODin1, then it is sufficient to note that in the interior of the support, $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) = f_i(\mathbf{Z}) + o(1)$ and therefore $c\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) = c\left(f_1(\mathbf{Z}), f_2(\mathbf{Z})\right) + o(1)$ for some functional c. The terms in (B.11) reduce to

$$c_{7,1,1,0}\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\frac{1}{N_1h_1^d}+c_{7,2,1,0}\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\frac{1}{N_2h_2^d}+o\left(\frac{1}{N_1h_1^d}+\frac{1}{N_2h_2^d}\right).$$

For ODin2, we need the higher order terms. To separate the dependence on h_i from the constants, we need more information about the functional g and its derivatives. Consider the special case where the functional $g(x,y)$ has derivatives of the form of $x^\alpha y^\beta$ with $\alpha,\beta < 0$. This includes the important cases of the KL divergence and the Renyi divergence. The generalized binomial theorem states that if $\binom{\alpha}{m} := \frac{\alpha(\alpha-1)\dots(\alpha-m+1)}{m!}$ and if q and t are real numbers with $|q| > |t|$, then for any complex number α,

$$(q+t)^\alpha = \sum_{m=0}^{\infty}\binom{\alpha}{m}q^{\alpha-m}t^m. \tag{B.12}$$

Since the densities are bounded away from zero, for sufficiently small h_i, we have that

$f_i(\mathbf{Z}) > \left|\sum_{j=\nu/2}^{\lfloor s/2\rfloor} c_{i,j}(\mathbf{Z})h_i^{2j} + O\left(h_i^s\right)\right|$. Applying the generalized binomial theorem and Lemma B.1 gives that

$$\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})\right)^{\alpha} = \sum_{m=0}^{\infty}\binom{\alpha}{m} f_i^{\alpha-m}(\mathbf{Z})\left(\sum_{j=\nu/2}^{\lfloor s/2\rfloor} c_{i,j}(\mathbf{Z})h_i^{2j} + O\left(h_i^s\right)\right)^m.$$

Since m is an integer, the exponents of the h_i terms are also integers. Thus (B.11) gives in this case

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) - g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\right] \\
&= \sum_{j=1}^{\lambda/2}\sum_{m=0}^{\lfloor s/2\rfloor}\left(c_{8,1,j,m}\left(\mathbf{Z}\right)\frac{h_1^{2m}}{\left(N_1h_1^d\right)^j} + c_{8,2,j,m}\left(\mathbf{Z}\right)\frac{h_2^{2m}}{\left(N_2h_2^d\right)^j}\right) \\
&\quad + \sum_{j=1}^{\lambda/2}\sum_{m=0}^{\lfloor s/2\rfloor}\sum_{i=1}^{\lambda/2}\sum_{n=0}^{\lfloor s/2\rfloor} c_{8,j,i,m,n}\left(\mathbf{Z}\right)\frac{h_1^{2m}h_2^{2n}}{\left(N_1h_1^d\right)^j\left(N_2h_2^d\right)^i} \\
&\quad + O\left(\frac{1}{\left(N_1h_1^d\right)^{\frac{\lambda}{2}}} + \frac{1}{\left(N_2h_2^d\right)^{\frac{\lambda}{2}}} + h_1^s + h_2^s\right). \qquad \text{(B.13)}
\end{aligned}$$

As before, the case for $\mathbf{Z}$ close to the boundary of the support is more complicated. However, by using a similar technique to the proof of Lemma B.2 for $\mathbf{Z}$ at the boundary and combining with the previous results, we find that for general g,

$$\begin{aligned}
\mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) - g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\right] = \quad & c_{9,1}\frac{1}{N_1h_1^d} + c_{9,2}\frac{1}{N_2h_2^d} \\
& + o\left(\frac{1}{N_1h_1^d} + \frac{1}{N_2h_2^d}\right).
\end{aligned} \qquad \text{(B.14)}$$

If $g(x,y)$ has derivatives of the form of $x^\alpha y^\beta$ with $\alpha,\beta < 0$, then we can similarly obtain

$$
\begin{aligned}
&\mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\right] \\
= &\sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\left(c_{9,1,j,m}\frac{h_1^m}{\left(N_1h_1^d\right)^j}+c_{9,2,j,m}\frac{h_2^m}{\left(N_2h_2^d\right)^j}\right) \\
&+\sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\sum_{i=1}^{\lambda/2}\sum_{n=0}^{r}c_{9,j,i,m,n}\frac{h_1^mh_2^n}{\left(N_1h_1^d\right)^j\left(N_2h_2^d\right)^i} \\
&+O\left(\frac{1}{\left(N_1h_1^d\right)^{\frac{\lambda}{2}}}+\frac{1}{\left(N_2h_2^d\right)^{\frac{\lambda}{2}}}+h_1^s+h_2^s\right).
\end{aligned}
\tag{B.15}
$$

Combining (B.8) with either (B.14) or (B.15) completes the proof.

B.2 Proof of Theorem II.3 (Variance)

To bound the variance of the plug-in estimator $\tilde{\mathbf{G}}_{h_1,h_2}$, we will use the Efron-Stein inequality [53]:

Lemma B.4 (Efron-Stein Inequality). *Let $\mathbf{X}_1,\dots,\mathbf{X}_n,\mathbf{X}_1',\dots,\mathbf{X}_n'$ be independent random variables on the space $\mathcal{S}$. Then if $f:\mathcal{S}\times\cdots\times\mathcal{S}\to\mathbb{R}$, we have that*

$$\mathbb{V}\left[f(\mathbf{X}_1,\dots,\mathbf{X}_n)\right]\le\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\left(f(\mathbf{X}_1,\dots,\mathbf{X}_n)-f(\mathbf{X}_1,\dots,\mathbf{X}_i',\dots,\mathbf{X}_n)\right)^2\right].$$

Suppose we have samples $\{\mathbf{X}_1,\dots,\mathbf{X}_{N_2},\mathbf{Y}_1,\dots,\mathbf{Y}_{N_1}\}$ and $\{\mathbf{X}_1',\dots,\mathbf{X}_{N_2},\mathbf{Y}_1,\dots,\mathbf{Y}_{N_1}\}$ and denote the respective estimators as $\tilde{\mathbf{G}}_{h_1,h_2}$ and $\tilde{\mathbf{G}}_{h_1,h_2}'$. We have that

$$
\begin{aligned}
\left|\tilde{\mathbf{G}}_{h_1,h_2}-\tilde{\mathbf{G}}_{h_1,h_2}'\right|\le\ &\frac{1}{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1'),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1')\right)\right| \\
&+\frac{1}{N_2}\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}'(\mathbf{X}_j)\right)\right|.
\end{aligned}
\tag{B.16}
$$

Since g is Lipschitz continuous with constant C_g, we have

$$\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right)\right| \leq C_g\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'})\right| +C_g\left|\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right|, \tag{B.17}$$

$$\begin{aligned}\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'})\right| &= \frac{1}{N_1h_1^d}\left|\sum_{i=1}^{N_1}\left(K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right)\right| \\ &\leq \frac{1}{N_1h_1^d}\sum_{i=1}^{N_1}\left|K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right| \\ \implies \mathbb{E}\left[\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'})\right|^2\right] &\leq \frac{1}{N_1h_1^{2d}}\sum_{i=1}^{N_1}\mathbb{E}\left[\left(K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right)^2\right],\end{aligned} \tag{B.18}$$

where the last step follows from Jensen's inequality. By making the substitution $\mathbf{u}_i=\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}$ and $\mathbf{u}_i^{'}=\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}$, this gives

$$\begin{aligned}&\frac{1}{h_1^{2d}}\mathbb{E}\left[\left(K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right)^2\right] \\ &= \frac{1}{h^{2d}}\int\left(K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right)^2 f_2(\mathbf{X}_1)f_2(\mathbf{X}_1^{'})f_1(\mathbf{Y}_i)d\mathbf{X}_1d\mathbf{X}_1^{'}d\mathbf{Y}_i \\ &\leq 2||K||_\infty^2.\end{aligned}$$

Combining this with (C.19) gives

$$\mathbb{E}\left[\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'})\right|^2\right]\leq 2||K||_\infty^2.$$

Similarly,

$$\mathbb{E}\left[\left|\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right|^2\right]\leq 2||K||_\infty^2.$$

Combining these results with (C.18) gives

$$\mathbb{E}\left[\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right)\right)^2\right]\leq 8C_g^2||K||_\infty^2. \tag{B.19}$$

The second term in (D.5) is controlled in a similar way. From the Lipschitz condition,

$$\begin{aligned}&\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|^2\\ &\leq\; C_g^2\left|\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)-\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right|^2\\ &=\;\frac{C_g^2}{M_2^2h_2^{2d}}\left(K\left(\frac{\mathbf{X}_j-\mathbf{X}_1}{h}\right)-K\left(\frac{\mathbf{X}_j-\mathbf{X}_1^{'}}{h}\right)\right)^2.\end{aligned}$$

The h_2^{2d} terms are eliminated by making the substitutions of $\mathbf{u}_j=\frac{\mathbf{X}_j-\mathbf{X}_1}{h_2}$ and $\mathbf{u}_j^{'}=\frac{\mathbf{X}_j-\mathbf{X}_1^{'}}{h_2}$ within the expectation to obtain

$$\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|^2\right]\;\leq\;\frac{2C_g^2||K||_\infty^2}{M_2^2} \tag{B.20}$$

$$\begin{aligned}&\implies\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|\right)^2\right] &\text{(B.21)}\\ &=\;\sum_{j=2}^{N_2}\sum_{i=2}^{N_2}\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|\right.\\ &\qquad\left.\times\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_i)\right)\right|\right]\\ &\leq\;M_2^2\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|^2\right]\\ &\leq\;2C_g^2||K||_\infty^2, &\text{(B.22)}\end{aligned}$$

where we use the Cauchy Schwarz inequality to bound the expectation within each summand. Finally, applying Jensen's inequality and (B.19) and (C.27) gives

$$\begin{aligned}\mathbb{E}\left[\left|\tilde{\mathbf{G}}_{h_1,h_2}-\tilde{\mathbf{G}}'_{h_1,h_2}\right|^2\right] &\le \frac{2}{N_2^2}\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1)\right)\right|^2\right]\\ &\quad +\frac{2}{N_2^2}\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j)\right)\right|\right)^2\right]\\ &\le \frac{20C_g^2||K||_\infty^2}{N_2^2}.\end{aligned}$$

Now suppose we have samples $\{\mathbf{X}_1,\dots,\mathbf{X}_{N_2},\mathbf{Y}_1,\dots,\mathbf{Y}_{N_1}\}$ and $\left\{\mathbf{X}_1,\dots,\mathbf{X}_{N_2},\mathbf{Y}'_1,\dots,\mathbf{Y}_{N_1}\right\}$ and denote the respective estimators as $\tilde{\mathbf{G}}_{h_1,h_2}$ and $\tilde{\mathbf{G}}'_{h_1,h_2}$. Then

$$\begin{aligned}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right| &\le C_g\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j)-\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j)\right|\\ &= \frac{C_g}{N_1h_1^d}\left|K\left(\frac{\mathbf{X}_j-\mathbf{Y}_1}{h_1}\right)-K\left(\frac{\mathbf{X}_j}{}\right.\right.\\ \implies \mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right|^2\right] &\le \frac{2C_g^2||K||_\infty^2}{N_1^2}.\end{aligned}$$

Thus using a similar argument as was used to obtain (C.27),

$$\begin{aligned}\mathbb{E}\left[\left|\tilde{\mathbf{G}}_{h_1,h_2}-\tilde{\mathbf{G}}'_{h_1,h_2}\right|^2\right] &\le \frac{1}{N_2^2}\mathbb{E}\left[\left(\sum_{j=1}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right|\right)^2\right]\\ &\le \frac{2C_g^2||K||_\infty^2}{N_2^2}.\end{aligned}$$

Applying the Efron-Stein inequality gives

$$\mathbb{V}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right]\le\frac{10C_g^2||K||_\infty^2}{N_2}+\frac{C_g^2||K||_\infty^2N_1}{N_2^2}.$$

B.3 Proof of Theorem II.6 (CLT)

We are interested in the asymptotic distribution of

$$\begin{aligned}\sqrt{N_2}\left(\tilde{\mathbf{G}}_{h_1,h_2} - \mathbb{E}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right]\right) &= \frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\Big(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - \mathbb{E}_{\mathbf{X}_j}\Big[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\Big]\\ &+\frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\Big(\mathbb{E}_{\mathbf{X}_j}\Big[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\Big] - \mathbb{E}\Big[g\Big(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\end{aligned}$$

Note that by the standard central limit theorem [51], the second term converges in distribution to a Gaussian random variable. If the first term converges in probability to a constant (specifically, 0), then we can use Slutsky's theorem [78] to find the asymptotic distribution. So now we focus on the first term which we denote as $\mathbf{V}_{N_2}$.

To prove convergence in probability, we will use Chebyshev's inequality. Note that $\mathbb{E}\left[\mathbf{V}_{N_2}\right] = 0$. To bound the variance of $\mathbf{V}_{N_2}$, we again use the Efron-Stein inequality. Let $\mathbf{X}_1^{'}$ be drawn from f_2 and denote $\mathbf{V}_{N_2}$ and $\mathbf{V}_{N_2}^{'}$ as the sequences using $\mathbf{X}_1$ and $\mathbf{X}_1^{'}$, respectively. Then

$$\begin{aligned}\mathbf{V}_{N_2} - \mathbf{V}_{N_2}^{'} &= \frac{1}{\sqrt{N_2}}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right) - \mathbb{E}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right)\\ &-\frac{1}{\sqrt{N_2}}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right) - \mathbb{E}_{\mathbf{X}_1^{'}}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right)\right]\right)\\ &+\frac{1}{\sqrt{N_2}}\sum_{j=2}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right)\end{aligned} \quad (B.23)$$

Note that

$$\mathbb{E}\left[\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right) - \mathbb{E}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right)^2\right] = \mathbb{E}\Big[\mathbb{V}_{\mathbf{X}_1}\Big[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\Big]$$

If we condition on $\mathbf{X}_1$, then by the standard central limit theorem $\sqrt{N_i h_i^d}\Big(\tilde{\mathbf{f}}_{i,h_i}(\mathbf{X}_1) - \mathbb{E}_{\mathbf{X}_1}\Big[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{X}_1)\Big]$ converges in distribution to a zero mean Gaussian random variable with variance $\sigma_i^2(\mathbf{X}_1) = O(1)$. This is true even if $\mathbf{X}_1$ is close to the boundary of the support of the densities. The KDEs $\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)$ and $\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)$ are conditionally independent given $\mathbf{X}_1$ as are their limiting distributions. Thus the KDEs converge jointly in distribution to a Gaussian random vector with zero mean, zero covariance, and their respective variances. By the delta method [103], we have that if $g(x,y)$ is continuously differentiable with respect to both x and y at $\mathbb{E}_{\mathbf{X}_1}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{X}_1)\right]$ for $i=1,2$, respectively, then

$$\mathbb{V}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]=O\left(\frac{1}{N_1h_1^d}+\frac{1}{N_2h_2^d}\right)=o(1),$$

provided that $N_ih_i^d\to\infty$. Thus $\mathbb{E}\left[\mathbb{V}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right]=o(1)$. A similar result holds when we replace $\mathbf{X}_1$ with $\mathbf{X}_1^{'}$.

For the third term in (B.23),

$$\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|\right)^2\right]$$

$$=\sum_{j=2}^{N_2}\sum_{i=2}^{N_2}\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}\right.\right.\right.$$

There are M_2 terms where $i=j$ and we have from Appendix D.2 (see (B.20)) that

$$\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|^2\right]\leq\frac{2C_g^2||K||_\infty^2}{M_2^2}.$$

Thus these terms are $O\left(\frac{1}{M_2}\right)$. There are $M_2^2-M_2$ terms when $i\neq j$. In this case, we can do four substitutions of the form $\mathbf{u}_j=\frac{\mathbf{X}_j-\mathbf{X}_1}{h_2}$ to obtain

$$\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_i\right.\right.\right.$$

Then since $h_2^d = o(1)$, we get

$$\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}^{'}_{2,h_2}(\mathbf{X}_j)\right)\right|\right)^2\right]=o(1), \qquad \text{(B.24)}$$

$$\begin{aligned}\implies \mathbb{E}\left[\left(\mathbf{V}_{N_2}-\mathbf{V}^{'}_{N_2}\right)^2\right] &\leq \frac{3}{N_2}\mathbb{E}\left[\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-\mathbb{E}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right)^2\right] \\ &\quad +\frac{3}{N_2}\mathbb{E}\left[\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}^{'}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}^{'}_1)\right)-\mathbb{E}_{\mathbf{X}^{'}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}^{'}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}^{'}_1)\right)\right]\right)^2\right] \\ &\quad +\frac{3}{N_2}\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}^{'}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}^{'}_{2,h_2}(\mathbf{X}_j)\right)\right)\right)^2\right. \\ &= o\left(\frac{1}{N_2}\right).\end{aligned}$$

Now consider samples $\{\mathbf{X}_1,\ldots,\mathbf{X}_{N_2},\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_1}\}$ and $\left\{\mathbf{X}_1,\ldots,\mathbf{X}_{N_2},\mathbf{Y}^{'}_1,\ldots,\mathbf{Y}_{N_1}\right\}$ and the respective sequences $\mathbf{V}_{N_2}$ and $\mathbf{V}^{'}_{N_2}$. Then

$$\mathbf{V}_{N_2}-\mathbf{V}^{'}_{N_2} \;=\; \frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}^{'}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right).$$

Using a similar argument as that used to obtain (B.24), we have that if $h_1^d = o(1)$ and $N_1 \to \infty$, then

$$\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}^{'}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right|\right)^2\right]=o(1)$$

$$\implies \mathbb{E}\left[\left(\mathbf{V}_{N_2}-\mathbf{V}^{'}_{N_2}\right)^2\right]=o\left(\frac{1}{N_2}\right).$$

Applying the Efron-Stein inequality gives

$$\mathbb{V}\left[\mathbf{V}_{N_2}\right]=o\left(\frac{N_2+N_1}{N_2}\right)=o(1).$$

Thus by Chebyshev's inequality,

$$\Pr\left(|\mathbf{V}_{N_2}| > \epsilon\right) \leq \frac{\mathbb{V}\left[\mathbf{V}_{N_2}\right]}{\epsilon^2} = o(1),$$

and therefore $\mathbf{V}_{N_2}$ converges to zero in probability. By Slutsky's theorem, $\sqrt{N_2}\bigg(\tilde{\mathbf{G}}_{h_1,h_2} - \mathbb{E}\Big[\tilde{\mathbf{G}}_{h_1,h_2}\Big]$ converges in distribution to a zero mean Gaussian random variable with variance

$$\mathbb{V}\left[\mathbb{E}_{\mathbf{X}}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X})\right)\right]\right],$$

where $\mathbf{X}$ is drawn from f_2.

For the weighted ensemble estimator, we wish to know the asymptotic distribution of $\sqrt{N_2}\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right]\right)$ where $\tilde{\mathbf{G}}_w = \sum_{l\in\bar{l}} w(l)\tilde{\mathbf{G}}_{h(l)}$. We have that

$$\begin{aligned}\sqrt{N_2}\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right]\right) &= \frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\sum_{l\in\bar{l}} w(l)\bigg(g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right) - \mathbb{E}_{\mathbf{X}_j}\bigg[g\bigg(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h(l)}(\\ &\quad + \frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\Bigg(\mathbb{E}_{\mathbf{X}_j}\left[\sum_{l\in\bar{l}} w(l) g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right)\right] - \mathbb{E}\Bigg[\sum_{l\in\bar{l}} w(l) g\bigg(\tilde{\mathbf{f}}\end{aligned}$$

The second term again converges in distribution to a Gaussian random variable by the central limit theorem. The mean and variance are, respectively, zero and

$$\mathbb{V}\left[\sum_{l\in\bar{l}} w(l)\mathbb{E}_{\mathbf{X}}\left[g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}), \tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X})\right)\right]\right].$$

The first term is equal to

$$\begin{aligned}\sum_{l\in\bar{l}} w(l)\left(\frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right) - \mathbb{E}_{\mathbf{X}_j}\left[g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right)\right]\right)\right) &= \sum_{l\in\bar{l}} w(l) o_P \\ &= o_P(1),\end{aligned}$$

where $o_P(1)$ denotes convergence to zero in probability. In the last step, we used the fact that if two random variables converge in probability to constants, then their linear combination converges in probability to the linear combination of the constants. Combining this result with Slutsky's theorem completes the proof.

B.4 Proof of Theorem II.7 (Uniform MSE)

Since the MSE is equal to the square of the bias plus the variance, we can upper bound the left hand side of (2.9) with

$$\begin{aligned}\sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)} \mathbb{E}\left[\left(\tilde{\mathbf{G}}_{w_0} - G(p,q)\right)^2\right] &= \sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)} \left(\text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right)^2 + \text{Var}\left(\tilde{\mathbf{G}}_{w_0}\right)\right) \\ &\leq \sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)} \text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right)^2 + \sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)} \text{Var}\left(\tilde{\mathbf{G}}_{w_0}\right).\end{aligned}$$

From the assumptions (lipschitz, kernel bounded, weight calculated from relaxed opt. prob), we have that

$$\begin{aligned}\sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)} \text{Var}\left(\tilde{\mathbf{G}}_{w_0}\right) &\leq \sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)} \frac{11C_g^2||w_0||_2^2||K||_\infty}{N} \\ &= \frac{11C_g^2||w_0||_2^2||K||_\infty}{N},\end{aligned}$$

where the last step follows from the fact that all of the terms are independent of p and q.

For the bias, recall that if g is infinitely differentiable and if the optimal weight w_0 is calculated using the relaxed convex optimization problem, then

$$\begin{aligned}\text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right) &= \sum_{i\in J} c_i(p,q)\epsilon N^{-1/2}, \\ \implies \text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right)^2 &= \frac{\epsilon^2}{N}\left(\sum_{i\in J} c_i(p,q)\right)^2. \qquad \text{(B.25)}\end{aligned}$$

We use a topology argument to bound the supremum of this term. We will use the

Extreme Value Theorem [151]:

Theorem B.5 (Extreme Value Theorem)**.** *Let $f : X \to \mathbb{R}$ be continuous. If X is compact, then for every $x \in X$, there exist points $c, d \in X$ s.t. $f(c) \leq f(x) \leq f(d)$.*

By this theorem, f achieves its minimum and maximum on X. Our approach is to first show that the functionals $c_i(p, q)$ are continuous wrt p and q in some appropriate norm. We will then show that the space $\Sigma(s, K, \epsilon_0, \epsilon_\infty)$ is compact wrt this norm. The Extreme Value Theorem can then be applied to bound the supremum of (B.25).

We first define the norm. Let $\alpha = s - r > 0$. We use the standard norm on the space $\Sigma(s, K)$ [55]:

$$\begin{aligned} ||f|| &= ||f||_{\Sigma(s,K)} \\ &= ||f||_{C^r} + \max_{|\beta|=r} |D^\beta f|_{C^{0,\alpha}} \end{aligned}$$

where

$$\begin{aligned} ||f||_{C^r} &= \max_{|\beta|\leq r} \sup_{x\in\mathcal{S}} |D^\beta f(x)|, \\ |f|_{C^{0,\alpha}} &= \sup_{x\neq y\in\mathcal{S}} \frac{|f(x) - f(y)|}{|x-y|^\alpha}. \end{aligned}$$

Lemma B.6. *The functionals $c_m(p, q)$ are continuous wrt the norm* $\max(||p||_{C^r}, ||q||_{C^r})$*.*

Proof. The functionals $c_m(p, q)$ depend on terms of the form

$$c(p, q) = \int \left(\left. \frac{\partial^{i+j} g(t_1, t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1 = p(x) \\ t_2 = q(x)}} \right) D^\beta p(x) D^\gamma q(x) q(x) dx. \tag{B.26}$$

It is sufficient to show that this is continuous. Let $\epsilon > 0$ and $\max\left(||p - p_0||_{C^r}, ||q - q_0||_{C^r}\right) <$

δ where $\delta > 0$ will be chosen later. Then by applying the triangle inequality for integration and adding and subtracting terms, we have that

$$|c(p,q) - c(p_0,q_0)|$$

$$\begin{aligned}
\leq \;& \int \left| \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1 = p(x) \\ t_2 = q(x)}} \right) D^\beta p(x) D^\gamma q(x) \left(q(x) - q_0(x)\right) \right| dx \\
&+ \int \left| \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1 = p(x) \\ t_2 = q(x)}} \right) D^\beta p(x) q_0(x) \left(D^\gamma q(x) - D^\gamma q_0(x)\right) \right| dx \\
&+ \int \left| D^\beta p_0(x) D^\gamma q_0(x) q_0(x) \left(\left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1 = p(x) \\ t_2 = q(x)}} \right) - \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1 = p_0(x) \\ t_2 = q_0(x)}} \right) \right) \right| \\
&+ \int \left| \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1 = p(x) \\ t_2 = q(x)}} \right) D^\gamma q_0(x) q_0(x) \left(D^\beta p(x) - D^\beta p_0(x)\right) \right| dx. \qquad \text{(B.2}
\end{aligned}$$

By Assumption $\mathcal{A}.4$, the absolute value of the mixed derivatives of g is bounded on the range defined for p and q by some constant $C_{i,j}$. Also, $q_0(x) \leq \epsilon_\infty$. Furthermore, since $D^\gamma q_0$ and $D^\beta p$ are continuous, and since $\mathcal{S} \subset \mathbb{R}^d$ is compact, then the absolute value of the derivatives $D^\gamma q_0$ and $D^\beta p$ is also bounded by a constant $\epsilon_\infty^{'}$. Let $\delta_0 > 0$. Then since the mixed derivatives of g are continuous on the interval $[\epsilon_0, \epsilon_\infty]$, they are uniformly continuous. Therefore, we can choose δ small enough s.t.

$$\left|\left(\left.\frac{\partial^{i+j}g(t_1,t_2)}{\partial t_1^i \partial t_2^j}\right|_{\begin{array}{l} t_1 = p(x) \\ t_2 = q(x)\end{array}}\right) - \left(\left.\frac{\partial^{i+j}g(t_1,t_2)}{\partial t_1^i \partial t_2^j}\right|_{\begin{array}{l} t_1 = p_0(x) \\ t_2 = q_0(x)\end{array}}\right)\right| < \delta_0. \tag{B.28}$$

Combining all of these results with (B.27) gives

$$\begin{aligned} |c(p,q) - c(p_0,q_0)| \;\; &\leq \;\; \lambda(\mathcal{S})\delta C_{ij}\epsilon_{\infty}^{'}\left(2+\epsilon_{\infty}\right) \\ &\quad + \lambda(\mathcal{S})\epsilon_{\infty}^{'}\epsilon_{\infty}\left(2\delta_0 + C_{ij}\delta\right), \end{aligned}$$

where $\lambda(\mathcal{S})$ is the Lebesgue measure of $\mathcal{S}$. This is bounded since $\mathcal{S}$ is compact. Let $\delta_0^{'} > 0$ be s.t. if $\max\left(||p-p_0||_{C^r}, ||q-q_0||_{C^r}\right) < \delta_0^{'}$, then (B.28) is less than $\frac{\epsilon}{4\lambda(\mathcal{S})\epsilon_{\infty}^{'}\epsilon_{\infty}}$. Let $\delta_1 = \frac{\epsilon}{4\lambda(\mathcal{S})C_{ij}\epsilon_{\infty}^{'}(1+\epsilon_{\infty})}$. Then if $\delta < \min(\delta_0^{'}, \delta_1)$, then

$$|c(p,q) - c(p_0,q_0)| < \epsilon.$$

□

Given that each $c_i(p,q)$ is continuous, then $\left(\sum_{i\in J} c_i(p,q)\right)^2$ is also continuous wrt p and q.

We now argue that $\Sigma(s,K)$ is compact. First, a set is relatively compact if its closure is compact. By the Arzela-Ascoli theorem [72], the space $\Sigma(s,K)$ is relatively compact in the topology induced by the $\|\cdot\|_{\Sigma(t,K)}$ norm for any $t < s$. We choose $t = r$. It can then be shown that under the $\|\cdot\|_{\Sigma(r,K)}$ norm, $\Sigma(s,K)$ is complete [55]. Since $\Sigma(s,K)$ is contained in a metric space, then it is also closed and therefore equal to its closure. Thus $\Sigma(s,K)$ is compact. Then since $\Sigma(s,K,\epsilon_0,\epsilon_{\infty})$ is closed in $\Sigma(s,K)$, it is also compact. Therefore, since for each $p,q \in \Sigma(s,K,\epsilon_0,\epsilon_{\infty})$, $\left(\sum_{i\in J} c_i(p,q)\right)^2 < \infty$, by the Extreme Value Theorem we have

$$\begin{aligned}\sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)} \text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right)^2 &= \sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)} \frac{\epsilon^2}{N}\left(\sum_{i\in J} c_i(p,q)\right)^2 \\ &= \frac{\epsilon^2}{N}C,\end{aligned}$$

where we use the fact that J is finite (see Section 2.2.3 for the set J when using ODin1 or ODin2).

APPENDIX C

Proofs for k-nn Plug-in Estimators

This appendix contains the proofs involving the k-nn plug-in estimators of divergence functionals.

C.1 Proof of Theorem III.1 (Bias)

In this section, we prove the bias results in Thm. III.1. The bias of the base k-nn plug-in estimator $\hat{\mathbf{G}}_{k_1,k_2}$ can be expressed as

$$\begin{aligned}\mathbb{B}\left[\hat{\mathbf{G}}_{k_1,k_2}\right] &= \mathbb{E}\left[g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}),\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right)-g\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\right]\\ &= \mathbb{E}\left[g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}),\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right)-g\left(\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}(\mathbf{Z})}\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z},\rho_{2,k_2}(\mathbf{Z})}\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right)\right]\\ &\quad +\mathbb{E}\left[g\left(\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}(\mathbf{Z})}\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z},\rho_{2,k_2}(\mathbf{Z})}\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right)-g\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\right]\end{aligned} \tag{C.1}$$

where $\mathbf{Z}$ is drawn from f_2 and $\rho_{i,k_i}(\mathbf{Z})$ is the k_ith nearest neighbor distance of $\mathbf{Z}$ in the respective samples. For notational simplicity, let $\rho_{i,k_i}(\mathbf{Z}) = \rho_{i,k_i}$. We take a similar approach to the bias proof for the KDE plug-in estimator. In fact, the k-nn density estimator can be viewed as a kernel density estimator. Let K be the uniform kernel on the unit ball. That is,

$$K(x) = \begin{cases} \frac{1}{c_d}, & ||x|| < 1 \\ 0, & \text{otherwise,} \end{cases}$$

where c_d is the volume of the unit ball in $\mathbb{R}^d$. Then we have that

$$\begin{aligned} \hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}) &= \frac{1}{N_1 \rho_{1,k_1}^d} \sum_{i=1}^{N_1} K\left(\frac{\mathbf{Z} - \mathbf{Y}_i}{\rho_{1,k_1}}\right), \\ \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z}) &= \frac{1}{N_2 \rho_{2,k_2}^d} \sum_{i=1}^{N_2} K\left(\frac{\mathbf{Z} - \mathbf{X}_i}{\rho_{2,k_2}}\right). \end{aligned}$$

The fact that the k-nn distances are random requires extra care. However, we can condition on these distances with these representations which enables us to use some of the same tools as in the KDE approach. Define

$$\begin{aligned} S_{k_i}(\mathbf{Z}) &= \left\{X \in \mathbb{R}^d : \|X - \mathbf{Z}\| < \rho_{i,k_i}\right\}, \\ \implies \Pr\left(S_{k_i}(\mathbf{Z})\right) &= \int\limits_{S_{k_i}(\mathbf{Z})} f_i(x)dx. \end{aligned}$$

Note that from [129], we have that

$$\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}} \hat{\mathbf{f}}_{i,k_i}(\mathbf{Z}) = \frac{k_i - 1}{N_i} \frac{1}{\rho_{i,k_i}^d} \frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)} \int\limits_{S_{k_i}(\mathbf{Z})} K\left(\frac{\mathbf{Z} - x}{\rho_{i,k_i}}\right) f_i(x)dx \tag{C.2}$$

The Taylor series expansion of $g\left(\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}} \hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}} \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right)$ around $f_1(\mathbf{Z})$ and $f_2(\mathbf{Z})$ is

$$g\left(\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}} \hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}} \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left(\left. \frac{\partial^{i+j} g(x,y)}{\partial x^i \partial y^j} \right|_{\substack{x=f_1(\mathbf{Z}) \\ y=f_2(\mathbf{Z})}} \right) \frac{\mathbb{B}^i_{\mathbf{Z},\rho_{1,k_1}} \left[\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z})\right] \mathbb{B}^j_{\mathbf{Z},\rho_{2,k_2}} \left[\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z}\right.}{i!j!} \tag{C.3}$$

where $\mathbb{B}^j_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right] = \left(\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z}) - f_i(\mathbf{Z})\right)^j$. We thus require an expression for $\mathbb{B}_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right]$. Since we are conditioning on ρ_{i,k_i}, we can consider separately the cases when $\mathbf{Z}$ is in the interior of the support $\mathcal{S}$ or when $\mathbf{Z}$ is near the boundary of the support. As before, A point $X \in \mathcal{S}$ is defined to be in the interior of$\mathcal{S}$ if for all $Y \notin \mathcal{S}$, $K\left(\frac{X-Y}{h_i}\right) = 0$. A point$X \in \mathcal{S}$ is near the boundary of the support if it is not in the interior. Denote the region in the interior and near the boundary wrt ρ_{i,k_i} as $\mathcal{S}_{I_i}$ and $\mathcal{S}_{B_i}$, respectively. Recall that we assume that $\mathcal{S} = [0,1]^d$, the unit cube.

Consider now $\int_{S_{k_i}(\mathbf{Z})} K\left(\frac{\mathbf{Z}-x}{\rho_{i,k_i}}\right) f_i(x)dx$. Substituting $u = \frac{x-\mathbf{Z}}{\rho_{i,k_i}}$ and then taking a Taylor series expansion of f_i using multi-index notation gives

$$\begin{aligned}\int\limits_{S_{k_i}(\mathbf{Z})} K\left(\frac{x-\mathbf{Z}}{\rho_{i,k_i}}\right) f_i(x)dx &= \rho^d_{i,k_i}\int\limits_{||u||<1} K(u)\, f_i(\mathbf{Z}+u\rho_{i,k_i})du \\ &= \sum_{|\alpha|\leq\lfloor s\rfloor}\frac{D^\alpha f_i(\mathbf{Z})}{\alpha!}\rho^{d+|\alpha|}_{i,k_i}\int\limits_{u:\mathbf{Z}+u\rho_{i,k_i}\in\mathcal{S}} u^\alpha K(u)du, +O\left(\rho^{d+s}_{i,k_i}\right)\end{aligned}$$

$$\implies \mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z}) = \frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\left(\sum_{|\alpha|\leq\lfloor s\rfloor}\frac{D^\alpha f_i(\mathbf{Z})}{\alpha!}\rho^{|\alpha|}_{i,k_i}\int\limits_{u:\mathbf{Z}+u\rho_{i,k_i}\in\mathcal{S}} u^\alpha K(u)du + O\left(\rho^s_{i,k_i}\right)\right). \tag{C.4}$$

Lemma C.1. *Let $\gamma(x,y)$ be an arbitrary function satisfying $\sup_{x,y}|\gamma(x,y)| < \infty$. Let $\mathcal{S} = [0,1]^d$ and let $f_1, f_2 \in \Sigma(s,L)$. Let $\mathbf{Z}$ be a realization of the density f_2 independent of $\hat{\mathbf{f}}_{i,k_i}$ for $i = 1,2$. Then for any integer $\lambda \geq 0$,*

$$\begin{aligned}\mathbb{E}\left[\gamma\left(f_1(\mathbf{Z}), f_2(\mathbf{Z})\right)\mathbb{B}^q_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right]\right] &= \sum_{j=1}^{\lfloor s\rfloor} c_{15,i,j,q}\left(\frac{k_i}{N_i}\right)^{\frac{j}{d}} + \sum_{m=0}^{\lambda}\sum_{\substack{j=0\\ m+j\neq 0}}^{\lfloor s\rfloor}\frac{c_{15,i,q,j,m}}{k_i^{\frac{1+m}{2}}}\left(\frac{k_i}{N_i}\right)^{\frac{j}{d}} \\ &\quad +O\left(\left(\frac{k_i}{N_i}\right)^{\frac{\min(s,d)}{d}} + \frac{1}{k_i^{\frac{2+\lambda}{2}}}\right).\end{aligned}$$

Proof. We use the substitution $\mathbf{T}_i = \Pr\left(S_{k_i}(\mathbf{Z})\right)$ which is the kth order statistic of a uniform random variable [129]. Therefore, $\mathbf{T}_i$ has a beta distribution with parameters

k_i and $N_i - k_i + 1$. This gives

$$\begin{aligned}\mathbb{E}\left[\gamma\left(f_1(\mathbf{Z}), f_2(\mathbf{Z})\right)\mathbb{B}^q_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right]\right] &= (k_i-1)\binom{N_i-1}{k_i-1}\int\limits_{\mathcal{S}}\int\limits_0^1 t^{k-1}(1-t)^{n-k}\mathbb{B}^q_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right] dt f \\ &= (k_i-1)\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k-1}(1-t)^{n-k}\int\limits_{\mathcal{S}}\mathbb{B}^q_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right] f_i(\\ &= (k_i-1)\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k-1}(1-t)^{n-k}\int\limits_{\mathcal{S}_{I_i}}\mathbb{B}^q_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right] f_i(\\ &\quad +(k_i-1)\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k-1}(1-t)^{n-k}\int\limits_{\mathcal{S}_{B_i}}\mathbb{B}^q_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right] f\end{aligned}$$

Note that $\mathbf{T}_i$ monotonically increases with ρ_{i,k_i} and is therefore invertible. Thus ρ_{i,k_i} and $\mathbf{T}_i$ are deterministically related and ρ_{i,k_i} can be viewed as a function of $\mathbf{T}_i$. Thus we can consider separately the cases where $\mathbf{Z}$ is in $\mathcal{S}_{I_i}$ and $\mathcal{S}_{B_i}$ even after making the change of variables.

We first consider $\mathbf{Z} \in \mathcal{S}_{I_i}$. It is clear in this case by (C.4) and the symmetry of $K(u)$ that

$$\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right] = \frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\left(f_i(\mathbf{Z}) + \sum_{j=1}^{\lfloor s/2\rfloor} c_{i,j}(\mathbf{Z})\rho_{i,k_i}^{2j} + O\left(\rho_{i,k_i}^s\right)\right).$$

For $q \geq 2$, we obtain by the binomial theorem,

$$\begin{aligned}\left(\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right)^j &= \left(\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\right)^j\left(f_i^j(\mathbf{Z}) + \sum_{n=1}^{\lfloor s/2\rfloor} c_{i,j,n}(\mathbf{Z})\rho_{i,k_i}^{2n} + O\left(\rho_{i,k_i}^s\right)\right), \\ \mathbb{B}^q_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right] &= \sum_{j=0}^{q}\binom{q}{j}\left(\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right)^j\left(f_i(\mathbf{Z})\right)^{q-j}(-1)^j \\ &= \sum_{j=0}^{q}\binom{q}{j}\left(\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\right)^j(-1)^j\left(f_i^q(\mathbf{Z}) + \sum_{n=1}^{\lfloor s/2\rfloor} c_{i,j,n}(\mathbf{Z})f_i(\mathbf{Z})^{q-j}\rho_{i,k_i}^{2n} + O\right.\end{aligned}$$

By applying concentration inequality arguments [184], it can be shown that with high probability,

$$\left(\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\right)^j = \theta\left(\frac{1}{\left(1+\frac{\sqrt{6}}{\sqrt{k_i}}\right)^j}\right). \tag{C.5}$$

Then applying the binomial theorem in reverse gives (with high probability)

$$\begin{aligned}\sum_{j=0}^{q}\binom{q}{j}\left(\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\right)^j(-1)^j &= \left(1-\frac{1}{1+\frac{\sqrt{6}}{\sqrt{k_i}}}\right)^q \\ &= \left(\frac{6}{k_i}\right)^{\frac{q}{2}}\frac{1}{\left(1+\sqrt{\frac{6}{k_i}}\right)^q} \\ &= \left(\frac{6}{k_i}\right)^{\frac{q}{2}}\sum_{j=0}^{\infty}\binom{-q}{j}(-1)^j\left(\frac{6}{k_i}\right)^{\frac{j}{2}} \\ &= \sum_{j=0}^{\lambda-1}\theta\left(\frac{1}{k_i^{\frac{q+j}{2}}}\right)+O\left(\frac{1}{k_i^{\frac{q+\lambda}{2}}}\right),\end{aligned}$$

where λ is any nonnegative integer. Thus

$$\mathbb{E}\left[\sum_{j=0}^{q}\binom{q}{j}\left(\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\right)^j(-1)^j f_i^q(\mathbf{Z})\right] = \sum_{j=0}^{\lambda-1}c_{3,i,j,q}\frac{1}{k_i^{\frac{q+j}{2}}}+O\left(\frac{1}{k_i^{\frac{q+\lambda}{2}}}\right).$$

For $q=1$, we have

$$(k_i-1)\binom{N_i-1}{k_i-1}\int_0^1 t^{k_i-2}(1-t)^{n-k_i}\int_{\mathcal{S}_{I_i}} f_i(Z)f_2(Z)dzdt - \int_{\mathcal{S}_{I_i}} f_i(Z)f_2(Z)dz = 0.$$

For the terms that include ρ_{i,k_i}^{λ} for some positive integer λ, we have for $\mathbf{Z}\in\mathcal{S}_{I_i}$ that

$$\mathbb{E}\left[\rho_{i,k_i}^{\lambda}\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\right] = (k_i-1)\binom{N_i-1}{k_i-1}\int_0^1 t^{k_i-2}(1-t)^{n-k_i}\int_{\mathcal{S}_{I_i}}\rho_{i,k_i}^{\lambda}f_2(Z)dZdt.$$

We now find an expression for ρ_{i,k_i} in terms of $\mathbf{T}_i$ when $\mathbf{Z}\in\mathcal{S}_{I_i}$. Recall that $\mathbf{T}_i =$

$\Pr\left(S_{k_i}(\mathbf{Z})\right)$. By Taylor series expansion,

$$\begin{aligned}
\mathbf{T}_i &= \int\limits_{S_{k_i}(\mathbf{Z})} f_i(x)dx \\
&= \rho_{i,k_i}^d\left(f_i(\mathbf{Z})c_d + \sum_{j=1}^{\lfloor s/2\rfloor} c_{4,i,j}(\mathbf{Z})\rho_{i,k_i}^{2j} + O\left(\rho_{i,k_i}^s\right)\right) \\
\implies \rho_{i,k_i} &= \frac{\mathbf{T}_i^{\frac{1}{d}}}{\left(f_i(\mathbf{Z})c_d + \sum_{j=1}^{\lfloor s/2\rfloor} c_{4,i,j}(\mathbf{Z})\rho_{i,k_i}^{2j} + O\left(\rho_{i,k_i}^s\right)\right)^{\frac{1}{d}}}. \qquad \text{(C.6)}
\end{aligned}$$

Note that as $\rho_{i,k_i} \downarrow 0$, we have that $\left|\sum_{j=1}^{\lfloor s/2\rfloor} c_{4,i,j}(\mathbf{Z})\rho_{i,k_i}^{2j} + O\left(\rho_{i,k_i}^s\right)\right| < f_i(\mathbf{Z})c_d$ for sufficiently small ρ_{i,k_i} since we assume that $f_i(x) \geq \epsilon_0 > 0$. Therefore, we can apply the generalized binomial theorem to obtain

$$\begin{aligned}
\left(f_i(\mathbf{Z})c_d + \sum_{j=1}^{\lfloor s/2\rfloor} c_{4,i,j}(\mathbf{Z})\rho_{i,k_i}^{2j} + O\left(\rho_{i,k_i}^s\right)\right)^{-\frac{1}{d}} &= \sum_{m=0}^{\infty}\binom{-1/d}{m}\left(f_i(\mathbf{Z})c_d\right)^{-1/d-m} \\
&\quad \times\left(\sum_{j=1}^{\lfloor s/2\rfloor} c_{4,i,j}(\mathbf{Z})\rho_{i,k_i}^{2j} + O\left(\rho_{i,k_i}^s\right)\right)^m \\
&= \left(f_i(\mathbf{Z})c_d\right)^{-1/d} + \sum_{j=1}^{\lfloor s/2\rfloor} c_{5,i,j}(\mathbf{Z})\rho_{i,k_i}^{2j} + O\left(\rho_{i,k_i}^s\right).
\end{aligned}$$

Using this expression in (C.6) and resubstituting the LHS into the RHS gives that

$$\begin{aligned}
\rho_{i,k_i} &= \left(\frac{\mathbf{T}_i}{f_i(\mathbf{Z})c_d}\right)^{\frac{1}{d}} + \sum_{j=1}^{\lfloor s/2\rfloor} c_{6,i,j}(\mathbf{Z})\mathbf{T}_i^{2j/d} + O\left(\mathbf{T}_i^{s/d}\right), \\
\implies \rho_{i,k_i}^{\lambda} &= \left(\frac{\mathbf{T}_i}{f_i(\mathbf{Z})c_d}\right)^{\frac{\lambda}{d}} + \sum_{j=1}^{\lfloor s/2\rfloor} c_{7,i,j}(\mathbf{Z})\mathbf{T}_i^{2j\lambda/d} + O\left(\mathbf{T}_i^{s\lambda/d}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\rho_{i,k_i}^{\lambda}\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\right] &= (k_i-1)\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k_i-2+\lambda/d}(1-t)^{n-k_i}\int\limits_{\mathcal{S}_{I_i}}\frac{f_2(Z)}{\left(f_i(Z)c_d\right)^{\lambda/d}}dZdt \\ &\quad+\sum_{j=1}^{\lfloor s/2\rfloor}(k_i-1)\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k_i-2+2j\lambda/d}(1-t)^{n-k_i}\int\limits_{\mathcal{S}_{I_i}} f_2(Z)c_{7,i,j}(Z)dZd \\ &= c_{7,i,0}\left(\frac{k_i}{N_i}\right)^{\lambda/d}+\sum_{j=1}^{\lfloor s/2\rfloor}c_{7,i,j}\left(\frac{k_i}{N_i}\right)^{2\lambda j/d}+O\left(\left(\frac{k_i}{N_i}\right)^{\frac{s}{d}}\right).\end{aligned}$$

Combining this result with (C.5) gives for $q\geq 2$ and any integer $\lambda\geq 0$

$$N_i\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k_i-2}(1-t)^{N_i-k_i}\int\limits_{\mathcal{S}_{I_i}}\mathbb{B}^q_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right]f_i(Z)\gamma\left(f_1(Z),f_2(Z)\right)dZdt$$

$$=\sum_{j=0}^{\lambda-1}c_{3,i,j,q}\frac{1}{k^{\frac{q+j}{2}}}+O\left(\frac{1}{k^{\frac{q+\lambda}{2}}}+\left(\frac{k_i}{N_i}\right)^{\frac{s}{d}}\right)+\sum_{m=0}^{\lambda-1}\sum_{j=1}^{\lfloor s/2\rfloor}c_{7,i,j,m,q}\left(\frac{k_i}{N_i}\right)^{2j/d}\frac{1}{k_i^{\frac{q-1+m}{2}}}.$$

Similarly, for $q=1$,

$$N_i\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k_i-2}(1-t)^{N_i-k_i}\int\limits_{\mathcal{S}_{I_i}}\mathbb{B}_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right]f_i(Z)\gamma\left(f_1(Z),f_2(Z)\right)dZdt$$

$$=\sum_{j=1}^{\lfloor s/2\rfloor}c_{7,i,j,m,1}\left(\frac{k_i}{N_i}\right)^{2j/d}+O\left(\left(\frac{k_i}{N_i}\right)^{\frac{s}{d}}\right).$$

We now consider the case where $\mathbf{Z}\in\mathcal{S}_{B_i}$. We take a similar approach as in the

KDE case where we extend the density beyond the boundary. This gives

$$\begin{aligned}
\mathbb{B}_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})\right] &= \frac{k_i-1}{N_i}\frac{1}{\rho_{i,k_i}^d}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\int\limits_{S_{k_i}(\mathbf{Z})\cap\mathcal{S}} K\left(\frac{\mathbf{Z}-x}{\rho_{i,k_i}}\right)f_i(x)dx \\
&= \frac{k_i-1}{N_i}\frac{1}{\rho_{i,k_i}^d}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\int\limits_{S_{k_i}(\mathbf{Z})} K\left(\frac{\mathbf{Z}-x}{\rho_{i,k_i}}\right)f_i(x)dx - f_i(\mathbf{Z}) \\
&\quad -\frac{k_i-1}{N_i}\frac{1}{\rho_{i,k_i}^d}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\int\limits_{x\notin\mathcal{S}} K\left(\frac{\mathbf{Z}-x}{\rho_{i,k_i}}\right)f_i(x)dx \\
&= T_1(\mathbf{Z},\rho_{i,k_i}) - T_2(\mathbf{Z},\rho_{i,k_i}).
\end{aligned}$$

The expression for $T_1(\mathbf{Z},\rho_{i,k_i})$ is identical to that when $\mathbf{Z}\in\mathcal{S}_{I_i}$ and so taking the expectation gives the same results. Therefore, we focus on $T_2(\mathbf{Z},\rho_{i,k_i})$. As before, we substitute $u=(\mathbf{Z}-x)/\rho_{i,k_i}$ inside the integral and take a Taylor series expansion of f_i to get

$$\sum_{|\alpha|\leq\lfloor s\rfloor}\frac{D^\alpha f_i(\mathbf{Z})}{\alpha!}\rho_{i,k_i}^{d+|\alpha|}\int\limits_{u:\mathbf{Z}+u\rho_{i,k_i}\notin\mathcal{S}} u^\alpha K(u)du + O\left(\rho_{i,k_i}^{d+s}\right).$$

As before, we can again substitute $\mathbf{T}_i=\Pr\left(S_{k_i}(\mathbf{Z})\right)$. However, we need to find an expression for ρ_{i,k_i} in terms of $\mathbf{T}_i$ for $\mathbf{Z}\in\mathcal{S}_{B_i}$. Note that

$$\begin{aligned}
\mathbf{T}_i &= \int\limits_{z\in S_{k_i}(\mathbf{Z})\cap\mathcal{S}} f_i(z)dz. \\
&= \int\limits_{z\in S_{k_i}(\mathbf{Z})} f(z)dz - \int\limits_{z\in S_{k_i}(\mathbf{Z})\cap\mathcal{S}^C} f(z)dz \\
&= \rho_{i,k_i}^d\left(f_i(\mathbf{Z})c_d + \sum_{j=1}^{\lfloor s/2\rfloor} c_{4,i,j}(\mathbf{Z}){\rho_{i,k_i}}^{2j} + O(\rho_{i,k_i}^s)\right) \\
&\quad - \int\limits_{z\in S_{k_i}(\mathbf{Z})\cap\mathcal{S}^C}\left(\sum_{|\alpha|\leq\lfloor s\rfloor}\frac{(z-\mathbf{Z})^\alpha}{\alpha!}D^\alpha f(\mathbf{Z}) + O\left((z-\mathbf{Z})^s\right)\right)dz. \qquad \text{(C.7)}
\end{aligned}$$

We need to simplify the second integral in (C.7) before solving for ρ_{i,k_i}. If we assume that the support $\mathcal{S}=[0,1]^d$, then we can use the techniques used in Appendix ????

to show that the unit cube satisfies the boundary conditions for the KDE plug-in estimators.

Assume that d is odd as as the case for even d will be similar. We first consider the case where only a single coordinate $\mathbf{Z}_{(1)}$ is close to the boundary. Without loss of generality, we assume that $\mathbf{Z}_{(1)}$ is close to 1. Then for a given α, we can use (???) (A.11) to obtain

$$\int\limits_{z\in S_{k_i}(\mathbf{Z})\cap\mathcal{S}^C} \frac{(z-\mathbf{Z})^\alpha}{\alpha!} D^\alpha f_i(\mathbf{Z})dz = \rho_{i,k_i}^{d+|\alpha|}\sum_{m=0}^{|\alpha|}\sum_{\ell=0}^{d+|\alpha|} p_{m,\ell,\alpha,i}(\mathbf{Z})\left(\sqrt{1-\left(\frac{1-\mathbf{Z}_{(1)}}{\rho_{i,k_i}}\right)^2}\right)^\ell\left(\frac{1-\mathbf{Z}_{(1)}}{\rho_{i,k_i}}\right)^m, \tag{C.8}$$

where $p_{m,\ell,\alpha,i}(\mathbf{Z})$ is $\lfloor s\rfloor - |\alpha|$ times differentiable wrt $\mathbf{Z}$. Now expand $p_{m,\ell,\alpha,i}(\mathbf{Z})$ only in the $\mathbf{Z}_{(1)}$ coordinate at $\mathbf{Z}_{(1)} = 1$ to get

$$p_{m,\ell,\alpha,i}(\mathbf{Z}) = \sum_{j=0}^{\lfloor s\rfloor-|\alpha|} \frac{\partial^j p_{m,\ell,\alpha,i}(1,\mathbf{Z}_{(2)},\dots,\mathbf{Z}_{(d)})}{\partial \mathbf{Z}_{(1)}^j}\frac{\left(1-\mathbf{Z}_{(1)}\right)^n}{j!}.$$

Substituting this into (C.8) and substituting $\mathbf{W} = \frac{1-\mathbf{Z}_{(1)}}{\rho_{i,k_i}}$ gives

$$\sum_{m=0}^{|\alpha|}\sum_{\ell=0}^{d+|\alpha|}\sum_{j=0}^{\lfloor s\rfloor-|\alpha|} \frac{\partial^j p_{m,\ell,\alpha,i}(1,\mathbf{Z}_{(2)},\dots,\mathbf{Z}_{(d)})}{\partial \mathbf{Z}_{(1)}^j}\frac{1}{j!}\left(\sqrt{1-\left(\frac{1-\mathbf{Z}_{(1)}}{\rho_{i,k_i}}\right)^2}\right)^\ell\left(\frac{1-\mathbf{Z}_{(1)}}{\rho_{i,k_i}}\right)^{m+j}\rho_{i,k_i}^{j+d+|\alpha|}$$

$$=\sum_{m=0}^{|\alpha|}\sum_{\ell=0}^{d+|\alpha|}\sum_{j=0}^{\lfloor s\rfloor-|\alpha|} p^{'}_{m,\ell,\alpha,i}(\mathbf{Z}^{'})\left(\sqrt{1-\mathbf{W}^2}\right)^\ell \mathbf{W}^{m+j}\rho_{i,k_i}^{j+d+|\alpha|},$$

where $\mathbf{Z}^{'} = (1,\mathbf{Z}_{(2)},\dots,\mathbf{Z}_{(d)})$ and $p^{'}_{m,\ell,\alpha,i}(\mathbf{Z}^{'}) = \frac{\partial^j p_{m,\ell,\alpha,i}(1,\mathbf{Z}_{(2)},\dots,\mathbf{Z}_{(d)})}{\partial \mathbf{Z}_{(1)}^j}\frac{1}{j!}$. The variable $\mathbf{W}$ ranges from 0 to 1. Thus we have separated the dependence on ρ_{i,k_i}. Substituting these results into (C.7) gives

$$\begin{aligned}\mathbf{T}_i &= \rho_{i,k_i}^d \left(f_i(\mathbf{Z})c_d + \sum_{j=1}^{\lfloor s/2 \rfloor} c_{4,i,j}(\mathbf{Z}) {\rho_{i,k_i}}^{2j} + O(\rho_{i,k_i}^s) \right) \\ &\quad -\rho_{i,k_i}^d \sum_{|\alpha|\leq\lfloor s\rfloor} \sum_{m=0}^{|\alpha|} \sum_{\ell=0}^{d+|\alpha|} \sum_{j=0}^{\lfloor s\rfloor-|\alpha|} p^{'}_{m,\ell,\alpha,i}(\mathbf{Z}^{'}) \left(\sqrt{1-\mathbf{W}^2}\right)^{\ell} \mathbf{W}^{m+j} \rho_{i,k_i}^{j+|\alpha|}.\end{aligned}$$

By substituting $\mathbf{Z}_{(1)} = 1 - \mathbf{W}\rho_{i,k_i}$ in the first term and taking a Taylor series expansion of $f_i(\mathbf{Z})^{-1/d}$ and $c_{4,i,j}(\mathbf{Z})$ at $\mathbf{Z}_{(1)} = 1$ gives

$$\sum_{j=0}^{\lfloor s\rfloor} c_{8,i,j}(\mathbf{Z}^{''}) \rho_{i,k_i}^{j+d} + O\left(\rho_{i,k_i}^{s+d}\right),$$

where $\mathbf{Z}^{''} = (\mathbf{W}, \mathbf{Z}_{(2)}, \ldots, \mathbf{Z}_{(d)})$. Thus we can write

$$\begin{aligned}\mathbf{T}_i &= \rho_{i,k_i}^d \left(\sum_{j=0}^{\lfloor s\rfloor} c_{9,i,j}(\mathbf{Z}^{''}) \rho_{i,k_i}^j + O\left(\rho_{i,k_i}^s\right) \right) \\ \Longrightarrow \rho_{i,k_i} &= \frac{t^{\frac{1}{d}}}{\left(\sum_{j=0}^{\lfloor s\rfloor} c_{9,i,j}(\mathbf{Z}^{''}) \rho_{i,k_i}^j + O\left(\rho_{i,k_i}^s\right) \right)^{\frac{1}{d}}}. \qquad \text{(C.9)}\end{aligned}$$

Then since $\rho_{i,k_i} \downarrow 0$, applying the generalized binomial theorem to the denominator gives

$$\begin{aligned}\left(\sum_{j=0}^{\lfloor s\rfloor} c_{9,i,j}(\mathbf{Z}^{''}) \rho_{i,k_i}^j + O\left(\rho_{i,k_i}^s\right) \right)^{-\frac{1}{d}} &= \sum_{m=0}^{\infty} \binom{-1/d}{m} c_{9,i,0}(\mathbf{Z}^{''})^{-\frac{1}{d}-j} \left(\sum_{j=1}^{\lfloor s\rfloor} c_{9,i,j}(\mathbf{Z}^{''}) \rho_{i,k_i}^j + O\left(\rho_{i,k_i}^s\right) \right)^m \\ &= c_{9,i,0}(\mathbf{Z}^{''})^{-\frac{1}{d}} + \sum_{j=1}^{\lfloor s\rfloor} c_{10,i,j}(\mathbf{Z}^{''}) \rho_{i,k_i}^j + O\left(\rho_{i,k_i}^s\right).\end{aligned}$$

Applying this result to (C.9) gives

$$\rho_{i,k_i} = \left(\frac{\mathbf{T}_i}{c_{9,i,0}(\mathbf{Z}^{''})} \right)^{\frac{1}{d}} + \mathbf{T}_i^{\frac{1}{d}} \sum_{j=1}^{\lfloor s\rfloor} c_{10,i,j}(\mathbf{Z}^{''}) \rho_{i,k_i}^j + O\left(\mathbf{T}_i^{\frac{1}{d}} \rho_{i,k_i}^s \right). \qquad \text{(C.10)}$$

Resubstituting the LHS of (C.10) into the RHS multiple times then gives

$$\begin{aligned} \rho_{i,k_i} &= \sum_{j=1}^{\lfloor s \rfloor} c_{11,i,j}(\mathbf{Z}'')\mathbf{T}_i^{\frac{j}{d}} + O\left(\mathbf{T}_i^{\frac{s}{d}}\right) \\ \implies \rho_{i,k_i}^{\lambda} &= \sum_{j=1}^{\lfloor s \rfloor} c_{12,i,j,\lambda}(\mathbf{Z}'')\mathbf{T}_i^{\frac{j\lambda}{d}} + O\left(\mathbf{T}_i^{\frac{s\lambda}{d}}\right). \end{aligned}$$

Given these results and the fact that $\mathbf{T}_i$ has a beta distribution, we have that

$$\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}_{B_i}\}}\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\left[\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\rho_{i,k_i}^{\lambda}\right]\right] = \frac{k_i-1}{N_i}\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k-2}(1-t)^{n-k}\int\limits_{\mathcal{S}}\rho_{i,k_i}^{\lambda}f_i(Z)dZd$$

Taking a Taylor series expansion of f_i at $Z_{(1)} = 1$ gives

$$f_i(Z) = \sum_{j=0}^{\lfloor s \rfloor}\frac{\partial^j f_i(Z')}{\partial Z_{(1)}^j}W^j\rho_{i,k_i}^j + O\left(\rho_{i,k_i}^s\right).$$

Combining all of these results gives that $\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}_{B_i}\}}\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\left[\frac{k_i-1}{N_i}\frac{1}{\Pr\left(S_{k_i}(\mathbf{Z})\right)}\rho_{i,k_i}^{\lambda}\right]\right]$ has terms of the form of

$$(k_i-1)\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k-2+\frac{\lambda+1}{d}}(1-t)^{n-k}dt = \left(\frac{k_i}{N_i}\right)^{\frac{\lambda+1}{d}} + o\left(\frac{k_i}{N_i}\right).$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[T_2(\mathbf{Z},\rho_{i,k_i})\right] &= (k_i-1)\binom{N_i-1}{k_i-1}\int\limits_0^1 t^{k_i-2}(1-t)^{N_i-k_i}\int\limits_{S_{B_i}}\left(\sum_{j=0}^{\lfloor s \rfloor}c_{13,i,j}(Z'')t^{\frac{j+1}{d}} + O\left(t^{\frac{s}{d}}\right)\right)dZ''dt \\ &= \sum_{j=1}^{\lfloor s \rfloor}c_{14,i,j}\left(\frac{k_i}{N_i}\right)^{\frac{j}{d}} + O\left(\left(\frac{k_i}{N_i}\right)^{\min(s,d)/d}\right). \end{aligned} \tag{C.11}$$

For $\mathbb{E}\left[(T_1(\mathbf{Z},\rho_{i,k_i}) - T_2(\mathbf{Z},\rho_{i,k_i}))^q\right]$, we have by the binomial theorem that

$$(T_1(\mathbf{Z},\rho_{i,k_i}) - T_2(\mathbf{Z},\rho_{i,k_i}))^q = \sum_{j=0}^{q} \binom{q}{j} T_1(\mathbf{Z},\rho_{i,k_i})^j T_2(\mathbf{Z},\rho_{i,k_i})^{q-j}.$$

Applying a similar analysis gives similar results.

For the case when $S_{k_i}(\mathbf{Z})$ intersects multiple boundary points, a similar approach can be used to prove the boundary conditions for the KDE plug-in estimator in Appendix ????. This will yield a similar expression to (C.11). Combining all results with the fact that $\gamma(x,y)$ is bounded finishes the proof. □

Lemma C.2. *Let $\gamma(x,y)$ be an arbitrary function satisfying $\sup_{x,y}|\gamma(x,y)| < \infty$. Let $\mathbf{Z}$ be a realization of the density f_2 independent of $\hat{\mathbf{f}}_{i,k_i}$ for $i = 1,2$. Then for any integer $\lambda \geq 0$*

$$\begin{aligned}\mathbb{E}\left[\gamma(f_1(\mathbf{Z}), f_2(\mathbf{Z}))\mathbb{B}^t_{\mathbf{Z},\rho_{1,k_1}}\left[\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z})\right]\mathbb{B}^q_{\mathbf{Z},\rho_{2,k_2}}\left[\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right]\right] \quad &= \quad \sum_{j=0}^{\lfloor s\rfloor}\sum_{\substack{i=0\\ i+j\neq 0}}^{\lfloor s\rfloor} c_{16,i,j,q,t}\left(\frac{k_1}{N_1}\right)^{\frac{i}{d}}\left(\frac{k_2}{N_2}\right)^{\frac{j}{d}} + O\Bigg(\\ &+\sum_{m=0}^{\lambda}\sum_{\substack{j=0\\ m+j\neq 0}}^{\lfloor s\rfloor}\sum_{n=0}^{\lambda}\sum_{\substack{i=0\\ n+i\neq 0}}^{\lfloor s\rfloor}\frac{c_{16,i,j,q,t,m,n}}{k_1^{\frac{1+m}{2}}k_2^{\frac{1+n}{2}}}\left(\frac{k_1}{N_1}\right)\end{aligned}$$

Proof. Note that ρ_{1,k_1} and ρ_{2,k_2} are conditionally independent of each other given $\mathbf{Z}$. Applying similar techniques as in the proof of Lemma C.1 yields the result. □

Applying Lemmas C.1 and C.2 to (C.3) gives

$$\mathbb{E}\left[g\left(\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}(\mathbf{Z})}\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}(\mathbf{Z})}\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right) - g\left(f_1(\mathbf{Z}), f_2(\mathbf{Z})\right)\right]$$

$$= \sum_{j=0}^{\lfloor s \rfloor} \sum_{\substack{i=0 \\ i+j\neq 0}}^{\lfloor s \rfloor} c_{18,i,j} \left(\frac{k_1}{N_1}\right)^{\frac{i}{d}} \left(\frac{k_2}{N_2}\right)^{\frac{j}{d}} + O\left(\max\left(\frac{k_1}{N_1}, \frac{k_2}{N_2}\right)^{\frac{\min(s,d)}{d}} + \frac{1}{\min(k_1,k_2)^{\frac{2+\lambda}{2}}}\right)$$

$$+ \sum_{m=0}^{\lambda} \sum_{\substack{j=0 \\ j+m\neq 0}}^{r} \left(\frac{c_{17,1,j,m}}{k_1^{\frac{1+m}{2}}} \left(\frac{k_1}{N_1}\right)^{\frac{j}{d}} + \frac{c_{17,2,j,m}}{k_2^{\frac{1+m}{2}}} \left(\frac{k_2}{N_2}\right)^{\frac{j}{d}}\right) + \sum_{j=1}^{r} \left(c_{17,1,j} \left(\frac{k_1}{N_1}\right)^{\frac{j}{d}} + c_{17,2,j} \left(\frac{k_2}{N_2}\right)^{\frac{j}{d}}\right)$$

$$+ \sum_{m=0}^{\lambda} \sum_{\substack{j=0 \\ m+j\neq 0}}^{\lfloor s \rfloor} \sum_{n=0}^{\lambda} \sum_{\substack{i=0 \\ n+i\neq 0}}^{\lfloor s \rfloor} \frac{c_{18,i,j,m,n}}{k_1^{\frac{1+m}{2}} k_2^{\frac{1+n}{2}}} \left(\frac{k_1}{N_1}\right)^{\frac{i}{d}} \left(\frac{k_2}{N_2}\right)^{\frac{j}{d}}. \tag{C.12}$$

We now focus on the first term in (D.1). The truncated Taylor series expansion of $g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}), \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right)$ around $\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}} \hat{\mathbf{f}}_{1,k_1}(\mathbf{Z})$ and $\mathbb{E}_{\mathbf{Z},\rho_{2,k_2}} \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})$ gives

$$g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}), \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right) = \sum_{i=0}^{\nu} \sum_{j=0}^{\nu} \left(\left.\frac{\partial^{i+j} g(x,y)}{\partial x^i \partial y^j}\right|_{\substack{x=\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}} \hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}) \\ y=\mathbb{E}_{\mathbf{Z},\rho_{2,k_2}} \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})}} \right) \frac{\hat{\mathbf{e}}_{1,k_1}^{i}(\mathbf{Z}) \hat{\mathbf{e}}_{2,k_2}^{j}(\mathbf{Z})}{i!j!} + o\left(\hat{\mathbf{e}}_{1,k_1}^{\nu}(\mathbf{Z}) + \hat{\mathbf{e}}_{2,k_2}^{\nu}(\mathbf{Z}\right. \tag{C.13}$$

where $\hat{\mathbf{e}}_{i,k_i} := \hat{\mathbf{f}}_{i,k_i}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z},\rho_{i,k_i}} \hat{\mathbf{f}}_{i,k_i}(\mathbf{Z})$. We thus require expressions for $\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{e}}_{i,k_i}^{j}(\mathbf{Z})\right]$ to control this expression.

Lemma C.3. *Let* $\mathbf{Z}$ *be a realization of the density* f_2 *that is in the interior of the support wrt* ρ_{i,k_i} *and is independent of* $\hat{\mathbf{f}}_{i,k_i}$ *for* $i = 1, 2$*. Let* $n(q)$ *be the set of integer divisors of* q *including 1 but excluding* q*. Then,*

$$\mathbb{E}_{\mathbf{Z},\rho_{i,k_i}}\left[\hat{\mathbf{e}}_{i,k_i}^{q}(\mathbf{Z})\right] = \begin{cases} \frac{k_i-1}{N_i \Pr\left(S_{k_i}(\mathbf{Z})\right)} \sum_{j\in n(q)} \frac{1}{\left(N_i \rho_{i,k_i}^d\right)^{q-j}} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{,i,q,j,m}(\mathbf{Z}) \rho_{i,k_i}^{2m}, & q \geq 2 \\ 0, & q = 1 \end{cases}$$

$$\mathbb{E}_{\mathbf{Z},\rho_{1,k_1},\rho_{2,k_2}}\left[\hat{\mathbf{e}}_{1,k_1}^{q}(\mathbf{Z}) \hat{\mathbf{e}}_{2,k_2}^{l}(\mathbf{Z})\right] = \begin{cases} \frac{k_i-1}{N_i \Pr\left(S_{k_i}(\mathbf{Z})\right)} \left(\sum_{j\in n(q)} \frac{1}{\left(N_1 \rho_{1,k_1}^d\right)^{q-j}} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{,1,q,j,m}(\mathbf{Z}) \rho_{1,k_1}^{2m}\right) \times & q, l \geq \\ \left(\sum_{i\in n(l)} \frac{1}{\left(N_2 \rho_{2,k_2}^d\right)^{l-i}} \sum_{t=0}^{\lfloor s/2 \rfloor} c_{,2,l,i,t}(\mathbf{Z}) \rho_{2,k_2}^{2t}\right) + O\left(\frac{1}{N_1} + \frac{1}{N_2}\right), & \\ 0, & q = \end{cases}$$

Proof. The proof is very similar to the analagous statement on the KDE results in Lemma ????. Define the random variable $\mathbf{V}_i(\mathbf{Z}) = K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{\rho_{2,k_2}}\right) - \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}} K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{\rho_{2,k_2}}\right)$. Then

$$\begin{aligned}
\hat{\mathbf{e}}_{2,k_2}(\mathbf{Z}) &= \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}} \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z}) \\
&= \frac{1}{N_2 \rho_{2,k_2}^d} \sum_{i=1}^{N_2} \mathbf{V}_i(\mathbf{Z}).
\end{aligned}$$

As before, $\mathbb{E}_{\mathbf{Z},\rho_{2,k_2}} \mathbf{V}_i(\mathbf{Z}) = 0$. From previous results, we have for $j \geq 1$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z},\rho_{2,k_2}}\left[K^j\left(\frac{\mathbf{X}_i-\mathbf{Z}}{\rho_{2,k_2}}\right)\right] &= \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}}\left[K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{\rho_{2,k_2}}\right)\right] \\
&= \frac{k_2-1}{N_2} \frac{\rho_{2,k_2}^d}{\Pr\left(S_{k_2}(\mathbf{Z})\right)} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{2,m}(\mathbf{Z}) \rho_{2,k_2}^{2m} + O\left(\rho_{2,k_2}^s\right).
\end{aligned}$$

By the binomial theorem,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z},\rho_{2,k_2}}\left[\mathbf{V}_i^j(\mathbf{Z})\right] &= \sum_{n=0}^{j} \binom{j}{n} \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}}\left[K^j\left(\frac{\mathbf{X}_i-\mathbf{Z}}{\rho_{2,k_2}}\right)\right] \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}}\left[K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{\rho_{2,k_2}}\right)\right]^{j-n} \\
&= \sum_{n=0}^{j} \binom{j}{n} \left(\frac{k_2-1}{N_2} \frac{\rho_{2,k_2}^d}{\Pr\left(S_{k_2}(\mathbf{Z})\right)} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{2,m}(\mathbf{Z}) \rho_{2,k_2}^{2m}\right) O\left(\left(\frac{\rho_{2,k_2}^d}{\Pr\left(S_{k_2}(\mathbf{Z})\right)} \frac{k_2-1}{N_2}\right)^{j-n}\right. \\
&= \frac{k_2-1}{N_2} \frac{\rho_{2,k_2}^d}{\Pr\left(S_{k_2}(\mathbf{Z})\right)} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{2,m}(\mathbf{Z}) \rho_{2,k_2}^{2m} + O\left(\left(\frac{\rho_{2,k_2}^d}{\Pr\left(S_{k_2}(\mathbf{Z})\right)} \frac{k_2-1}{N_2}\right)^2\right).
\end{aligned}$$

Given these results and the fact that with high probability

$$\left(\frac{1}{\Pr\left(S_{k_2}(\mathbf{Z})\right)} \frac{k_2-1}{N_2}\right)^2 = O\left(\frac{1}{k}\right),$$

a similar procedure as in the proof of Lemma ???? gives the result. □

For general g, we can only say that

$$\left.\frac{\partial^{i+j} g(x,y)}{\partial x^i \partial y^j}\right|_{\substack{x=\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}}\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}) \\ y=\mathbb{E}_{\mathbf{Z},\rho_{2,k_2}}\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})}} = O(1).$$

By applying similar techniques as in the proofs of Lemmas C.1 and C.2, it can then be shown with the application of Lemma C.3 that the expected value of (C.13) reduces to

$$\mathbb{E}\left[g\left(\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}(\mathbf{Z})}\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}(\mathbf{Z})}\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right)\right] + O\left(\frac{1}{k_1}+\frac{1}{k_2}\right). \tag{C.14}$$

If $g(x,y)$ has mixed derivatives of the form of $x^\alpha y^\beta$ for $\alpha, \beta \in \mathbb{R}$, we can apply the generalized binomial theorem prior to taking the expectation as for the KDE plug-in estimator to show that

$$\begin{aligned}
&\mathbb{E}\left[g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}), \hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right) - g\left(\mathbb{E}_{\mathbf{Z},\rho_{1,k_1}(\mathbf{Z})}\hat{\mathbf{f}}_{1,k_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z},\rho_{2,k_2}(\mathbf{Z})}\hat{\mathbf{f}}_{2,k_2}(\mathbf{Z})\right)\right] \\
= &\sum_{j=1}^{\nu/2}\sum_{m=0}^{r}\sum_{i=1}^{\nu/2}\sum_{n=0}^{r}\frac{c_{19,j,i,m,n}}{k_1^j k_2^i}\left(\frac{k_1}{N_1}\right)^{\frac{m}{d}}\left(\frac{k_2}{N_2}\right)^{\frac{n}{d}} + O\left(\frac{1}{k_1^{\nu/2}}+\frac{1}{k_2^{\nu/2}}+\left(\frac{k_1}{N_1}\right)^{\frac{s}{d}}+\left(\frac{k_2}{N_2}\right)^{\frac{s}{d}}\right) \\
&+\sum_{j=1}^{\nu/2}\sum_{m=0}^{r}\left(\frac{c_{19,1,j,m}}{k_1^j}\left(\frac{k_1}{N_1}\right)^{\frac{m}{d}}+\frac{c_{19,2,j,m}}{k_2^j}\left(\frac{k_2}{N_2}\right)^{\frac{m}{d}}\right).
\end{aligned} \tag{C.15}$$

Combining (C.3) with either (C.14) or (C.15) completes the proof.

C.2 Proof of Theorem III.2 (Variance)

To bound the variance of the plug-in estimator $\hat{\mathbf{G}}_{k_1,k_2}$, we will again use the Efron-Stein inequality [53] (Lemma B.4). Suppose we have samples $\{\mathbf{X}_1,\ldots,\mathbf{X}_{N_2},\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_1}\}$ and $\left\{\mathbf{X}_1',\ldots,\mathbf{X}_{N_2},\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_1}\right\}$ and denote the respective estimators as $\hat{\mathbf{G}}_{k_1,k_2}$ and

$\hat{\mathbf{G}}'_{k_1,k_2}$. We have that

$$\begin{aligned}\left|\hat{\mathbf{G}}_{k_1,k_2}-\hat{\mathbf{G}}'_{k_1,k_2}\right| &\leq \frac{1}{N_2}\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_1),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_1)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}'_1),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}'_1)\right)\right| \\ &\quad +\frac{1}{N_2}\sum_{j=2}^{N_2}\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_j)\right)\right| \qquad (C.16)\end{aligned}$$

Define $\mathbf{P}_{k_i}(\mathbf{X}_j)=\Pr\left(S_{k_i}(\mathbf{X}_j)\right)$. This is a random variable denoting the probability that a point drawn from f_i falls into the k_ith nearest neighbor ball of $\mathbf{X}_j$. As mentioned in Appendix C.1, the distribution of $\mathbf{P}_{k_i}(\mathbf{X}_j)$ is independent of $\mathbf{X}_j$ and f_i and is a beta random variable [62] with density

$$f_{k_i}\left(p_{k_i}\right)=\frac{M_i!}{(k_i-1)!(M_i-k_i)!}p_{k_i}^{k_i-1}\left(1-p_{k_i}\right)^{M_i-k_i}.$$

Define

$$\bar{\mathbf{f}}_{i,k_i}(\mathbf{X}_j)=f_i(\mathbf{X}_j)\frac{k_i-1}{M_i\mathbf{P}_{k_i}(\mathbf{X}_j)}.$$

We then have that with high probability [184],

$$\hat{\mathbf{f}}_{i,k_i}(\mathbf{X}_j)=\bar{\mathbf{f}}_{i,k_i}(\mathbf{X}_j)+O\left(\left(\frac{k_i}{M_i}\right)^{\frac{2}{d}}\right). \qquad (C.17)$$

The following lemma can be used to control the first term in (D.5)

Lemma C.4.

$$\mathbb{E}\left[\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_1),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_1)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}'_1),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}'_1)\right)\right|^2\right]=O(1).$$

Proof. Since g is Lipschitz continuous with constant C_g, we have

$$\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_1),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_1)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}'_1),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}'_1)\right)\right| \leq C_g\left(\left|\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_1)-\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}'_1)\right|+\left|\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_1)-\hat{\mathbf{f}}_{2,\right.\right.$$

From the triangle inequality, Jensen's inequality, and (C.17), we get

$$\begin{aligned}\mathbb{E}\left[\left|\tilde{\mathbf{f}}_{i,h_i}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{i,h_i}(\mathbf{X}_1')\right|^2\right] &\leq 2\mathbb{E}\left[\left(\hat{\mathbf{f}}_{i,k_i}(\mathbf{X}_1)\right)^2\right] \\ &\leq 4\mathbb{E}\left[\left(\bar{\mathbf{f}}_{i,k_i}(\mathbf{X}_1)\right)^2\right]+O\left(\left(\frac{k_i}{M_i}\right)^{\frac{4}{d}}\right) \\ &= 4\mathbb{E}\left[f_i^2(\mathbf{X}_1)\right]\frac{(k_i-1)^2}{M_i^2}\cdot\frac{M_i(M_i-1)}{(k_i-1)(k_i-2)}+O\left(\left(\frac{k_i}{M_i}\right)^{\frac{4}{d}}\right) \qquad (C.19)\end{aligned}$$

Combining (C.19) with (C.18) after applying Jensen's inequality gives the result. □

To control the second term in (D.5), consider the following events:

- $A_1(\mathbf{X}_i)$: $\mathbf{X}_1$ is strictly within the k_2-nn ball around $\mathbf{X}_i$ wrt the sample $\{\mathbf{X}_1,\dots,\mathbf{X}_{N_2}\}\setminus\{\mathbf{X}_i\}$

  .

- $A_2(\mathbf{X}_i)$: $\mathbf{X}_1$ is the k_2th nearest neighbor of $\mathbf{X}_i$ wrt the sample $\{\mathbf{X}_1,\dots,\mathbf{X}_{N_2}\}\setminus\{\mathbf{X}_i\}$

  .

- $A_3(\mathbf{X}_i)$: $\mathbf{X}_1$ is strictly outside of the k_2-nn ball around $\mathbf{X}_i$ wrt the sample $\{\mathbf{X}_1,\dots,\mathbf{X}_{N_2}\}\setminus\{\mathbf{X}_i\}$.

- $B_1(\mathbf{X}_i)$: $\mathbf{X}_1'$ is strictly within the k_2-nn ball around $\mathbf{X}_i$ wrt the sample $\left\{\mathbf{X}_1',\mathbf{X}_2,\dots,\mathbf{X}_{N_2}\right\}\setminus\{\mathbf{X}_i$

  .

- $B_2(\mathbf{X}_i)$: $\mathbf{X}_1'$ is the k_2th nearest neighbor of $\mathbf{X}_i$ wrt the sample $\left\{\mathbf{X}_1',\mathbf{X}_2,\dots,\mathbf{X}_{N_2}\right\}\setminus\{\mathbf{X}_i\}$.

- $B_3(\mathbf{X}_i)$: $\mathbf{X}_1'$ is strictly outside the k_2-nn ball around $\mathbf{X}_i$ wrt the sample $\left\{\mathbf{X}_1',\mathbf{X}_2,\dots,\mathbf{X}_{N_2}\right\}\setminus\{\mathbf{X}$

- $BE(\mathbf{X}_i)=(A_1(\mathbf{X}_i)\cap B_3(\mathbf{X}_i))\cup(A_3(\mathbf{X}_i)\cap B_1(\mathbf{X}_i))$.

- $BE_1(\mathbf{X}_i,\mathbf{X}_j)=BE(\mathbf{X}_i)\cap[BE(\mathbf{X}_j)\cup A_2(\mathbf{X}_j)\cup B_2(\mathbf{X}_j)]$.

- $BE_2(\mathbf{X}_i,\mathbf{X}_j)=A_2(\mathbf{X}_i)\cap[A_2(\mathbf{X}_j)\cup B_2(\mathbf{X}_j)]$.

- $BE_3(\mathbf{X}_i,\mathbf{X}_j)=B_2(\mathbf{X}_i)\cap B_2(\mathbf{X}_j)$.

Note that if neither $BE_1(\mathbf{X}_i, \mathbf{X}_j)$, $BE_2(\mathbf{X}_i, \mathbf{X}_j)$, nor $BE_3(\mathbf{X}_i, \mathbf{X}_j)$ hold, then

$$\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_i)\right) - g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i), \hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_i)\right)\right| \left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_i)\right) - g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i), \hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_i)\right)\right| \tag{C.20}$$

since either $\hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_i) = \hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_i)$ or $\hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_j) = \hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)$. The same result holds if $\mathbf{X}_i$ or $\mathbf{X}_j$ are switched. Thus we only need to focus on the cases where these events are true. Note that since the samples are iid, the probability that $A_2(\mathbf{X}_i)$ occurs is $1/N_2$. Similarly, the probability of $B_2(\mathbf{X}_i)$ is $1/N_2$.

Claim C.5. *The following hold:*

1. $\Pr\left(BE_1(\mathbf{X}_i, \mathbf{X}_j)\right) = O\left(\left(\frac{k_2}{M_2}\right)^2\right)$
2. $\Pr\left(BE_2(\mathbf{X}_i, \mathbf{X}_j)\right) = O\left(\frac{1}{N_2^2}\right)$
3. $\Pr\left(BE_3(\mathbf{X}_i, \mathbf{X}_j)\right) = O\left(\frac{1}{N_2^2}\right)$

Proof. For the first expression, consider first the case $BE(\mathbf{X}_i) \cap BE(\mathbf{X}_j)$. If $\mathbf{X}_i$ and $\mathbf{X}_j$ are far apart with disjoint k_2-nn balls, we can treat the probability of $BE(\mathbf{X}_i)$ and $BE(\mathbf{X}_j)$ separately within each ball which is $O\left(\frac{k_2}{M_2}\right)$ in each case. This gives a combined probability of $O\left(\left(\frac{k_2}{M_2}\right)^2\right)$ when the balls are disjoint. On the other hand, the probability that the k_2-nn balls intersect is $O\left(\frac{k_2}{M_2}\right)$. In this case, the probability of the event $BE(\mathbf{X}_i) \cap BE(\mathbf{X}_j)$ is $O\left(\frac{k_2}{M_2}\right)$. Combining these facts proves the claim for $BE(\mathbf{X}_i) \cap BE(\mathbf{X}_j)$.

Now consider $BE(\mathbf{X}_i) \cap A_2(\mathbf{X}_j)$. In a similar manner as above, if the two k_2-nn balls are disjoint, we treat the probability of the two events separately within each ball separately giving a combined probability of $O\left(\frac{k_2}{M_2^2}\right)$. Again, the probability that the k_2-nn balls intersect is $O\left(\frac{k_2}{M_2}\right)$ and the resulting probability of $BE(\mathbf{X}_i) \cap A_2(\mathbf{X}_j)$ is $O\left(\frac{k_2}{M_2}\right)$ giving a combined probability of $O\left(\left(\frac{k_2}{M_2}\right)^2\right)$. Similarly, $\Pr\left(BE(\mathbf{X}_i) \cap B_2(\mathbf{X}_j)\right) = O\left(\left(\frac{k_2}{M_2}\right)^2\right)$ which completes the proof for the first expression.

For the second and third expressions, note that since the points $\left\{\mathbf{X}_1', \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N_2}\right\}$ are all iid, $A_2(\mathbf{X}_i)$ is independent of $A_2(\mathbf{X}_j)$ and $B_2(\mathbf{X}_j)$ and $B_2(\mathbf{X}_i)$ is independent of $B_2(\mathbf{X}_j)$. Thus the probability of each of the intersecting events is $1/N_2^2$ which completes the proof. □

From the Lipschitz condition,

$$\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j), \hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right) - g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j), \hat{\mathbf{f}}_{2,k_2}'(\mathbf{X}_j)\right)\right|^2 \leq C_g^2 \left|\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j) - \hat{\mathbf{f}}_{2,k_2}'(\mathbf{X}_j)\right|^2 \tag{C.21}$$

Now suppose that $A_1(\mathbf{X}_j) \cap B_3(\mathbf{X}_j)$ occurs. In this case, $\hat{\mathbf{f}}_{2,k_2}'(\mathbf{X}_j) = \frac{k_2-1}{k_2}\bar{\mathbf{f}}_{2,k_2+1}(\mathbf{X}_j)$. To obtain a bound for $\mathbb{E}\left[\left|\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j) - \hat{\mathbf{f}}_{2,k_2}'(\mathbf{X}_j)\right|^2\right]$, we need the joint distribution of $\bar{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)$ and $\bar{\mathbf{f}}_{2,k_2+1}(\mathbf{X}_j)$ as

$$\left|\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j) - \hat{\mathbf{f}}_{2,k_2}'(\mathbf{X}_j)\right|^2 \leq 2\left|\bar{\mathbf{f}}_{2,k_2}(\mathbf{X}_j) - \frac{k_2-1}{k_2}\bar{\mathbf{f}}_{2,k_2+1}(\mathbf{X}_j)\right|^2 + O\left(\left(\frac{k_2}{M_2}\right)^{\frac{4}{d}}\right). \tag{C.22}$$

Lemma C.6. *The density function of the joint distribution of* $\mathbf{P}_{k_2}$ *and* $\mathbf{P}_{k_2+1}$ *is*

$$f_{P_{k_2},P_{k_2+1}}(p,q) = 1_{\{p\leq q\}}\frac{M_2!}{(k_2-1)!\,(M_2-k_2-1)!}p^{k_2-1}\,(1-q)^{M_2-k_2-1}. \tag{C.23}$$

Proof. For $\mathbf{P}_{k_2}$, let $\mathbf{r}_{k_2}$ be the corresponding k-nn radius. Let δ_p, $\delta_q > 0$. We are interested in the event $\{p \leq \mathbf{P}_{k_2} \leq p+\delta_p, q \leq \mathbf{P}_{k_2+1} \leq q+\delta_q\}$. Consider the following events:

- C_1: There are $k_2 - 1$ points within the radius $\mathbf{r}_{k_2}$.
- C_2: The k_2th point is in the interval $[\mathbf{r}_{k_2}, \mathbf{r}_{k_2} + \epsilon(\delta_p)]$.
- C_3: The $k_2 + 1$th point is in the interval $[\mathbf{r}_{k_2+1}, \mathbf{r}_{k_2+1} + \epsilon(\delta_q)]$.
- C_4: The remaining $M_2 - k_2 - 1$ points are outside the radius $\mathbf{r}_{k_2+1} + \epsilon(\delta_q)$.
- C_5: $\mathbf{r}_{k_2} \leq \mathbf{r}_{k_2+1}$

We have that

$$\Pr\left(p \leq \mathbf{P}_{k_2} \leq p+\delta_p, q \leq \mathbf{P}_{k_2+1} \leq q+\delta_q\right) = \Pr\left(\bigcap_{i=1}^{5} C_i\right).$$

Of the $M_2!$ different ways to permute the M_2 points, there are $(k_2-1)!$ permutations for the points inside the k_2-nn ball and $(M_2-k_2-1)!$ permutations for the points outside the (k_2+1)-nn ball. So the number of different point configurations with k_2-1 points inside $\mathbf{r}_{k_2}$ and M_2-k_2-1 points outside $\mathbf{r}_{k_2+1}$ is $\frac{M_2!}{(k_2-1)!(M_2-k_2-1)!}$. This gives

$$\Pr\left(p \leq \mathbf{P}_{k_2} \leq p+\delta_p, q \leq \mathbf{P}_{k_2+1} \leq q+\delta_q\right) = 1_{\{p\leq q\}} \frac{M_2!}{(k_2-1)!(M_2-k_2-1)!} p^{k_2-1}(1-q)^{M_2-k_2-1}\delta_p\delta_q. \tag{C.24}$$

The p^{k_2-1} term is the probability that k_2-1 points fall within a ball of radius p (the coverage probability). The $(1-q)^{M_2-k_2-1}$ term is the probability that M_2-k_2-1 points fall outside a ball of radius with coverage probability q. The δ_q and δ_p terms correspond to the events that one point falls exactly at radius p and another point falls exactly at radius q. The LHS of (C.24) is equal to the probability of these events. The combinatorial term then accurately accounts for the different possible combinations. From (C.24), we get the density in (C.23). □

From Lemma C.6,

$$\begin{aligned}\mathbb{E}\left[\bar{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\bar{\mathbf{f}}_{2,k_2+1}(\mathbf{X}_j)\right] &= \mathbb{E}\left[f_2^2(\mathbf{X}_j)\frac{k_2(k_2-1)}{M_2^2\mathbf{P}_{k_2}(\mathbf{X}_j)\mathbf{P}_{k_2+1}(\mathbf{X}_j)}\right] \\ &= \mathbb{E}\left[f_2^2(\mathbf{X}_j)\right]\frac{k_2(M_2-1)}{(k_2-1)M_2}.\end{aligned}$$

Then since $\mathbb{E}\left[\mathbf{P}_{k_2}^{-2}\right]=\frac{M_2(M_2-1)}{(k_2-1)(k_2-2)}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left|\bar{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)-\frac{k_2-1}{k_2}\bar{\mathbf{f}}_{2,k_2+1}(\mathbf{X}_j)\right|^2\right] &= \mathbb{E}\left[f_2^2(\mathbf{X}_j)\right]\frac{M_2-1}{M_2}\cdot\frac{2}{k_2\left(k_2-2\right)} \\ &= O\left(\frac{1}{k_2^2}\right). \qquad \text{(C.25)}\end{aligned}$$

A similar result follows if $A_3(\mathbf{X}_i)\cap B_1(\mathbf{X}_i)$ holds instead. Then (C.20) gives

$$\mathbb{E}\Bigg[\Bigg(\sum_{j=2}^{N_2}\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)-g\Big(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}'_{2,}\right.$$

$$=\sum_{i=2}^{N_2}\sum_{j=2}^{N_2}\mathbb{E}\Big[\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_i)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i),\hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_i)\right)\right|\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)-g\Big(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j\right.$$

$$\leq\sum_{i=2}^{N_2}\sum_{j=2}^{N_2}2\mathbb{E}\Bigg[\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_i)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i),\hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_i)\right)\right|\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)-g\Big(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}\right.$$

$$\text{(C.26)}$$

Combining the results from (C.26), (C.21), (C.22), (C.25), and Claim C.5 with the Cauchy-Schwarz inequality gives

$$\begin{aligned}\text{LHS (C.26)} &\leq 2M_2^2\mathbb{E}\left[\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_i)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_i),\hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_i)\right)\right|^2\middle|BE_1(\mathbf{X}_i,\mathbf{X}_j)\right]\Pr\left(BE_1\right. \\ &\leq 2M_2^2C_g^2\mathbb{E}\left[\left|\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)-\hat{\mathbf{f}}'_{2,k_2}(\mathbf{X}_j)\right|^2\middle|BE_1(\mathbf{X}_i,\mathbf{X}_j)\right]\Pr\left(BE_1(\mathbf{X}_i,\mathbf{X}_j)\right)+O(1) \\ &\leq 4M_2^2C_g^2\mathbb{E}\left[\left|\bar{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)-\frac{k_2-1}{k_2}\bar{\mathbf{f}}_{2,k_2+1}(\mathbf{X}_j)\right|^2\middle|BE_1(\mathbf{X}_i,\mathbf{X}_j)\right]\Pr\left(BE_1(\mathbf{X}_i,\mathbf{X}_j)\right)+O\Bigg(\\ &= O\left(M_2^2\cdot\frac{1}{k_2^2}\cdot\left(\frac{k_2}{M_2}\right)^2\right)+O\left(\left(\frac{k_2}{M_2}\right)^{\frac{4}{d}}+1\right) \\ &= O(1).\end{aligned}$$

Applying Jensen's inequality to (D.5) and applying (C.27) and Lemma C.4 gives

$$\begin{aligned}
\mathbb{E}\left[\left|\hat{\mathbf{G}}_{k_1,k_2}-\hat{\mathbf{G}}_{k_1,k_2}^{'}\right|^2\right] &\leq \frac{2}{N_2^2}\mathbb{E}\left[\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_1),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_1)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_1^{'}),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_1^{'})\right)\right|^2\right] \\
&\quad +\frac{2}{N_2^2}\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}^{'}(\mathbf{X}_j)\right)\right|\right)^2\right] \\
&= O\left(\frac{1}{N_2^2}\right).
\end{aligned}$$

Now suppose we have samples $\left\{\mathbf{X}_1,\dots,\mathbf{X}_{N_2},\mathbf{Y}_1,\dots,\mathbf{Y}_{N_1}\right\}$ and $\left\{\mathbf{X}_1,\dots,\mathbf{X}_{N_2},\mathbf{Y}_1^{'},\dots,\mathbf{Y}_{N_1}\right\}$ and denote the respective estimators as $\hat{\mathbf{G}}_{k_1,k_2}$ and $\hat{\mathbf{G}}_{k_1,k_2}^{'}$. Then

$$\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}^{'}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)\right| \leq C_g\left|\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j)-\hat{\mathbf{f}}_{1,k_1}^{'}(\mathbf{X}_j)\right|$$

Thus by similar arguments as was used to obtain (C.27),

$$\begin{aligned}
\mathbb{E}\left[\left|\hat{\mathbf{G}}_{k_1,k_2}-\hat{\mathbf{G}}_{k_1,k_2}^{'}\right|^2\right] &\leq \frac{1}{N_2^2}\mathbb{E}\left[\left(\sum_{j=1}^{N_2}\left|g\left(\hat{\mathbf{f}}_{1,k_1}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)-g\left(\hat{\mathbf{f}}_{1,k_1}^{'}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k_2}(\mathbf{X}_j)\right)\right|\right)^2\right] \\
&= O\left(\frac{1}{N_2^2}\right).
\end{aligned}$$

Applying the Efron-Stein inequality gives

$$\mathbb{V}\left[\hat{\mathbf{G}}_{k_1,k_2}\right]=O\left(\frac{1}{N_2}+\frac{N_1}{N_2^2}\right).$$

C.3 Proof of Theorem III.4 (CLT)

We use Lemma C.7 which is adapted from [186]:

Lemma C.7. *Let the random variables $\{\mathbf{Y}_{M,i}\}_{i=1}^{N}$ belong to a zero mean, unit variance, interchangeable process for all values of M. Assume that $Cov(\mathbf{Y}_{M,1},\mathbf{Y}_{M,2})$ and*

$Cov(\mathbf{Y}_{M,1}^2, \mathbf{Y}_{M,2}^2)$ are $o(1)$ as $M \to \infty$. Then the random variable

$$\mathbf{S}_{N,M} = \frac{\sum_{i=1}^N \mathbf{Y}_{M,i}}{\sqrt{\mathbb{V}\left[\sum_{i=1}^N \mathbf{Y}_{M,i}\right]}} \tag{C.28}$$

converges in distribution to a standard normal random variable.

The proof of this lemma is identical to that in [186] (See "Proof of Theorem 3.3 and Theorem 5.3" in [186]). The relaxed assumptions in Lemma C.7 enable us to prove the central limit theorem under more relaxed conditions on the densities. Assume for simplicity that $N_1 = M_2 = M$ and $k_1(l) = k_2(l) = k(l)$. Define

$$\mathbf{Y}_{M,i} = \frac{\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{f}}_{1,k(l)}\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}\mathbf{X}_i\right) - \mathbb{E}\left[\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{f}}_{1,k(l)}\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}\mathbf{X}_i\right)\right]}{\sqrt{\mathbb{V}\left[\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{f}}_{1,k(l)}\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}\mathbf{X}_i\right)\right]}}. \tag{C.29}$$

This gives

$$\mathbf{S}_{N,M} = \frac{\hat{\mathbf{G}}_w - \mathbb{E}\left[\hat{\mathbf{G}}_w\right]}{\sqrt{\mathbb{V}\left[\hat{\mathbf{G}}_w\right]}}.$$

To bound the covariance between $\mathbf{Y}_{M,1}$ and $\mathbf{Y}_{M,2}$ and between $\mathbf{Y}_{M,1}^2$ and $\mathbf{Y}_{M,2}^2$, it is necessary to show that the denominator of $\mathbf{Y}_{M,i}$ converges to a nonzero constant or to zero sufficiently slowly. The numerator and denominator of $\mathbf{Y}_{M,i}$ are, respectively,

$$\begin{aligned}
&\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right) - \mathbb{E}\left[\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right)\right] \\
&= \sum_{l\in\bar{l}} w(l) \left(g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right) - \mathbb{E}\left[g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right)\right]\right),
\end{aligned}$$

$$\sqrt{\mathbb{V}\left[\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right)\right]}$$

$$= \sqrt{\sum_{l\in\bar{l}}\sum_{l'\in\bar{l}} w(l)w(l')Cov\left(g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i),\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right), g\left(\hat{\mathbf{f}}_{1,k(l')}(\mathbf{X}_i),\hat{\mathbf{f}}_{2,k(l')}(\mathbf{X}_i)\right)\right)}. \tag{C.30}$$

Thus we require bounds on $Cov\left(g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i),\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right), g\left(\hat{\mathbf{f}}_{1,k(l')}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k(l')}(\mathbf{X}_j)\right)\right)$ to bound the covariance between $\mathbf{Y}_{M,1}$ and $\mathbf{Y}_{M,2}$.

Define $\mathcal{M}(\mathbf{Z}) := \mathbf{Z} - \mathbb{E}\mathbf{Z}$ and $\bar{\mathbf{e}}_{i,k(l)}(\mathbf{Z}) := \hat{\mathbf{f}}_{i,k(l)}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}\hat{\mathbf{f}}_{i,k(l)}(\mathbf{Z})$. A Taylor series expansion of $g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_n),\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_n)\right)$ around $\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_n)$ and $\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_n)$ gives

$$\begin{aligned} g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_n),\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_n)\right) &= \sum_{i=0}^{1}\sum_{j=0}^{1}\left(\left.\frac{\partial^{i+j}g(x,y)}{\partial x^i\partial y^j}\right|_{\substack{x=\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_n)\\ y=\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_n)}}\right)\frac{\bar{\mathbf{e}}^i_{1,k(l)}(\mathbf{X}_n)\bar{\mathbf{e}}^j_{2,k(l)}(\mathbf{X}_n)}{i!j!} \\ &\quad + o\left(\bar{\mathbf{e}}_{1,k(l)}(\mathbf{X}_n)+\bar{\mathbf{e}}_{2,k(l)}(\mathbf{X}_n)+\bar{\mathbf{e}}_{1,k(l)}(\mathbf{X}_n)\bar{\mathbf{e}}_{2,k(l)}(\mathbf{X}_n)\right)\end{aligned}$$

Define

$$\begin{aligned} \mathbf{p}_n^{(l)} &:= \mathcal{M}\left(g\left(\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_n),\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_n)\right)\right), \\ \mathbf{q}_n^{(l)} &:= \mathcal{M}\left(\frac{\partial}{\partial x}g\left(\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_n),\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_n)\right)\bar{\mathbf{e}}_{1,k(l)}(\mathbf{X}_n)\right), \\ \mathbf{r}_n^{(l)} &:= \mathcal{M}\left(\frac{\partial}{\partial y}g\left(\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_n),\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_n)\right)\bar{\mathbf{e}}_{2,k(l)}(\mathbf{X}_n)\right), \\ \mathbf{s}_n^{(l)} &:= \mathcal{M}\left(\frac{\partial^2}{\partial x\partial y}g\left(\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_n),\mathbb{E}_{\mathbf{X}_n}\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_n)\right)\bar{\mathbf{e}}_{1,k(l)}(\mathbf{X}_n)\bar{\mathbf{e}}_{2,k(l)}(\mathbf{X}_n)\right), \\ \mathbf{t}_n^{(l)} &:= \mathcal{M}\left(o\left(\bar{\mathbf{e}}_{1,k(l)}(\mathbf{X}_n)+\bar{\mathbf{e}}_{2,k(l)}(\mathbf{X}_n)+\bar{\mathbf{e}}_{1,k(l)}(\mathbf{X}_n)\bar{\mathbf{e}}_{2,k(l)}(\mathbf{X}_n)\right)\right). \end{aligned}$$

This gives

$$Cov\left(g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i),\hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right), g\left(\hat{\mathbf{f}}_{1,k(l')}(\mathbf{X}_j),\hat{\mathbf{f}}_{2,k(l')}(\mathbf{X}_j)\right)\right)$$

$$= \mathbb{E}\left[\left(\mathbf{p}_i^{(l)}+\mathbf{q}_i^{(l)}+\mathbf{r}_i^{(l)}+\mathbf{s}_i^{(l)}+\mathbf{t}_i^{(l)}\right)\left(\mathbf{p}_j^{(l')}+\mathbf{q}_j^{(l')}+\mathbf{r}_j^{(l')}+\mathbf{s}_j^{(l')}+\mathbf{t}_j^{(l')}\right)\right]. \tag{C.31}$$

Lemma C.8. *Let $l, l' \in \bar{l}$ be fixed and $k(l) \to \infty$ as $M \to \infty$ for each $l \in \bar{l}$. Let $\gamma_1(x)$ and $\gamma_2(x)$ be arbitrary functions with* $\sup_x |\gamma_i(x)| < \infty$, $i = 1, 2$. *Then if $q + r \geq 1$*

and $q' + r' \geq 1$,

$$Cov\left(\gamma_1(\mathbf{X}_i)\bar{\mathbf{e}}_{i,k(l)}(\mathbf{X}_i), \gamma_2(\mathbf{X}_j)\bar{\mathbf{e}}_{i,k(l')}(\mathbf{X}_j)\right) = O\left(\frac{1}{\sqrt{k(l)k(l')}}\right),$$

$$Cov\left(\gamma_1(\mathbf{X}_i)\bar{\mathbf{e}}^q_{1,k(l)}(\mathbf{X}_i)\bar{\mathbf{e}}^r_{2,k(l)}(\mathbf{X}_i), \gamma_2(\mathbf{X}_j)\bar{\mathbf{e}}^{q'}_{1,k(l')}(\mathbf{X}_j)\bar{\mathbf{e}}^{r'}_{2,k(l')}(\mathbf{X}_j)\right) = O\left(\frac{1}{\sqrt{k(l)^{q+r}k(l')^{q'+r'}}}\right).$$

Proof. These results follow from an application of Cauchy-Schwarz and Lemma C.3.

□

Lemma C.9. *Let* $l, l' \in \bar{l}$ *be fixed and* $k(l) \to \infty$ *as* $M \to \infty$ *for each* $l \in \bar{l}$*. Then*

$$Cov\left(g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right), g\left(\hat{\mathbf{f}}_{1,k(l')}(\mathbf{X}_j), \hat{\mathbf{f}}_{2,k(l')}(\mathbf{X}_j)\right)\right)$$

$$= \begin{cases} \mathbb{E}\left[\mathbf{p}_i^{(l)}\mathbf{p}_i^{(l')}\right] + O\left(\frac{1}{\sqrt{k(l)k(l')}}\right), & i = j \\ O\left(\frac{1}{\sqrt{k(l)k(l')}}\right) + o\left(\frac{1}{k(l')}\right), & i \neq j. \end{cases}$$

Proof. Consider first $i = j$. Applying Lemma C.8 to (C.31) gives

$$Cov\left(g\left(\hat{\mathbf{f}}_{1,k(l)}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l)}(\mathbf{X}_i)\right), g\left(\hat{\mathbf{f}}_{1,k(l')}(\mathbf{X}_i), \hat{\mathbf{f}}_{2,k(l')}(\mathbf{X}_i)\right)\right) = \mathbb{E}\left[\mathbf{p}_i^{(l)}\mathbf{p}_i^{(l')}\right] + O\left(\frac{1}{\sqrt{k(l)k(l')}}\right).$$

When $i \neq j$, $\mathbb{E}\left[\mathbf{p}_i^{(l)}\left(\mathbf{p}_j^{(l')} + \mathbf{q}_j^{(l')} + \mathbf{r}_j^{(l')} + \mathbf{s}_j^{(l')} + \mathbf{t}_j^{(l')}\right)\right] = 0$ since $\mathbf{X}_i$ and $\mathbf{X}_j$ are independent. A direct application of Lemma C.8 gives

$$\begin{aligned}
\mathbb{E}\left[\mathbf{q}_i^{(l)}\mathbf{q}_j^{(l')}\right] &= O\left(\frac{1}{\sqrt{k(l)k(l')}}\right), \\
\mathbb{E}\left[\mathbf{q}_i^{(l)}\mathbf{r}_j^{(l')}\right] &= O\left(\frac{1}{\sqrt{k(l)k(l')}}\right), \\
\mathbb{E}\left[\mathbf{q}_i^{(l)}\mathbf{s}_j^{(l')}\right] &= O\left(\frac{1}{\sqrt{k(l)k(l')^2}}\right), \\
\mathbb{E}\left[\mathbf{s}_i^{(l)}\mathbf{s}_j^{(l')}\right] &= O\left(\frac{1}{k(l)k(l')}\right), \\
\mathbb{E}\left[\mathbf{s}_i^{(l)}\mathbf{r}_j^{(l')}\right] &= O\left(\frac{1}{\sqrt{k(l)^2k(l')}}\right), \\
\mathbb{E}\left[\mathbf{r}_i^{(l)}\mathbf{r}_j^{(l')}\right] &= O\left(\frac{1}{\sqrt{k(l)k(l')}}\right).
\end{aligned}$$

To handle the implicit constants in the $\mathbf{t}_i^{(l)}$ terms, Cauchy-Schwarz can be applied with Lemma C.8 to get

$$\begin{aligned}
\mathbb{E}\left[\mathbf{q}_i^{(l)}\mathbf{t}_j^{(l')}\right] &= o\left(\frac{1}{k(l')}\right), \\
\mathbb{E}\left[\mathbf{r}_i^{(l)}\mathbf{t}_j^{(l')}\right] &= o\left(\frac{1}{k(l')}\right), \\
\mathbb{E}\left[\mathbf{s}_i^{(l)}\mathbf{t}_j^{(l')}\right] &= o\left(\frac{1}{k(l')}\right), \\
\mathbb{E}\left[\mathbf{t}_i^{(l)}\mathbf{t}_j^{(l')}\right] &= o\left(\frac{1}{k(l')}\right).
\end{aligned}$$

Combining these results with (C.31) completes the proof. □

Since $\mathbf{p}_i(l) = \mathcal{M}\left(g\left(f_1(\mathbf{X}_i), f_2(\mathbf{X}_i)\right)\right) + o(1)$, $\mathbb{E}\left[\mathbf{p}_i^{(l)}\mathbf{p}_i^{(l')}\right]$ is guaranteed to be a nonzero constant if

$$\mathbb{E}\left[g\left(f_1(\mathbf{X}_i), f_2(\mathbf{X}_i)\right)^2\right] \neq \mathbb{E}\left[g\left(f_1(\mathbf{X}_i), f_2(\mathbf{X}_i)\right)\right]^2. \tag{C.32}$$

In this case, applying Lemma C.9 to (C.29) gives $Cov\left(\mathbf{Y}_{M,1}, \mathbf{Y}_{M,2}\right) = o(1)$ as long as

$k(l) \to \infty$ as $M \to \infty$ for each $l \in \bar{l}$. Unfortunately, the condition in (C.32) does not hold for the important case of f-divergence functionals when the densities f_1 and f_2 are equal almost everywhere. However, we still have that the denominator in (C.29) converges more slowly to zero than the numerator as long as $k(l), k(l') \to \infty$ at the same rate for each $l, l' \in \bar{l}$ as the $o\left(\frac{1}{k(l')}\right)$ goes to zero faster than $O\left(\frac{1}{\sqrt{k(l)k(l')}}\right)$. Thus we still get $Cov\left(\mathbf{Y}_{M,1}, \mathbf{Y}_{M,2}\right) = o(1)$ in this case.

For the covariance between $\mathbf{Y}^2_{M,1}$ and $\mathbf{Y}^2_{M,2}$, we only need to focus on the numerator terms as the denominator terms will be similar as before. Thus the numerator of the covariance is

$$\sum_{l\in\bar{l}}\sum_{l'\in\bar{l}}\sum_{j\in\bar{l}}\sum_{j'\in\bar{l}} Cov\left[\left(\mathbf{p}_1^{(l)} + \mathbf{q}_1^{(l)} + \mathbf{r}_1^{(l)} + \mathbf{s}_1^{(l)}\right)\left(\mathbf{p}_1^{(l')} + \mathbf{q}_1^{(l')} + \mathbf{r}_1^{(l')} + \mathbf{s}_1^{(l')}\right),\right.$$

$$\left.\left(\mathbf{p}_2^{(j)} + \mathbf{q}_2^{(j)} + \mathbf{r}_2^{(j)} + \mathbf{s}_2^{(j)}\right)\left(\mathbf{p}_2^{(j')} + \mathbf{q}_2^{(j')} + \mathbf{r}_2^{(j')} + \mathbf{s}_2^{(j')}\right)\right].$$

If $l = l'$ and $j = j'$, then the previous results apply and we get $O\left(\frac{1}{M}\right) + o\left(\frac{1}{k(l')}\right)$. For the general case, the terms with either $\mathbf{p}_1^{(l)}\mathbf{p}_1^{(l')}$ in the left hand side or $\mathbf{p}_2^{(j)}\mathbf{p}_2^{(j')}$ in the right hand side are zero due to independence. For the remaining terms, note that the proof of Lemma 10 in [141] gives under certain conditions that for functions $\gamma_1(x)$ and $\gamma_2(x)$ under the same assumptions as in Lemma C.8,

$$Cov\left[\gamma_1(\mathbf{X}_1)\bar{\mathbf{e}}^s_{1,k(l)}(\mathbf{X}_1)\bar{\mathbf{e}}^q_{2,k(l)}(\mathbf{X}_1)\bar{\mathbf{e}}^{s'}_{1,k(l')}(\mathbf{X}_1)\bar{\mathbf{e}}^{q'}_{2,k(l')}(\mathbf{X}_1),\right.$$

$$\left.\gamma_2(\mathbf{X}_2)\bar{\mathbf{e}}^t_{1,k(j)}(\mathbf{X}_2)\bar{\mathbf{e}}^r_{2,k(j)}(\mathbf{X}_2)\bar{\mathbf{e}}^{t'}_{1,k(j')}(\mathbf{X}_2)\bar{\mathbf{e}}^{r'}_{1,k(j')}(\mathbf{X}_2)\right]$$

$$= O\left(\frac{1}{k(l)^{\frac{s+q}{2}}k(l')^{\frac{s'+q'}{2}}k(j)^{\frac{t+r}{2}}k(j')^{\frac{t'+r'}{2}}}\right). \tag{C.33}$$

As stated in [141], the conditions required for this expression to hold are "(1) There must be at least one positive exponent on both sides of the arguments in the covariance. (2) $\{s + s' + t + t' \neq 1\} \cap \{q + q' + r + r' \neq 1\}$." If neither of the conditions holds in condition (2), then the covariance in (C.33) reduces to the covariance with only one error term on each side. If only one of the conditions holds, then the covariance is zero. This means that if $k(l), k(l') \to \infty$ at the same rate for each $l, l' \in \bar{l}$, then (C.33) reduces to $o\left(\frac{1}{k(l)}\right)$. Combining this result with the previous result on the denominator of $\mathbf{Y}_{M,i}$ gives that $Cov\left(\mathbf{Y}_{M,1}^2, \mathbf{Y}_{M,2}^2\right) = o(1)$. Then by Lemma C.7, $\frac{\hat{\mathbf{G}}_w - \mathbb{E}\left[\hat{\mathbf{G}}_w\right]}{\sqrt{\mathbb{V}\left[\hat{\mathbf{G}}_w\right]}}$ converges in distribution to a standard normal random variable.

APPENDIX D

Proofs for Mutual Information Extension

In this appendix, we prove the convergence results given in Chapter IV.

D.1 Proof of Theorem IV.1 (Bias)

The proof of the bias results in Theorem IV.1 is similar to the proof of the bias results for the divergence functional estimators in Chapter II and so we only sketch it here. The primary differences deal with the product of the marginal KDEs.

The bias of $\tilde{\mathbf{G}}_{h_X,h_Y}$ can be expressed as

$$\begin{aligned}
\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_Y}\right] &= \mathbb{E}\left[g\left(\frac{\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})}{\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})}\right)-g\left(\frac{f_X(\mathbf{X})f_Y(\mathbf{Y})}{f_{XY}(\mathbf{X},\mathbf{Y})}\right)\right] \\
&= \mathbb{E}\left[g\left(\frac{\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})}{\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})}\right)-g\left(\frac{\mathbb{E}_{\mathbf{Z}}X\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right]\mathbb{E}_{\mathbf{Z}}Y\left[\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right]}{\mathbb{E}_{\mathbf{Z}}X,Y\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})}\right)\right] \\
&\quad +\mathbb{E}\left[g\left(\frac{\mathbb{E}_{\mathbf{Z}}X\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right]\mathbb{E}_{\mathbf{Z}}Y\left[\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right]}{\mathbb{E}_{\mathbf{Z}}X,Y\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})}\right)-g\left(\frac{f_X(\mathbf{X})f_Y(\mathbf{Y})}{f_{XY}(\mathbf{X},\mathbf{Y})}\right)\right]
\end{aligned} \tag{D.1}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are drawn jointly from f_{XY}. We can view these terms as a variance-like component (the first term) and a bias-like component, where the respective Taylor series expansions depend on variance-like or bias-like terms of the KDEs.

We first consider the bias-like term, i.e. the second term in (D.1). The Taylor series expansion of $g\left(\frac{\mathbb{E}_{\mathbf{Z}}X\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right]\mathbb{E}_{\mathbf{Z}}Y\left[\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right]}{\mathbb{E}_{\mathbf{Z}}X,Y\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})}\right)$ around $f_X(\mathbf{X})f_Y(\mathbf{Y})$ and $f_{XY}(\mathbf{X},\mathbf{Y})$ gives an expansion with terms of the form of

$$\begin{aligned}
\mathbb{B}^i_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right] &= \Big(\mathbb{E}_{\mathbf{Z}}X\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right]\mathbb{E}_{\mathbf{Z}}Y\left[\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right] - f_X(\mathbf{X})f_Y(\mathbf{Y})\Big)^i, \\
\mathbb{B}^i_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})\right] &= \Big(\mathbb{E}_{\mathbf{Z}}X,Y\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y}) - f_{XY}(\mathbf{X},\mathbf{Y})\Big)^i.
\end{aligned}$$

Since we are not doing boundary correction, we need to consider separately the cases when $\mathbf{Z}$ is in the interior of the support $\mathcal{S}_X \times \mathcal{S}_Y$ and when $\mathbf{Z}$ is close to the boundary of the support. For precise definitions, a point $Z = (X,Y) \in \mathcal{S}_X \times \mathcal{S}_Y$ is in the interior of $\mathcal{S}_X \times \mathcal{S}_Y$ if for all $Z' \notin \mathcal{S}_X \times \mathcal{S}_Y$, $K_X\left(\frac{X-X'}{h_X}\right)K_Y\left(\frac{Y-Y'}{h_Y}\right) = 0$, and a point $Z \in \mathcal{S}_X \times \mathcal{S}_Y$ is near the boundary of the support if it is not in the interior.

It can be shown by Taylor series expansions of the probability densities that for $\mathbf{Z} = (\mathbf{X},\mathbf{Y})$ drawn from f_{XY} in the interior of $\mathcal{S}_X \times \mathcal{S}_Y$, then

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}X\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right] &= f_X(\mathbf{X}) + \sum_{j=1}^{\lfloor s/2\rfloor} c_{X,j}(\mathbf{X})h_X^{2j} + O\left(h_X^s\right), \qquad (D.2)\\
\mathbb{E}_{\mathbf{Z}}Y\left[\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right] &= f_Y(\mathbf{Y}) + \sum_{j=1}^{\lfloor s/2\rfloor} c_{Y,j}(\mathbf{Y})h_Y^{2j} + O\left(h_Y^s\right), \\
\mathbb{E}_{\mathbf{Z}}X,Y\left[\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z})\right] &= f_{XY}(\mathbf{X},\mathbf{Y}) + \sum_{\substack{i=0\\ i+j\neq 0}}^{\lfloor s/2\rfloor}\sum_{j=0}^{\lfloor s/2\rfloor} c_{XY,i,j}(\mathbf{X},\mathbf{Y})h_X^{2i}h_Y^{2j} + O\left(h_X^s + h_Y^s\right).
\end{aligned}$$

For a point near the boundary of the support, we extend the expectation beyond the support of the density. As an example if $\mathbf{X}$ is near the boundary of $\mathcal{S}_X$, then we get

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{X})\right]-f_i(\mathbf{X}) &= \frac{1}{h_X^{d_X}}\int\limits_{V:V\in\mathcal{S}_X} K_X\left(\frac{\mathbf{X}-V}{h_X}\right)f_X(V)dV-f_X(\mathbf{X}) \\ &= \left[\frac{1}{h_X^{d_X}}\int\limits_{V:K_X\left(\frac{\mathbf{X}-V}{h_X}\right)>0} K_X\left(\frac{\mathbf{X}-V}{h_X}\right)f_X(V)dV-f_X(\mathbf{X})\right] \\ &\quad -\left[\frac{1}{h_X^{d_X}}\int\limits_{V:V\notin\mathcal{S}_X} K_X\left(\frac{\mathbf{X}-V}{h_X}\right)f_X(V)dV\right] \\ &= T_{1,X}(\mathbf{X})-T_{2,X}(\mathbf{X}). \qquad (D.3)\end{aligned}$$

As in [145], we only evaulate the density f_X and its derivatives at points within the support when we take its Taylor series expansion. Thus the exact manner in which we define the extension of f_X does not matter as long as the Taylor series remains the same and as long as the extension is smooth. Thus the expected value of $T_{1,X}(\mathbf{X})$ gives an expression of the form of (D.2). For the $T_{2,X}(\mathbf{X})$ term, we perform a similar Taylor series expansion and then apply the condition in assumption $\mathcal{A}.5$ to obtain

$$\mathbb{E}\left[T_{2,X}(\mathbf{X})\right]=\sum_{i=1}^{r}e_ih_X^i+o\left(h_X^r\right).$$

Similar expressions can be found for $\tilde{\mathbf{f}}_{Y,h_Y}$ and $\tilde{\mathbf{f}}_{Z,h_Z}$ and for when (D.3) is raised to a power t. Applying this result gives for the second term in (D.1),

$$\sum_{\substack{j=0\\ i+j\neq 0}}^{r}\sum_{i=0}^{r}c_{10,i,j}h_X^ih_Y^j+O\left(h_X^s+h_Y^s\right).$$

For the first term in (D.1), a Taylor series expansion of $g\left(\frac{\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})}{\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})}\right)$ around $\mathbb{E}_{\mathbf{Z}}X\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right]\mathbb{E}_{\mathbf{Z}}Y\left[\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right]$ and $\mathbb{E}_{\mathbf{Z}}X,Y\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})$ gives an expansion with terms of the form of

$$\begin{aligned}
\tilde{\mathbf{e}}^q_{Z,h_Z}(\mathbf{Z}) &= \left(\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}Z\left[\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z})\right]\right)^q, \\
\tilde{\mathbf{e}}^q_{XY,h_X,h_Y}(\mathbf{Z}) &= \left(\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}) - \mathbb{E}_{\mathbf{Z}}X\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right]\mathbb{E}_{\mathbf{Z}}Y\left[\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right]\right)^q.
\end{aligned}$$

By following a procedure similar to that in [145], we can take the expected value of these expressions to obtain terms of the form

$$\frac{1}{Nh_X^{d_X}}, \frac{1}{Nh_Y^{d_Y}}, \frac{1}{N^2h_X^{d_X}h_Y^{d_Y}}, \tag{D.4}$$

and their respective powers. For general functionals g, we can only guarantee that the mixed derivatives of g evaluated at $\mathbb{E}_{\mathbf{Z}}X\left[\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right]\mathbb{E}_{\mathbf{Z}}Y\left[\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right]$ and $\mathbb{E}_{\mathbf{Z}}X,Y\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X},\mathbf{Y})$ converge to the mixed derivative evaluated at $f_X(\mathbf{X})f_Y(\mathbf{Y})$ and $f_{XY}(\mathbf{X},\mathbf{Y})$ at some rate $o(1)$. Thus we are left with the following terms in the bias:

$$o\left(\frac{1}{Nh_X^{d_X}} + \frac{1}{Nh_Y^{d_Y}}\right)$$

However, if we know that $g\left(t_1,t_2\right)$ has j,l-th order mixed derivatives $\frac{\partial^{j+l}}{\partial t_1^j \partial t_2^l}$ that depend on t_1 and t_2 only through $t_1^\alpha t_2^\beta$ for some $\alpha,\beta\in\mathbb{R}$, then by the generalized binomial theorem, we find that

$$\left(\mathbb{E}_{\mathbf{Z}}\mathbf{X}\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X})\right)^\alpha = \sum_{m=0}^{\infty}\binom{\alpha}{m}f_X^{\alpha-m}(\mathbf{X})\left(\sum_{j=1}^{\lfloor s/2\rfloor}c_{i,j}(\mathbf{X})h_X^{2j} + O\left(h_X^s\right)\right)^m.$$

A similar result holds for $\left(\mathbb{E}_{\mathbf{Z}}\mathbf{Y}\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y})\right)^\alpha$ and $\left(\mathbb{E}_{\mathbf{Z}}Z\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z})\right)^\alpha$. Combining these expressions with (D.4) completes the proof.

D.2 Proof of Theorem IV.2 (Variance)

As for the bias, the proof of the variance result in Theorem IV.2 is similar to the proof of the variance result in Chapter II and so we do not present all of the details. The primary differences again deal with the product of the marginal KDEs. The proof uses the Efron-Stein inequality as for the divergence functional estimators.

In this case we consider the samples $\{\mathbf{Z}_1, \dots, \mathbf{Z}_N\}$ and $\{\mathbf{Z}'_1, \dots, \mathbf{Z}'_N\}$ and the respective estimators $\tilde{\mathbf{G}}_{h_X,h_Y}$ and $\tilde{\mathbf{G}}'_{h_X,h_Y}$. By the triangle inequality,

$$\begin{aligned}\left|\tilde{\mathbf{G}}_{h_X,h_Y} - \tilde{\mathbf{G}}'_{h_X,h_Y}\right| &\leq \frac{1}{N}\left|g\left(\frac{\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_1)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_1)}{\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X}_1,\mathbf{Y}_1)}\right) - g\left(\frac{\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}'_1)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}'_1)}{\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X}'_1,\mathbf{Y}'_1)}\right)\right| \\ &+\frac{1}{N}\sum_{j=2}^{N_2}\left|g\left(\frac{\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_1)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_1)}{\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X}_1,\mathbf{Y}_1)}\right) - g\left(\frac{\tilde{\mathbf{f}}'_{X,h_X}(\mathbf{X}_1)\tilde{\mathbf{f}}'_{Y,h_Y}(\mathbf{Y}_1)}{\tilde{\mathbf{f}}'_{Z,h_Z}(\mathbf{X}_1,\mathbf{Y}_1)}\right)\right|\end{aligned}\tag{D.5}$$

By the Lipschitz condition on g, the first term in (D.5) can be decomposed into terms of the form of

$$\left|\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}_1) - \tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}'_1)\right|,$$

$$\left|\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_1)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_1) - \tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}'_1)\tilde{\mathbf{f}}'_{Y,h_Y}(\mathbf{Y}_1)\right|.$$

By making a substitution in the expectation, it can be shown that

$$\mathbb{E}\left[\left|\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}_1) - \tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}'_1)\right|^2\right] \leq 2||K_X \cdot K_Y||_\infty^2.$$

For the product of the marginal KDEs, we have that

$$\begin{aligned}\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_1)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_1) &= \frac{1}{M^2 h_X^{d_X} h_Y^{d_Y}}\sum_{i=2}^{N}\sum_{j=2}^{N} K_X\left(\frac{\mathbf{X}_1-\mathbf{X}_i}{h_X}\right)K_Y\left(\frac{\mathbf{Y}_1-\mathbf{Y}_j}{h_Y}\right) \\ &= \frac{1}{M}\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}_1) + \frac{1}{M^2 h_X^{d_X} h_Y^{d_Y}}\sum_{i\neq j} K_X\left(\frac{\mathbf{X}_1-\mathbf{X}_i}{h_X}\right)K_Y\left(\frac{\mathbf{Y}_1-\mathbf{Y}_j}{h_Y}\right).\end{aligned}$$

By applying the triangle inequality, Jensen's inequality, and similar substitutions, we get

$$\begin{aligned}
\mathbb{E}\left[\left|\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_1)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_1) - \tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_1^{'})\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_1^{'})\right|^2\right] &\leq \mathbb{E}\left[\frac{2}{M^2}\left|\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}_1) - \tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}_1^{'})\right|^2\right] \\
&\quad + \frac{2(M-1)}{M^3 h_X^{2d_X} h_Y^{2d_Y}} \times \\
&\quad \sum_{i\neq j}\mathbb{E}\left[\left(K_X\left(\frac{\mathbf{X}_1-\mathbf{X}_i}{h_X}\right)K_Y\left(\frac{\mathbf{Y}_1-\mathbf{Y}_j}{h_Y}\right)\right.\right. \\
&\quad \left.\left. -K_X\left(\frac{\mathbf{X}_1^{'}-\mathbf{X}_i}{h_X}\right)K_Y\left(\frac{\mathbf{Y}_1^{'}-\mathbf{Y}_j}{h_Y}\right)\right)^2\right] \\
&\leq \frac{4+2(M-1)^2}{M^2}||K_X\cdot K_Y||^2.
\end{aligned}$$

For the second term in (D.5), it can be shown that

$$\begin{aligned}
\mathbb{E}\left[\left|\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}_i) - \tilde{\mathbf{f}}_{Z,h_Z}^{'}(\mathbf{Z}_i)\right|^2\right] &= \frac{1}{M^2 h_X^{2d_X} h_Y^{2d_Y}}\mathbb{E}\left[\left(K_X\left(\frac{\mathbf{X}_1-\mathbf{X}_i}{h_X}\right)K_Y\left(\frac{\mathbf{Y}_1-\mathbf{Y}_j}{h_Y}\right)\right.\right. \\
&\quad \left.\left. -K_X\left(\frac{\mathbf{X}_1^{'}-\mathbf{X}_i}{h_X}\right)K_Y\left(\frac{\mathbf{Y}_1^{'}-\mathbf{Y}_j}{h_Y}\right)\right)^2\right] \\
&\leq \frac{2||K_X\cdot K_Y||_\infty^2}{M^2}.
\end{aligned}$$

By a similar approach,

$$\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_i)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_i) - \tilde{\mathbf{f}}_{X,h_X}^{'}(\mathbf{X}_i)\tilde{\mathbf{f}}_{Y,h_Y}^{'}(\mathbf{Y}_i)$$

$$\begin{aligned}
&= \tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{Z}_i) - \tilde{\mathbf{f}}_{Z,h_Z}^{'}(\mathbf{Z}_i) + \frac{1}{M^2 h_X^{d_X} h_Y^{d_Y}}\left(\sum_{\substack{n=2\\ n\neq i}} K_Y\left(\frac{\mathbf{Y}_i-\mathbf{Y}_n}{h_Y}\right)\left(K_X\left(\frac{\mathbf{X}_i-\mathbf{X}_1}{h_X}\right) - K_X\left(\frac{\mathbf{X}_i-\mathbf{X}_1^{'}}{h_X}\right)\right)\right. \\
&\quad \left. + \sum_{\substack{n=2\\ n\neq i}} K_X\left(\frac{\mathbf{X}_i-\mathbf{X}_n}{h_X}\right)\left(K_Y\left(\frac{\mathbf{Y}_i-\mathbf{Y}_1}{h_Y}\right) - K_Y\left(\frac{\mathbf{Y}_i-\mathbf{Y}_1^{'}}{h_Y}\right)\right)\right),
\end{aligned}$$

$$\implies \mathbb{E}\left[\left|\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_i)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_i) - \tilde{\mathbf{f}}_{X,h_X}^{'}(\mathbf{X}_i)\tilde{\mathbf{f}}_{Y,h_Y}^{'}(\mathbf{Y}_i)\right|^2\right] \leq 6||K_X\cdot K_Y||_\infty^2\left(\frac{1}{M^2} + \frac{(M-2)^2}{M^4}\right)$$

We can then apply the Cauchy Schwarz inequality to bound the square of the second term in (D.5) to get

$$\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\frac{\tilde{\mathbf{f}}_{X,h_X}(\mathbf{X}_1)\tilde{\mathbf{f}}_{Y,h_Y}(\mathbf{Y}_1)}{\tilde{\mathbf{f}}_{Z,h_Z}(\mathbf{X}_1,\mathbf{Y}_1)}\right)-g\left(\frac{\tilde{\mathbf{f}}'_{X,h_X}(\mathbf{X}_1)\tilde{\mathbf{f}}'_{Y,h_Y}(\mathbf{Y}_1)}{\tilde{\mathbf{f}}'_{Z,h_Z}(\mathbf{X}_1,\mathbf{Y}_1)}\right)\right|\right)^2\right] \leq 14C_g^2||K_X{\cdot}K_Y||_\infty^2.$$

Applying Jensen's inequality in conjunction with these results gives

$$\mathbb{E}\left[\left|\tilde{\mathbf{G}}_{h_X,h_Y}-\tilde{\mathbf{G}}'_{h_X,h_Y}\right|^2\right] \leq \frac{44C_g^2||K_X\cdot K_Y||_\infty^2}{N^2}.$$

Applying the Efron-Stein inequality finishes the proof.

D.3 Theory for Mixed Random Variables

D.3.1 Proof of Theorem IV.4 (Bias)

Let $\mathbf{h}_{X|y} = l\mathbf{N}_y^{-\beta}$ for some positive l and $0 < \beta < \frac{1}{d_X}$. Under assumptions $\mathcal{A}.0 - \mathcal{A}.5$, we prove that for general g, the bias of the plug-in estimator $\tilde{\mathbf{G}}_{h_X,h_{X|Y}}$

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\right] &= \sum_{\substack{j=0\\ i+j\neq 0}}^{r}\sum_{i=0}^{r} c_{13,i,j}h_X^i l^j N^{-j\beta}+\frac{c_{14,X}}{Nh_X^{d_X}}+\frac{c_{14,y}}{l^{d_X}N^{1-\beta d_X}} \\ &\quad +O\left(h_X^s+N^{-s\beta}+\frac{1}{Nh_X^{d_X}}+\frac{1}{N^{1-\beta d_X}}+\frac{1}{N}\right). \qquad \text{(D.6)}\end{aligned}$$

Furthermore, if $g(t_1,t_2)$ has j,l-th order mixed derivatives $\frac{\partial^{j+l}}{\partial t_1^j \partial t_2^l}$ that depend on t_1 and t_2 only through $t_1^\alpha t_2^\beta$ for some $\alpha,\beta\in\mathbb{R}$, then for any positive integer $\lambda \geq 2$, the bias is

$$\begin{aligned}
\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\right] &= \sum_{\substack{j=0\\ i+j\neq 0}}^{r}\sum_{i=0}^{r} c_{13,i,j} h_X^i l^j N^{-j\beta} + \sum_{j=1}^{\lambda/2}\sum_{i=1}^{\lambda/2}\sum_{m=0}^{r}\sum_{n=0}^{r} c_{14,j,i,m,n}\frac{h_X^m l^n N^{-n\beta}}{\left(Nh_X^{d_X}\right)^j\left(l^{d_X}N^{1-\beta d_X}\right)^i} \\
&\quad + \sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\sum_{n=0}^{r}\left(c_{14,m,n,j,X}\frac{h_X^m l^n N^{-n\beta}}{\left(Nh_X^{d_X}\right)^j} + c_{14,m,n,j,Y}\frac{h_X^m l^n N^{-n\beta}}{\left(l^{d_X}N^{1-\beta d_X}\right)^j}\right) \\
&\quad + O\left(h_X^s + N^{-s\beta} + \frac{1}{\left(Nh_X^{d_X}\right)^{\lambda/2}} + \frac{1}{\left(N^{1-\beta d_X}\right)^{\lambda/2}} + \frac{1}{N}\right). \qquad \text{(D.7)}
\end{aligned}$$

We only prove (D.6) as the proof of (D.7) is identical. The bias of $\tilde{\mathbf{G}}_{h_X,h_{X|Y}}$ is

$$\begin{aligned}
\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\right] &= \mathbb{E}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\right] - G(\mathbf{X};\mathbf{Y}) \\
&= \mathbb{E}\left[\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y}{N}\tilde{\mathbf{G}}_{h_X,h_{X|y}} - g\left(\frac{f_X(\mathbf{X})}{f_{X|Y}(\mathbf{X}|\mathbf{Y})}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y}{N}\tilde{\mathbf{G}}_{h_X,h_{X|y}} - g\left(\frac{f_X(\mathbf{X})}{f_{X|Y}(\mathbf{X}|\mathbf{Y})}\right)\right|\mathbf{Y},\mathbf{Y}_1,\dots,\mathbf{Y}_N\right]\right] \\
&= \mathbb{E}\left[\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y}{N}\mathbb{E}\left[\left(\tilde{\mathbf{G}}_{h_X,h_{X|y}} - g\left(\frac{f_X(\mathbf{X})}{f_{X|Y}(\mathbf{X}|\mathbf{Y})}\right)\right)\right|\mathbf{Y},\mathbf{Y}_1,\dots,\mathbf{Y}_N\right]\right] \\
&= \mathbb{E}\left[\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y}{N}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_{X|y}}\right|\mathbf{Y}_1,\dots,\mathbf{Y}_N\right]\right],
\end{aligned}$$

where we use the law of total expectation and the fact that $\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y}{N} = 1$. Let $\mathbf{h}_{X|y} = l\mathbf{N}_y^{-\beta}$ for some positive l and $0 < \beta < \frac{1}{d_X}$. From Theorem 1, the conditional bias of $\tilde{\mathbf{G}}_{h_X,h_{X|y}}$ given $\mathbf{Y}_1,\dots,\mathbf{Y}_N$ is

$$\begin{aligned}
\mathbb{B}\left[\tilde{\mathbf{G}}_{h_X,h_{X|y}}\middle|\mathbf{Y}_1,\dots,\mathbf{Y}_N\right] &= \sum_{\substack{j=0\\ i+j\neq 0}}^{r}\sum_{i=0}^{r} c_{10,i,j}h_X^i\mathbf{h}_{X|y}^j + \frac{c_{11,X}}{\mathbf{N}_y h_X^{d_X}} + \frac{c_{11,y}}{\mathbf{N}_y \mathbf{h}_{X|y}^{d_X}} \\
&\quad +O\left(h_X^s + \mathbf{h}_{X|y}^s + \frac{1}{\mathbf{N}_y h_X^{d_X}} + \frac{1}{\mathbf{N}_y \mathbf{h}_{X|y}^{d_X}}\right) \\
&= \sum_{\substack{j=0\\ i+j\neq 0}}^{r}\sum_{i=0}^{r} c_{10,i,j}h_X^i l^j \mathbf{N}_y^{-j\beta} + \frac{c_{11,X}}{\mathbf{N}_y h_X^{d_X}} + \frac{c_{11,y}}{l^{d_X}\mathbf{N}_y^{1-\beta d_X}} \\
&\quad +O\left(h_X^s + \mathbf{N}_y^{-s\beta} + \frac{1}{\mathbf{N}_y h_X^{d_X}} + \frac{1}{\mathbf{N}_y^{1-\beta d_X}}\right). \qquad \text{(D.8)}
\end{aligned}$$

$\mathbf{N}_y$ is a binomial random variable Multiplying (D.8) by $\mathbf{N}_y$ results in terms of the form of $\mathbf{N}_y^{1-\gamma}$ with $\gamma \geq 0$. $\mathbf{N}_y$ is a binomial random variable with parameter $f_Y(y)$,N trials, and mean $Nf_Y(y)$. We can compute the fractional moments of a binomial random variable by using the generalized binomial theorem to obtain

$$\begin{aligned}
\mathbb{E}\left[\mathbf{N}_y^\alpha\right] &= \sum_{i=0}^{\infty}\begin{pmatrix}\alpha\\ i\end{pmatrix}(Nf_Y(y))^{\alpha-i}\,\mathbb{E}\left[(\mathbf{N}_Y - Nf_Y(y))^i\right] \\
&= \sum_{i=0}^{\infty}\begin{pmatrix}\alpha\\ i\end{pmatrix}(Nf_Y(y))^{\alpha-i}\sum_{n=0}^{\lfloor i/2\rfloor} c_{n,i}(f_Y(y))N^n \\
&= \sum_{i=0}^{\infty}\begin{pmatrix}\alpha\\ i\end{pmatrix}f_Y(y)^{\alpha-i}\sum_{n=0}^{\lfloor i/2\rfloor} c_{n,i}(f_Y(y))N^{\alpha-i+n},
\end{aligned}$$

where we use the following expression for the i-th central moment of a binomial random variable derived by Riordan [170]:

$$\mathbb{E}\left[(\mathbf{N}_Y - Nf_Y(y))^i\right] = \sum_{n=0}^{\lfloor i/2\rfloor} c_{n,i}(f_Y(y))N^n.$$

If $\alpha = 1-\gamma$, then dividing by N results in terms of the form of $N^{-\gamma-i+n}$. Since

$n \leq \lfloor i/2 \rfloor$, $-\gamma - i + n$ is always less than zero and is only greater than -1 if $i = 0$. This completes the proof.

D.3.2 Proof of Theorem IV.5 (Variance)

As for the bias, we assume that $\mathbf{h}_{X|y} = l\mathbf{N}_y^{-\beta}$ for some positive l and $0 < \beta < \frac{1}{d_X}$. By the law of total variance, we have

$$\mathbb{V}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\right] = \mathbb{E}\left[\mathbb{V}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\middle|\mathbf{Y}_1,\dots,\mathbf{Y}_N\right]\right] + \mathbb{V}\left[\mathbb{E}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\middle|\mathbf{Y}_1,\dots,\mathbf{Y}_N\right]\right]. \tag{D.9}$$

Note that given all of the $\mathbf{Y}_i$'s, the estimators $\tilde{\mathbf{G}}_{h_X,h_{X|y}}$ are all independent since they use different sets of $\mathbf{X}_i$'s for each y. By Theorem 2, we have

$$\begin{aligned}\mathbb{V}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\middle|\mathbf{Y}_1,\dots,\mathbf{Y}_N\right] &= O\left(\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y^2}{N^2}\cdot\frac{1}{\mathbf{N}_y}\right)\\ &= O\left(\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y}{N^2}\right).\end{aligned}$$

Taking the expectation wrt $\mathbf{Y}_1, \dots \mathbf{Y}_N$ then gives $O\left(\frac{1}{N}\right)$ for the first term in (D.9).

For the second term in (D.9), from (D.8) we have that for general g

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathbf{G}}_{h_X,h_{X|y}}\middle|\mathbf{Y}_1,\dots,\mathbf{Y}_N\right] &= O\left(\sum_{j=0}^{r}\mathbf{N}_y^{-j\beta} + \frac{1}{\mathbf{N}_y} + \mathbf{N}_y^{-s\beta} + \mathbf{N}_y^{1-\beta d_X}\right)\\ &= O\left(f\left(\mathbf{N}_y\right)\right).\end{aligned}$$

By the Efron-Stein inequality, we have that if $\mathbf{N}_y'$ is an independent and identically distributed realization of $\mathbf{N}_y$, then

$$\begin{aligned}
\mathbb{V}\left[\sum_{y\in\mathcal{S}_Y}\frac{\mathbf{N}_y}{N}f\left(\mathbf{N}_y\right)\right] &\leq \frac{1}{2N^2}\sum_{y\in\mathcal{S}_Y}\mathbb{E}\left[\left(\mathbf{N}_y f\left(\mathbf{N}_y\right)-\mathbf{N}_y^{'}f\left(\mathbf{N}_y^{'}\right)\right)^2\right] \\
&= O\left(\frac{1}{N^2}\mathbb{E}\left[\left(\mathbf{N}_y f\left(\mathbf{N}_y\right)-\mathbf{N}_y^{'}f\left(\mathbf{N}_y^{'}\right)\right)^2\right]\right) \\
&= O\left(\frac{1}{N^2}\mathbb{V}\left[\mathbf{N}_y f\left(\mathbf{N}_y\right)\right]\right), \qquad \text{(D.10)}
\end{aligned}$$

where the second step follows from the fact that $\mathcal{S}_Y$ is finite and the last step follows from the fact that $\mathbf{N}_y$ and $\mathbf{N}_y^{'}$ are iid. The expression $\mathbb{V}\left[\mathbf{N}_y f\left(\mathbf{N}_y\right)\right]$ is simply a sum of terms of the form of $\mathbb{V}\left[\mathbf{N}_y^\gamma\right]$ where $0 < \gamma \leq 1$. Even the covariance terms can be bounded by the square root of the product of these terms by the Cauchy Schwarz inequality.

Let $p_y = f_Y(y)$. Consider the Taylor series expansion of the function $h(x) = x^\gamma$ at the point Np_y. This is

$$\begin{aligned}
h(x) &= \left(Np_y\right)^\gamma + \gamma\left(Np_y\right)^{\gamma-1}\left(x-Np_y\right)+\frac{\gamma(\gamma-1)}{2}\left(Np_y\right)^{\gamma-2}\left(x-Np_y\right)^2 \\
&\quad +\sum_{k=3}^{\infty}\frac{\gamma(\gamma-1)\dots(\gamma-k+1)}{k!}\left(Np_y\right)^{\gamma-k}\left(x-Np_y\right)^2. \qquad \text{(D.11)}
\end{aligned}$$

From Riordan [170], we know that the ith central moment of $\mathbf{N}_y$ is $O\left(N^{\lfloor i/2\rfloor}\right)$. Then since $\gamma \leq 1$, the last terms in (D.11) are $O\left(N^{-1}\right)$ when $x = \mathbf{N}_y$ and we take the expectation. Thus

$$\begin{aligned}
\mathbb{E}\left[\mathbf{N}_y^\gamma\right] &= \left(Np_y\right)^\gamma + \frac{\gamma(\gamma-1)}{2}\left(Np_y\right)^{\gamma-1}\left(1-p_y\right)+O\left(N^{-1}\right) \\
\implies \mathbb{E}\left[\mathbf{N}_y^\gamma\right]^2 &= \left(Np_y\right)^{2\gamma}+\gamma(\gamma-1)(1-p_y)\left(Np_y\right)^{2\gamma-1}+\left(\frac{\gamma(\gamma-1)}{2}\right)^2\left(Np_y\right)^{2\gamma-2} \\
&\quad +O\left(N^{-1}\right).
\end{aligned}$$

By a similar Taylor series expansion, we have that

$$\mathbb{E}\left[\mathbf{N}_y^{2\gamma}\right] = \left(Np_y\right)^{2\gamma} + \gamma(2\gamma-1)(1-p_y)\left(Np_y\right)^{2\gamma-1} + O\left(N^{-1}\right).$$

Combining these results gives

$$\begin{aligned}\mathbb{V}\left[\mathbf{N}_y^{\gamma}\right] &= \mathbb{E}\left[\mathbf{N}_y^{2\gamma}\right] - \mathbb{E}\left[\mathbf{N}_y^{\gamma}\right]^2 \\ &= O\left(N^{2\gamma-1} + N^{2\gamma-2} + N^{-1}\right) \\ &= O\left(N\right),\end{aligned}$$

where the last step follows from the fact that $\gamma \leq 1$. Combining this result with (D.10) gives

$$\mathbb{V}\left[\mathbb{E}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\middle|\mathbf{Y}_1,\dots,\mathbf{Y}_N\right]\right] = O\left(\frac{1}{N}\right).$$

By the law of total variance, $\mathbb{V}\left[\tilde{\mathbf{G}}_{h_X,h_{X|Y}}\right] = O\left(\frac{1}{N}\right)$.

APPENDIX E

Details on Methods Applied to Sunspot Images and HFO Data

This appendix contains details on some of the methods used to analyze the sunspot images.

E.1 Intrinsic Dimension Estimation of Manifolds

Consider data that are described in an extrinsic Euclidean space of d dimensions. However, suppose the data actually lie on a lower dimensional manifold $\mathcal{M}$. Thus the intrinsic dimension m of the data corresponds to the dimension of $\mathcal{M}$. For example, data may be given to us in a 3 dimensional space but lie on the surface of a sphere. Thus the intrinsic dimension of the data would be 2.

In some cases, data points from the same data set may lie on different manifolds. For example, part of the data with an extrinsic dimension of 3 could lie on the surface of a sphere ($m = 2$) while another part may lie on a circle ($m = 1$). We then say that data points from these different manifolds have a different *local* intrinsic dimension. The local intrinsic dimension gives some measure of the local complexity of the image. Additionally, the local intrinsic dimension is useful for dictionary learning because we can use it to determine whether different-sized dictionaries should be used for different regions, e.g. within the sunspots and outside of the sunspots.

We now describe the k-NN estimator of intrinsic dimension in more detail. For a set of independently identically distributed random vectors $\mathbf{Z}_n = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ with values in a compact subset of $\mathbb{R}^d$, the k-nearest neighbors of $\mathbf{z}_i$ in $\mathbf{Z}_n$ are the k points in $\mathbf{Z}_n \backslash \{\mathbf{z}_i\}$ closest to $\mathbf{z}_i$ as measured by the Euclidean distance $|| \cdot ||$. The k-NN graph is then formed by assigning edges between a point in $\mathbf{Z}_n$ and its k-nearest neighbors. The intrinsic dimension is related to the total edge length of the k-NN graph and can be estimated based on this relationship. The k-NN graph is then formed by assigning edges between a point in $\mathbf{Z}_n$ and its k-nearest neighbors and has total edge length defined as

$$L_{\gamma,k}(\mathbf{Z}_n) = \sum_{i=1}^{n} \sum_{\mathbf{z} \in \mathcal{N}_{k,i}} ||\mathbf{z} - \mathbf{z}_i||^{\gamma},$$

where $\gamma > 0$ is a power weighting constant and $\mathcal{N}_{k,i}$ is the set of k nearest neighbors of $\mathbf{z}_i$. It has been shown that for large n,

$$L_{\gamma,k}(\mathbf{Z}_n) = n^{\alpha(m)} c + \epsilon_n,$$

where $\alpha = (m - \gamma)/m$, c is a constant with respect to $\alpha(m)$, and ϵ_n is an error term that decreases to zero a.s. as $n \to \infty$ [40]. A global intrinsic dimension estimate $\hat{m}$ is found based on this relationship using non-linear least squares over different values of n [31].

A local estimate of intrinsic dimension at a point $\mathbf{z}_i$ can be found by running the algorithm over a smaller neighborhood about $\mathbf{z}_i$. The variance of this local estimate is then reduced by smoothing via majority voting in a neighborhood of $\mathbf{z}_i$ [31].

E.2 Matrix Factorization

As mentioned in Section 6.2, the goal of matrix factorization is to accurately decompose the $2m^2 \times n$ data matrix $\mathbf{Z}$ into the product of two matrices $\mathbf{A}$ (with size $2m^2 \times r$) and $\mathbf{H}$ (with size $r \times n$), where $\mathbf{A}$ has fewer columns than rows ($r < 2m^2$). The matrix $\mathbf{A}$ is the dictionary and the matrix $\mathbf{H}$ is the coefficient matrix. The columns of $\mathbf{A}$ form a basis for the data in $\mathbf{Z}$.

The two matrix factorization methods we use are singular value decomposition (SVD) and nonnegative matrix factorization (NMF). These two methods can be viewed as solving two different optimization problems where the objective function is the same but the constraints differ. Let $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r]$ and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots \mathbf{h}_n]$. For SVD, the optimization problem is

$$\begin{array}{rl} \min_{\mathbf{A},\mathbf{H}} & ||\mathbf{Z} - \mathbf{AH}||_F^2 \\ \text{subject to} & \mathbf{a}_i^T \mathbf{a}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} . \end{array}$$

In words, SVD requires the columns of $\mathbf{A}$ to be orthonormal.

For standard NMF, the optimization problem is

$$\begin{array}{rll} \min_{\mathbf{A},\mathbf{H}} & ||\mathbf{Z} - \mathbf{AH}||_F^2 & \\ \text{subject to} & \mathbf{a}_i \geq 0, & \forall i = 1, \ldots, r \\ & \mathbf{h}_i \geq 0, & \forall i = 1, \ldots, n \end{array} ,$$

where $\mathbf{a} \geq 0$ applied to a vector $\mathbf{a}$ implies that all of $\mathbf{a}$'s entries are greater than or equal to 0. In our problem, only the continuum is nonnegative so we only apply the constraint to the continuum part of the matrix $\mathbf{A}$. So if $\mathbf{a}_i$ and $\mathbf{b}_i$ are both vectors with length m^2 corresponding to the continuum and magnetogram parts, respectively, then we have $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_r \\ \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_r \end{bmatrix}$. The NMF method we use also constrains the columns of $\mathbf{H}$ to lie on a simplex, i.e. $\sum_{j=1}^{r} \mathbf{h}_i(j) = 1$. Thus the optimization problem for our approach to NMF is

$$\begin{array}{rrl} \min_{\mathbf{A},\mathbf{H}} & ||\mathbf{Z} - \mathbf{AH}||_F^2 & \\ \text{subject to} & \mathbf{a}_i \geq 0, & \forall i = 1, \dots, r \\ & \mathbf{h}_i \geq 0, & \forall i = 1, \dots, n \\ & \sum_{j=1}^{r} \mathbf{h}_i(j) = 1, & \forall i = 1, \dots, n \end{array}.$$

This problem is not convex and is solved in an alternating manner by fixing $\mathbf{H}$, finding the matrix $\mathbf{A}$ that solves the problem assuming $\mathbf{H}$ is fixed, and then solving for $\mathbf{H}$ while $\mathbf{A}$ is fixed. This process is repeated until the algorithm converges to a local minimum. See [125] for more details on the convergence analysis.

E.3 The EAC-DC Clustering Method

Let $V = \{v_1, v_2, \dots, v_N\}$ be a set of vertices and let $E = \{e_{ij}\}$, where e_{ij} denotes an edge between vertices v_i, v_j, $i, j \in \{1, \dots, N\}$, be a set of undirected edges between them. The pair $(V, E) = G$ is the corresponding undirected graph. In our application, V corresponds to the set of AR image pairs being clustered and E contains all possible edges between the vertices. The weight of an edge e_{ij} is defined as w_{ij} and measures the base dissimilarity between two vertices v_i and v_j. In many applications, the base dissimilarity is the Euclidean distance. In our case, we use the Hellinger distance as the base dissimilarity measure.

A spanning tree T of the graph G is a connected acyclic subgraph that passes through all N vertices of the graph and the weight of T is the sum of all the edge weights used to construct the tree, $\sum_{e_{ij} \in T} w_{ij}$. A minimal spanning tree of G is a spanning tree which has the minimal weight $\min_T \sum_{e_{ij}\in T} w_{ij}$.

Prim's algorithm [164] is used by [67] to construct the dual rooted MST. In Prim's algorithm, the MST is grown sequentially where at each step, a single edge is added. This edge corresponds to the edge with minimal weight that connects a previously unconnected vertex to the existing tree. The root of the MST corresponds to the beginning vertex. For the dual rooted MST, we begin with two vertices v_i and v_j and construct the minimal spanning trees T_i and T_j. At each step, the two edges that would grow both trees T_i and T_j using Prim's algorithm are proposed and the edge with minimal weight is added. This continues until T_i and T_j connect. The weight of the final edge added in this algorithm defines a new metric between the vertices v_i and v_j. This process is repeated for all pairs of vertices and this new metric is used as input to spectral clustering [67].

A primary advantage of this metric based on the hitting time of the two MSTs is that it depends on the MST topology of the data. Thus if two vertices belong to the same cluster, then the MST distance between them will be small since cluster points will be close together. This is the case even if the vertices are far away from each other (e.g. on opposite ends of the cluster). However, if the two vertices are in different clusters that are well separated, then the MST distance between them will be large. See Figure E.1 for an example. Thus this method of clustering is very robust to the shape of the clusters. [67] contains many more examples.

The MST based metric can be computationally intensive to compute as Prim's algorithm must be run as many times as there are pairs of vertices. To counter this, [67] proposed the EAC-DC algorithm which uses the information from only a subset of the dual rooted MSTs. This is done by calculating the dual rooted MSTs for a random pair of vertices. Three clusters are defined for each run: all vertices that are connected to one of the roots in the MSTs form two of the clusters (one for each root) while all points that are not connected to either of the MSTs are assigned to a third

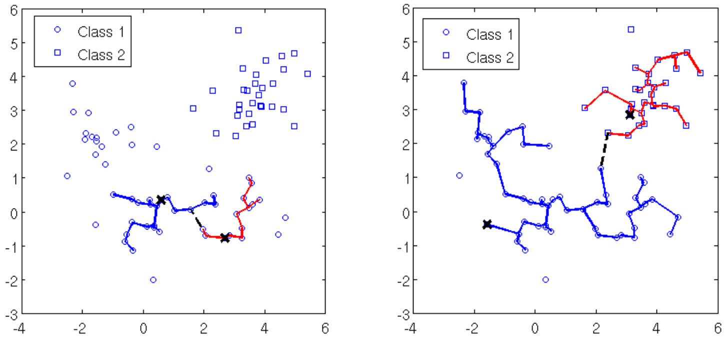

Figure E.1: Dual rooted Prim tree built on a 2-dimensional data set when the roots are chosen from the same class (left) and different classes (right). The X's mark the roots of the trees and the dashed line is the last connected edge. The length of the last connected edge is greater when the roots belong to clusters that are more separated.

"rejection" cluster. A co-association measure for two vertices is then defined as the number of times those vertices are contained in the same non-rejection cluster divided by the total number of runs (dual rooted MSTs). This co-association measure forms a similarity measure to which spectral clustering is applied.

BIBLIOGRAPHY

[1] Abramenko, V. I. (2005), Multifractal Analysis Of Solar Magnetograms, *Sol. Phys.*, *228*, 29–42, doi:10.1007/s11207-005-3525-9.

[2] Aha, D. W. (1992), Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms, *International Journal of Man-Machine Studies*, *36*(2), 267–287.

[3] Ahmed, O. W., R. Qahwaji, T. Colak, P. A. Higgins, P. T. Gallagher, and D. S. Bloomfield (2011), Solar Flare Prediction Using Advanced Feature Extraction, Machine Learning, and Feature Selection, *Sol. Phys.*, p. 404, doi:10.1007/s11207-011-9896-1.

[4] Antos, A., and I. Kontoyiannis (2001), Convergence properties of functional estimates for discrete distributions, *Random Structures & Algorithms*, *19*(3-4), 163–193.

[5] Antos, A., L. Devroye, and L. Györfi (1999), Lower bounds for Bayes error estimation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *21*(7), 643–645.

[6] Avi-Itzhak, H., and T. Diep (1996), Arbitrarily tight upper and lower bounds on the Bayesian probability of error, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(1), 89–91.

[7] Bache, K., and M. Lichman (2013), UCI machine learning repository.

[8] Banerjee, A., S. Merugu, I. S. Dhillon, and J. Ghosh (2005), Clustering with Bregman divergences, *The Journal of Machine Learning Research*, *6*, 1705–1749.

[9] Barkman, J. (1958), Phytosociology and ecology of cryptogamic epiphytes, including a taxonomic survey and description of their vegetation units in europe, *Assen. Van Gorcum.*

[10] Barnes, G., K. D. Leka, E. A. Schumer, and D. J. Della-Rose (2007), Probabilistic forecasting of solar flares from vector magnetogram data, *Space Weather*, *5*, S09002, doi:10.1029/2007SW000317.

[11] Basseville, M. (2013), Divergence measures for statistical data processing–An annotated bibliography, *Signal Processing*, *93*(4), 621–633.

[12] Bazot, C., N. Dobigeon, J.-Y. Tourneret, A. Zaas, G. Ginsburg, and A. O Hero III (2013), Unsupervised bayesian linear unmixing of gene expression microarrays, *BMC Bioinformatics*, *14*(1), 99, doi:10.1186/1471-2105-14-99.

[13] Bellman, R. E. (1961), *Adaptive Control Processes*, Princeton University Press.

[14] Berisha, V., and A. O. Hero III (2015), Empirical non-parametric estimation of the fisher information, *IEEE Signal Processing Letters*, *22*(7), 988–992.

[15] Berisha, V., A. Wisler, A. O. Hero III, and A. Spanias (2015), Empirically estimable classification bounds based on a new divergence measure, *IEEE Transactions on Signal Processing*.

[16] Berlinet, A., L. Devroye, and L. Györfi (1995), Asymptotic normality of L1 error in density estimation, *Statistics*, *26*, 329–343.

[17] Berlinet, A., L. Györfi, and I. Dénes (1997), Asymptotic normality of relative entropy in multivariate density estimation, *Publications de l'Institut de Statistique de l'Université de Paris*, *41*, 3–27.

[18] Bhattacharyya, A. (1946), On a measure of divergence between two multinomial populations, *Sankhyā: The Indian Journal of Statistics*, pp. 401–406.

[19] Bickel, P. J., and M. Rosenblatt (1973), On some global measures of the deviations of density function estimates, *The Annals of Statistics*, pp. 1071–1095.

[20] Bioucas-Dias, J. M., A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot (2012), Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches, *IEEE J. Sel. Topics Appl. Earth Observations Remote Sensing*, *5*(2), 354–379.

[21] Birgé, L., and P. Massart (1995), Estimation of integral functionals of a density, *The Annals of Statistics*, *23*(1), 11–29.

[22] Blanco, J., M. Stead, A. Krieger, J. Viventi, W. Marsh, K. Lee, G. Worrell, and B. Litt (2010), Unsupervised classification of high-frequency oscillations in human neocortical epilepsy and control patients, *J Neurophysiol*, *104*, 2900–2912.

[23] Blanco, J., et al. (2011), Data mining neocortical high-frequency oscillations in epilepsy and controls, *Brain*, *134*, 2948–2959.

[24] Bloomfield, D. S., P. A. Higgins, R. T. J. McAteer, and P. T. Gallagher (2012), Toward Reliable Benchmarking of Solar Flare Forecasting Methods, *ApJ*, *747*, L41, doi:10.1088/2041-8205/747/2/L41.

[25] Bobra, M. G., and S. Couvidat (2015), Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-learning Algorithm, *ApJ*, *798*, 135, doi:10.1088/0004-637X/798/2/135.

[26] Bornmann, P. L., and D. Shaw (1994), Flare rates and the mcintosh active-region classifications, *Solar physics*, *150*(1-2), 127–146.

[27] Borovsky, J. E. (2014), Canonical correlation analysis of the combined solar wind and geomagnetic index data sets, *Journal of Geophysical Research (Space Physics)*, *119*, 5364–5381, doi:10.1002/2013JA019607.

[28] Bruzzone, L., F. Roli, and S. B. Serpico (1995), An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection, *Geoscience and Remote Sensing, IEEE Transactions on*, *33*(6), 1318–1321.

[29] Bühlmann, P., and S. Van De Geer (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

[30] Cadavid, A. C., J. K. Lawrence, and A. Ruzmaikin (2008), Principal Components and Independent Component Analysis of Solar and Space Data, *Sol. Phys.*, *248*, 247–261, doi:10.1007/s11207-007-9026-2.

[31] Carter, K. M., R. Raich, and A. O. Hero III (2010), On local intrinsic dimension estimation and its applications, *Signal Processing, IEEE Transactions on*, *58*(2), 650–663.

[32] Chai, B., D. Walther, D. Beck, and L. Fei-Fei (2009), Exploring functional connectivities of the human brain using multivariate information analysis, in *Advances in neural information processing systems*, pp. 270–278.

[33] Chernoff, H. (1952), A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *The Annals of Mathematical Statistics*, pp. 493–507.

[34] Cho, J., D. Koo, E. Joo, D. Seo, S. Hong, P. Jiruska, and S. Hong (2014), Resection of individually identified high-rate high-frequency oscillations region is associated with favorable outcome in neocortical epilepsy, *Epilepsia*, *55*, 1872–83.

[35] Colak, T., and R. Qahwaji (2008), Automated McIntosh-Based Classification of Sunspot Groups Using MDI Images, *Solar Physics*, *248*, 277–296, doi:10.1007/s11207-007-9094-3.

[36] Colak, T., and R. Qahwaji (2009), Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares, *Space Weather*, *7*, S06001, doi:10.1029/2008SW000401.

[37] Comon, P., and C. Jutten (2010), *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*, Academic Press, Oxford.

[38] Conlon, P. A., P. T. Gallagher, R. T. J. McAteer, J. Ireland, C. A. Young, P. Kestener, R. J. Hewett, and K. Maguire (2008), Multifractal Properties of Evolving Active Regions, *Sol. Phys.*, *248*, 297–309, doi:10.1007/s11207-007-9074-7.

[39] Conlon, P. A., R. T. J. McAteer, P. T. Gallagher, and L. Fennell (2010), Quantifying the Evolving Magnetic Structure of Active Regions, *ApJ*, *722*, 577–585, doi:10.1088/0004-637X/722/1/577.

[40] Costa, J. A., and A. O. Hero III (2006), Determining intrinsic dimension and entropy of high-dimensional shape spaces, in *Statistics and Analysis of Shapes*, pp. 231–252, Springer.

[41] Cover, T. M., and J. A. Thomas (2012), *Elements of information theory*, John Wiley & Sons.

[42] Csiszar, I. (1967), Information-type measures of difference of probability distributions and indirect observations, *Studia Sci. MAth. Hungar.*, *2*, 299–318.

[43] Darbellay, G. A., I. Vajda, et al. (1999), Estimation of the information by an adaptive partitioning of the observation space, *IEEE Trans. Information Theory*, *45*(4), 1315–1321.

[44] DeForest, C. (2004), On re-sampling of solar images, *Solar Physics*, *219*(1), 3–23.

[45] Dhillon, I. S., S. Mallela, and R. Kumar (2003), A divisive information theoretic feature clustering algorithm for text classification, *The Journal of Machine Learning Research*, *3*, 1265–1287.

[46] Ding, C., T. Li, and M. I. Jordan (2010), Convex and semi-nonnegative matrix factorizations, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(1), 45–55.

[47] Dobigeon, N., J.-Y. Tourneret, C. Richard, J. Bermudez, S. Mclaughlin, and A. O. Hero (2014), Nonlinear unmixing of hyperspectral images: Models and algorithms, *Signal Processing Magazine, IEEE*, *31*(1), 82–94.

[48] Doquire, G., B. Frénay, M. Verleysen, et al. (2013), Risk estimation and feature selection, in *European Symposium on Artificial Neural Networks (ESANN 2013)*.

[49] Dudok de Wit, T., S. Moussaoui, C. Guennou, F. Auchère, G. Cessateur, M. Kretzschmar, L. A. Vieira, and F. F. Goryaev (2013), Coronal Temperature Maps from Solar EUV Images: A Blind Source Separation Approach, *Sol. Phys.*, *283*, 31–47, doi:10.1007/s11207-012-0142-2.

[50] Dudok DeWit, T., and F. Auchère (2007), Multispectral analysis of solar EUV images: linking temperature to morphology, *A&A*, *466*, 347–355, doi:10.1051/0004-6361:20066764.

[51] Durrett, R. (2010), *Probability: Theory and Examples*, Cambridge University Press.

[52] Edelman, A., T. A. Arias, and S. T. Smith (1998), The geometry of algorithms with orthogonality constraints, *SIAM journal on Matrix Analysis and Applications*, *20*(2), 303–353.

[53] Efron, B., and C. Stein (1981), The jackknife estimate of variance, *The Annals of Statistics*, pp. 586–596.

[54] Elad, M., and M. Aharon (2006), Image denoising via sparse and redundant representations over learned dictionaries, *Image Processing, IEEE Transactions on*, *15*(12), 3736–3745, doi:10.1109/TIP.2006.881969.

[55] Evans, L. C. (2010), *Partial differential equations*, American Mathematical Society.

[56] Falconer, D. A., R. L. Moore, and G. A. Gary (2008), Magnetogram Measures of Total Nonpotentiality for Prediction of Solar Coronal Mass Ejections from Active Regions of Any Degree of Magnetic Complexity, *ApJ*, *689*, 1433–1442, doi:10.1086/591045.

[57] Fano, R. M. (1968), *Transmission of Information: A Statistical Theory of Communications*, Massachusetts Institute of technology.

[58] Fisher, R. (1936), The use of multiple measurements in taxonomical problems, *Annals of Eugenics*, *7*(2), 179–188.

[59] Fisher, R. A. (1936), The use of multiple measurements in taxonomic problems, *Annals of eugenics*, *7*(2), 179–188.

[60] Friedman, J. H., and L. C. Rafsky (1979), Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests, *The Annals of Statistics*, pp. 697–717.

[61] Frigyik, B., M. R. Gupta, et al. (2012), Bounds on the bayes error given moments, *Information Theory, IEEE Transactions on*, *58*(6), 3606–3612.

[62] Fukunaga, K., and L. D. Hostetler (1973), Optimization of k nearest neighbor density estimates, *Information Theory, IEEE Transactions on*, *19*(3), 320–326.

[63] Fukunaga, K., and D. M. Hummels (1987), Bayes error estimation using parzen and k-nn procedures, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (5), 634–643.

[64] Fukunaga, K., and D. M. Hummels (1987), Bias of nearest neighbor error estimates, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1), 103–112.

[65] Fukunaga, K., and D. M. Hummels (1989), Leave-one-out procedures for nonparametric error estimates, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *11*(4), 421–423.

[66] Gallagher, P. T., Y.-J. Moon, and H. Wang (2002), Active-Region Monitoring and Flare Forecasting I. Data Processing and First Results, *Sol. Phys.*, *209*, 171–183, doi:10.1023/A:1020950221179.

[67] Galluccio, L., O. Michel, P. Comon, M. Kliger, and A. O. Hero III (2013), Clustering with a new distance measure based on a dual-rooted tree, *Information Sciences*, *251*, 96–113.

[68] Gao, S., G. Ver Steeg, and A. Galstyan (2015), Efficient estimation of mutual information for strongly dependent variables, in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 277–286.

[69] Gao, W., S. Oh, and P. Viswanath (2016), Demystifying fixed k-nearest neighbor information estimators, *arXiv preprint arXiv:1604.03006*.

[70] Georgoulis, M. K. (2005), Turbulence In The Solar Atmosphere: Manifestations And Diagnostics Via Solar Image Processing, *Sol. Phys.*, *228*, 5–27, doi:10.1007/s11207-005-2513-4.

[71] Georgoulis, M. K., and D. M. Rust (2007), Quantitative Forecasting of Major Solar Flares, *ApJ*, *661*, L109–L112, doi:10.1086/518718.

[72] Gilbarg, D., and N. S. Trudinger (2001), *Elliptic partial differential equations of second order*, Springer.

[73] 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.

[74] Gliske, S., Z. Irwin, C. Chestek, and W. Stacey (2015), Automated identification of seizure onset zone with a universal detector of high frequency oscillations, *Clin. Neurophysiol.*, *INSERT*.

[75] Gliske, S. V., K. R. Moon, W. C. Stacey, and A. O. Hero III (2016), The intrinsic value of HFO features as a biomarker of epileptic activity, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 6290–6294.

[76] Goria, M. N., N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi (2005), A new class of random vector entropy estimators and its applications in testing statistical hypotheses, *Nonparametric Statistics*, *17*(3), 277–297.

[77] Guo, J., H. Zhang, O. V. Chumak, and Y. Liu (2006), A Quantitative Study on Magnetic Configuration for Active Regions, *Sol. Phys.*, *237*, 25–43, doi: 10.1007/s11207-006-2081-2.

[78] Gut, A. (2012), *Probability: A Graduate Course*, Springer Science & Business Media.

[79] Győri, L., T. Baranyi, and A. Ludmány (2010), Photospheric data programs at the Debrecen Observatory, *Proceedings of the International Astronomical Union*, *6*(S273), 403–407.

[80] Haegelen, C., et al. (2013), Highfrequency oscillations, extent of surgical resection, and surgical outcome in drugresistant focal epilepsy, *Epilepsia*, *54*, 848–57.

[81] Hale, G. E., F. Ellerman, S. B. Nicholson, and A. H. Joy (1919), The Magnetic Polarity of Sun-Spots, *ApJ*, *49*, 153, doi:10.1086/142452.

[82] Hamza, A. B., and H. Krim (2003), Image registration and segmentation by maximizing the Jensen-Rényi divergence, in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 147–163, Springer.

[83] Hansen, B. E. (2009), Lecture notes on nonparametrics.

[84] Härdle, W., and L. Simar (2007), *Applied multivariate statistical analysis*, Springer.

[85] Hashlamoun, W. A., P. K. Varshney, and V. Samarasooriya (1994), A tight upper bound on the Bayesian probability of error, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(2), 220–224.

[86] Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning*, vol. 2, Springer.

[87] Hellinger, E. (1909), Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen., *Journal für die reine und angewandte Mathematik*, *136*, 210–271.

[88] Hellman, M., and J. Raviv (1970), Probability of error, equivocation, and the chernoff bound, *IEEE Transactions on Information Theory*, *16*(4).

[89] Hero, A., and B. Rajaratnam (2011), Large-scale correlation screening, *Journal of the American Statistical Association*, *106*(496), 1540–1552.

[90] Hero III, A. O., B. Ma, O. Michel, and J. Gorman (2002), Applications of entropic spanning graphs, *Signal Processing Magazine, IEEE*, *19*(5), 85–95.

[91] Hewett, R. J., P. T. Gallagher, R. T. J. McAteer, C. A. Young, J. Ireland, P. A. Conlon, and K. Maguire (2008), Multiscale Analysis of Active Region Evolution, *Sol. Phys.*, *248*, 311–322, doi:10.1007/s11207-007-9028-0.

[92] Higgins, P. A., P. T. Gallagher, R. McAteer, and D. S. Bloomfield (2011), Solar magnetic feature detection and tracking for space weather monitoring, *Advances in Space Research*, *47*(12), 2105–2117.

[93] Hild, K. E., D. Erdogmus, and J. C. Principe (2001), Blind source separation using Renyi's mutual information, *Signal Processing Letters, IEEE*, *8*(6), 174–176.

[94] Holappa, L., K. Mursula, T. Asikainen, and I. G. Richardson (2014), Annual fractions of high-speed streams from principal component analysis of local geomagnetic activity, *Journal of Geophysical Research (Space Physics)*, *119*, 4544–4555, doi:10.1002/2014JA019958.

[95] Hotelling, H. (1936), Relations between two sets of variates, *Biometrika*, *28*, 321–377.

[96] Huang, X., D. Yu, Q. Hu, H. Wang, and Y. Cui (2010), Short-Term Solar Flare Prediction Using Predictor Teams, *Sol. Phys.*, *263*, 175–184, doi:10.1007/s11207-010-9542-3.

[97] Inglada, J. (2003), Change detection on sar images by using a parametric estimation of the kullback-leibler divergence, in *Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International*, vol. 6, pp. 4104–4106, IEEE.

[98] Ireland, J., C. A. Young, R. T. J. McAteer, C. Whelan, R. J. Hewett, and P. T. Gallagher (2008), Multiresolution Analysis of Active Region Magnetic Structure and its Correlation with the Mount Wilson Classification and Flaring Activity, *Solar Physics*, *252*, 121–137, doi:10.1007/s11207-008-9233-5.

[99] Jebara, T., R. Kondor, and A. Howard (2004), Probability product kernels, *The Journal of Machine Learning Research*, *5*, 819–844.

[100] Jolliffe, I. T. (2002), *Principal Component Analysis, 2nd edition*, 487 p pp., Springer-Verlag New York, Inc., New York.

[101] Kaiser, J. (1990), On a simple algorithm to calculate the energy of a signal, in *IEEE Int. Conf. Acoustic Speech Signal Process*, IEEE.

[102] Kandasamy, K., A. Krishnamurthy, B. Poczos, L. Wasserman, and J. Robins (2015), Nonparametric von mises estimators for entropies, divergences and mutual informations, in *Advances in Neural Information Processing Systems*, pp. 397–405.

[103] Keener, R. (2010), *Theoretical Statistics: Topics for a Core Course*, Springer Science & Business Media.

[104] Kerber, K., M. Dmpelmann, B. Schelter, P. Le Van, R. Korinthenberg, A. Schulze-Bonhage, and J. Jacobs (2014), Differentiation of specific ripple patterns helps to identify epileptogenic areas for surgical procedures, *Clin. Neurophysiol.*, *125*, 1339–1345.

[105] Kestener, P., P. A. Conlon, A. Khalil, L. Fennell, R. T. J. McAteer, P. T. Gallagher, and A. Arneodo (2010), Characterizing Complexity in Solar Magnetogram Data Using a Wavelet-based Segmentation Method, *ApJ*, *717*, 995–1005, doi:10.1088/0004-637X/717/2/995.

[106] Khan, S., S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov (2007), Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data, *Physical Review E*, *76*(2), 026,209.

[107] Kohavi, R., and G. John (1997), Wrappers for feature subset selection, *Artificial intelligence*, *97*(1), 273–324.

[108] Korzhik, V., and I. Fedyanin (2015), Steganographic applications of the nearest-neighbor approach to Kullback-Leibler divergence estimation, in *Digital Information, Networking, and Wireless Communications (DINWC), 2015 Third International Conference on*, pp. 133–138, IEEE.

[109] Kozachenko, L., and N. N. Leonenko (1987), Sample estimate of the entropy of a random vector, *Problemy Peredachi Informatsii*, *23*(2), 9–16.

[110] Kraskov, A., H. Stögbauer, and P. Grassberger (2004), Estimating mutual information, *Physical review E*, *69*(6), 066,138.

[111] Krishnamurthy, A., K. Kandasamy, B. Poczos, and L. Wasserman (2014), Nonparametric estimation of renyi divergence and friends, in *Proceedings of The 31st International Conference on Machine Learning*, pp. 919–927.

[112] Kruskal, J. B., and M. Wish (1978), *Multidimensional Scaling*, vol. 11, Sage.

[113] Kullback, S., and R. A. Leibler (1951), On information and sufficiency, *The Annals of Mathematical Statistics*, *22*(1), 79–86.

[114] Kwak, N., and C.-H. Choi (2002), Input feature selection by mutual information based on parzen window, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(12), 1667–1671.

[115] Langville, A. N., C. D. Meyer, R. Albright, J. Cox, and D. Duling (2006), Initializations for the nonnegative matrix factorization, in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 23–26, Citeseer.

[116] Laurent, B. (1996), Efficient estimation of integral functionals of a density, *The Annals of Statistics*, *24*(2), 659–681.

[117] Lawrence, J. K., A. Cadavid, and A. Ruzmaikin (2004), Principal Component Analysis of the Solar Magnetic Field I: The Axisymmetric Field at the Photosphere, *Sol. Phys.*, *225*, 1–19, doi:10.1007/s11207-004-3257-2.

[118] Le, T. K. (2013), Information dependency: Strong consistency of Darbellay–Vajda partition estimators, *Journal of Statistical Planning and Inference*, *143*(12), 2089–2100.

[119] Lee, D. D., and H. S. Seung (2001), Algorithms for non-negative matrix factorization, in *Advances in Neural Information Processing Systems (NIPS)*, pp. 556–562.

[120] Lee, K., Y.-J. Moon, J.-Y. Lee, K.-S. Lee, and H. Na (2012), Solar Flare Occurrence Rate and Probability in Terms of the Sunspot Classification Supplemented with Sunspot Area and Its Changes, *Sol. Phys.*, *281*, 639–650, doi: 10.1007/s11207-012-0091-9.

[121] Leka, K. D., and G. Barnes (2004), Photospheric Magnetic Field Properties of Flaring vs. Flare-Quiet Active Regions III: Discriminant Analysis of a Statistically Significant Database, in *American Astronomical Society Meeting Abstracts #204*, *Bulletin of the American Astronomical Society*, vol. 36, p. 715.

[122] Levina, E., and P. J. Bickel (2004), Maximum likelihood estimation of intrinsic dimension, in *Advances in Neural Information Processing Systems*, pp. 777–784.

[123] Lewi, J., R. Butera, and L. Paninski (2006), Real-time adaptive information-theoretic optimization of neurophysiology experiments, in *Advances in Neural Information Processing Systems*, pp. 857–864.

[124] Li, S. (2011), Concise formulas for the area and volume of a hyperspherical cap, *Asian Journal of Mathematics and Statistics*, *4*(1), 66–70.

[125] Lin, C.-J. (2007), Projected gradient methods for nonnegative matrix factorization, *Neural computation*, *19*(10), 2756–2779.

[126] Liu, G., G. Xia, W. Yang, and N. Xue (2014), SAR image segmentation via nonlocal active contours, in *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, pp. 3730–3733, IEEE.

[127] Liu, H., L. Wasserman, and J. D. Lafferty (2012), Exponential concentration for mutual information estimation with application to forests, in *Advances in Neural Information Processing Systems*, pp. 2537–2545.

[128] Loftsgaarden, D. O., and C. P. Quesenberry (1965), A nonparametric estimate of a multivariate density function, *The Annals of Mathematical Statistics*, *36*(3), 1049–1051.

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

[130] Mallat, S., and Z. Zhang (1993), Matching pursuits with time-frequency dictionaries, *Signal Processing, IEEE Transactions on*, *41*(12), 3397–3415, doi:10.1109/78.258082.

[131] Matsumoto, A., B. Brinkmann, S. Stead, J. Matsumoto, M. Kucewicz, W. Marsh, F. Meyer, and G. Worrell (2013), Pathological and physiological high-frequency oscillations in focal human epilepsy, *J. Neurophysiol.*, *110*, 1958–64.

[132] Mayfield, E. B., and J. K. Lawrence (1985), The correlation of solar flare production with magnetic energy in active regions, *Sol. Phys.*, *96*, 293–305, doi:10.1007/BF00149685.

[133] McAteer, R. T. J., P. T. Gallagher, and J. Ireland (2005), Statistics of Active Region Complexity: A Large-Scale Fractal Dimension Survey, *ApJ*, *631*, 628–635, doi:10.1086/432412.

[134] McAteer, R. T. J., P. T. Gallagher, and P. A. Conlon (2010), Turbulence, complexity, and solar flares, *Advances in Space Research*, *45*, 1067–1074, doi:10.1016/j.asr.2009.08.026.

[135] McIntosh, P. S. (1990), The classification of sunspot groups, *Sol. Phys.*, *125*, 251–267, doi:10.1007/BF00158405.

[136] Mihoko, M., and S. Eguchi (2002), Robust blind source separation by beta divergence, *Neural computation*, *14*(8), 1859–1886.

[137] Mittelman, R., N. Dobigeon, and A. Hero (2012), Hyperspectral image unmixing using a multiresolution sticky hdp, *Signal Processing, IEEE Transactions on*, *60*(4), 1656–1671, doi:10.1109/TSP.2011.2180718.

[138] Mohamed, S., and D. J. Rezende (2015), Variational information maximisation for intrinsically motivated reinforcement learning, in *Advances in Neural Information Processing Systems*, pp. 2116–2124.

[139] Moon, K., V. Delouille, and A. O. Hero III (2015), Meta learning of bounds on the Bayes classifier error, in *IEEE Signal Processing and SP Education Workshop*, pp. 13–18, IEEE.

[140] Moon, K. R., and A. O. Hero III (2014), Ensemble estimation of multivariate f-divergence, in *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 356–360, IEEE.

[141] Moon, K. R., and A. O. Hero III (2014), Multivariate f-divergence estimation with confidence, in *Advances in Neural Information Processing Systems*, pp. 2420–2428.

[142] Moon, K. R., J. J. Li, V. Delouille, F. Watson, and A. O. Hero III (2014), Image patch analysis and clustering of sunspots: A dimensionality reduction approach, in *IEEE International Conference on Image Processing*, pp. 1623–1627, IEEE.

[143] Moon, K. R., V. Delouille, J. J. Li, R. De Visscher, F. Watson, and A. O. Hero III (2016), Image patch analysis of sunspots and active regions. II. Clustering via matrix factorization, *Journal of Space Weather and Space Climate*, *6*(A3).

[144] Moon, K. R., J. J. Li, V. Delouille, R. De Visscher, F. Watson, and A. O. Hero III (2016), Image patch analysis of sunspots and active regions. I. Intrinsic dimension and correlation analysis, *Journal of Space Weather and Space Climate*, *6*(A2).

[145] Moon, K. R., K. Sricharan, K. Greenewald, and A. O. Hero III (2016), Nonparametric ensemble estimation of distributional functionals, *arXiv preprint arXiv:1601.06884v2*.

[146] Moon, K. R., K. Sricharan, K. Greenewald, and A. O. Hero III (2016), Improving convergence of divergence functional ensemble estimators, in *2016 IEEE International Symposium on Information Theory (ISIT)*.

[147] Moon, T. K., and W. C. Stirling (2000), *Mathematical Methods and Algorithms for Signal Processing*, Prentice hall New York.

[148] Moreno, P. J., P. P. Ho, and N. Vasconcelos (2003), A kullback-leibler divergence based kernel for svm classification in multimedia applications, in *Advances in neural information processing systems*.

[149] Muandet, K., K. Fukumizu, F. Dinuzzo, and B. Schölkopf (2012), Learning from distributions via support measure machines, in *Advances in neural information processing systems*, pp. 10–18.

[150] Muller, K. E. (1982), Understanding canonical correlation through the general linear model and principal components, *American Statistics*, *36*, 342?354.

[151] Munkres, J. (2000), *Topology*, Prentice Hall, Inc.

[152] 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.

[153] Nimon, K., R. Henson, and M. Gates (2010), Revisiting interpretation of canonical correlation analysis: A tutorial and demonstration of canonical commonality analysis, *Multivar. Behav.*, *45*, 702?724.

[154] Ochiai, A. (1957), Zoogeographical studies on the soleoid fishes found japan and its neighboring regions. ii, *Bull. Jap. Soc. Sci. Fish*, *22*(9), 526–530.

[155] Oliva, J., B. Póczos, and J. Schneider (2013), Distribution to distribution regression, in *Proceedings of The 30th International Conference on Machine Learning*, pp. 1049–1057.

[156] Pál, D., B. Póczos, and C. Szepesvári (2010), Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs, in *Advances in Neural Information Processing Systems*, pp. 1849–1857.

[157] Park, S., S. Lee, H. Che, and C. CK (2012), Ictal high-gamma oscillation (60-99 hz) in intracranial electroencephalography and postoperative seizure outcome in neocortical epilepsy, *Clin. Neurophysiol.*, *123*(6), 1100–10.

[158] Pearce, A., D. Wulsin, J. Blanco, A. Krieger, B. Litt, and W. Stacey (2013), Temporal changes of neocortical high-frequency oscillations in epilepsy, *J. Neurophysiol*, *110*, 1167–1179.

[159] Peng, H., F. Long, and C. Ding (2005), Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *27*(8), 1226–1238.

[160] Phillips, K. J. H. (1991), Solar flares - A review, *Vistas in Astronomy*, *34*, 353–365, doi:10.1016/0083-6656(91)90014-J.

[161] Póczos, B., and J. G. Schneider (2011), On the estimation of alpha-divergences, in *International Conference on Artificial Intelligence and Statistics*, pp. 609–617.

[162] Póczos, B., L. Xiong, and J. Schneider (2011), Nonparametric divergence estimation with applications to machine learning on distributions, in *UAI*.

[163] Póczos, B., A. Rinaldo, A. Singh, and L. Wasserman (2012), Distribution-free distribution regression, *AISTATS*.

[164] Prim, R. C. (1957), Shortest connection networks and some generalizations, *Bell system technical journal*, *36*(6), 1389–1401.

[165] Principe, J. C. (2010), *Information theoretic learning: Renyi's entropy and kernel perspectives*, Springer Science & Business Media.

[166] Ramírez, I., and G. Sapiro (2012), An mdl framework for sparse coding and dictionary learning, *Signal Processing, IEEE Transactions on*, *60*(6), 2913–2927.

[167] Rand, W. M. (1971), Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, *66*(336), 846–850.

[168] Rast, M. P. (2003), The scales of granulation, mesogranulation, and supergranulation, *The Astrophysical Journal*, *597*(2), 1200.

[169] Rieutord, M., T. Roudier, J. Malherbe, and F. Rincon (2000), On mesogranulation, network formation and supergranulation, *Astronomy and Astrophysics*, *357*, 1063–1072.

[170] Riordan, J. (1937), Moment recurrence relations for binomial, poisson and hypergeometric frequency distributions, *The Annals of Mathematical Statistics*, *8*(2), 103–111.

[171] Rousseeuw, P. J. (1987), Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, *20*, 53–65.

[172] Salge, C., C. Glackin, and D. Polani (2014), Changing the environment based on empowerment as intrinsic motivation, *Entropy*, *16*(5), 2789–2819.

[173] Sammis, I., F. Tang, and H. Zirin (2000), The dependence of large flare occurrence on the magnetic structure of sunspots, *The Astrophysical Journal*, *540*(1), 583.

[174] Scherrer, P. H., et al. (1995), The Solar Oscillations Investigation - Michelson Doppler Imager, *Solar Physics*, *162*, 129–188, doi:10.1007/BF00733429.

[175] Schneidman, E., W. Bialek, and M. J. B. II (2003), An information theoretic approach to the functional classification of neurons, *Advances in Neural Information Processing Systems*, *15*, 197–204.

[176] Schrijver, C. J. (2007), A Characteristic Magnetic Field Pattern Associated with All Major Solar Flares and Its Use in Flare Forecasting, *ApJ*, *655*, L117–L120, doi:10.1086/511857.

[177] Seichepine, N., S. Essid, C. Févotte, and O. Cappé (2014), Soft nonnegative matrix co-factorization, *Signal Processing, IEEE Transactions on*, *62*(22), 5940–5949.

[178] Sethian, J. A. (1995), A fast marching level set method for monotonically advancing fronts, in *Proc. Nat. Acad. Sci*, pp. 1591–1595.

[179] Silva, J., and S. S. Narayanan (2010), Information divergence estimation based on data-dependent partitions, *Journal of Statistical Planning and Inference*, *140*(11), 3180–3198.

[180] Singh, S., and B. Póczos (2014), Exponential concentration of a density functional estimator, in *Advances in Neural Information Processing Systems*, pp. 3032–3040.

[181] Singh, S., and B. Póczos (2014), Generalized exponential concentration inequality for rényi divergence estimation, in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 333–341.

[182] Smola, A., A. Gretton, L. Song, and B. Schölkopf (2007), A hilbert space embedding for distributions, in *Algorithmic Learning Theory*, pp. 13–31, Springer.

[183] Song, H., C. Tan, J. Jing, H. Wang, V. Yurchyshyn, and V. Abramenko (2009), Statistical Assessment of Photospheric Magnetic Features in Imminent Solar Flare Predictions, *Sol. Phys.*, *254*, 101–125, doi:10.1007/s11207-008-9288-3.

[184] Sricharan, K. (2012), Neighborhood graphs for estimation of density functionals, Ph.D. thesis, UNIVERSITY OF MICHIGAN.

[185] Sricharan, K., and A. O. Hero (2011), Efficient anomaly detection using bipartite k-nn graphs, in *Advances in Neural Information Processing Systems*, pp. 478–486.

[186] Sricharan, K., R. Raich, and A. O. Hero (2012), Estimation of nonlinear functionals of densities with confidence, *IEEE Trans. Information Theory*, *58*(7), 4135–4159.

[187] Sricharan, K., D. Wei, and A. O. Hero (2013), Ensemble estimators for multivariate entropy estimation, *Information Theory, IEEE Transactions on*, *59*(7), 4374–4388.

[188] Stenning, D. C., T. C. M. Lee, D. A. van Dyk, V. Kashyap, J. Sandell, and C. A. Young (2013), Morphological feature extraction for statistical learning with applications to solar image data, *Statistical Analysis and Data Mining*, *6*(4), 329–345, doi:10.1002/sam.11200.

[189] Stewart, G. W. (1973), Error and perturbation bounds for subspaces associated with certain eigenvalue problems, *SIAM Review*, *15*(4), 727–764.

[190] Stone, C. J. (1977), Consistent nonparametric regression, *The annals of statistics*, pp. 595–620.

[191] Suzuki, T., M. Sugiyama, J. Sese, and T. Kanamori (2008), Approximating mutual information by maximum likelihood density ratio estimation., *FSDM*, *4*, 5–20.

[192] Tiwari, S. K., M. van Noort, A. Lagg, and S. K. Solanki (2013), Structure of sunspot penumbral filaments: a remarkable uniformity of properties, *Astronomy & Astrophysics*, *557*, A25.

[193] Torkkola, K. (2003), Feature extraction by non parametric mutual information maximization, *The Journal of Machine Learning Research*, *3*, 1415–1438.

[194] Tumer, K., and J. Ghosh (2003), Bayes error rate estimation using classifier ensembles, *International Journal of Smart Engineering System Design*, *5*(2), 95–109.

[195] Vemuri, B. C., M. Liu, S. Amari, and F. Nielsen (2011), Total Bregman divergence and its applications to DTI analysis, *Medical Imaging, IEEE Transactions on*, *30*(2), 475–483.

[196] Vergara, J. R., and P. A. Estévez (2014), A review of feature selection methods based on mutual information, *Neural Computing and Applications*, *24*(1), 175–186.

[197] Wagenaar, J., G. Worrell, Z. Ives, D. Matthias, B. Litt, and A. Schulze-Bonhage (2015), Collaborating and sharing data in epilepsy research, *J. Clin Neurophysiol.*, *32*, 235–9.

[198] Wang, J., Y. Xu, D. Zhang, and J. You (2010), An efficient method for computing orthogonal discriminant vectors, *Neurocomputing*, *73*, 2168–76.

[199] Wang, Q., S. R. Kulkarni, and S. Verdú (2005), Divergence estimation of continuous distributions based on data-dependent partitions, *IEEE Trans. Information Theory*, *51*(9), 3064–3074.

[200] Wang, Q., S. R. Kulkarni, and S. Verdú (2009), Divergence estimation for multidimensional densities via k-nearest-neighbor distances, *IEEE Trans. Information Theory*, *55*(5), 2392–2405.

[201] Watson, F. T., L. Fletcher, and S. Marshall (2011), Evolution of sunspot properties during solar cycle 23, *Astronomy & Astrophysics*, *533*, A14, doi: 10.1051/0004-6361/201116655.

[202] Worrell, G., K. Jerbi, K. Kobayashi, J. Lina, R. Zelmann, and M. Le Van Quyen (2012), Recording and analysis techniques for high-frequency oscillations, *Prog. Neurobiol.*, *98*, 265–278.

[203] Xuan, G., X. Zhu, P. Chai, Z. Zhang, Y. Q. Shi, and D. Fu (2006), Feature selection based on the bhattacharyya distance, in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3, pp. 1232–1235, IEEE.

[204] Yaghoobi, M., T. Blumensath, and M. E. Davies (2009), Dictionary learning for sparse approximations with the majorization method, *Signal Processing, IEEE Transactions on*, *57*(6), 2178–2191.

[205] Ye, K., and L.-H. Lim (2014. arXiv:1407.0900), Distance between subspaces of different dimensions, *Preprint*.

[206] Yokoya, N., T. Yairi, and A. Iwasaki (2012), Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion, *Geoscience and Remote Sensing, IEEE Transactions on*, *50*(2), 528–537.

[207] Yu, D., X. Huang, H. Wang, Y. Cui, Q. Hu, and R. Zhou (2010), Short-term Solar Flare Level Prediction Using a Bayesian Network Approach, *ApJ*, *710*, 869–877, doi:10.1088/0004-637X/710/1/869.

[208] Zhang, J., and H. Deng (2007), Gene selection for classification of microarray data based on the bayes error, *BMC bioinformatics*, *8*(1), 370.

[209] Zharkova, V. V., S. J. Shepherd, and S. I. Zharkov (2012), Principal component analysis of background and sunspot magnetic field variations during solar cycles 21-23, *MNRAS*, *424*, 2943–2953, doi:10.1111/j.1365-2966.2012.21436.x.