entropy

MDPI

Article

Geometric Estimation of Multivariate Dependency

Salimeh Yasaei Sekeh *,† **and Alfred O. Hero**

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

* Correspondence: salimehy@umich.edu or salimeh.yasaei@maine.edu

† Current address: School of Computing and Information Science, University of Maine, Orono, ME 04469, USA.

Received: 21 May 2019; Accepted: 8 August 2019; Published: 12 August 2019

Abstract: This paper proposes a geometric estimator of dependency between a pair of multivariate random variables. The proposed estimator of dependency is based on a randomly permuted geometric graph (the minimal spanning tree) over the two multivariate samples. This estimator converges to a quantity that we call the geometric mutual information (GMI), which is equivalent to the Henze–Penrose divergence. between the joint distribution of the multivariate samples and the product of the marginals. The GMI has many of the same properties as standard MI but can be estimated from empirical data without density estimation; making it scalable to large datasets. The proposed empirical estimator of GMI is simple to implement, involving the construction of an minimal spanning tree (MST) spanning over both the original data and a randomly permuted version of this data. We establish asymptotic convergence of the estimator and convergence rates of the bias and variance for smooth multivariate density functions belonging to a Hölder class. We demonstrate the advantages of our proposed geometric dependency estimator in a series of experiments.

Keywords: Henze–Penrose mutual information; Friedman–Rafsky test statistic; geometric mutual information; convergence rates; bias and variance tradeoff; optimization; minimal spanning trees

1. Introduction

Estimation of multivariate dependency has many applications in fields such as information theory, clustering, structure learning, data processing, feature selection, time series prediction, and reinforcement learning, see [1–10], respectively. It is difficult to accurately estimate the mutual information in high-dimensional settings, specially where the data is multivariate with an absolutely continuous density with respect to Lebesgue measure—the setting considered in this paper. An important and regular measure of dependency is the Shannon mutual information (MI), which has seen extensive use across many application domains. However, the estimation of mutual information can often be challenging. In this paper, we focus on a measure of MI that we call the Geometric MI (GMI). This MI measure is defined as the asymptotic large sample limit of a randomized minimal spanning tree (MST) statistic spanning the multivariate sample realizations. The GMI is related to a divergence measure called the Henze–Penrose divergence [11,12], and related to the multivariate runs test [13]. In [14,15], it was shown that this divergence measure can be used to specify a tighter bound for the Bayes error rate for testing if a random sample comes from one of two distributions the bound in [14,15] is tighter than previous divergence-type bounds such as the Bhattacharrya bound [16]. Furthermore, the authors of [17] proposed a non-parametric bound on multi-class classification Bayes error rate using a global MST graph.

Let $\mathbf{X}$ and $\mathbf{Y}$ be random variables with unknown joint density f_{XY} and marginal densities f_X and f_Y, respectively, and consider two hypotheses: H_0, $\mathbf{X}$ and $\mathbf{Y}$ are independent and H_1, $\mathbf{X}$ and $\mathbf{Y}$ are dependent,

$$H_0 : f_{XY} = f_X f_Y, \quad \text{versus} \quad H_1 : f_{XY} \neq f_X f_Y.$$

The GMI is defined as the Henze–Penrose divergence between f_{XY} and $f_X f_Y$ which can be used as a dependency measure. In this paper, we prove that for large sample size n the randomized MST statistic spanning the original multivariate sample realizations and a randomly shuffled data set converges almost surely to the GMI measure. A direct implication of [14,15] is that the GMI provides a tighter bound on the Bayes misclassification rate for the optimal test of independence. In this paper, we propose an estimator based on a random permutation modification of the Friedman–Rafsky multivariate test statistic and show that under certain conditions the GMI estimator achieves the parametric mean square error (MSE) rate when the joint density is bounded and smooth. Importantly unlike other measures of MI, our proposed GMI estimator does not require explicit estimation of the joint and marginal densities.

Computational complexity is an important challenge in machine learning and data science. Most plug-in-based estimators, such as the kernel density estimator (KDE) or the K-nearest-neighbor (KNN) estimator with known convergence rate, require runtime complexity of $O(n^2)$, which is not suitable for large scale applications. Noshad et al. proposed a graph theoretic direct estimation method based on nearest-neighbor ratios (NNR) [18]. The NNR estimator is based on k-NN graph and computationally more tractable than other competing estimators with complexity $O(kn \log n)$. The construction of the minimal spanning tree lies at the heart of the GMI estimator proposed in this paper. Since the GMI estimator is based on the Euclidean MST the dual-tree algorithm by March et al. [19] can be applied. This algorithm is based on the construction of Borůvka [20] and implements the Euclidean MST in approximately $O(nlogn)$ time. In this paper, we experimentally show that for large sample size the proposed GMI estimator has faster runtime than the KDE plug-in method.

1.1. Related Work

Estimation of mutual information has a rich history. The most common estimators of MI are based on plug-in density estimation, e.g., using the histogram, kernel density or kNN density estimators [21,22]. Motivated by ensemble methods applied to divergence estimation [23,24], in [22] an ensemble method for combining multiple KDE bandwidths was proposed for estimating MI. Under certain smoothness conditions this ensemble MI estimator was shown to achieve parametric convergence rates.

Another class of estimators of multivariate dependency bypasses the difficult density estimation task. This class includes the statistically consistent estimators of Rényi-α and KL mutual information which are motivated by the asymptotic limit of the length of the KNN graph, [25,26] when joint density is smooth. The estimator of [27] builds on KNN methods for Rényi entropy estimation. The authors of [26], showed that when MI is large the KNN and KDE approaches are ill-suited for estimating MI since the joint density may be insufficiently smooth when there are strong dependencies. To overcome this issue an assumption on the smoothness of the density is required, see [28,29], and [23,24]. For all these methods, the optimal parametric rate of MSE convergence is achieved when the densities are either d, $(d+1)/2$ or $d/2$ times differentiable [30]. In this paper, we assume that joint and marginal densities are smooth in the sense that they belong to Hölder continuous classes of densities $\Sigma_d(\eta, K)$, where the smoothness parameter $\eta \in (0, 1]$ and the Lipschitz constant $K > 0$.

A MI measure based on the Pearson chi-square divergence was considered in [31] that is computational efficient and numerically stable. The authors of [27,32] used nearest-neighbor graph and minimal spanning tree approaches, respectively, to estimate Rényi mutual information. In [22], a non-parametric mutual information estimator was proposed using a weighted ensemble method with $O(1/n)$ parametric convergence rate. This estimator was based on plug-in density estimation, which is challenging in high dimension.

Our proposed dependency estimator differs from previous methods in the following ways. First, it estimates a different measure of mutual information, the GMI. Second, instead of using the KNN graph the estimator of GMI uses a randomized minimal spanning tree that spans the multivariate realizations. The proposed GMI estimator is motivated by the multivariate runs test of Friedman and

Rafsky (FR) [33] which is a multivariate generalization of the univariate Smirnov maximum deviation test [34] and the Wald-Wolfowitz [35] runs test in one dimension. We also emphasize that the proposed GMI estimator does not require boundary correction, in contrast to other graph-based estimators, such as, the NNR estimator [18], scalable MI estimator [36], or cross match statistic [37].

1.2. Contribution

The contribution of this paper has three components

(1) We propose a novel non-parametric multivariate dependency measure, referred to as geometric mutual information (GMI), which is based on graph-based divergence estimation. The geometric mutual information is constructed using a minimal spanning tree and is a function of the Friedman–Rafsky multivariate test statistic.

(2) We establish properties of the proposed dependency measure analogous to those of Shannon mutual information, such as, convexity, concavity, chain rule, and a type of data-processing inequality.

(3) We derive a bound on the MSE rate for the proposed geometric estimator. An advantage of the estimator is that it achieves the optimal MSE rate without the need for boundary correction, which is required for most plug-in estimators.

1.3. Organization

The rest of the paper is organized as follows. In Section 2, we define the geometric mutual information and establish some of its mathematical properties. In Sections 2.2 and 2.3, we introduce a statistically consistent GMI estimator and derive a bound on its mean square error convergence rate. In Section 3 we verify the theory through experiments.

Throughout the paper, we denote statistical expectation by $\mathbb{E}$ and the variance by abbreviation Var. Bold face type indicates random vectors. All densities are assumed to be absolutely continuous with respect to non-atomic Lebesgue measure.

2. The Geometric Mutual Information (GMI)

In this section, we first review the definition of the Henze–Penrose (HP) divergence measure defined by Berisha and Hero in [13,14]. The Henze–Penrose divergence between densities f and g with domain $\mathbb{R}^d$ for parameter $p \in (0,1)$ is defined as follows (see [13–15]):

$$D_p(f,g) = \frac{1}{4pq}\left[\int \frac{\left(pf(\mathbf{x}) - qg(\mathbf{x})\right)^2}{pf(\mathbf{x}) + qg(\mathbf{x})}\,\mathrm{d}\mathbf{x} - (p-q)^2\right], \tag{1}$$

where $q = 1 - p$. This functional is an f-divergence [38], equivalently, as an Ali-Silvey distance [39], i.e., it satisfies the properties of non-negativity, monotonicity, and joint convexity [15]. The measure (1) takes values in $[0,1]$ and $D_p(f,g) = 0$ if and only if $f = g$ almost surely.

The mutual information measure is defined as follows. Let f_X, f_Y, and f_{XY} be the marginal and joint distributions, respectively, of random vectors $\mathbf{X} \in \mathbb{R}^{d_x}$, $\mathbf{Y} \in \mathbb{R}^{d_y}$ where d_x and d_y are positive integers. Then by using (1), a Henze–Penrose generalization of the mutual information between $\mathbf{X}$ and $\mathbf{Y}$, is defined by

$$\begin{aligned} I_p(\mathbf{X};\mathbf{Y}) &= D_p(f_{XY}, f_X f_Y) \\ &= \frac{1}{4pq}\left[\iint \frac{\left(pf_{XY}(\mathbf{x},\mathbf{y}) - qf_X(\mathbf{y})f_Y(\mathbf{y})\right)^2}{pf_{XY}(\mathbf{x},\mathbf{y}) + qf_X(\mathbf{x})f_Y(\mathbf{y})}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} - (p-q)^2\right]. \end{aligned} \tag{2}$$

We will show below that $I_p(\mathbf{X};\mathbf{Y})$ has a geometric interpretation in terms of the large sample limit of a minimal spanning tree spanning n sample realizations of the merged labeled samples $\mathbf{X} \cup \mathbf{Y}$. Thus,

we call $I_p(\mathbf{X};\mathbf{Y})$ the GMI between $\mathbf{X}$ and $\mathbf{Y}$. The GMI satisfies similar properties to other definitions of mutual information, such as Shannon and Rényi mutual information. Recalling (3) in [14], an alternative form of I_p is given by

$$I_p(\mathbf{X};\mathbf{Y}) = 1 - A_p(\mathbf{X};\mathbf{Y}) = \frac{u_p(\mathbf{X};\mathbf{Y})}{4pq} - \frac{(p-q)^2}{4pq}, \tag{3}$$

where

$$\begin{aligned} A_p(\mathbf{X};\mathbf{Y}) &= \iint \frac{f_{XY}(\mathbf{x},\mathbf{y}) f_X(\mathbf{x}) f_Y(\mathbf{y})}{p f_{XY}(\mathbf{x},\mathbf{y}) + q f_X(\mathbf{x}) f_Y(\mathbf{y})}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = \mathbb{E}_{XY}\left[\left(p\, \frac{f_{XY}(\mathbf{X},\mathbf{Y})}{f_X(\mathbf{X}) f_Y(\mathbf{Y})} + q\right)^{-1}\right], \text{ and} \\ u_p(\mathbf{X};\mathbf{Y}) &= \iint \frac{\left(p f_{XY}(\mathbf{x},\mathbf{y}) - q f_X(\mathbf{x}) f_Y(\mathbf{y})\right)^2}{p f_{XY}(\mathbf{x},\mathbf{y}) + q f_X(\mathbf{x}) f_Y(\mathbf{y})}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = 1 - 4pq\; A_p(\mathbf{X};\mathbf{Y}). \end{aligned} \tag{4}$$

The function $A_p(\mathbf{X};\mathbf{Y})$ was defined in [13] and is called the geometric affinity between $\mathbf{X}$ and $\mathbf{Y}$. The next subsection of the paper is dedicated to the basic inequalities and properties of the proposed GMI measure (2).

2.1. Properties of the Geometric Mutual Information

In this subsection we establish basic inequalities and properties of the GMI, I_p, given in (2). The following theorem shows that $I_p(\mathbf{X};\mathbf{Y})$ is a concave function in f_X and a convex function in $f_{Y|X}$. The proof is given in Appendix A.1.

Theorem 1. *Denote by $\tilde{I}_p(f_{XY})$ the GMI $I_p(\mathbf{X};\mathbf{Y})$ when $\mathbf{X} \in \mathbb{R}^{d_x}$ and $\mathbf{Y} \in \mathbb{R}^{d_y}$ have joint density f_{XY}. Then the GMI satisfies*

(i) *Concavity in f_X: Let $f_{Y|X}$ be conditional density of $\mathbf{Y}$ given $\mathbf{X}$ and let g_X and h_X be densities on $\mathbb{R}^{d_x}$. Then for $\lambda_1, \lambda_2 \in [0,1]$, $\lambda_1 + \lambda_2 = 1$*

$$\tilde{I}_p\left(\lambda_1 f_{Y|X} g_X + \lambda_2 f_{Y|X} h_X\right) \geq \lambda_1 \tilde{I}_p(f_{Y|X} g_X) + \lambda_2 \tilde{I}_p(f_{Y|X} h_X). \tag{5}$$

The inequality is strict unless either λ_1 or λ_2 are zero or $h_X = g_X$.

(ii) *Convexity in $f_{Y|X}$: Let $g_{Y|X}$ and $h_{Y|X}$ be conditional densities of $\mathbf{Y}$ given $\mathbf{X}$ and let f_X be marginal density. Then for $\lambda_1, \lambda_2 \in [0,1]$, $\lambda_1 + \lambda_2 = 1$*

$$\tilde{I}_p\left(\lambda_1 g_{Y|X} f_X + \lambda_2 h_{Y|X} f_X\right) \leq \lambda_1 \tilde{I}_p(g_{Y|X} f_X) + \lambda_2 \tilde{I}_p(h_{Y|X} f_X). \tag{6}$$

The inequality is strict unless either λ_1 or λ_2 are zero or $h_{Y|X} = g_{Y|X}$.

The GMI, $I_p(\mathbf{X};\mathbf{Y})$, satisfies properties analogous to the standard chain rule and the data-processing inequality [40]. For random variables $\mathbf{X} \in \mathbb{R}^{d_x}$, $\mathbf{Y} \in \mathbb{R}^{d_y}$, and $\mathbf{Z} \in \mathbb{R}^{d_z}$ with conditional density $f_{XY|Z}$ we define the conditional GMI

$$\begin{aligned} I_p(\mathbf{X};\mathbf{Y}|\mathbf{Z}) &= \mathbb{E}_{\mathbf{Z}}\left[\tilde{I}_p(f_{XY|Z})\right], \quad \text{where} \\ \tilde{I}_p(f_{XY|Z}) &= 1 - \iint \frac{f_{XY|Z}(\mathbf{x},\mathbf{y}|\mathbf{z}) f_{X|Z}(\mathbf{x}|\mathbf{z}) f_{Y|Z}(\mathbf{y}|\mathbf{z})}{p\, f_{XY|Z}(\mathbf{x},\mathbf{y}|\mathbf{z}) + q\, f_{X|Z}(\mathbf{x}|\mathbf{z}) f_{Y|Z}(\mathbf{y}|\mathbf{z})}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}. \end{aligned} \tag{7}$$

The next theorem establishes a relation between the joint and conditional GMI.

Theorem 2. *For given d-dimensional random vector* $\mathbf{X}$ *with components* $X_1, X_2, \ldots, X_d$ *and random variable* Y,

$$I_p(\mathbf{X};Y) \geq I_p(X_1;Y) - \sum_{i=1}^{d-1} \Big(1 - I_p(X_i;Y|\mathbf{X}^{i-1})\Big), \tag{8}$$

where $\mathbf{X}^i := X_1, X_2, \ldots, X_i$ *and the conditional GMI* $I_p(X_i;Y|\mathbf{X}^{i-1})$ *is defined in (7).*

For $d = 2$ Theorem 2 reduces to

$$I_p(X_1, X_2; Y) \geq I_p(X_1;Y) - \big(1 - I_p(X_2;Y|X_1)\big), \tag{9}$$

Please note that when $\sum_{i=1}^{d-1} \Big(1 - I_p(X_i;Y|\mathbf{X}^{i-1})\Big) \geq 1$, the inequality (8) is trivial since $0 \leq I_p(X_1;Y) \leq 1$. The proof of Theorem 2 is given in Appendix A.2. Theorem 2 is next applied to the case where $\mathbf{X}$ and $\mathbf{Y}$ form a Markov chain. The proof of the following "leany" data-processing inequality (Proposition 1) is provided in Appendices section, Appendix A.3.

Proposition 1. *Suppose random vectors* $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ *form a Markov chain denoted,* $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$, *in the sense that* $f_{XYZ} = f_{X|Y} f_{Y|Z} f_Z$. *Then for* $p \in (0,1)$

$$I_p(\mathbf{Y};\mathbf{X}) \geq I_p(\mathbf{Z};\mathbf{X}) - \Big(p\, \mathbb{E}_{XY}\big[\delta_{X,Y}\big] + (1-p)\Big)^{-1}, \tag{10}$$

where

$$\delta_{X,Y} = \int \frac{f_{X|Y}(\mathbf{X}|\mathbf{Y})\, f_{Z|Y}(\mathbf{z}|\mathbf{Y})}{f_{X|Z}(\mathbf{X}|\mathbf{z})}\, \mathrm{d}\mathbf{z}.$$

Furthermore, if both $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ and $\mathbf{X} \to \mathbf{Z} \to \mathbf{Y}$ together hold true, we have $I_p(\mathbf{Y};\mathbf{X}) = I_p(\mathbf{Z};\mathbf{X})$.

The inequality in (10) becomes interpretable as the standard data-processing inequality $I_p(\mathbf{Y};\mathbf{X}) \geq I_p(\mathbf{Z};\mathbf{X})$, when

$$\mathbb{E}_Z\left[\frac{f(\mathbf{Z}|\mathbf{Y})}{f(\mathbf{Z}|\mathbf{X})}\right] = \infty,$$

since

$$\mathbb{E}_{XY}\big[\delta_{X,Y}\big] = \mathbb{E}_{XY}\left(\frac{f(\mathbf{X}|\mathbf{Y})}{f(\mathbf{X})}\mathbb{E}_Z\left[\frac{f(\mathbf{Z}|\mathbf{Y})}{f(\mathbf{Z}|\mathbf{X})}\right]\right).$$

2.2. The Friedman–Rafsky Estimator

Let a random sample $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ from $f_{XY}(\mathbf{x},\mathbf{y})$ be available. Here we show that the GMI $I_p(\mathbf{X};\mathbf{Y})$ can be directly estimated without estimating the densities. The estimator is inspired by the MST construction of [33] that provides a consistent estimate of the Henze–Penrose divergence [14,15]. We denote by $\mathbf{z}_i$ the i-th joint sample $\mathbf{x}_i, \mathbf{y}_i$ and by $\mathcal{Z}_n$ the sample set $\{\mathbf{z}_i\}_{i=1}^n$. Divide the sample set $\mathcal{Z}_n$ into two subsets $\mathcal{Z}'_{n'}$ and $\mathcal{Z}''_{n''}$ with the proportion $\alpha = n'/n$ and $\beta = n''/n$, where $\alpha + \beta = 1$.

Denote by $\widetilde{\mathcal{Z}}_{n''}$ the set

$$\big\{(\mathbf{x}_{i_k}, \mathbf{y}_{j_k}),\ k = 1, \ldots, n'',\ \text{selected at random from } \mathcal{Z}''_{n''}\big\}:$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n''} \end{pmatrix} \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n''} \end{pmatrix}$$

This means that for each $\mathbf{z}_{ik} = (\mathbf{x}_{ik}, \mathbf{y}_{ik}) \in \mathcal{Z}''_{n''}$ given the first element $\mathbf{x}_{ik}$ the second element $\mathbf{y}_{ik}$ is replaced by a randomly selected $y \in \{y_{jk}\}_{j=1}^{n''}$. This results in a random shuffling of the binary relation

relating y_{ik} in y_{jk}. The estimator of $I_p(\mathbf{X};\mathbf{Y})$ is derived based on the Friedman–Rafsky (FR) multivariate runs test statistic [33] on the concatenated data set , $\mathcal{Z}'_{n'} \cup \widetilde{\mathcal{Z}}_{n''}$. The FR test statistic is defined as the number of edges in the MST spanning the merged data set that connect a point in $\mathcal{Z}'_{n'}$ to a point in $\widetilde{\mathcal{Z}}_{n''}$. This test statistic is denoted by $\mathfrak{R}_{n',n''} := \mathfrak{R}_{n',n''}(\mathcal{Z}'_{n'}, \widetilde{\mathcal{Z}}_{n''})$. Please note that since the MST is unique with probability one (under the assumption that all density functions are Lebesgue continuous) then all inter point distances between nodes are distinct. This estimator converges to $I_p(\mathbf{X};\mathbf{Y})$ almost surely as $n \to \infty$. The procedure is summarized in Algorithm 1.

Algorithm 1: MST-based estimator of GMI

Input: Data set $\mathcal{Z}_n := \left\{(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n\right\}$

1: Find $\tilde{\alpha}$ using arguments in Section 2.4
2: $n' \leftarrow \tilde{\alpha} n$, $n'' \leftarrow (1-\tilde{\alpha})n$
3: Divide $\mathcal{Z}_n$ into two subsets $\mathcal{Z}'_{n'}$ and $\mathcal{Z}''_{n''}$
4: $\widetilde{\mathcal{Z}}_{n''} \leftarrow \left\{(\mathbf{x}_{ik}, \mathbf{y}_{jk})_{k=1}^{n''}\right.$: shuffle first and second elements of pairs in $\left.\mathcal{Z}''_{n''}\right\}$
5: $\widehat{\mathcal{Z}} \leftarrow \mathcal{Z}'_{n'} \cup \widetilde{\mathcal{Z}}''_{n''}$
6: Construct MST on $\widehat{\mathcal{Z}}$
7: $\mathfrak{R}_{n',n''} \leftarrow$ # edges connecting a node in $\mathcal{Z}'_{n'}$ to a node of $\widetilde{\mathcal{Z}}_{n''}$
8: $\widehat{I}_p \leftarrow 1 - \mathfrak{R}_{n',n''} \dfrac{n'+n''}{2n'n''}$

Output: $\widehat{I}_p$, where $p = \tilde{\alpha}$

Theorem 3 shows that the output in Algorithm 1 estimates the GMI with parameter $p = \alpha$. The proof is provided in Appendix A.4.

Theorem 3. *For given proportionality parameter $\alpha \in (0,1)$, choose n', n'' such that $n' + n'' = n$ and, as $n \to \infty$, we have $n'/n \to \alpha$ and $n''/n \to \beta = 1-\alpha$. Then*

$$1 - \mathfrak{R}_{n',n''}(\mathcal{Z}'_{n'}, \widetilde{\mathcal{Z}}_{n''}) \frac{n}{2n'\,n''} \to I_\alpha(\mathbf{X};\mathbf{Y}), \quad a.s. \tag{11}$$

Please note that the asymptotic limit in (11) depends on the proportionality parameter α. Later in Section 2.4, we discuss the choice of an optimal parameter $\tilde{\alpha}$. In Figure 1, we illustrate the MST constructed over merged independent ($\rho = 0$) and highly dependent ($\rho = 0.9$) data sets drawn from two-dimensional normal distributions with correlation coefficients ρ. Notice that the edges of the MST connecting samples with different colors, corresponding to independent and dependent samples, respectively, are indicated in green. The total number of green edges is the FR test statistic $\mathfrak{R}_{n',n''}(\mathcal{Z}'_{n'}, \widetilde{\mathcal{Z}}_{n''})$.

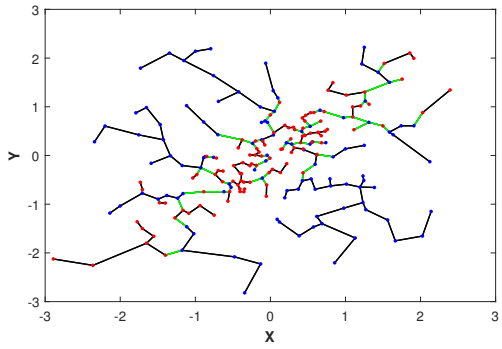

Figure 1. The MST and FR statistic of spanning the merged set of normal points when **X** and **Y** are independent (denoted in blue points) and when **X** and **Y** are highly dependent (denoted in red points). The FR test statistic is the number of edges in the MST that connect samples from different color nodes (denoted in green) and it is used to estimate the GMI I_p.

2.3. Convergence Rates

In this subsection we characterize the MSE convergence rates of the GMI estimator of Section 2.2 in the form of upper bounds on the bias and the variance. This MSE bound is given in terms of the sample size n, the dimension d, and the proportionality parameter α. Deriving convergence rates for mutual information estimators has been of interest in information theory and machine learning [22,27]. The rates are typically derived in terms of a smoothness condition on the densities, such as the Hölder condition [41]. Here we assume f_X, f_Y and f_{XY} have support sets $\mathbb{S}_X$, $\mathbb{S}_Y$, and $\mathbb{S}_{XY} := \mathbb{S}_X \times \mathbb{S}_Y$, respectively, and are smooth in the sense that they belong to Hölder continuous classes of densities $\Sigma_d^s(\eta, \mathrm{K})$, $0 < \eta \leq 1$ [42,43]:

Definition 1. *(Hölder class): Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space. The Hölder class of functions $\Sigma_d(\eta, K)$, with Hölder parameters η and K, consists of functions g that satisfy*

$$\left\{ g : \left\| g(\mathbf{z}) - p_{\mathbf{x}}^{\lfloor \eta \rfloor}(\mathbf{z}) \right\|_d \leq K \left\| \mathbf{x} - \mathbf{z} \right\|_d^{\eta}, \;\; \mathbf{x}, \, \mathbf{z} \in \mathcal{X} \right\}, \tag{12}$$

where $p_{\mathbf{x}}^k(\mathbf{z})$ is the Taylor polynomial (multinomial) of g of order k expanded about the point $\mathbf{x}$ and $\lfloor \eta \rfloor$ is defined as the greatest integer strictly less than η.

To explore the optimal choice of parameter α we require bounds on the bias and variance bounds, provided in Appendix A.5. To obtain such bounds, we will make several assumptions on the absolutely continuous densities f_X, f_Y, f_{XY} and support sets $\mathbb{S}_X$, $\mathbb{S}_Y$, $\mathbb{S}_{XY}$:

(**A**.1) Each of the densities belong to $\Sigma_d(\eta, K)$ with smoothness parameters η and Lipschitz constant K.

(**A**.2) The volumes of the support sets are finite, i.e., $0 < \mathbb{V}(\mathbb{S}_X) < \infty, \quad 0 < \mathbb{V}(\mathbb{S}_Y) < \infty$.

(**A**.3) All densities are bounded i.e., there exist two sets of constants C_X^L, C_Y^L, C_{XY}^L and C_X^U, C_Y^U, C_{XY}^U such that $0 < C_X^L \leq f_X \leq C_X^U < \infty$, $0 < C_Y^L \leq f_Y \leq C_Y^U < \infty$ and $0 < C_{XY}^L \leq f_{XY} \leq C_{XY}^U < \infty$.

The following theorem on the bias follows under assumptions (**A**.1) and (**A**.3):

Theorem 4. *For given $\alpha \in (0,1)$, $\beta = 1 - \alpha$, $d \geq 2$, and $0 < \eta \leq 1$ the bias of the $\mathfrak{R}_{n',n''} := \mathfrak{R}_{n',n''}(\mathcal{Z}'_{n'}, \widetilde{\mathcal{Z}}_{n''})$ satisfies*

$$\begin{aligned} &\left| \frac{\mathbb{E}\left[\mathfrak{R}_{n',n''}\right]}{n} - 2\alpha\beta \iint \frac{f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})}{\alpha f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) + \beta f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})} \, \mathrm{dxdy} \right| \\ &\qquad \leq O\left(\max\left\{ n^{-\eta^2/(d(1+\eta))}, \; (\beta n)^{-\eta/(1+\eta)}, \; c_d n^{-1} \right\} \right), \end{aligned} \tag{13}$$

where c_d is the largest possible degree of any vertex of MST on $\mathcal{Z}'_{n'} \cup \widetilde{\mathcal{Z}}_{n''}$. The explicit form of (13) is provided in Appendix A.5.

Please note that according to Theorem 13 in [44], the constant c_d is lower bounded by $\Omega\left(\sqrt{d} 2^{n(1-H(\gamma))}\right)$, $\gamma = 2^{-d}$ and $H(\gamma)$ is the binary entropy i.e.,

$$H(\gamma) = -\gamma \log \gamma - (1-\gamma) \log(1-\gamma).$$

A proof of Theorem 4 is given in Appendix A.5. The next theorem gives an upper bound on the variance of the FR estimator $\mathfrak{R}_{n',n''}$. The proof of the variance result requires a different approach than the bias bound (the Efron–Stein inequality [45]). It is similar to arguments in ([46], Appendix A.3), and is omitted. In Theorem 5 we assume that the densities f_X, f_Y, and f_{XY} are absolutely continuous and bounded (**A**.3).

Theorem 5. *Given $\alpha \in (0,1)$, the variance of the estimator $\mathfrak{R}_{n',n''} := \mathfrak{R}_{n',n''}(\mathcal{Z}'_{n'}, \widetilde{\mathcal{Z}}_{n''})$ is bounded by*

$$Var\left(\frac{\mathfrak{R}_{n',n''}}{n}\right) \leq \frac{(1-\alpha)\, c_d}{n}, \quad \alpha = n'/n, \tag{14}$$

where c_d is a constant depending only on the dimension d.

2.4. Minimax Parameter α

Recall assumptions (**A**.1), (**A**.2), and (**A**.3) in Section 2.3. The constant α can be chosen to minimize the maximum the MSE converges rate where the maximum is taken over the space of Hölder smooth joint densities f_{XY}.

Throughout this subsection we use the following notations:

- $\epsilon_{XY} := f_{XY}/f_X f_Y$,
- $C^L_\epsilon := C^L_{XY}/C^U_X C^U_Y$ and $C^U_\epsilon := C^U_{XY}/C^L_X C^L_Y$,
- $C_n := C^L_{XY}\, n/2$,
- $\alpha^L_0 := \dfrac{2}{C_n}$ and $\alpha^U_0 := \min\left\{\dfrac{1}{4}, \dfrac{1+1/C_n}{4+2C^U_\epsilon}, 1-n^{\eta/d-1}\right\}$, where η is the smoothness parameter,
- $l_n := \left\lfloor n^{\eta/(d^2(1+\eta))}\right\rfloor$.

Now define $\widetilde{G}^{\alpha,\beta}_{\epsilon_{XY},n}(\mathbf{x},\mathbf{y})$ by

$$\frac{(\epsilon_{XY}(\mathbf{x},\mathbf{y}) + 1/(\beta C_n))(1+\epsilon_{XY}(\mathbf{xy}) + 1/(\beta C_n))}{(\alpha+\beta\epsilon_{XY}(\mathbf{x},\mathbf{y}))^2}, \quad \beta = 1-\alpha. \tag{15}$$

Consider the following optimization problem:

$$\begin{aligned} \min_{\alpha}\ \max_{\epsilon_{XY}}\quad & \widetilde{\Delta}(\alpha,\epsilon_{XY}) + c_d(1-\alpha)\, n^{-1} \\ \text{subject to}\quad & C^L_\epsilon \leq \epsilon_{XY} \leq C^U_\epsilon, \\ & \alpha^L_0 \leq \alpha \leq \alpha^U_0, \end{aligned} \tag{16}$$

where

$$\widetilde{\Delta}(\alpha,\epsilon_{XY}) := D(n,l_n,d,\eta) + \widetilde{D}(n,l_n,d) C^U_{XY} \iint_{\mathbb{S}_{XY}} \widetilde{G}^{\alpha,\beta}_{\epsilon_{XY},n}(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}, \tag{17}$$

and

$$D(n,l_n,d,\eta) = c_2 l_n^d n^{-1} + c_d 2^d n^{-1} + c' l_n^d n^{-\eta/d} + c l_n^d n^{-1/d} + 2c_1 l_n^{d-1} n^{1/d-1} + c_3 l_n^{-d\eta}, \tag{18}$$

$$\begin{aligned} \widetilde{D}(n,l_n,d) = {} & 2 + n^{-1} 2c'' \sum_{i=1}^{M} l_n\, l_n^d a_i^{-1} + n^{-3/2} 2c'_1 \sum_{i=1}^{M} l_n\, l_n^{d/2} \sqrt{b_i} a_i^2 \\ & + n^{-1} \sum_{i=1}^{M} 2n^{-3/2} l_n^{-d/2} \frac{\sqrt{b_i}}{a_i^2} \left(n a_i l_n^d + n^2 a_i^2\right)^{1/2} \left(n b_i l_n^d + n^2 b_i^2\right)^{1/2}. \end{aligned} \tag{19}$$

Please note that in (18), c, c', c_1, c_2 are constants, and c_d only depends on the dimension d. Also, in (19), a_i and b_i are constants. Let ϵ^*_{XY} be the optimal ϵ_{XY} i.e., ϵ^*_{XY} be the solution of the optimization problem (16). Set

$$\Xi(\alpha) := \frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\widetilde{\Delta}(\alpha,\epsilon^*_{XY}) + c_d(1-\alpha)\, n^{-1}\right), \tag{20}$$

such that $\widetilde{\Delta}(\alpha, \epsilon^*_{XY})$ is (17) when $\epsilon_{XY} = \epsilon^*_{XY}$. For $\alpha \in [\alpha_0^L, \alpha_0^U]$, the optimal choice of ϵ_{XY} in terms of maximizing the MSE is $\epsilon^*_{XY} = C_\epsilon^U$ and the saddle point for the parameter α, denoted by $\tilde{\alpha}$, is given as follows:

- $\tilde{\alpha} = \alpha_0^U$, if $\Xi(\alpha_0^U) < 0$.
- $\tilde{\alpha} = \alpha_0^L$, if $\Xi(\alpha_0^L) > 0$.
- $\tilde{\alpha} = \Xi^{-1}(0)$, if $\alpha_0^L \leq \Xi^{-1}(0) \leq \alpha_0^U$.

Further details are given in Appendix A.6.

3. Simulation Study

In this section, numerical simulations are presented that illustrate the theory in Section 2. We perform multiple experiments to demonstrate the utility of the proposed GMI estimator of the HP-divergence in terms of the dimension d and the sample size n. Our proposed MST-based estimator of the GMI is compared to density plug-in estimators of the GMI, in particular the standard KDE density plug-in estimator of [22], where the convergence rates of Theorems 4 and 5 are validated. We use multivariate normal simulated data in the experiments. In this section, we also discuss the choice of the proportionality parameter α and compare runtime of the proposed GMI estimator approach with KDE method.

Here we perform four sets of experiments to illustrate the proposed GMI estimator. For the first set of experiments the MSE of the GMI estimator in Algorithm 1 is shown in Figure 2-left. The samples were drawn from d-dimensional normal distribution, with various sample sizes and dimensions $d = 6, 10, 12$. We selected the proportionality parameter $\alpha = 0.3$ and computed the MSE in terms of the sample size n. We show the log–log plot of MSE when n varies in $[100, 1500]$. Please note that the empirically optimal proportion α depends on n, so to avoid the computational complexity we fixed α for this experiment. The experimental result shown in Figure 2-left validates the theoretical MSE growth rates derived from (13) and (14), i.e., decreasing sub-linearly in n and increasing exponentially in d.

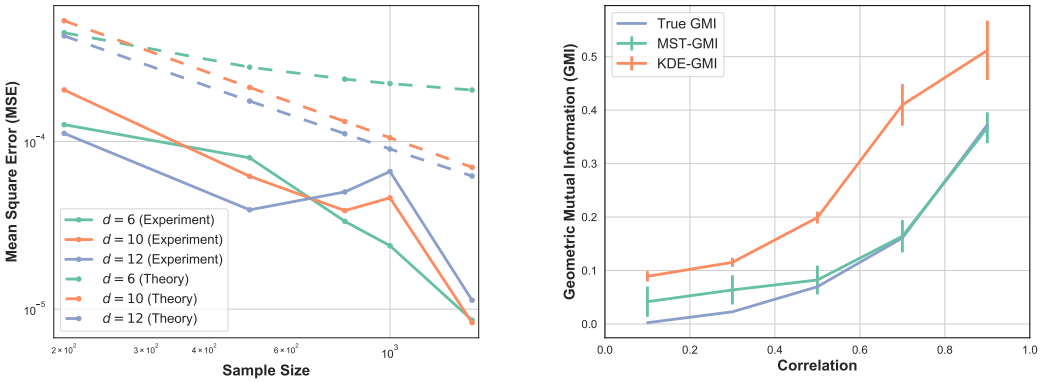

Figure 2. (**left**) Log–log plot of theoretical and experimental MSE of the proposed MST-based GMI estimator as a function of sample size n for $d = 6, 10, 12$ and fixed smoothness parameter η. (**right**) The GMI estimator was implemented using two approaches, Algorithm 1 and KDE method where the KDE-GMI used KDE density estimators in the formula (2). In this experiment, samples are generated from the two-dimensional normal distribution with zero mean and covariance matrix (21) for various value of $\rho \in [0.1, 0.9]$.

In Figure 2-right, we compare the proposed MST-based GMI estimator with the KDE-GMI estimator [22]. For the KDE approach, we estimated the joint and marginal densities and then plugged them into the proposed expression (2). The bandwidth h used for the KDE plug-in estimator was set as $h = n^{-1/(d+1)}$. The choice of h minimizes the bound on the MSE of the plug-in estimator. We generated data from the two-dimensional normal distribution with zero mean and covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{21}$$

The coefficient ρ is varied in range $[0.1, 0.9]$. The true GMI was computed by the Monte Carlo approximation to the integral (2). Please note that as ρ increases, the MST-GMI outperforms the KDE-GMI approach. In this set of experiments $\alpha = 0.6$.

Figure 3 again compares the MST-GMI estimator with the KDE-GMI estimator. samples are drawn from the multivariate standard normal distribution with dimensions $d = 4$ and $d = 12$. In both cases the proportionality parameter $\alpha = 0.5$. The left plots in Figure 3 show the MSE (100 trials) of the GMI estimator implemented with an KDE estimator (with bandwidth as in Figure 2 i.e., $h = n^{-1/(d+1)}$) for dimensions $d = 4, 12$ and various sample sizes. For all dimensions and sample sizes the MST-GMI estimator also outperforms the plug-in KDE-GMI estimator based on the estimated log–log MSE slope given in Figure 3 (left plots). The right plots in Figure 3 compares the MST-GMI with the KDE-GMI. In this experiment, the error bars denote standard deviations with 100 trials. We observe that for higher dimension $d = 12$ and larger sample size n, the KDE-GMI approaches the true GMI at a slower rate than the MST-GMI estimator. This reflects the power of the proposed graph-based approach to estimating GMI.

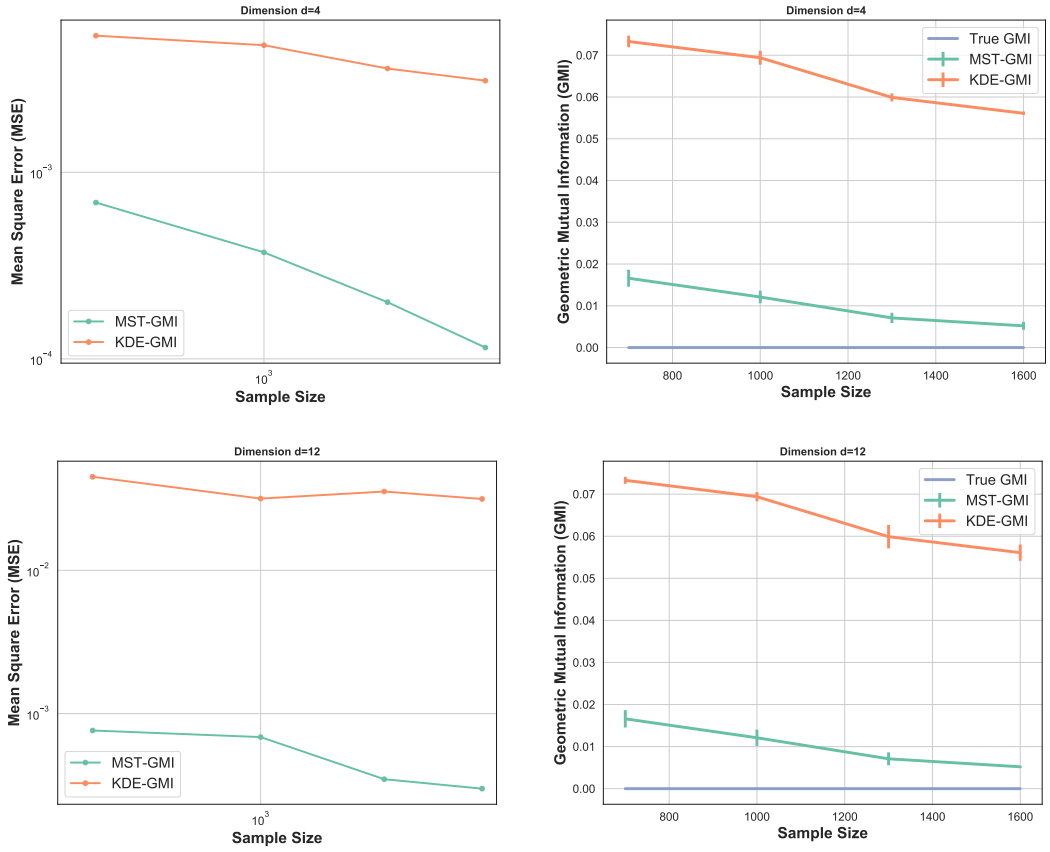

Figure 3. MSE log–log plots as a function of sample size n (**left**) for the proposed MST-GMI estimator ("Estimated GMI") and the standard KDE-GMI plug-in estimator of GMI. The right column of plots correspond to the GMI estimated for dimension $d = 4$ (**top**) and $d = 12$ (**bottom**). In both cases the proportionality parameter α is 0.5. The MST-GMI estimator in both plots for sample size n in $[700, 1600]$ outperforms the KDE-GMI estimator, especially for larger dimensions.

The comparison between MSEs for various dimension d is shown in Figure 4 (left). This experiment highlights the impact of higher dimension on the GMI estimators. As expected, for

larger sample size n, MSE decreases while for higher dimension it increases. In this setting, we have generated samples from standard normal distribution with size $n \in [10^2, 4 \times 10^3]$ and $\alpha = 0.5$. From Figure 4 (left) we observe that for larger sample size, MSE curves are ordered based on their corresponding dimensions. Results in Section 2.4 strongly depend on the lower bounds $C_X^L,\ C_Y^L,\ C_{XY}^L$ and upper bounds $C_X^U,\ C_Y^U,\ C_{XY}^U$ and provide optimal parameter α in the range $[\alpha_0^L, \alpha_0^U]$, therefore in the experiment section we only analyze one case where the lower bounds $C_X^L,\ C_Y^L,\ C_{XY}^L$ and upper bounds $C_X^U,\ C_Y^U,\ C_{XY}^U$ are known and the optimal α becomes α_0^L. Figure 4 (right) illustrates the MSE vs proportion parameter α when $n = 500, 10^4$ samples are generated from truncated normal distribution with $\rho = 0.7, 0.5$. First, following Section 2.4, we compute the bound $[\alpha_0^L, \alpha_0^U]$ and then derive the optimal α in this range. Therefore, each experiment with different sample size and ρ provides different range $[\alpha_0^L, \alpha_0^U]$. We observe that the MSE does not appeared a monotonic function in α and its behavior strongly depends on sample size n, d, and density functions' bounds. Additional study of the dependency is described in Appendix A.6. In this set of experiments $\Xi(\alpha_0^L) > 0$, therefore following the results in Section 2.4, we have $\tilde{\alpha} = \alpha_0^L$. In this experiment the optimal value of α is always the lower bound α_0^L and indicated in the Figure 4 (right).

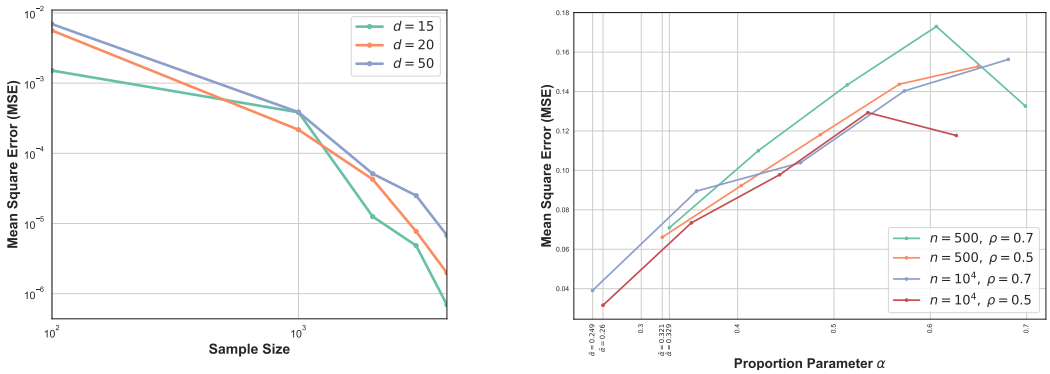

Figure 4. MSE log–log plots as a function of sample size n for the proposed FR estimator. We compare the MSE of our proposed FR estimator for various dimensions $d = 15, 20, 50$ (**left**). As d increases, the blue curve takes larger values than green and orange curves i.e., MSE increases as d grows. However, this is more evidential for large sample size n. The second experiment (**right**) focuses on optimal proportion α for $n = 500, 10^4$ and $\rho = 0.7, 0.5$. $\tilde{\alpha}$ is the optimal α for $\alpha \in [\alpha_0^L, \alpha_0^U]$.

The parameter α is studied further for three scenarios where the lower bounds $C_X^L,\ C_Y^L,\ C_{XY}^L$ and upper bounds $C_X^U,\ C_Y^U,\ C_{XY}^U$ are assumed unknown, therefore results in Section 2.4 are not applicable. In this set of experiments, we varied α in the range $(0, 1)$ to divide our original sample. We generated sample from an isotropic multivariate standard normal distribution $(\rho = 0)$ in all three scenarios (all features are independent). Therefore, the true GMI is zero and in all scenarios the GMI column, corresponding to the MST-GMI, is compared with zero. In each scenario we fixed dimension d and sample size n and varied $\alpha = 0.2, 0.5, 0.8$. The dimension and sample size in Scenarios 1,2, and 3 are $d = 6, 8, 10$ and $n = 1000, 1500, 2000$, respectively. In Table 1 the last column (α) stars the parameter $\alpha \in \{0.2, 0.5, 0.8\}$ with the minimum MSE and GMI (I_α) in each scenario. Table 1 shows that in these sets of experiments when $\alpha = 0.5$, the GMI estimator has less MSE (i.e., is more accurate) than when $\alpha = 0.2$ or $\alpha = 0.8$. This experimentally demonstrates that if we split our training data, the proposed Algorithm 1 performs better with $\alpha = 0.5$.

Table 1. Comparison between different scenarios of various dimensions and sample sizes in terms of parameter α. We applied the MST-GMI estimator to estimate the GMI (I_α) with $\alpha = 0.2, 0.5, 0.8$. We varied dimension $d = 6, 8, 10$ and sample size $n = 1000, 1500, 2000$ in each scenario. We observe that for $\alpha = \{0.2, 0.5, 0.8\}$, the MST-GMI estimator provides lowest MSE when $\alpha = 0.5$ indicated by star (*).

Overview Table for Different d, n, and α					
Experiments	**Dimension (d)**	**Sample Size (n)**	**GMI (I_α)**	**MSE ($\times 10^{-4}$)**	**Parameter (α)**
Scenario 1–1	6	1000	0.0229	12	0.2
Scenario 1–2	6	1000	0.0143	4.7944	0.5 *
Scenario 1–3	6	1000	0.0176	6.3867	0.8
Scenario 2–1	8	1500	0.0246	11	0.2
Scenario 2–2	8	1500	0.0074	1.6053	0.5 *
Scenario 2–3	8	1500	0.0137	5.3863	0.8
Scenario 3–1	10	2000	0.0074	2.3604	0.2
Scenario 3–2	10	2000	0.0029	0.54180	0.5 *
Scenario 3–3	10	2000	0.0262	11	0.8

Finally, Figure 5 shows the runtime as a function of sample size n. We vary sample size in the range $[10^3, 10^4]$. Observe that for smaller number of samples the KDE-GMI method is slightly faster but as n becomes large we see significant relative speedup of the proposed MST-GMI method.

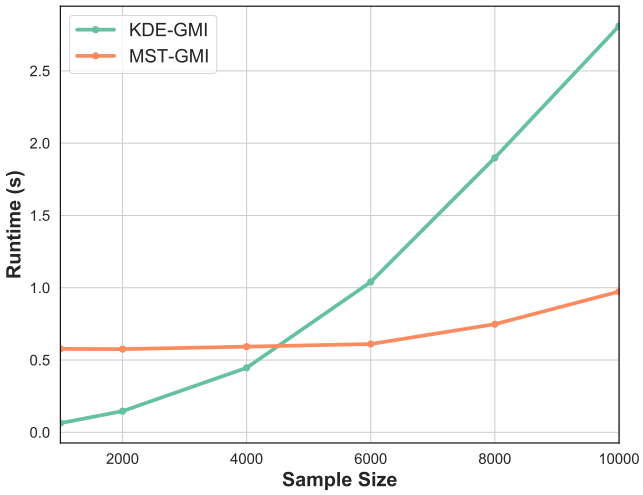

Figure 5. Runtime of KDE approach and proposed MST-based estimator of GMI vs sample size. The proposed GMI estimator achieves significant speedup, while for small sample size, the KDE method becomes overly fast. Please note that in this experiment the sample is generated from the Gaussian distribution in dimension $d = 2$.

4. Conclusions

In this paper, we have proposed a new measure of mutual information, called Geometric MI (GMI), which is related to the Henze–Penrose divergence. The GMI can be viewed as dependency measure that is the limit of the Friedman–Rafsky test statistic, which depends on the MST over all data points. We established some properties of the GMI in terms of convexity/concavity, chain rule, and a type of data-processing inequality. A direct estimator of the GMI, called the MST-GMI, was introduced that uses random permutations of observed relationships between variables in the multivariate samples. An explicit form for the MSE convergence rate bound was derived that depends on a free parameter called the proportionality parameter. An asymptotically optimal form for this free parameter was given that minimizes the MSE convergence rate. Simulation studies were performed that illustrate and verify the theory.

Author Contributions: S.Y.S. wrote this article primarily under the supervision of A.O.H. as principle investigator (PI), and A.O.H. edited the paper. S.Y.S. provided the primary contributions for the proofs of all theorems and performed all experiments.

Funding: The work presented in this paper was partially supported by ARO grant W911NF-15-1-0479 and DOE grant DE-NA0002534.

Acknowledgments: The authors would like to thank Brandon Oselio for the helpful comments.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A.

We organize the appendices as follows: Theorem 1 which establishes convexity/concavity is proved in Appendix A.1. Appendixs A.2 and A.3 establish the inequality (9) and (10) for given $p \in (0,1)$, respectively. In Appendix A.4, we first prove that the set $\widetilde{\mathcal{Z}}_{n''}$ which is randomly generated from original dependent data, contains asymptotically independent samples. Later by using the generated independent sample $\widetilde{\mathcal{Z}}_{n''}$ we show that for given α the FR estimator of the GMI given in Algorithm 1 tends to I_α. Appendix A.5 dedicates proof of Theorem 4. The proportionality parameter (α) optimization strategy is presented in Appendix A.6.

Appendix A.1. Theorem 1

Proof. The proof is similar to the result for standard (Shannon) mutual information. However, we require the following lemma, proven in analogous manner to the log-sum inequality:

Lemma A1. *For non-negative real numbers $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$, given $p \in (0,1)$, $q = 1-p$,*

$$\sum_{i=1}^{n} \alpha_i \left(p\left(\frac{\beta_i}{\alpha_i}\right) + q \right)^{-1} \geq \sum_{i=1}^{n} \alpha_i \left(p\left(\frac{\sum_{i=1}^{n} \beta_i}{\sum_{i=1}^{n} \alpha_i} \right) + q \right)^{-1}.$$

Notice this follows by using the convex function $u(y) = y^2/(p + q\,y)$ for any $p \in (0,1)$, $q = 1-p$, and the Jensen inequality.

Define the shorthand $\int_{\mathbf{x}}$, $\int_{\mathbf{y}}$, and $\int_{\mathbf{xy}}$ for $\int d\mathbf{x}$, $\int d\mathbf{y}$ and $\iint d\mathbf{x}d\mathbf{y}$, respectively. To prove part (i) of Theorem 1, we represent the LHS of (5) as:

$$\tilde{I}_p(\lambda_1 f_{Y|X} g_X + \lambda_2 f_{Y|X} h_X) = 1 - \int_{\mathbf{xy}} \left(\lambda_1 f_{Y|X} g_X + \lambda_2 f_{Y|X} h_X\right) \times$$

$$\left[p \frac{\lambda_1 f_{Y|X} g_X + \lambda_2 f_{Y|X} h_X}{\left(\int_{\mathbf{x}} \lambda_1 f_{Y|X} g_X + \lambda_2 f_{Y|X} h_X\right)\left(\int_{\mathbf{y}} \lambda_1 f_{Y|X} g_X + \lambda_2 f_{Y|X} h_X\right)} + q \right]^{-1}$$

$$= 1 - \int_{\mathbf{xy}} \left(\lambda_1 f_{Y|X} g_X + \lambda_2 f_{Y|X} h_X\right) \left[p \frac{f_{Y|X}}{\left(\int_{\mathbf{x}} \lambda_1 f_{Y|X} g_X + \lambda_2 f_{Y|X} h_X\right)} + q \right]^{-1}.$$

Furthermore, the RHS of (5) can be rewritten as

$$\begin{aligned}
&\lambda_1 \tilde{I}_p(f_{Y|X}g_X) + \lambda_2 \tilde{I}_p(f_{Y|X}h_X) \\
&= 1 - \int_{\mathbf{xy}} \left(\lambda_1 f_{Y|X}g_X \left[p \frac{f_{Y|X}g_X}{(\int_{\mathbf{x}} f_{Y|X}g_X)(\int_{\mathbf{y}} f_{Y|X}g_X)} + q \right]^{-1} + \lambda_2 f_{Y|X}h_X \left[p \frac{f_{Y|X}h_X}{(\int_{\mathbf{x}} f_{Y|X}h_X)(\int_{\mathbf{y}} f_{Y|X}h_X)} + q \right]^{-1} \right) \\
&= 1 - \int_{\mathbf{xy}} \left(\lambda_1 f_{Y|X}g_X \left[p \frac{f_{Y|X}}{\int_{\mathbf{x}} f_{Y|X}g_X} + q \right]^{-1} + \lambda_2 f_{Y|X}h_X \left[p \frac{f_{Y|X}}{\int_{\mathbf{x}} f_{Y|X}h_X} + q \right]^{-1} \right).
\end{aligned}$$

Thus, to prove LHS $\geq$ RHS, we use the inequality below:

$$\begin{aligned}
&\left(\lambda_1 f_{Y|X}g_X + \lambda_2 f_{Y|X}h_X\right) \left[p \frac{f_{Y|X}}{\int_{\mathbf{x}} \lambda_1 f_{Y|X}g_X + \lambda_2 f_{Y|X}h_X} + q \right]^{-1} \\
&\leq \left(\lambda_1 f_{Y|X}g_X\right) \left[p \frac{\pi}{\int_{\mathbf{x}} f_{Y|X}g_X} + q \right]^{-1} + \left(\lambda_2 f_{Y|X}h_X\right) \left[p \frac{\pi}{\int_{\mathbf{x}} f_{Y|X}h_X} + q \right]^{-1}.
\end{aligned}$$

In Lemma A1, let

$$\alpha_1 = \frac{\lambda_1 \left(\int_{\mathbf{x}} f_{Y|X}g_X\right) \left(\lambda_1 f_{Y|X}g_X + \lambda_2 f_{Y|X}h_X\right)}{\int_{\mathbf{x}} \lambda_1 f_{Y|X}g_X + \lambda_2 f_{Y|X}h_X},$$

$$\alpha_2 = \frac{\lambda_2 \left(\int_{\mathbf{x}} f_{Y|X}h_X\right) \left(\lambda_1 f_{Y|X}g_X + \lambda_2 f_{Y|X}h_X\right)}{\int_{\mathbf{x}} \lambda_1 f_{Y|X}g_X + \lambda_2 f_{Y|X}h_X},$$

and for $i = 1, 2$,

$$\beta_i = \frac{\lambda_i f_{Y|X} \left(\lambda_1 f_{Y|X}g_X + \lambda_2 f_{Y|X}h_X\right)}{\int_{\mathbf{x}} \lambda_1 f_{Y|X}g_X + \lambda_2 f_{Y|X}h_X}.$$

Then the claimed assertion (i) is obtained. Part (ii) follows by convexity of D_p and the following expression:

$$\begin{aligned}
&\tilde{I}_p(\lambda_1 g_{Y|X}f_X + \lambda_2 h_{Y|X}f_X) \\
&= D_p\left(\lambda_1 f_X g_{Y|X} + \lambda_2 f_X h_{Y|X}\,, \left(\int_{\mathbf{x}} \lambda_1 f_X g_{Y|X} + \lambda_2 f_X h_{Y|X}\right)\left(\int_{\mathbf{y}} \lambda_1 \phi\pi_1 + \lambda_2 \phi\pi_2\right)\right) \\
&= D_p\left(\lambda_1 f_X g_{Y|X} + \lambda_2 f_X h_{Y|X}\,, f_X\left(\int_{\mathbf{x}} \lambda_1 f_X g_{Y|X} + \lambda_2 f_X h_{Y|X}\right)\right) \\
&= D_p\left(\lambda_1 f_X g_{Y|X} + \lambda_2 f_X h_{Y|X}\,, \lambda_1\left(\int_{\mathbf{x}} f_X g_{Y|X}\right)\left(\int_{\mathbf{y}} f_X g_{Y|X}\right) + \lambda_2\left(\int_{\mathbf{x}} f_X h_{Y|X}\right)\left(\int_{\mathbf{y}} f_X h_{Y|X}\right)\right).
\end{aligned}$$

Therefore, the claim in (6) is proved. □

Appendix A.2. Theorem 2

Proof. We start with (9). Given $p \in (0,1)$ and $q = 1-p$, we can easily check that for positive $t > q$, $s > q$, such that $t, s \neq 1$:

$$(t+s)\left[(ts) + q(1-t-s)\right] - p(ts) \geq 0.$$

This implies

$$p\Big(\frac{t-q}{p}\Big)\Big(\frac{s-q}{p}\Big) + q \geq \frac{t\,s}{t+s}.$$

By substituting

$$\frac{f_{X_1Y}(x_1,y)}{f_{X_1}(x_1)\,f_Y(y)} = \frac{t-q}{p}, \qquad \frac{f_{X_2Y|X_1}(x_2,y|x_1)}{f_{X_2|X_1}(x_2|x_1)\,f_{Y|X_1}(y|x_1)} = \frac{s-q}{p},$$

we get

$$\begin{aligned}\left(p\,\frac{f_{X_1X_2Y}(x_1,x_2,y)}{f_{X_1X_2}(x_1,x_2)\,f_Y(y)} + q\right)^{-1} \leq \left(p\,\frac{f_{X_1Y}(x_1,y)}{f_{X_1}(x_1)\,f_Y(y)} + q\right)^{-1} \\ + \left(p\,\frac{f_{X_2Y|X_1}(x_2,y|x_1)}{f_{X_2|X_1}(x_2|x_1)\,f_{Y|X_1}(y|x_1)} + q\right)^{-1}.\end{aligned} \tag{A1}$$

Consequently

$$I_p(X_1,X_2;Y) \geq I_p(X_1;Y) - \mathbb{E}_f\left[\left(p\,\frac{f_{X_2Y|X_1}(x_2,y|x_1)}{f_{X_2|X_1}(x_2|x_1)\,f_{Y|X_1}(y|x_1)} + q\right)^{-1}\right]. \tag{A2}$$

Here f is the joint PDF of random vector (X_1, X_2, Y). From the conditional GMI definition in (7) the expectation term in (A2) is equivalent to $1 - I_p(X_2;Y|X_1)$. This completes the proof. □

Appendix A.3. Proposition 1

Proof. Recall the Theorem 2, part (i). First from $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ we have $f_{XYZ} = f_{XY} f_{Z|Y}$ and then by applying the Jensen inequality,

$$\begin{aligned} &I_p(\mathbf{X};\mathbf{Y}) = I_p(\mathbf{X};\mathbf{Y},\mathbf{Z}) \qquad \text{and} \\ &I_p(\mathbf{X};\mathbf{Y},\mathbf{Z}) \geq I_p(\mathbf{Z};\mathbf{X}) - \mathbb{E}\Big[\big(p\,\pi(\mathbf{X},\mathbf{Y},\mathbf{Z}) + q\big)^{-1}\Big] \\ &\qquad\qquad \geq I_p(\mathbf{Z};\mathbf{X}) - \Big(p\,\mathbb{E}\big[\pi(\mathbf{X},\mathbf{Y},\mathbf{Z})\big] + q\Big)^{-1}, \end{aligned} \tag{A3}$$

where

$$\pi(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{f_{YX|Z}(\mathbf{y},\mathbf{x}|\mathbf{z})}{f_{Y|Z}(\mathbf{y}|\mathbf{z}) f_{X|Z}(\mathbf{x}|\mathbf{z})}.$$

Now by Markovian property we can immediately simplify the RHS in (A3) to the RHS in (10).

Furthermore, we can easily show that if $\mathbf{X} \to \mathbf{Z} \to \mathbf{Y}$, we have $f_{XYZ} = f_{ZX} f_{Y|Z}$ and therefore $I_p(\mathbf{Z};\mathbf{X}) = I_p(\mathbf{X};\mathbf{Y},\mathbf{Z})$. This together with (A3) proves that under both conditions $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ and $\mathbf{X} \to \mathbf{Z} \to \mathbf{Y}$, the equality $I_p(\mathbf{X};\mathbf{Y}) = I_p(\mathbf{Z};\mathbf{X})$ holds true. □

Appendix A.4. Theorem 3

Proof. We first derive two required Lemmas A2 and A3 below:

Lemma A2. *Consider random vector* $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ *with joint probability density function (pdf)* f_{XY}. *Let* $\mathfrak{Z}_n = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ *be a set of samples with pdf* f_{XY}. *Let* $\mathfrak{Z}'_{n'}$ *and* $\mathfrak{Z}''_{n''}$ *be two distinct subsets of* $\mathfrak{Z}_n$ *such that* $n' + n'' = n$ *and sample proportion is* $\alpha = n'/n$ *and* $\beta = 1 - \alpha$. *Next, let* $\widetilde{\mathfrak{Z}}_{n''} = \{\widetilde{\mathbf{z}}_1, \ldots, \widetilde{\mathbf{z}}_{n''}\}$ *be a set of pairs such that* $\widetilde{\mathbf{z}}_k = (\mathbf{x}_{i_k}, \mathbf{y}_{j_k})$, $k = 1, \ldots, n''$ *are selected at random from* $\mathfrak{Z}'_{n''}$. *Denote* $\widetilde{\mathbf{Z}} = (\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ *as the random vector corresponding to samples in* $\widetilde{\mathfrak{Z}}_{n''}$. *Then as* $n \to \infty$ *such that* n'' *also grows in a linked manner that* $\beta \neq 0$ *then the distribution of* $\widetilde{\mathbf{Z}}$ *convergences to* $f_X \times f_Y$ *i.e., random vectors* $\widetilde{\mathbf{X}}$ *and* $\widetilde{\mathbf{Y}}$ *become mutually independent.*

Proof. Consider two subsets $\mathbb{A}$, $\mathbb{B} \subset \mathbb{R}^n$, then we have

$$P(\widetilde{\mathbf{X}} \in \mathbb{A}, \widetilde{\mathbf{Y}} \in \mathbb{B}) = \mathbb{E}\left[\mathbf{I}_{\mathbb{A}}(\widetilde{\mathbf{X}}).\, \mathbf{I}_{\mathbb{B}}(\widetilde{\mathbf{Y}})\right] = \mathbb{E}\left[\sum_{i,j} \mathbf{I}_{\mathbb{A}}(\mathbf{X}_i).\, \mathbf{I}_{\mathbb{B}}(\mathbf{Y}_j).\, P\big((\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) = (\mathbf{X}_i, \mathbf{Y}_j) \big| \mathbf{Z}_n\big)\right].$$

Here $\mathbf{I}_{\mathbb{A}}$ stands for the indicator function. Please note that

$$P\Big((\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) = (\mathbf{X}_i, \mathbf{Y}_j) \big| \mathbf{Z}_n\Big) = \frac{1}{n''^2},$$

and $\mathbf{X}_i$ and $\mathbf{Y}_j$, $i \neq j$ are independent, therefore

$$\begin{aligned} P(\widetilde{\mathbf{X}} \in \mathbb{A}, \widetilde{\mathbf{Y}} \in \mathbb{B}) &= \frac{1}{n''^2} \sum_{i \neq j} P(\mathbf{X}_i \in \mathbb{A}) P(\mathbf{Y}_j \in \mathbb{B}) + \frac{1}{n''^2} \sum_{i=1}^{n} P(\mathbf{X}_i \in \mathbb{A}, \mathbf{Y}_i \in \mathbb{B}) \\ &= P(\mathbf{X}_i \in \mathbb{A}) P(\mathbf{Y}_j \in \mathbb{B}) + \frac{1}{n''} \Big\{ P(\mathbf{X}_i \in \mathbb{A}, \mathbf{Y}_i \in \mathbb{B}) - P(\mathbf{X}_i \in \mathbb{A}) P(\mathbf{Y}_i \in \mathbb{B}) \Big\}, \end{aligned}$$

this implies that

$$\Big| P(\widetilde{\mathbf{X}} \in \mathbb{A}, \widetilde{\mathbf{Y}} \in \mathbb{B}) - P(\widetilde{\mathbf{X}} \in \mathbb{A}) P(\widetilde{\mathbf{Y}} \in \mathbb{B}) \Big| \leq \frac{1}{n''} \iint \Big| f_{XY}(\mathbf{x}, \mathbf{y}) - f_X(\mathbf{x}) f_Y(\mathbf{y}) \Big| \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}. \tag{A4}$$

On the other hand, we know that $n'' = \beta\, n$, so we get

$$\Big| P(\widetilde{\mathbf{X}} \in \mathbb{A}, \widetilde{\mathbf{Y}} \in \mathbb{B}) - P(\widetilde{\mathbf{X}} \in \mathbb{A}) P(\widetilde{\mathbf{Y}} \in \mathbb{B}) \Big| \leq \frac{1}{\beta\, n} \iint \Big| f_{XY}(\mathbf{x}, \mathbf{y}) - f_X(\mathbf{x}) f_Y(\mathbf{y}) \Big| \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}. \tag{A5}$$

From (A5), we observe that when β takes larger values the bound becomes tighter. So, if $n \to \infty$ such that n'' also becomes large enough in a linked manner so that $\beta =$ constant then the RHS in (A5) tends to zero. This implies that $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ become independent when $n \to \infty$. □

An immediate result of Lemma A2 is the following:

Lemma A3. *For given random vector* $\mathbf{Z}_n = (\mathbf{X}_n, \mathbf{Y}_n)$ *from joint density function* f_{XY} *and with marginal density functions* f_X *and* f_Y *let* $\widetilde{\mathfrak{Z}}_{n''} = \{\widetilde{\mathbf{z}}_1, \ldots, \widetilde{\mathbf{z}}_{n''}\}$ *be realization of random vector* $\widetilde{\mathbf{Z}}$ *as in Lemma A2 with parameter* $\beta = n''/n$. *Then for given points of* $\widetilde{\mathfrak{Z}}_{n''}$ *at* $\widetilde{\mathbf{z}} = (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$, *we have*

$$\Big| f_{\widetilde{Z}}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) - f_X(\widetilde{\mathbf{x}}) f_Y(\widetilde{\mathbf{y}}) \Big| = O\left(\frac{1}{\beta n}\right). \tag{A6}$$

Now, we want to provide a proof of assertion (11). Consider two subsets $\mathbf{Z}'_{n'}$ and $\widetilde{\mathbf{Z}}_{n''}$ as described in Section 2.2 . Assume that the components of sample $\widetilde{\mathbf{Z}}_{n''}$ follow density function $\widetilde{f}_{\widetilde{X}\widetilde{Y}}$. Therefore by owing to Lemmas A2 and A3, when $n \to \infty$ then $\widetilde{f}_{\widetilde{X}\widetilde{Y}} \to f_X f_Y$. Let $M_{n'}$ and $N_{n''}$ be Poisson variables with mean n' and n'' independent of one another and $\{\mathbf{Z}'_i\}$ and $\{\widetilde{\mathbf{Z}}_j\}$. Assume two Poisson processes $\mathfrak{Z}'_{n'} = \{\mathbf{Z}'_1, \ldots, \mathbf{Z}'_{M_{n'}}\}$ and $\widetilde{\mathfrak{Z}}_{n''} = \{\widetilde{\mathbf{Z}}_1, \ldots, \widetilde{\mathbf{Z}}_{N_{n''}}\}$, and denote the FR statistic $\mathfrak{R}'_{n',n''}$ on these processes. Following the arguments in [13,46] we shall prove the following:

$$\frac{\mathbb{E}\left[\mathfrak{R}'_{n',n''}\right]}{n'+n''} \to 2\alpha\beta \iint \frac{f_{X,Y}(\mathbf{x},\mathbf{y}) f_X(\mathbf{x}) f_Y(\mathbf{y})}{\alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta f_X(\mathbf{x}) f_Y(\mathbf{y})}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}.$$

This follows due to $\left|\mathfrak{R}'_{n',n''} - \mathfrak{R}_{n',n''}\right| \leq K_d\Big(|M_{n'} - n'| + |N_{n''} - n''|\Big)$, where K_d is a constant defined in Lemma 1, [13] and $n'+n''=n$. Thus, $(n'+n'')^{-1}\,\mathbb{E}\left|\mathfrak{R}'_{n',n''} - \mathfrak{R}_{n',n''}\right| \to 0$ as $n \to \infty$. Let $\mathbf{W}_1^{n',n''}, \mathbf{W}_2^{n',n''}, \ldots$ be independent variables with common density

$$\phi_{n',n''}(\mathbf{x},\mathbf{y}) = \Big(n' f_{XY}(\mathbf{x},\mathbf{y}) + n'' \widetilde{f}_{\widetilde{X}\widetilde{Y}}(\mathbf{x},\mathbf{y})\Big)/(n'+n''),$$

for $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d$. Let $L_{n',n''}$ be an independent Poisson variable with mean $n'+n''$. Let $\mathfrak{F}'_{n',n''} = \left\{\mathbf{W}_1^{n',n''}, \ldots, \mathbf{W}_{L_{n',n''}}^{n',n''}\right\}$ a non-homogeneous Poisson process of rate $n'\, f_{XY} + n''\, \widetilde{f}_{\widetilde{X}\widetilde{Y}}$. Assign mark 1 to a point in $\mathfrak{F}'_{n',n''}$ with probability

$$n'\, f_{XY}(\mathbf{x},\mathbf{y}) \Big/ \Big(n'\, f_{XY}(\mathbf{x},\mathbf{y}) + n''\, \widetilde{f}_{\widetilde{X}\widetilde{Y}}(\mathbf{x},\mathbf{y})\Big),$$

and mark 2 otherwise. By the marking theorem [13,47], the FR test statistic $\widetilde{\mathfrak{R}}_{n',n''}$ has the same distribution as $\mathfrak{R}'_{n',n''}$. Given points of $\mathfrak{F}'_{n',n''}$ at $\mathbf{z}' = (\mathbf{x}',\mathbf{y}')$ and $\mathbf{z}'' = (\mathbf{x}'',\mathbf{y}'')$, the probability that they have different marks is given by (A7).

$$g_{n',n''}(\mathbf{z}',\mathbf{z}'') = \frac{n'\, f_{XY}(\mathbf{x}',\mathbf{y}')\; n''\, \widetilde{f}_{\widetilde{X},\widetilde{Y}}(\mathbf{x}'',\mathbf{y}'') + n''\, \widetilde{f}_{\widetilde{X},\widetilde{Y}}(\mathbf{x}',\mathbf{y}')\; n'\, f_{XY}(\mathbf{x}'',\mathbf{y}'')}{\Big(n'\, f_{XY}(\mathbf{x}',\mathbf{y}') + n''\, \widetilde{f}_{\widetilde{X},\widetilde{Y}}(\mathbf{x}',\mathbf{y}')\Big)\Big(n'\, f_{XY}(\mathbf{x}'',\mathbf{y}'') + n''\, \widetilde{f}_{\widetilde{X},\widetilde{Y}}(\mathbf{x}'',\mathbf{y}'')\Big)}, \tag{A7}$$

define

$$g(\mathbf{z}',\mathbf{z}'') = \frac{\alpha\beta\left(f_{XY}(\mathbf{x}',\mathbf{y}') f_X(\mathbf{x}'') f_Y(\mathbf{y}'') + f_X(\mathbf{x}') f_Y(\mathbf{y}') f_{XY}(\mathbf{x}'',\mathbf{y}'')\right)}{\Big(\alpha\, f_{XY}(\mathbf{x}'',\mathbf{y}'') + \beta f_X(\mathbf{x}'') f_Y(\mathbf{y}'')\Big)\Big(\alpha f_{XY}(\mathbf{x}',\mathbf{y}') + \beta f_X(\mathbf{x}') f_Y(\mathbf{y}')\Big)}, \tag{A8}$$

then

$$\mathbb{E}\Big[\widetilde{\mathfrak{R}}'_{n',n''} | \mathfrak{F}'_{n',n''}\Big] = \sum_{i<j\leq L_{n',n''}}\sum g_n(\mathbf{W}_i^{n',n''}, \mathbf{W}_j^{n',n''}) \mathbf{I}_{\mathfrak{F}'_{n',n''}}(\mathbf{W}_i^{n',n''}, \mathbf{W}_j^{n',n''}). \tag{A9}$$

Now recall (A8). We observe that $g_{n',n''}(\mathbf{z}',\mathbf{z}'') \to g(\mathbf{z}',\mathbf{z}'')$. Going back to (A9), we can write

$$\mathbb{E}\Big[\widetilde{\mathfrak{R}}'_{n',n''}\Big] = \sum_{i<j\leq L_{n',n''}}\sum g_{n',n''}(\mathbf{W}_i^{n',n''}, \mathbf{W}_j^{n',n''}) \mathbf{I}_{\mathfrak{F}'_{n',n''}}(\mathbf{W}_i^{n',n''}, \mathbf{W}_j^{n',n''}) + o(n'+n''). \tag{A10}$$

For fixed n', n'' consider the collection:

$$\mathfrak{F}_{n',n''} = \left\{\mathbf{W}_1^{n',n''}, \ldots, \mathbf{W}_{n',n''}^{n'+n''}\right\}.$$

By the fact that $\mathbb{E}\Big[M_{n'} + N_{n''} - (n'+n'')\Big] = o(n'+n'')$, we have

$$\mathbb{E}\Big[\widetilde{\mathfrak{R}}'_{n',n''}\Big] = \sum_{i<j\leq n'+n''}\sum g_{n',n''}(\mathbf{W}_i^{n',n''}, \mathbf{W}_j^{n',n''}) \mathbf{I}_{\mathfrak{F}_{n',n''}}(\mathbf{W}_i^{n'}, \mathbf{W}_j^{n''}) + o(n'+n''). \tag{A11}$$

Introduce

$$\phi(\mathbf{x},\mathbf{y}) = \alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta f_X(\mathbf{x}) f_Y(\mathbf{y}).$$

Then $\phi_{n',n''}(\mathbf{x},\mathbf{y}) \to \phi(\mathbf{x},\mathbf{y})$ uniformly as $n'/n \to \alpha$ and $n''/n \to \beta$. Thus, using Proposition 1 in [13], we get

$$\frac{\mathbb{E}\left[\widetilde{\mathfrak{R}}'_{n',n''}\right]}{n} \to \int g(\mathbf{z},\mathbf{z})\phi(\mathbf{z})\mathrm{d}\mathbf{z} = \iint \frac{2\alpha\beta f_{XY}(\mathbf{x},\mathbf{y})f_X(\mathbf{x})f_Y(\mathbf{y})}{\alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta f_X(\mathbf{x})f_Y(\mathbf{y})}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}. \tag{A12}$$

□

Appendix A.5. Theorem 4

Proof. We begin by providing a family of bias rate bounds for the FR test statistic $\mathfrak{R}_{n',n''}$ in terms of a parameter l. Assume f_{XY}, f_X, and f_Y are in $\Sigma_d(\eta, K)$. Then by plugging the optimal l, we prove the bias rate bound given in (13).

Theorem A1. *Let $\mathfrak{R}_{n',n''} := \mathfrak{R}(\mathfrak{Z}_{n'}, \mathfrak{Z}_{n''})$ be the FR test statistic. Then a bound on the bias rate of the $\mathfrak{R}_{n',n''}$ estimator for $0 < \eta \leq 1$, $d \geq 2$ is given by*

$$\begin{aligned}&\left|\frac{\mathbb{E}\left[\mathfrak{R}_{n',n''}\right]}{n} - 2\alpha\beta\iint \frac{f_{XY}(\mathbf{x},\mathbf{y})f_X(\mathbf{x})f_Y(\mathbf{y})}{\alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta f_X(\mathbf{x})f_Y(\mathbf{y})}\,\mathrm{d}\mathbf{x}\right| \\ &\leq O\left(l^d(n)^{-\eta/d}\right) + O\left(l^{-d\eta}\right) + O\left(l^d\beta^{-1}n^{-1}\right) + O\left(c_d n^{-1}\right),\end{aligned} \tag{A13}$$

where $0 < \eta \leq 1$ is the Hölder smoothness parameter and c_d is the largest possible degree of any vertex of MST. Set

$$\alpha_i = \alpha n a_i l^d \left(1 - a_i l^{-d}\right) + (\alpha n)^2 a_i^2,$$

$$\beta_i = \beta n b_i l^d \left(1 - b_i l^{-d}\right) + (\beta n)^2 b_i^2.$$

and

$$\mathcal{A}^{\beta,\alpha}_{f,n}(\mathbf{x},\mathbf{y}) = \frac{2f_{XY}(\mathbf{x},\mathbf{y})\left(f_X(\mathbf{x})f_Y(\mathbf{y}) + \delta_f/(\beta n)\right)\left(f_{XY}(\mathbf{x},\mathbf{y})\sqrt{\alpha} + \left(f_X(\mathbf{x})f_Y(\mathbf{y}) + \delta_f/(\beta n)\right)\sqrt{\beta}\right)}{a_i^2 l^{-d}\left(\alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta\left(f_X(\mathbf{x})f_Y(\mathbf{y}) + \delta_f/(\beta n)\right)\right)^2}, \tag{A14}$$

where

$$\delta_f = \iint \left|f_{XY}(\mathbf{x},\mathbf{y}) - f_X(\mathbf{x})f_Y(\mathbf{y})\right| \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}, \tag{A15}$$

A more explicit form for the bound on the RHS is given below:

$$\begin{aligned}&\Delta(\alpha, f_{XY}, f_X f_Y) := c_2 l^d (n)^{-1} + c_d 2^d (n)^{-1} + O\left(l^d (n)^{-\eta/d}\right) + O\left(l^d (n)^{-1/2}\right) \\ &+O\left(c_d (n)^{-1/2}\right) + 2c_1 l^{d-1}(n)^{(1/d)-1} + \delta_f((\beta n)^{-1}) \iint \frac{2\alpha\beta f_{XY}(\mathbf{x},\mathbf{y})}{\alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta f_X(\mathbf{x})f_Y(\mathbf{y})}\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ &+(n)^{-1}\sum_{i=1}^{M} 2\iint f_{XY}(\mathbf{x},\mathbf{y})\left(f_X(\mathbf{x})f_Y(\mathbf{y}) + \delta_f/(\beta n)\right)\left(\alpha_i\beta_i\left(\alpha n a_i l^{-d} f^2_{XY}(\mathbf{x},\mathbf{y})\right.\right. \\ &\left.\left.+\beta n b_i l^{-d}\left(f_X(\mathbf{x})f_Y(\mathbf{y}) + \delta_f/(\beta n)\right)^2\right)\right)^{1/2} \Big/ \left(\alpha n a_i f_{XY}(\mathbf{x},\mathbf{y}) + \beta n b_i f_X(\mathbf{x})f_Y(\mathbf{y})\right)^2 \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ &+(n)^{-1}\sum_{i=1}^{M} O(l)\iint l^d (a_i)^{-1}\frac{2f_{XY}(\mathbf{x},\mathbf{y})\left(f_X(\mathbf{x})f_Y(\mathbf{y}) + \delta_f/(\beta n)\right)}{\alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta f_X(\mathbf{x})f_Y(\mathbf{y})}\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + O\left(l^{-d\eta}\right) \\ &+(n)^{-3/2}\sum_{i=1}^{M} O(l)\iint l^{-d/2}\sqrt{b_i}\,\mathcal{A}^{\beta,\alpha}_{f,n}(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}.\end{aligned} \tag{A16}$$

Proof. Consider two Poisson variables $M_{n'}$ and $N_{n''}$ with mean n' and n'' respectively and independent of one another and $\{\mathbf{Z}_i'\}$ and $\{\widetilde{\mathbf{Z}}_j\}$. Let $\mathfrak{Z}_{n'}'$ and $\widetilde{\mathfrak{Z}}_{n''}$ be the Poisson processes $\{\mathbf{Z}_1', \ldots, \mathbf{Z}_{M_{n'}}'\}$ and $\{\widetilde{\mathbf{Z}}_1, \ldots, \widetilde{\mathbf{Z}}_{N_{n''}}\}$. Likewise Appendix A.4, set $\mathfrak{R}_{n',n''}' = \mathfrak{R}(\mathfrak{Z}_{n'}', \widetilde{\mathfrak{Z}}_{n''})$. Applying Lemma 1, and (12) in [13], we can write

$$\left|\mathfrak{R}_{n',n''}' - \mathfrak{R}_{n',n''}\right| \leq c_d\big(|M_{n'} - n'| + |N_{n''} - n''|\big). \tag{A17}$$

Here c_d denotes the largest possible degree of any vertex of the MST in $\mathbb{R}^d$. Following the arguments in [46], we have $\mathbb{E}\big[|M_{n'} - n'|\big] = O(\,n'^{1/2})$ and $\mathbb{E}\big[|N_{n''} - n''|\big] = O(\,n''^{1/2})$. Hence

$$\frac{\mathbb{E}\Big[\mathfrak{R}_{n',n''}\Big]}{n'+n''} = \frac{\mathbb{E}\Big[\mathfrak{R}_{n',n''}'\Big]}{n'+n''} + O\left(c_d(n'+n'')^{-1/2}\right). \tag{A18}$$

Next let n_i' and n_i'' be independent binomial random variables with marginal densities $B(n', a_i l^{-d})$ and $B(n'', b_i l^{-d})$ such that a_i, b_i are non-negative constants $a_i \leq b_i$ and $\sum_{i=1}^{l^d} a_i l^{-d} = \sum_{i=1}^{l^d} b_i l^{-d} = 1$. Therefore, using the subadditivity property in Lemma 2.2, [46], we can write

$$\mathbb{E}\Big[\mathfrak{R}_{n',n''}'\Big] \leq \sum_{i=1}^{M} \mathbb{E}\Big[\mathbb{E}\Big[\mathfrak{R}_{n_i',n_i''}'|n_i', n_i''\Big]\Big] + 2\,c_1\,l^{d-1}(n'+n'')^{1/d}, \tag{A19}$$

where $M = l^d$, and $\eta > 0$ is the Hölder smoothness parameter. Furthermore, for given n_i', n_i'', let $\mathbf{W}_1^{n_i',n_i''}, \mathbf{W}_2^{n_i',n_i''}, \ldots$ be independent variables with common densities for $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d$:

$$g_{n_i',n_i''}(\mathbf{x}, \mathbf{y}) = \left(n_i' f_{XY}(\mathbf{x}, \mathbf{y}) + n_i'' \widetilde{f}_{\widetilde{X}\widetilde{Y}}(\mathbf{x}, \mathbf{y})\right) / (n_i' + n_i'').$$

Denote $L_{n_i',n_i''}$ be an independent Poisson variable with mean $n_i' + n_i''$ and $\mathfrak{F}_{n_i',n_i''}' = \{\mathbf{W}_1^{n_i',n_i''}, \ldots, \mathbf{W}_{L_{n_i',n_i''}}^{n_i',n_i''}\}$ a non-homogeneous Poisson of rate $n_i' f_{XY} + n_i'' \widetilde{f}_{\widetilde{X}\widetilde{Y}}$. Let $\mathfrak{F}_{n_i',n_i''}$ be the non-Poisson point process $\{\mathbf{W}_1^{n_i',n_i''}, \ldots \mathbf{W}_{n_i'+n_i''}^{n_i',n_i''}\}$. Assign a mark from the set $\{1, 2\}$ to each point of $\mathfrak{F}_{n_i',n_i''}'$. Let $\widetilde{\mathfrak{Z}}_{n_i'}'$ be the sets of points marked 1 with each probability $n_i' f_{XY}(\mathbf{x}, \mathbf{y}) / \big(n_i' f_{XY}(\mathbf{x}, \mathbf{y}) + n_i'' \widetilde{f}_{\widetilde{X}\widetilde{Y}}(\mathbf{x}, \mathbf{y})\big)$ and let $\widetilde{\mathfrak{Z}}_{n_i''}''$ be the set points with mark 2. Please note that owing to the marking theorem [47], $\widetilde{\mathfrak{Z}}_{n_i'}'$ and $\widetilde{\mathfrak{Z}}_{n_i''}''$ are independent Poisson processes with the same distribution as $\mathfrak{Z}_{n_i'}'$ and $\widetilde{\mathfrak{Z}}_{n_i''}$, respectively. Considering $\widetilde{\mathfrak{R}}_{n_i',n_i''}'$ as FR test statistic on nodes in $\widetilde{\mathfrak{Z}}_{n_i'}' \cup \widetilde{\mathfrak{Z}}_{n_i''}''$, we have

$$\mathbb{E}\Big[\mathfrak{R}_{n_i',n_i''}'|n_i', n_i''\Big] = \mathbb{E}\Big[\widetilde{\mathfrak{R}}_{n_i',n_i''}'|n_i', n_i''\Big].$$

By the fact that $\mathbb{E}\big[|M_{n'} + N_{n''} - n' - n''|\big] = O((n'+n'')^{1/2})$, we have

$$\begin{aligned}
\mathbb{E}\Big[\widetilde{\mathfrak{R}}_{n_i',n_i''}'|n_i', n_i''\Big] &= \mathbb{E}\Big[\mathbb{E}\big[\widetilde{\mathfrak{R}}_{n_i',n_i''}'|\mathfrak{F}_{n_i',n_i''}'\big]\Big] \\
&= \mathbb{E}\left[\sum_{s<j<n_i'+n_i''}\sum P_{n_i',n_i''}(\mathbf{W}_s^{n_i',n_i''}, \mathbf{W}_j^{n_i',n_i''})\,\mathbf{1}\left\{(\mathbf{W}_s^{n_i',n_i''}, \mathbf{W}_j^{n_i',n_i''}) \in \mathfrak{F}_{n_i',n_i''}\right\}\right] + O\left((n_i'+n_i'')^{1/2})\right).
\end{aligned}$$

Here $\mathbf{z}' = (\mathbf{x}', \mathbf{y}')$, $\mathbf{z}'' = (\mathbf{x}'', \mathbf{y}'')$, and $P_{n_i',n_i''}(\mathbf{z}', \mathbf{z}'')$ is given in below:

$$P_{n_i',n_i''}(\mathbf{z}',\mathbf{z}'') := P_r\left\{\text{mark } \mathbf{z}' \neq \text{mark } \mathbf{z}'', (\mathbf{z}',\mathbf{z}'') \in \mathfrak{F}_{n_i',n_i''}\right\}$$

$$= \frac{n_i' f_{XY}(\mathbf{x}',\mathbf{y}') n_i'' \widetilde{f}_{\widetilde{X}\widetilde{Y}}(\mathbf{x}'',\mathbf{y}'') + n_i' \widetilde{f}_{\widetilde{X}\widetilde{Y}}(\mathbf{x}',\mathbf{y}') n_i'' f_{X,Y}(\mathbf{x}'',\mathbf{y}'')}{\left(n_i'' f_{XY}(\mathbf{x}',\mathbf{y}') + n_i'' \widetilde{f}_{\widetilde{X}\widetilde{Y}}(\mathbf{x}',\mathbf{y}')\right)\left(n_1' f_{XY}(\mathbf{x}'',\mathbf{y}'') + n_i'' \widetilde{f}_{\widetilde{X}\widetilde{Y}}(\mathbf{x}'',\mathbf{y}'')\right)}.$$

Next set

$$\alpha_i = n' a_i l^d\ (1 - a_i l^{-d}) + n'^2 a_i^2, \quad \beta_i = n'' b_i l^d\ (1 - b_i l^{-d}) + n''^2 b_i^2.$$

By owing to the Lemma B.6 in [46] and applying analogous arguments, we can rewrite the expression in (A20):

$$\begin{aligned}\mathbb{E}\left[\mathfrak{R}'_{n',n''}\right] \leq \sum_{i=1}^{M} a_i b_i l^{-d} \iint \frac{2\, n' n'' f_{XY}(\mathbf{x},\mathbf{y}) f_{\widetilde{X}\widetilde{Y}}(\mathbf{x},\mathbf{y})}{n' a_i f_{X,Y}(\mathbf{x},\mathbf{y}) + n'' b_i f_{\widetilde{X}\widetilde{Y}}(\mathbf{x},\mathbf{y})}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} + 2c_1\, l^{d-1}(n'+n'')^{1/d} \\ + \sum_{i=1}^{M} 2 \iint \frac{f_{XY}(\mathbf{x},\mathbf{y}) f_{\widetilde{X}\widetilde{Y}}(\mathbf{x},\mathbf{y}) \left(\alpha_i \beta_i \left(n' a_i l^{-d} f_{XY}^2(\mathbf{x},\mathbf{y}) + n'' b_i l^{-d} f_{\widetilde{X}\widetilde{Y}}^2(\mathbf{x},\mathbf{y})\right)\right)^{1/2}}{\left(n' a_i f_{XY}(\mathbf{x},\mathbf{y}) + n'' b_i f_{\widetilde{X}\widetilde{Y}}(\mathbf{x},\mathbf{y})\right)^2} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ + \sum_{i=1}^{M} \mathbb{E}_{n_i',n_i''}\left[(n_i'+n_i'')\, \varsigma_\eta(l,n_i',n_i'')\right] + O\left(l^d (n'+n'')^{1-\eta/d}\right) + O\left(l^d (n'+n'')^{1/2}\right),\end{aligned} \tag{A20}$$

where

$$\varsigma_\eta(l,n_i',n_i'') = \left(O\Big(\frac{l}{n_i'+n_i''}\Big) - \frac{2\, l^d}{n_i'+n_i''}\right) \int g_{n_i',n_i''}(\mathbf{z}') P_{n_i',n_i''}(\mathbf{z}',\mathbf{z}')\, \mathrm{d}\mathbf{z}' + O(l^{-d\eta}).$$

Going back to Lemma A3, we know that

$$f_{\widetilde{X}\widetilde{Y}}(\mathbf{x},\mathbf{y}) = f_X(\mathbf{x}) f_Y(\mathbf{y}) + O\left(\frac{1}{\beta n}\right).$$

Therefore, the first term on the RHS of (A20) is less and equal to

$$\sum_{i=1}^{M} a_i b_i l^{-d} \iint \frac{2\, n' n'' f_{XY}(\mathbf{x},\mathbf{y}) f_X(\mathbf{x}) f_Y(\mathbf{y})}{n' a_i f_{X,Y}(\mathbf{x},\mathbf{y}) + n'' b_i f_X(\mathbf{x}) f_Y(\mathbf{y})}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}$$

$$+ \left(\frac{\delta_f}{\beta n}\right) \sum_{i=1}^{M} a_i b_i l^{-d} \iint \frac{2\, n' n'' f_{XY}(\mathbf{x},\mathbf{y})}{n' a_i f_{XY}(\mathbf{x},\mathbf{y}) + n'' b_i f_X(\mathbf{x}) f_Y(\mathbf{y})}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y},$$

and the second term is less and equal to

$$\sum_{i=1}^{M} 2 \iint f_{XY}(\mathbf{x},\mathbf{y}) \Big(f_X(\mathbf{x}) f_Y(\mathbf{y}) + \delta_f/(\beta n)\Big) \Big(\alpha_i \beta_i \Big(n' a_i l^{-d} f_{XY}^2(\mathbf{x},\mathbf{y}) + n'' b_i l^{-d} \big(f_X(\mathbf{x}) f_Y(\mathbf{y}) + \delta_f/(\beta n)\big)^2\Big)\Big)^{1/2} \Big/ \left(n' a_i f_{XY}(\mathbf{x},\mathbf{y}) + n'' b_i f_X(\mathbf{x}) f_Y(\mathbf{y})\right)^2\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y},$$

where

$$\delta_f = \iint \Big| f_{XY}(\mathbf{x},\mathbf{y}) - f_X(\mathbf{x}) f_Y(\mathbf{y}) \Big|\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}.$$

Recall the definition of the dual MST and FR statistic denoted by $\mathfrak{R}^*_{n',n''}$ following [46]:

Definition A1. (Dual MST, MST* and dual FR statistic $\mathfrak{R}^*_{m,n}$) *Let $\mathbb{F}_i$ be the set of corner points associated with a particular subsection Q_i, $1 \leq i \leq l^d$ of $[0,1]^d$. Define the dual* MST$^*(\mathfrak{X}_m \cup \mathfrak{Y}_n \cap Q_i)$ *as the boundary MST graph in partition Q_i [48], which contains $\mathfrak{X}_m$ and $\mathfrak{Y}_n$ points falling inside partition cell Q_i and those corner points in $\mathbb{F}_i$ which minimize total MST length. Please note that it is allowed to connect the MSTs in Q_i and Q_j through points strictly contained in Q_i and Q_j and therefore corner points. Thus, the dual MST can connect the points in $Q_i \cup Q_j$ by direct edges to pair to another point in $Q_i \cup Q_j$ or by passing through the corner the corner points which are all connected in order to minimize the total weights. To clarify, assume that there are two points in $Q_i \cup Q_j$, then the dual MST consists of the two edges connecting these points to the corner if they are closer to a corner point otherwise the dual MST connects them to each other.*

Furthermore, $\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n \cap Q_i)$ is defined as the number of edges in the MST* graph connecting nodes from different samples and number of edges connecting to the corner points. Please note that the edges connected to the corner nodes (regardless of the type of points) are always counted in the dual FR test statistic $\mathfrak{R}^*_{m,n}$.

Similarly, consider the Poisson processes samples and the FR test statistic over these samples, denoted by $\mathfrak{R}'^*_{n',n''}$. By superadditivity of the dual $\mathfrak{R}^*_{n',n''}$ [46], we have

$$\mathbb{E}\left[\mathfrak{R}'^*_{n',n''}\right] \geq \sum_{i=1}^{M} a_i\, l^{-d} \iint \frac{2n'n'' f_{XY}(\mathbf{x},\mathbf{y}) \left(f_X(\mathbf{x}) f_Y(\mathbf{y}) - \delta_f/(\beta n)\right)}{n' f_{XY}(\mathbf{x},\mathbf{y}) + n'' \left(f_X(\mathbf{x}) f_Y(\mathbf{y}) - \delta_f/(\beta n)\right)} \,\mathbf{dxdy} \tag{A21}$$

$$- \sum_{i=1}^{M} \mathbb{E}_{n'_i,n''_i}\left[(n'_i + n''_i)\, \varsigma_\eta(l, n'_i, n''_i)\right] - O\left(l^d (n'+n'')^{1-\eta/d}\right) - O\left(l^d (n'+n'')^{1/2}\right) - c_2\, l^d.$$

The first term of RHS in (A21) is greater or equal to

$$\iint \frac{2n'n'' f_{XY}(\mathbf{x},\mathbf{y}) f_X(\mathbf{x}) f_Y(\mathbf{y})}{n' f_{XY}(\mathbf{x},\mathbf{y}) + n'' f_X(\mathbf{x}) f_Y(\mathbf{y})} \,\mathbf{dxdy} - \frac{\delta_f}{\beta n} \iint \frac{2n'n'' f_{XY}(\mathbf{x},\mathbf{y})}{n' f_{XY}(\mathbf{x},\mathbf{y}) + n'' f_X(\mathbf{x}) f_Y(\mathbf{y})} \,\mathbf{dxdy}.$$

Furthermore,

$$\frac{\mathbb{E}\left[\mathfrak{R}'_{n',n''}\right]}{n} + \frac{c_d 2^d}{n} \geq \frac{\mathbb{E}\left[\mathfrak{R}'^*_{n',n''}\right]}{n},$$

where c_d is the largest possible degree of any vertex of the MST in $\mathbb{R}^d$, as before. Consequently, we have

$$\left| \frac{\mathbb{E}\left[\mathfrak{R}'_{n',n''}\right]}{n} - \iint \frac{2\alpha\beta f_{XY}(\mathbf{x},\mathbf{y}) f_X(\mathbf{x}) f_Y(\mathbf{y})}{\alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta f_X(\mathbf{x}) f_Y(\mathbf{y})} \,\mathbf{dxdy} \right| \leq \mathcal{B}(\alpha, f_{XY}, f_X f_Y), \tag{A22}$$

where $\mathcal{B}$ is defined in (A16) and $\mathcal{A}^{\beta,\alpha}_{f,n}(\mathbf{x},\mathbf{y})$ has been introduced in (A14). The last line in (A22) follows from the fact that

$$\sum_{i=1}^{M} \mathbb{E}_{n'_i,n''_i}\left[(n'_i + n''_i)\, \varsigma_\eta(l, n'_i, n''_i)\right] \leq \sum_{i=1}^{M} O(l) \iint l^{-d/2} \sqrt{b_i}\, \mathcal{A}^{\beta,n'/n}_{f,n}(\mathbf{x},\mathbf{y}) \mathbf{dxdy}$$

$$+ \sum_{i=1}^{M} O(l) \iint l^d (a_i)^{-1} \frac{2 f_{XY}(\mathbf{x},\mathbf{y}) \left(f_X(\mathbf{x}) f_Y(\mathbf{y}) + O(\delta_f/(\beta n))\right)}{n' f_{XY}(\mathbf{x},\mathbf{y}) + n'' f_X(\mathbf{x}) f_Y(\mathbf{y})} \,\mathbf{dxdy}.$$

Here $\mathcal{A}^{\beta,n'/n}_{f,n}(\mathbf{x},\mathbf{y})$ is given as (A14) by substituting n'/n in α such that $\beta = 1-\alpha$. Hence, the proof of Theorem A1 is completed. □

Going back to the proof of (13), without loss of generality assume that $(n)l^{-d} > 1$, for $d \geq 2$ and $0 < \eta \leq 1$. We select l as a function of n and β to be the sequence increasing in n which minimizes the maximum of these rates:

$$l(n,\beta) = arg \min_{l} \max \left\{ l^d (n)^{-\eta/d}, l^{-\eta d}, l^d \beta^{-1} n^{-1}, c_d 2^d n^{-1} \right\}. \tag{A23}$$

The solution $l = l(n,\beta)$ is obtained when $l^d (n)^{-\eta/d} = l^{-\eta d}$, or equivalently $l = \lfloor (n)^{\eta/(d^2(\eta+1))} \rfloor$ or when $l^d \beta^{-1} n^{-1} = l^{-\eta d}$, which implies $l = \lfloor (\beta\, n)^{1/(d(1+\eta))} \rfloor$. Substitute this l in the bound (A13) to obtain the RHS expression in (13) for $d \geq 2$. □

Appendix A.6.

Our main goal in Section 2.4 was to find proportion α such that the parametric MSE rate depending on the joint density f_{XY} and marginal densities f_X, f_Y is minimized. Recalling the explicit bias bound in (A16), it can be seen that this function will be a complicated function of f_{XY}, $f_X f_Y$ and α. By rearrangement of terms in (A13), we first find an upper bound for Δ in (A16), denoted by $\overline{\Delta}$, as follows:

$$\overline{\Delta}(\alpha, f_{XY}, f_X f_Y) = D(n, l_n, d, \eta) + \widetilde{D}(n, l_n, d) \mathbb{E}_{XY} \left[G^{\alpha,\beta}_{f,n}(X,Y) \right], \tag{A24}$$

where $l_n := \lfloor n^{\eta/(d^2(1+\eta))} \rfloor$. From Appendix A.5 we know that optimal l is given by (A23). One can check that for $\alpha \leq 1 - n^{(\eta/d)-1}$, the optimal $l = \lfloor n^{\eta/(d^2(1+\eta))} \rfloor$ provides a tighter bound. In (A24), the constants D and $\overline{D}$ are

$$D(n, l_n, d, \eta) = c_2 l_n^d n^{-1} + c_d 2^d n^{-1} + c' l_n^d n^{-\eta/d} + c l_n^d n^{-1/d} + 2c_1 l_n^{d-1} n^{1/d-1} + c_3 l_n^{-d\eta}, \tag{A25}$$

$$\begin{aligned} \widetilde{D}(n, l_n, d) = 2 + n^{-1} 2c'' \sum_{i=1}^{M} l_n\, l_n^d a_i^{-1} + n^{-3/2} 2c_1' \sum_{i=1}^{M} l_n\, l_n^{d/2} \sqrt{b_i} a_i^2 \\ + n^{-1} \sum_{i=1}^{M} 2 n^{-3/2} l_n^{-d/2} \frac{\sqrt{b_i}}{a_i^2} \left(n a_i l_n^d + n^2 a_i^2 \right)^{1/2} \left(n b_i l_n^d + n^2 b_i^2 \right)^{1/2}. \end{aligned} \tag{A26}$$

And the function $G^{\alpha,\beta}_{f,n}(\mathbf{x},\mathbf{y})$ is given as the following:

$$\begin{aligned} G^{\alpha,\beta}_{f,n}(\mathbf{x},\mathbf{y}) = \left(f_X(\mathbf{x}) f_Y(\mathbf{y}) + \delta_f/(n\beta) \right) \left(\sqrt{\alpha} f_{XY}(\mathbf{x},\mathbf{y}) \right. \\ \left. + \sqrt{\beta} \left(f_X(\mathbf{x}) f_Y(\mathbf{y}) + \delta_f/(\beta n) \right) \right) \Big/ \left(\alpha f_{XY}(\mathbf{x},\mathbf{y}) + \beta f_X(\mathbf{x}) f_Y(\mathbf{y}) \right)^2, \end{aligned} \tag{A27}$$

where δ_f is given in (A15). Next After all still the expression (A27) is complicated to optimize therefore we use the fact that $0 \leq \alpha, \beta \leq 1$ to bound the function $\mathbf{G}^{\alpha,\beta}_{f,n}(\mathbf{x},\mathbf{y})$. Define the set Γ

$$\Gamma := \left\{ \epsilon_{XY} : \left| \epsilon_{XY}(\mathbf{t}) - \epsilon_{XY}(\mathbf{t}') \right| \leq \overline{K} \|\mathbf{t} - \mathbf{t}'\|_d^{\eta} \right\},$$

where

$$\overline{K} = C_\epsilon^U K \left\{ C_{XY}^L + C_X^L + C_Y^L C_X^L C_X^U \right\}.$$

Here K is the smoothness constant in the Hölder class. Notice that set Γ is a convex set. We bound $\overline{\Delta}$ by

$$\widetilde{\Delta}(\alpha, \epsilon_{XY}) = D(n, l_n, d, \eta) + \widetilde{D}(n, l_n, d)\, C_{XY}^U \iint_{\mathbb{S}_{XY}} \widetilde{G}^{\alpha,\beta}_{\epsilon_{XY},n}(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}. \tag{A28}$$

Set $C_n = C_{XY}^L\, n/2$,

$$\widetilde{G}_n^{\alpha,\beta}(\epsilon_{XY}) = \frac{\left(\epsilon_{XY}^{-1}(\mathbf{x},\mathbf{y}) + (\beta C_n)^{-1})(1 + \epsilon_{XY}^{-1}(\mathbf{xy}) + (\beta C_n)^{-1}\right)}{\left(\alpha + \beta\epsilon_{XY}^{-1}(\mathbf{x},\mathbf{y})\right)^2}. \tag{A29}$$

This simplifies to

$$\widetilde{G}_n^{\alpha,\beta}(\epsilon_{XY}) = \frac{\left(1 + (\beta C_n)^{-1}\epsilon_{XY}\right)\left(1 + \epsilon_{XY} + (\beta C_n)^{-1}\epsilon_{XY}\right)}{\left(\alpha\epsilon_{XY} + \beta\right)^2}. \tag{A30}$$

Under the condition

$$\frac{2}{C_n} \le \alpha \le \min\left\{\frac{1}{2} + \frac{1}{2C_n}, \frac{1}{3} + \frac{2}{3C_n}\right\}, \tag{A31}$$

$\widetilde{G}_n^{\alpha,\beta}(\epsilon_{XY})$ is an increasing function in ϵ. Furthermore, for $\alpha \le \frac{1}{4}$ and

$$C_\epsilon^L \le \epsilon_{XY} \le \min\left\{C_\epsilon^U, \theta^U(\alpha)\right\}, \quad \text{where } \theta^U(\alpha) = \frac{1 - 4\alpha + 1/C_n}{2\alpha}, \tag{A32}$$

the function $\widetilde{G}_n^{\alpha,\beta}(\epsilon_{XY})$ is strictly concave. Next, to find an optimal α we consider the following optimization problem:

$$\begin{aligned} \min_{\alpha} \max_{\epsilon_{XY}\in\Gamma} \quad & \widetilde{\Delta}(\alpha, \epsilon_{XY}) + c_d(1-\alpha)n^{-1} \\ \text{subject to} \quad & C_\epsilon^L \le \epsilon_{XY} \le C_\epsilon^U, \end{aligned} \tag{A33}$$

here $\epsilon_{XY} = f_{XY}/f_X f_Y$, $C_\epsilon^U = C_{XY}^U/C_X^L C_Y^L$ and $C_\epsilon^L = C_{XY}^L/C_X^U C_Y^U$, such that $C_\epsilon^L \le 1$. We know that under conditions (A31) and (A32), the function $\widetilde{G}_n^{\alpha,\beta}$ is strictly concave and increasing in ϵ_{XY}. We first solve the optimization problem:

$$\begin{aligned} \max_{\epsilon_{XY}\in\Gamma} \quad & \iint_{\mathbb{S}_{XY}} \widetilde{G}_n^{\alpha,\beta}(\epsilon_{XY}(\mathbf{x},\mathbf{y}))\, \mathbf{dxdy} \\ \text{subject to} \quad & \theta_\epsilon^L(\alpha)\mathbb{V}(\mathbb{S}_{XY}) \le \iint_{\mathbb{S}_{XY}} \epsilon_{XY}(\mathbf{x},\mathbf{y})\mathbf{dxdy} \\ & \le \theta_\epsilon^U(\alpha)\mathbb{V}(\mathbb{S}_{XY}), \end{aligned} \tag{A34}$$

where

$$\theta_\epsilon^L(\alpha) := C_\epsilon^L, \quad \theta_\epsilon^U(\alpha) := \min\{C_\epsilon^U, \theta^U(\alpha)\}. \tag{A35}$$

The Lagrangian for this problem is

$$L(\epsilon_{XY}, \lambda_1, \lambda_2) = \iint_{\mathbb{S}_{XY}} \widetilde{G}_n^{\alpha,\beta}(\epsilon_{XY}(\mathbf{x},\mathbf{y}))\, \mathbf{dxdy} - \lambda_1\left(\iint_{\mathbb{S}_{XY}} \epsilon_{XY}(\mathbf{x},\mathbf{y})\, \mathbf{dxdy} - \theta_\epsilon^U(\alpha)\mathbb{V}(\mathbb{S}_{XY})\right)$$

$$-\lambda_2\left(\theta_\epsilon^L(\alpha)\mathbb{V}(\mathbb{S}_{XY}) - \iint_{\mathbb{S}_{XY}} \epsilon_{XY}(\mathbf{x},\mathbf{y})\, \mathbf{dxdy}\right).$$

In this case, the optimum ϵ_{XY}^* is bounded, $\theta_\epsilon^L(\alpha) \le \epsilon_{XY}^* \le \theta_\epsilon^U(\alpha)$, and the Lagrangian multiplier $\lambda_1^*,\ \lambda_2^* \ge 0$ is such that

$$\min_{\lambda_1,\lambda_2\ge 0} \max_{\epsilon_{XY}\in\Gamma} L(\epsilon_{XY}, \lambda_1, \lambda_2) = L(\epsilon_{XY}^*, \lambda_1^*, \lambda_2^*).$$

Set $G'_n(\epsilon_{XY}) = \dfrac{\mathrm{d}}{\mathrm{d}\epsilon_{XY}}\widetilde{G}_n^{\alpha,\beta}(\epsilon_{XY})$. In view of the concavity of $\widetilde{G}_{\epsilon_{XY},n}^{\alpha,\beta}$ and Lemma 1, page 227 in [49], maximizing $L(\epsilon_{XY}, \lambda_1^*, \lambda_2^*)$ over ϵ_{XY} is equivalent to

$$\iint_{\mathbb{S}_{XY}} \Big\{ G'_n\big(\epsilon^*_{XY}(\mathbf{x},\mathbf{y})\big) - (\lambda_1^* - \lambda_2^*) \Big\} \epsilon_{XY}(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \le 0, \tag{A36}$$

for all $\theta_\epsilon^L(\alpha) \le \epsilon^*_{XY} \le \theta_\epsilon^U(\alpha)$, and

$$\iint_{\mathbb{S}_{XY}} \Big\{ G'_n\big(\epsilon^*_{XY}(\mathbf{x},\mathbf{y})\big) - (\lambda_1^* - \lambda_2^*) \Big\} \epsilon^*_{XY}(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = 0. \tag{A37}$$

Denote ${G'}_n^{-1}$ the inverse function of G'_n. Since G'_n is strictly decreasing in ϵ^*_{XY} (this is because $\tilde{G}_n^{\alpha,\beta}(\epsilon_{XY})$ is strictly concave, so that ${G'}_n^{-1}$ is continuous and strictly decreasing in ϵ^*_{XY}). From (A36) and (A37), we see immediately that on any interval $\theta_\epsilon^L(\alpha) \le \epsilon^*_{XY} \le \theta_\epsilon^U(\alpha)$, we have $\epsilon^*_{XY} = {G'}_n^{-1}(\lambda_1^* - \lambda_2^*)$. We can write then

$$G'_n\big(\theta_\epsilon^U(\alpha)\big) \le \lambda_1^* - \lambda_2^* \le G'_n\big(\theta_\epsilon^L(\alpha)\big),$$

and $\lambda_1^*, \lambda_2^* \ge 0$. Next, we find the solution of

$$\min_{\lambda_1,\lambda_2 \ge 0} \ \overline{G}_n^{\alpha,\beta}(\lambda_1,\lambda_2), \quad \text{where}$$

$$\overline{G}_n^{\alpha,\beta}(\lambda_1,\lambda_2) = \mathbb{V}(\mathbb{S}_{XY}) \Big\{ \tilde{G}_n^{\alpha,\beta}\big({G'}_n^{-1}(\lambda_1-\lambda_2)\big) - (\lambda_1-\lambda_2){G'}_n^{-1}(\lambda_1-\lambda_2) + \lambda_1\theta_\epsilon^U(\alpha) - \lambda_2\theta_\epsilon^L(\alpha) \Big\}.$$

The function $\overline{G}_n^{\alpha,\beta}(\lambda_1,\lambda_2)$ is increasing in λ_1 and λ_2, and therefore it takes its minimum at $(\lambda_1^*, \lambda_2^*) = \big(G'_n(\theta_\epsilon^U(\alpha)), 0\big)$. This implies that $\epsilon^*_{XY} = \theta_\epsilon^U(\alpha)$. Returning to our primary minimization over α:

$$\begin{aligned} \min_{\alpha} \quad & \widetilde{\Delta}(\alpha, \epsilon^*_{XY}) + c_d(1-\alpha)n^{-1} \\ \text{subject to} \quad & \alpha_0^L \le \alpha \le \alpha_0^U, \end{aligned} \tag{A38}$$

where $\alpha_0^L = \dfrac{2}{C_n}$ and $\alpha_0^U = \min\Big\{\dfrac{1}{4}, 1 - n^{\eta/d-1}\Big\}$. We know that $\dfrac{1}{4} \le \dfrac{1}{3} + \dfrac{2}{3C_n}$ and $\dfrac{1}{4} \le \dfrac{1}{2} + \dfrac{1}{2C_n}$, therefore the condition below

$$\frac{2}{C_n} \le \alpha \le \min\Big\{\frac{1}{4}, 1 - n^{\eta/d-1}\Big\},$$

implies the constraint $\alpha_0^L \le \alpha \le \alpha_0^U$. Since the objective function (A38) is a complicated function in α, it is not feasible to determine whether it is a convex function in α. For this reason, let us solve the optimization problem in (A38) in a special case when $C_\epsilon^U \le \theta^U(\alpha)$. This implies $\epsilon^*_{XY} = C_\epsilon^U$. Under assumption C_ϵ^U the objective function in (A38) is convex in α. Also, the case $C_\epsilon^U \le \theta^U(\alpha)$ is equivalent to $\alpha \le \dfrac{1 + 1/C_n}{4 + 2C_\epsilon^U}$. Therefore, in the optimization problem we have constraint

$$\frac{2}{C_n} \le \alpha \le \min\Big\{\frac{1}{4}, \frac{1+1/C_n}{4+2C_\epsilon^U}, 1 - n^{\eta/d-1}\Big\}.$$

We know that $\widetilde{\Delta}(\alpha, \epsilon^*_{XY}) + c_d(1-\alpha)n^{-1}$ is convex over $\alpha \in [\alpha_0^L, \alpha_0^U]$. Therefore, the problem becomes ordinary convex optimization problem. Let $\tilde{\alpha}$, $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ be any points that satisfy the KKT conditions for this problem:

$$\begin{aligned} &\alpha_0^L - \tilde{\alpha} \leq 0, \quad \tilde{\alpha} - \alpha_0^U \leq 0, \quad \tilde{\lambda}_1, \tilde{\lambda}_2 \geq 0, \\ &\tilde{\lambda}_1(\alpha_0^L - \tilde{\alpha}) = 0, \quad \tilde{\lambda}_2(\tilde{\alpha} - \alpha_0^U) = 0, \\ &\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\widetilde{\Delta}(\tilde{\alpha}, \epsilon_{XY}^*) + c_d(1-\tilde{\alpha})\, n^{-1}\right) - \tilde{\lambda}_1 + \tilde{\lambda}_2 = 0. \end{aligned} \tag{A39}$$

Recall $\Xi(\alpha)$ from (20):

$$\Xi(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\widetilde{\Delta}(\alpha, \epsilon_{XY}^*) + c_d(1-\alpha)n^{-1}\right),$$

where $\widetilde{\Delta}$ is given in (A28). So, the last condition in (A39) becomes $\Xi(\tilde{\alpha}) = \tilde{\lambda}_1 - \tilde{\lambda}_2$. We then have

$$\alpha_0^L \leq \Xi^{-1}(\tilde{\lambda}_1 - \tilde{\lambda}_2) \leq \alpha_0^U,$$

where Ξ^{-1} is inverse function of Ξ. Since $\alpha_0^L \neq \alpha_0^U$, at least one of $\tilde{\lambda}_1$ or $\tilde{\lambda}_2$ should be zero:

- $\tilde{\lambda}_1 = 0, \tilde{\lambda}_2 \neq 0$. Then $\tilde{\alpha} = \alpha_0^U$ and implies $\tilde{\lambda}_2 = -\Xi(\alpha_0^U)$. Since $\tilde{\lambda}_2 > 0$, so this leads to $\Xi(\alpha_0^U) < 0$.
- $\tilde{\lambda}_2 = 0, \tilde{\lambda}_1 \neq 0$. Then $\tilde{\alpha} = \alpha_0^L$ and implies $\tilde{\lambda}_1 = \Xi(\alpha_0^L)$. We know that $\tilde{\lambda}_1 > 0$, hence $\Xi(\alpha_0^L) > 0$.
- $\tilde{\lambda}_1 = 0, \tilde{\lambda}_2 = 0$. Then $\tilde{\alpha} = \Xi^{-1}(0)$ and so $\alpha_0^L \leq \Xi^{-1}(0) \leq \alpha_0^U$.

Consequently, by following the behavior of $\Xi(\alpha)$ with respect to α_0^L and α_0^U, we can often find the optimal $\tilde{\alpha}$, $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$. For instance, if $\Xi(\alpha)$ is positive for all $\alpha \in [\alpha_0^L, \alpha_0^U]$ then we conclude that $\tilde{\alpha} = \alpha_0^L$.

References

1. Lewi, J.; Butera, R.; Paninski, L. Real-time adaptive information theoretic optimization of neurophysiology experiments. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2006; pp. 857–864.
2. Peng, H.C.; Herskovits, E.H.; Davatzikos, C. Bayesian Clustering Methods for Morphological Analysis of MR Images. In Proceedings of the IEEE International Symposium on Biomedical Imaging, Washington, DC, USA, 7–10 July 2002; pp. 485–488.
3. Moon, K.R.; Noshad, M.; Yasaei Sekeh, S.; Hero, A.O. Information theoretic structure learning with confidence. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.
4. Brillinger, D.R. Some data analyses using mutual information. *Braz. J. Probab. Stat.* **2004**, *18*, 163–183.
5. Torkkola, K. Feature extraction by non parametric mutual information maximization. *J. Mach. Learn. Res.* **2003**, *3*, 1415–1438.
6. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [CrossRef]
7. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
8. Peng, H.; Long, F.; Ding, C. Evaluation, Application, and Small Sample Performance. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 153–158.
9. Sorjamaa, A.; Hao, J.; Lendasse, A. Mutual information and knearest neighbors approximator for time series prediction. *Lecture Notes Comput. Sci.* **2005**, *3697*, 553–558.
10. Mohamed, S.; Rezende, D.J. Variational information maximization for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2015; pp. 2116–2124.
11. Neemuchwala, H.; Hero, A.O. Entropic graphs for registration. In *Multi-Sensor Image Fusion and its Applications*; CRC Press Book: Boca Raton, FL, USA, 2005; pp. 185–235.
12. Neemuchwala, H.; Hero, A.O.; Zabuawala, S.; Carson, P. Image registration methods in high-dimensional space. *Int. J. Imaging Syst. Technol.* **2006**, *16*, 130–145. [CrossRef]

13. Henze, N.; Penrose, M.D. On the multivarite runs test. *Ann. Stat.* **1999**, *27*, 290–298.
14. Berisha, V.; Hero, A.O. Empirical non-parametric estimation of the Fisher information. *IEEE Signal Process. Lett.* **2015**, *22*, 988–992. [CrossRef]
15. Berisha, V.; Wisler, A.; Hero, A.O.; Spanias, A. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Trans. Signal Process.* **2016**, *64*, 580–591. [CrossRef]
16. Kailath, T. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [CrossRef]
17. Yasaei Sekeh, S.; Oselio, B.; Hero, A.O. Multi-class Bayes error estimation with a global minimal spanning tree. In Proceedings of the 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2–5 October 2018.
18. Noshad, M.; Moon, K.R.; Yasaei Sekeh, S.; Hero, A.O. Direct Estimation of Information Divergence Using Nearest Neighbor Ratios. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017.
19. March, W.; Ram, P.; Gray, A. Fast Euclidean minimum spanning tree: Algorithm, analysis, and applications. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, 20–23 August 2010; pp. 603–612.
20. Borůvka, O. O jistém problému minimálním. *Práce Moravské Pridovedecké Spolecnosti* **1926**, *3*, 37–58.
21. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066–138. [CrossRef] [PubMed]
22. Moon, K.R.; Sricharan, K.; Hero, A.O. Ensemble Estimation of Mutual Information. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 3030–3034.
23. Moon, K.R.; Hero, A.O. Multivariate f-divergence estimation with confidence. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 2420–2428.
24. Moon, K.R.; Sricharan, K. Greenewald, K.; Hero, A.O. Improving convergence of divergence functional ensemble estimators. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016.
25. Leonenko, N.; Pronzato, L.; Savani, V. A class of Rényi information estimators for multidimensional densities. *Ann. Stat.* **2008**, *36*, 2153–2182. [CrossRef]
26. Gao, S.; Ver Steeg, G.; Galstyan, A. Efficient estimation of mutual information for strongly dependent variables. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 277–286.
27. Pál, D.; Póczos, B.; Szapesvári, C. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In Proceedings of the 24th Annual Conference on Neural Information Processing Systems 2010, Vancouver, BC, Canada, 6–9 December 2010.
28. Krishnamurthy, A.; Kandasamy, K.; Póczos, B.; Wasserman, L. Nonparametric estimation of Rényi divergence and friends. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 919–927.
29. Kandasamy, K.; Krishnamurthy, A.; Póczos, B.; Wasserman, L.; Robins, J. Nonparametric von mises estimators for entropies, divergences and mutual informations. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 397–405.
30. Singh, S.; Póczos, B. Analysis of k nearest neighbor distances with application to entropy estimation. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 1217–1225.
31. Sugiyama, M. Machine learning with squared-loss mutual information. *Entropy* **2012**, *15*, 80–112. [CrossRef]
32. Costa, A.; Hero, A.O. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. Signal Process.* **2004**, *52*, 2210–2221. [CrossRef]
33. Friedman, J.H.; Rafsky, L.C. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Stat.* **1979**, *7*, 697–717. [CrossRef]
34. Smirnov, N.V. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Moscow Univ.* **1939**, *2*, 3–6.

35. Wald, A.; Wolfowitz, J. On a test whether two samples are from the same population. *Ann. Math. Stat.* **1940**, *11*, 147–162. [CrossRef]
36. Noshad, M.; Hero, A.O. Scalable mutual information estimation using dependence graphs. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTAT), Brighton, UK, 12–17 May 2019.
37. Yasaei Sekeh, S.; Oselio, B.; Hero, A.O. A Dimension-Independent discriminant between distributions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
38. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **1967**, *2*, 299–318.
39. Ali, S.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B (Methodol.)* **1966**, *28*, 131–142. [CrossRef]
40. Cover, T.; Thomas, J. *Elements of Information Theory*, 1st ed.; John Wiley & Sons: Chichester, UK, 1991.
41. Härdle, W. *Applied Nonparametric Regression*; Cambridge University Press: Cambridge, UK, 1991.
42. Lorentz, G.G. *Approximation of Functions*; Holt, Rinehart and Winston: New York, NY, USA; Chicago, IL, USA; Toronto, ON, Canada, 1966.
43. Andersson, P. Characterization of pointwise Hölder regularity. *Appl. Comput. Harmon. Anal.* **1997**, *4*, 429–443. [CrossRef]
44. Robins, G.; Salowe, J.S. On the maximum degree of minimum spanning trees. In Proceedings of the SCG 94 Tenth Annual Symposium on Computational Geometry, Stony Brook, NY, USA, 6–8 June 1994; pp. 250–258.
45. Efron, B.; Stein, C. The jackknife estimate of variance. *Ann. Stat.* **1981**, *9*, 586–596. [CrossRef]
46. Yasaei Sekeh, S.; Noshad, M.; Moon, K.R.; Hero, A.O. Convergence Rates for Empirical Estimation of Binary Classification Bounds. *arXiv* **2018**, arXiv:1810.01015.
47. Kingman, J.F.C. *Poisson Processes*; Oxford Univ. Press: Oxford, UK, 1993.
48. Yukich, J.E. *Probability Theory of Classical Euclidean Optimization*; Vol. 1675 of Lecture Notes in Mathematics; Springer: Berlin, Germany, 1998.
49. Luenberger, D.G. *Optimization by Vector Space Methods*; Wiley Professional Paperback Series; Wiley-Interscience: Hoboken, NJ, USA, 1969.