Abstract

In this work, we leverage the intuitive geometry of the Voronoi diagram to develop a new class of methods for quantifying and testing for statistical dependence between real-valued random variables. The contributions of this work are threefold: First, we introduce the utility of probability integral transforms of the datapoints to produce Voronoi diagrams which exhibit desirable properties provided by the theory of copula distributions. Second, we propose a novel nonparametric statistical test (VORTI) to determine whether two variables are independent based on the areas of the cells in the Voronoi diagram. Based on similar ideas, we finally introduce a new method (SVORMI) for estimating mutual information which relies on a novel regularization technique we found no mention of in the literature on density estimation via Voronoi tessellation. Through extensive simulation experiments, we demonstrate that our proposed estimator is competitive with the state-of-the-art \( k \)-nearest neighbor mutual information estimator of Kraskov et al. (2004).

Introduction

Detecting various forms of deviations from statistical independence between random variables is a cornerstone of statistics. A wide range of approaches to this problem have been explored, ranging from correlation coefficients based on data values/ranks/distances [Pearson (1896); Hazewinkel (2001); Kendall (1938); Szekely and Rizzo (2009)], Fourier or RKHS representations [Stein and Weiss (1971); Gretton et al. (2005, 2007)], and information-theoretic tools [Cover and Thomas (2006)]. Each class of methods has its advantages and disadvantages and relies on different assumptions on relationships present in the data.

An extremely popular measure of dependence in machine learning is the mutual information (see Definition 1). Intuitively, mutual information (MI) measures the degree to which knowing the value of one variable gives information about another. It has several properties that make it a desirable basis for a measure of dependence: it is non-negative and symmetric, it is invariant under order-preserving transformations of random variables \( X \) and \( Y \), and, most importantly, it equals zero precisely when \( X \) and \( Y \) are statistically independent. These properties—which are not shared by simpler measures of correlation such as Pearson correlation—have led to its widespread use in diverse applications [Elemento et al. (2007)].
The challenge in applying mutual information to quantify relationships between continuous variables lies in the fact that it is defined as a functional of their unknown probability density function (p.d.f.), which in practice must be estimated from finite data samples from this distribution. This task has proven difficult, leading to the development of a wide range of mutual information estimators that have been based on different techniques. The simplest estimators rely on discretizing the data via gridding to estimate its joint distribution within each grid [Steuer et al. (2002)]. More sophisticated methods utilize kernel density estimation (KDE) to estimate a smooth joint distribution [Moon et al. (1995)]. Compared with gridding-based approaches, KDE mutual information estimators exhibit a better mean square error rate of convergence of the estimate to the underlying density, an insensitivity to the choice of origin, and the ability to specify more flexible window shapes than the rectangular window used in grid-based approaches [Steuer et al. (2002); Moon et al. (1995)].

Exploring the geometric properties of randomly sampled data points from some bivariate distribution, we find that the Voronoi diagram exhibits different characteristics in the case where the two random variables are statistically dependent than if the variables were independent. By exploiting these properties, we are able to develop new methods for assessing dependence and measuring mutual information. Our approaches are both intuitive and highly successful in simulation studies.

Notation

In this paper, all logarithms are assumed to be natural (base e), so MI is estimated in nats rather than bits (as is common practice in the continuous setting). Also, all geometry in this work is assumed to be Euclidean. To keep our exposition clean and intuitive, we limit our discussion to univariate real-valued random variables X, Y ∈ R (unless explicitly stated otherwise), although all of our methods may be applied to higher-dimensional variables in practice. Furthermore, the primary goal of many data analyses is to identify and measure pairwise dependencies between single features which are commonly measured on a continuous scale in R, so our restriction to the 1-D setting is widely applicable.

Let $F_{XY}$ denote the joint distribution of $(X,Y)$, let $F_X, F_Y$ denote the respective marginal distributions (as cumulative distribution functions), and let $f_{XY}, f_X, f_Y$ the corresponding probability density functions. Given n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ from $F_{XY}$, we wish to measure the statistical dependence between X and Y by estimating $I(X,Y)$, the continuous mutual information of X and Y.

**Definition 1** The mutual information of X and Y is given by

$$I(X,Y) := \int_{\mathbb{R}^2} f_{XY}(x,y) \log \left( \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} \right) \, d(x,y)$$

Copulas

In our methods, we first apply the probability integral transforms $U = F_X(X), V = F_Y(Y)$ so that the random vector $(U, V)$ has uniformly distributed marginals. By Sklar’s Theorem,
any distribution \( F_{XY} \) can be expressed in terms of univariate marginal distributions and a copula in which the dependence between \( X \) and \( Y \) is fully encoded Sklar (1959).

**Definition 2** The copula of \((X, Y)\) is defined as the joint c.d.f. of \((U, V)\)

\[
C(u, v) := \Pr[U \leq u, V \leq v] \Rightarrow C(u, v) = \Pr[X \leq F_X^{-1}(u), Y \leq F_Y^{-1}(v)]
\]  

(1)

Letting \( c(u, v) \) denote the p.d.f. corresponding to the copula c.d.f., a simple calculation demonstrates how the original joint density may be expressed in terms of the copula density of the transformed variables:

\[
f(x, y) = c(F_X(x), F_Y(y)) f_X(x) f_Y(y)
\]  

(2)

For independent random variables \( X, Y \): \( C(u, v) = u \cdot v \), while if \( X \) and \( Y \) are perfectly correlated, then \( C(u, v) = \min\{u, v\} \). This property of the copula is highly desirable for measuring dependence: no matter what form the joint distribution of \( X, Y \) takes (assuming it is sufficiently “nice”, e.g. smooth, supported on set of nonzero measure), if \( X \) and \( Y \) are independent, then after probability integral transform, the resulting joint distribution of \( U, V \) is always identical to the product distribution of standard Uniform\([0, 1]\) variables.

Since the true distributions are unknown, empirical c.d.f.s \( \hat{F}_X, \hat{F}_Y \) are used in practice to transform each sample \((x_i, y_i) \rightarrow (u_i = \hat{F}_X(x_i), v_i = \hat{F}_Y(y_i))\), where \( \hat{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[x_i \leq x] \). As the sample size \( n \) grows, these one-dimensional empirical c.d.f.s rapidly converge to their population counterparts pointwise at the standard \( \sqrt{n} \) rate given by the central limit theorem (Glivenko-Cantelli theorem says the convergence is in fact uniform), and a number of weak convergence results have also been established for the empirical copula process.

Voronoi diagrams for data analysis

After we have transformed our data via the probability integral transform so that all points lie in the 2-D unit square \( B = [0, 1]^2 \subset \mathbb{R}^2 \) (or the unit cube if dimension \( d > 2 \)). The Voronoi diagram/tesselation constructed from points \( x_1, \ldots, x_n \in B \), which we denote \( \text{Vor}(x_1, \ldots, x_n) \), is the unique partitioning of \( B \) into cells \( C_1, \ldots, C_n \) so that each polygon \( C_i = \{x \in B : ||x - x_i||_2 \leq ||x - x_j||_2 \quad \forall j \neq i\} \) is the region of \( B \) which is nearer to \( x_i \) than any of the other \( n - 1 \) points (see Figure 1 for an example). We include the boundary of \( B \) as faces of the cells which lie at the border.

Via the sweep-line algorithm of Fortune (1986), the Voronoi diagram of bivariate real-valued data points can be efficiently constructed in \( O(n \log n) \) time and \( O(n) \) space. For higher dimensional data, one may use the Bowyer-Watson incremental insertion algorithm [Bowyer (1981); Watson (1981)] or the method of Dwyer (1991) which is proven to run in linear expected time for random points (for a fixed \( d \)). These space-partitioning structures have a long history in computational geometry, but they have only drawn limited interest from statisticians, predominantly in the context of spatial data analysis [Barr (2008); Chui (2003)]. A major advantage of the Voronoi methods in data analysis are their adaptivity to the spatial structure of the sampled points. Given \( n \) points from an unknown distribution,
the standard Voronoi density estimator (VDE) produces the following step-function approximation to the probability density function (where $1(\cdot)$ denotes the indicator function):

$$
\hat{f}(p) = \frac{1}{n} \sum_{i=1}^{n} \frac{1[p \in C_i]}{\text{Vol}(C_i)} \quad \forall \ p \in \mathbb{R}^d
$$

The VDE may be derived from the fact that if we assume the underlying density is truly constant in each cell, i.e. $f(p) = c_i \ \forall p \in C_i$, then we must have:

$$
\text{Pr}[p \in C_i] = \int_{C_i} f(p) \, dp = \int_{C_i} c_i \, dp \implies c_i = \frac{\text{Pr}[p \in C_i]}{\text{Vol}(C_i)}
$$

where we have observed exactly one of our samples in each Voronoi cell, leading to the following empirical estimate: $\hat{\text{Pr}}[x \in C_i] = 1/n$. Intuitively, we expect more data from high density regions to appear in our sample and therefore the Voronoi cells in these regions will tend to have smaller volumes (in our 2-D setting, we use the terms volume and area interchangeably). For data sampled from a homogeneous Poisson process of rate $\lambda$, Meijering (1953) showed the expected Voronoi cell area of points sampled from a homogeneous Poisson process is $\lambda^{-1}$, a result which Barr (2008) have extended, demonstrating approximate unbiasedness of the Voronoi cell areas as estimators for the intensity of an inhomogeneous
Poisson process. Although Browne (2007) find that the bias of the VDE is remarkably low, they state that its variance is too large (as each Voronoi cell only contains a single point) to produce useful density estimates in practice. A number of variance-reducing modifications have been suggested to produce better estimators based on these geometric objects by imposing additional smoothness [Browne (2007, 2012); Liu (2007); Barr (2008)]. Because we are interested in measuring statistical dependence via the mutual information which involves averaging a function of the joint density over the sample space, we hypothesize that the effect of the highly variable Voronoi-based density estimates is mitigated by the concentration properties of the average (although the density estimates in neighboring cells are certainly not independent, the dependence between the estimates diminishes rapidly for cells which are farther apart in the Voronoi diagram, so we can still expect the mean estimate over all cells to exhibit some concentration).

Testing for statistical dependence

We have so far established two fundamental facts:

1. The cell-areas of the Voronoi tessellation for a random set of points offer an unbiased approximation of the local underlying density from which the points are drawn.

2. After probability integral transform, any independent \(X,Y\) follow the independent-bivariate uniform distribution over the unit square (with equal probability mass at each point in this region).

Integrating these two ideas, we expect the areas of the cells in the Voronoi diagram over \(\{(u_i, v_i)\}_{i=1}^n\) to exhibit more uniformity if \(X\) and \(Y\) are independent than otherwise. To empirically investigate this assumption, we generate the following 2 datasets of independent variables with very different marginal distributions and 3 datasets of different types of dependent variables:

- 1000 i.i.d. points where \(X_i \sim \text{Uniform}[0, 1], Y_i \sim \text{Uniform}[0, 1]\)
- 1000 i.i.d. points where \(X_i \sim N(5, 10), Y_i \sim N(2, 3)\)
- 1000 i.i.d. points where \((X_i, Y_i)\) follow a bivariate Gaussian with correlation = 0.9
- 1000 i.i.d. points where \((X_i, Y_i)\) follow a bivariate Gaussian with correlation = 0.2
- 1000 i.i.d. points where \(X_i \sim N(0, 2)\) and \(Y_i \sim \sin(x_i) + \epsilon\) where \(\epsilon \sim \text{Uniform}[-\frac{1}{2}, \frac{1}{2}]\) (see Figure 6 for an example dataset)

For each of the distributions of the i.i.d. points, we generate the corresponding Voronoi diagram and compute the Voronoi test for independence (VORTI) 10 times and we plot the resulting distribution of Voronoi cell areas in Figure 2.

**VORTI (Voronoi Test for Independence)**

\[H_0 : X \text{ and } Y \text{ are independent}
\]
\[H_A : X \text{ and } Y \text{ are not independent (i.e. } P(Y \mid X) \neq P(Y))\]
Figure 2: Violin plots the distribution of cell areas in the Voronoi diagrams corresponding to 1000 randomly sampled points from different distributions. Each plot depicts a kernel density estimate in the vertical direction fitted to the Voronoi cell areas from 10 datasets of size 1000 sampled from each distribution. The median Voronoi cell area is represented by the white dot, which is in turn incorporated into a boxplot of the cell areas (shown in black).

1. Compute empirical c.d.f.s $\hat{F}_X, \hat{F}_Y$ (where $\hat{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} 1[x_i \leq x]$)

2. For $i \in \{1, \ldots, n\}$: transform data $u_i = \hat{F}_X(x_i), v_i = \hat{F}_Y(y_i)$

3. Construct Voronoi diagram $Vor((u_1, v_1), \ldots, (u_n, v_n))$

4. Compute empirical distribution $Areas$ which is a list of the areas of each cell in the Voronoi diagram.

5. Lookup the appropriate approximate null distribution $\hat{F}_0$ for this sample size $n$ from a pre-computed table

6. Perform one-sample Kolmogorov-Smirnov test whether $Areas$ is a sample from $\hat{F}_0$ and return the $p$-value of this test as the $p$-value for this independence test

VORTI involves a pre-computation stage which is only done one time ever to obtain simulated null-distributions of the Voronoi-cell areas under independence for large sample sizes $n$. In this pre-computation: we generate $B$ simulated datasets of size $n$ in which both variables are i.i.d. samples from the $Uniform[0,1]$ distribution. For each dataset, we construct the Voronoi diagram (after empirical copula transform) and calculate the areas of each Voronoi cell. Since the joint distribution of the probability-integral-transformed $U, V$ follows the independent-uniform distribution for any independent $X, Y$ (regardless of their
Voronoi Methods for Assessing Statistical Dependence

distribution), the distribution of these pre-computed areas converges (as \( B \to \infty \)) to \( F_0 \), the true distribution of \( \text{Areas} \) under the null hypothesis \( H_0 \) (we set \( B = 1000 \) in our experiments). Note that \( n \) must be very large; we cannot obtain an accurate null distribution of our test statistic this way in the small sample-size regime because there is no guarantee that the empirical copula process which describes our finite sampled data is close to the independent uniform joint distribution. For small sample sizes, the underlying joint distribution of \( X \) and \( Y \) thus cannot be ignored (even under independence) as it will heavily influence the distribution of the Voronoi-cell \( \text{Areas} \) statistic computed from the probability-integral-transformed data.

Nevertheless, assuming \( n \) is sufficiently large, we have: if \( X \) and \( Y \) are independent, then \( \text{Areas} \) is approximately a sample from this null distribution, so we can test whether the latter is possible to determine whether the former holds. Furthermore, if \( X \) and \( Y \) are not independent, then the corresponding \( U,V \) will not be uniformly distributed in the unit square, so there will be regions of unit square where the joint density of \( (U,V) \) is \( > 1 \) from which Voronoi cell areas are expected to be smaller than under a uniform density, as well as regions of the square where the density is less than \( 1 \) in which Voronoi cell areas will be larger on average. Thus, in the case of dependent \( X,Y \), the distribution from which \( \text{Areas} \) is a sampled will necessarily be distinct from the null distribution. By adopting Voronoi cell-areas as a test-statistic, we manage to convert a 2-D goodness-of-fit test of the empirical 2-D copula distribution against the independent-uniform joint distribution (which is equivalent to our test of \( H_0 \) vs \( H_A \)) to a 1-D goodness-of-fit test of the empirical distribution of \( \text{Areas} \) against \( F_0 \). This reduction is likely even more useful if we are testing higher-dimensional \( X \) and \( Y \) for independence as the performance of goodness-of-fit tests heavily suffers with increasing dimensionality.

The final step of VORTI employs the Kolmogorov-Smirnov (KS) test, a popular non-parametric method for determining whether one-dimensional observed data is sampled from some continuous reference distribution \( F_0 \) [Bickel and Doksum (2003)]. This test requires minimal assumptions regarding the true data-generating distribution and compares both location and shape of the empirical distribution with \( F_0 \). As a result of its extreme flexibility, the KS test does not exhibit much statistical power, generally requiring fairly large \( n \) to reject the null hypothesis. This is not a problem for our method as we anyway require large \( n \) to establish an accurate approximation to \( F_0 \) in the first place. While we have not yet explored the statistical power of VORTI in practice, we believe the proposal of a novel nonparametric test to detect arbitrary forms of dependence from real-valued data is a significant contribution in itself, considering the dearth of methods currently available to tackle this basic problem.

Estimating mutual information

The simplest adaptation of the Voronoi diagram to estimate mutual information is to simply replace the rectangular grid (of user-specified bin size) in the discretization-based estimators with the data-adaptive tessellation produced by the Voronoi method and the corresponding
density estimates of the VDE. We improve upon this approach by leveraging properties of the copula.

**Theorem 1** The mutual information between $X$ and $Y$ is equal to the negative continuous entropy of the corresponding copula distribution:

$$I(X, Y) = \int_0^1 \int_0^1 c(u, v) \log |c(u, v)| \, du \, dv$$

**Proof**

$$I(X, Y) = \int_{\mathbb{R}^2} f_{XY}(x, y) \log \left( \frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} \right) \, d(x, y)$$

$$= \int_{\mathbb{R}^2} c(F_X(x), F_Y(y)) f_X(x) f_Y(y) \log [c(F_X(x), F_Y(y))] \, d(x, y) \quad \text{(applying (2))}$$

$$= \int_0^1 \int_0^1 c(u, v) \log |c(u, v)| \, du \, dv \quad \text{(by change-of-variables transform)}$$

Combining the result of Theorem 1 and the definition of the VDE (3), we obtain a Voronoi-based estimator of mutual information (VORMI) by computing the following statistic from the cells of the Voronoi diagram constructed over the probability integral transformed data points:

$$VORMI = \sum_{i=1}^n \frac{1}{n \cdot \text{Area}(C_i)} \log \left[ \frac{1}{n \cdot \text{Area}(C_i)} \right] \text{Area}(C_i)$$

$$= \frac{1}{n} \sum_{i=1}^n \log \left[ \frac{1}{n \cdot \text{Area}(C_i)} \right] \quad (4)$$

Interested in estimating multivariate entropies, Miller (2003) have previously considered this same estimator, and like Browne (2007, 2012); Barr (2008), they present a method to reduce the variance of the resulting estimator. We now describe all of the previously proposed variance reduction strategies in turn and discuss why each one is inadequate for accurate mutual information estimation.

To remedy the high-variance of the VORMI estimator, Miller (2003) suggest agglomeratively merging neighboring Voronoi cells into “hyper-regions” (starting with $m$ random seed cells) and performing the estimation as if these hyper-regions were the original Voronoi cells, but they themselves acknowledge that this Voronoi-cell-binning clearly biases the resulting entropy estimates upwards (since the probability mass is assumed to uniformly distributed within each hyper-region). We find that even without the merging of cells into hyper-regions, VORMI tends to significantly underestimate mutual information in the case where $X$ and $Y$ are highly dependent (see Figure 3) due to its assumption of uniformly distributed probability mass within each cell, and merging cells into even larger supposedly uniformly-distributed regions will only exacerbate this problem. Furthermore, such
a merging approach negates a major advantage the Voronoi tessellation posses over kernel/griding based methods, namely the fact that it probes the data at the finest resolution possible. Because Miller (2003) do not provide any empirical results demonstrating that use of the hyper-region areas leads to improved entropy estimates, we opt to forgo this approach.

To improve upon the VDE, Browne (2012) propose subsampling the data (only retaining \( m \) out of the \( n \) points where \( m \ll n \)) and averaging the Voronoi regions over numerous random subsamples to obtain a smoother final estimate of the density. Although this is a reasonable approach for density estimation, it produces significantly downwardly biased MI estimates. Because each subsample only consists of a small fraction of the data, the Voronoi diagram over the subsample is thus far less fine-grained than the Voronoi diagram over the full data, which as in the case of the hyper-region approach leads to larger regions which are assigned uniform probability. In the case of dependent random variables (in which case the underlying copula is not uniform over the unit square), this again leads to the appearance of greater copula entropy than is actually present in the underlying distribution and thus underestimates of mutual information from this subsample. Finally, averaging over downwardly biased subsamples cannot resolve this issue, so the resulting overall MI estimator is also heavily biased. We implemented a subsampling-based variant of VORMI and confirmed empirically that this is indeed the case (especially in the small sample-size regime of 30-100 which is typical for datasets in the sciences).

Another improvement to reduce the variance of the VDE proposed by both Browne (2007) and Barr (2008) is the regularization of the Voronoi diagram over the observed data so that it more closely resembles a centroidal Voronoi tessellation (CVT). Given a set of points and their Voronoi diagram, the CVT is produced by iteratively updating the locations of each point to be the centroid of its Voronoi cell and recomputing the Voronoi diagram until we have a “regular” Voronoi tessellation in which each point lies exactly at the centroid of its Voronoi cell. However, theoretical justifications of this approach are unclear, the method is unfaithful to the observed data (points are moved away from their original locations), and the resulting distribution of Voronoi cell areas after regularization is biased toward uniformity.

From Figure 3, we find that VORMI in fact does not even exhibit the problem of significant variance (it is no more variable that the \( k \)-NN estimator), suggesting that sole emphasis on variance-reduction will only produce limited improvements in MI estimates. While the individual probability density estimate at a specific location is highly variable (it entirely depends on the area of the Voronoi cell encompassing this location which may fluctuate greatly depending on the geometric arrangement of the sampled data points), VORMI sums functions of these cell-areas over all \( n \) Voronoi cells and therefore exhibits much greater stability presumably as a result of the concentration of measure phenomenon (despite the fact that estimates from nearby cells are not necessarily independent).

Clearly evident from Figure 3, the primary factor impairing the performance of VORMI is its high bias rather than a large variance: From independent (or low-dependence) data, VORMI heavily overestimates MI on average due to the fact that numerous small Voronoi
cells appear with high probability, even if the data are truly uniformly distributed throughout the unit square (see Figure 1 for an example). In the opposite situation where $X$ and $Y$ are highly dependent, VORMI tends to significantly underestimate MI because the assumption of uniform density within each Voronoi cell is a significant overestimate in the cells which touch the border of the $[0,1]^2$ bounding square assuming $X,Y$ are each supported on the entire real line (because then their densities must fall off near the boundaries of the space in which all probability integral transformed data points are constrained to lie). Furthermore, the number of these boundary-adjacent cells is typically much greater when the variables are highly dependent; see Figure 1). This issue is most apparent in the extreme case where $X$ and $Y$ are comonotonic or countermonotonic (i.e. $Y = g(X)$ for some monotonic function $g$), then all empirical probability integral transformed data points in will lie on a line $L$. Thus, every cell of the Voronoi diagram over these points will touch the unit square boundary and the areas of these cells do not accurately reflect the fact that the underlying joint distribution is only supported on $L$.

Smoothed Voronoi diagram for mutual information estimation

To amend the previously highlighted faults of VORMI, we propose a bias-correction technique which drastically improves MI estimates. The resulting estimator, which we refer to as SVORMI, is very simple and balances the following two separate objectives while remaining faithful to the original Voronoi diagram, and hence the data:

1. Smoothness in the distribution corresponding to Voronoi cell pseudo-areas
2. Mitigation of the effects of the boundary-bordering cell-areas which may be very poor approximations of the local density.

Maintaining a set $B$ which contains the indices of Voronoi cells that intersect the unit square boundary, SVORMI also maintains a pseudo-area $A_i$ for each Voronoi cell $C_i$ which is a smoothed approximation to the cell’s true area. Initializing all $A_i$ with the areas of the corresponding Voronoi cells in the tessellation constructed over our probability-integral-transformed data points, SVORMI simultaneously applies a border-sensitive local-averaging update to all $A_i$ in the $t^{th}$ iteration.

More specifically, in each iteration of SVORMI, the pseudo-area of a non-boundary-bordering cell $C_i$ is locally averaged with the pseudo-areas of its non-boundary-bordering neighboring Voronoi cells to produce the new pseudo area $A_i$, while the pseudo-area of a boundary-bordering cell is locally averaged with the pseudo-areas of all neighboring cells, regardless whether they’re boundary-bordering or not. In a single-iteration, these updates are simultaneously applied to all cells using the pseudo-area values from the previous iteration. Finally, the mutual information is estimated using the VORMI formula (4), but with pseudo-areas used in place of the original Voronoi cell areas.

The rationale behind our algorithm is that the non-boundary cell update will correct for the presence of a few Voronoi cells with minuscule area in the case of independent random variables by averaging their areas with the (in high probability larger-area) neighboring cells. At the same time, the areas of boundary cells, which are more likely to lead to inaccurate density/MI estimates are not allowed to affect the pseudo-areas of non-boundary
SVORMI Iterative update

\[ N(C_i) = \text{indices of neighboring Voronoi cells which share a face with } C_i \]
\[ B = \text{indices of Voronoi cells which intersect unit square boundary} \]

If \( i \notin B \) and \( |N(C_i) \setminus B| > 0 \):
\[ A_i^{(t+1)} \leftarrow \frac{1}{|N(C_i) \setminus B| + 1} \left[ A_i^{(t)} + \sum_{j \in N(C_i) \setminus B} A_j^{(t)} \right] \]

If \( i \in B \):
\[ A_i^{(t+1)} \leftarrow \frac{1}{|N(C_i)| + 1} \left[ A_i^{(t)} + \sum_{j \in N(C_i)} A_j^{(t)} \right] \]

Otherwise:
\[ A_i^{(t+1)} \leftarrow A_i^{(t)} \]

We find SVORMI to be fairly insensitive to the choice of \( T \) (the total number of border-sensitive neighborhood averaging updates) which in addition to controlling the bias due to the border Voronoi cells and cells of minuscule size also controls the overall smoothness of the resulting estimate. The resulting pseudo-areas produced after \( T \) SVORMI updates may also be viewed as solutions to the following modified graph-Laplacian optimization problem for a discrete sequence of increasing values of the regularization parameter \( \lambda \):

\[
\min_{A_1, \ldots, A_n} \sum_{i=1}^{n} (A_i - \text{Area}(C_i))^2 + \lambda \left( \sum_{(i,j) \in E, i,j \notin B} (A_i - A_j)^2 + \sum_{(i,j) \in E, i,j \in B} (A_i - A_j)^2 + \sum_{(i,j) \in E} (A_i - A_j)^2 \right)
\]

subject to \( \sum_{i=1}^{n} A_i = 1 \) and \( A_i > 0 \quad \forall i \) \hspace{1cm} (5)

where \( E \) is the set of \( O(n \log n) \) edges of the triangulation graph corresponding to the Delaunay triangulation of our data points (two data points are connected by an edge in this triangulation iff their Voronoi cells share a face). The Laplacian-optimization perspective of SVORMI makes explicit the objective of overall smoothness of the VDE-estimated density while simultaneously remaining faithful to the observed data. Solving this quadratic programming problem under different values of \( \lambda \) allows us to control the regularization of the SVORMI method at finer scales than permitted by the discrete nature of \( T \).

A principled way to select the parameter \( \lambda \) or \( T \) is leave-one-out likelihood maximization: Each individual data point \((u_i, v_i)\) is in turn left-out of the dataset and the Voronoi diagram construction proceeds over the other \( n-1 \) points. After we have computed the corresponding \( n-1 \) pseudo-areas, we evaluate the likelihood of the held-out \((u_i, v_i)\) as the probability density estimate produced for the Voronoi cell containing \((u_i, v_i)\) by the VDE with Voronoi
cell areas replaced by their smoothed pseudo-counterparts. Subsequently, we select the parameter setting which produces the best average held-out likelihood across each of the n points being omitted as the held-out point. We finally note that this leave-one-out method can be efficiently implemented if we have already computed the full Voronoi diagram by deleting a single point and updating the appropriate Voronoi cells from the data structure, rather than constructing the Voronoi diagram over the remaining n − 1 points from scratch for each of the n iterations of leave-one-out likelihood evaluation.

Results

One of the leading mutual information estimators is the k-nearest neighbor method (k-NN) introduced by Kraskov et al. (2004). This popular state-of-the-art method uses similar principles to the KDE-based approach, but rather than Gaussian kernels, k-nearest neighbor distances are used to obtain entropy estimates. The k-NN method exhibits a number of advantages over KDE mutual information estimators: it is more efficient, it is adaptive (resolution is higher where data are more numerous), and tends to have less bias [Kraskov et al. (2004)].

In our experiments, we find that this k-NN method significantly outperforms both the discretized and KDE-based mutual information estimates in applications to continuous data without additional parameter optimization. We thus omit comparisons of the methods presented in this paper with discretized / KDE-based mutual information estimates, focusing solely on how our methods fare against Kraskov’s presumably superior method.

SVORMI and the k-NN estimator are similar in their reliance on nearest neighbors, although the nearest neighbors considered in SVORMI are the set of points in the space which are closest to each data point, while the nearest neighbors in the k-NN method are simply k other samples which lie closer to a given data point than any other of the samples. Both methods are data efficient (considering structures down to the smallest scale possible for the given data), adaptive, and have minimal bias, although for extremely highly dependent variables, the MI estimates of both SVORMI and the k-NN estimator tend to remain biased downwards. However, this downward bias is of little concern since such great degrees of dependence are virtually nonexistent in real-world datasets, and furthermore, the continuous form of MI itself is an increasing unstable measure as Y approaches a function of X (it tends toward infinity; see Fact 1).

The number of local averaging updates used in SVORMI (T) is conceptually similar to the number of neighbors (k) in the k-NN MI estimation algorithm in that it regulates the resolution at which we treat the data (with T = 0, i.e. no neighborhood pseudo-area averaging, corresponding to the case in which the data are treated at the finest possible scale). In our experiments, we perform no parameter tuning, setting both T and k equal to \( \log_2(n) \) to obtain fair comparisons (the theory behind k-NN methods suggests \( O(\log n) \) to be optimal choice of k as sample size grows, but we have no theoretical justification for setting T to this value and it may be that SVORMI performs even better under different
Figure 3: Mutual information estimates from VORMI, SVORMI, and the $k$-NN method of Kraskov et al. (2004) applied to bivariate Gaussian data (of the indicated sample sizes $n$) and different levels of correlation. For each sample size and each correlation value, we sampled 100 datasets of size $n$ and computed estimates for each dataset. The colored dots indicate the average estimate over these 100 repetitions and the vertical bars denote the standard deviation, while the black dots illustrate the true mutual information of the underlying distribution.
Figures 3 and 4 clearly demonstrate that for samples of various sizes from a bivariate Gaussian distribution, the performance of SVORMI is competitive to that of the $k$-NN.
Voronoi Methods for Assessing Statistical Dependence

<table>
<thead>
<tr>
<th>Function Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear+Periodic, High Freq</td>
<td>$y = \frac{1}{10} \sin(10.6(2x - 1)) + \frac{11}{10}(2x - 1) \quad x \in [0, 1]$</td>
</tr>
<tr>
<td>Exponential $[2^x]$</td>
<td>$y = 2^x \quad x \in [0, 10]$</td>
</tr>
<tr>
<td>Parabola</td>
<td>$y = 4x^2 \quad x \in [-\frac{1}{2}, \frac{1}{2}]$</td>
</tr>
<tr>
<td>Sine, High Freq</td>
<td>$y = \sin(16\pi x) \quad x \in [0, 1]$</td>
</tr>
<tr>
<td>Varying Freq [Medium] Cosine</td>
<td>$y = \sin(5\pi x(1 + x)) \quad x \in [0, 1]$</td>
</tr>
</tbody>
</table>

Table 1: Definitions of the functions used to analyze the various mutual information estimators in this work.

estimator for all but the highest correlation values which are of little interest in practical settings. Verifying that our method matches the current leading estimator on Gaussian data is important, but the primary advantage of mutual information as a measure of dependence stems from its ability to detect dependence in arbitrary irregular distributions beyond Gaussians. Therefore, we compare SVORMI against the $k$-NN estimator over datasets sampled from the 5 different functional forms detailed in Table 1, where we add increasing amounts of noise in the $y$ coordinates (measured by the coefficient of determination, $R^2$) to determine how these methods fare in different signal-to-noise regimes. For each functional form, at a given sample size and noise-level, we generate 50 datasets to obtain stable measurements of the estimators’ performances (see Figure 7 for an example of one dataset used in our evaluations). Figures 5, 9, 10 demonstrate that for certain distributions / sample-sizes / noise-levels, SVORMI may be a better MI estimator than the $k$-NN approach (under the squared-loss criterion which is the standard decision-theoretic measure by which continuous-valued estimators are judged).

**Future Work**

The most immediate extension of this work is an investigation of how our methods cope with the curse of dimensionality. The volumes of Voronoi cells will inevitably produce worse estimates of the underlying distribution in high-dimensional settings, and it is important to establish to what extent the statistical power of VORTI and the convergence-rate of SVORMI to the true mutual information are weakened for large $d$. Furthermore, computation of Voronoi diagrams and volumes of cells also becomes expensive in the high dimensional setting, possibly requiring efficient approximation algorithms to maintain the feasibility of applying our methods to large datasets. It may be preferable in higher dimensions to adapt our methods for the dual of the Voronoi Diagram, the Delaunay triangulation. Many applied domains would also benefit from extensions of these methods to metrics beyond the Euclidean distance. In general, we believe there are many more promising results to be found by bringing tools from computational geometry to problems in statistics and data analysis.
Figure 5: Mean-squared errors of the MI estimates obtained via nearest-neighbor methods of Kraskov et al. (2004) and our Voronoi diagram based approaches from data sampled from the functional form described in Table 1 with an amount of noise added to $Y$ which is specified by the coefficients of determi-
nation ($R^2$ values). For ease of comparison we have inversely scaled each MSE by the square of the true underlying mutual information (where if $R^2 = 0$, i.e. $MI = 0$, we scale by $0.01^2$). We have run SVORMI under two choices for the number of local-averaging updates $T$ ($1$ and $\log_2(n)$) and the $k$-NN method with both $k = 3$ (the default in the parmigene $R$ package) and $k = \log_2(n)$.
References


Supplementary Figures and Facts

Figure 6: An i.i.d. sample of 1000 points where $X \sim N(0, 2)$ and $Y_i \mid x_i \sim \sin(x_i) + \text{Uniform}[-\frac{1}{2}, \frac{1}{2}]$

Figure 7: Two different samples drawn from the Linear + Periodic functional form described in Table 1 with no noise introduced on the dataset depicted on the right and a small amount of noise added in the dataset with coefficient of determination $= 0.9$. 
Figure 8: Mean-squared errors of the MI estimates obtained via nearest-neighbor methods of Kraskov et al. (2004) and our Voronoi diagram based approaches from data sampled from the functional form described in Table 1 with an amount of noise added to $Y$ which is specified by the coefficients of determination ($R^2$ values). For ease of comparison we have inversely scaled each MSE by the square of the true underlying mutual information (where for $R^2 = 0$, i.e. $MI = 0$, we scale by $0.01^2$). We have run SVORMI under two choices for the number of local-averaging updates $T$ ($1$ and $\log_2(n)$) and the $k$-NN method with $k = 3$ (the default in the parmigene $R$ package) and $k = \log_2(n)$. 

---

**Table 1**

<table>
<thead>
<tr>
<th>Function Form</th>
<th>Coefficients of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sin(n)$</td>
<td>$0.5$</td>
</tr>
<tr>
<td>$\cos(n)$</td>
<td>$0.3$</td>
</tr>
<tr>
<td>$n^2$</td>
<td>$0.7$</td>
</tr>
<tr>
<td>$n^3$</td>
<td>$0.9$</td>
</tr>
</tbody>
</table>

---

**Notes**

- The $R^2$ values are based on the coefficient of determination.
- We have run SVORMI under two choices for the number of local-averaging updates $T$ ($1$ and $\log_2(n)$) and the $k$-NN method with $k = 3$ (the default in the parmigene $R$ package) and $k = \log_2(n)$.
Figure 9: Mean-squared errors of the MI estimates obtained via nearest-neighbor methods of Kraskov et al. (2004) and our Voronoi diagram based approaches from data sampled from the functional form described in Table 1 with an amount of noise added to $Y$ which is specified by the coefficients of determination ($R^2$ values). For ease of comparison we have inversely scaled each MSE by the square of the true underlying mutual information (where for $R^2 = 0$, i.e. $MI = 0$, we scale by $0.01^2$). We have run SVORMI under two choices for the number of local-averaging updates $T$ (1 and $\log_2(n)$) and the $k$-NN method with $k = 3$ (the default in the parmigene $R$ package) and $k = \log_2(n)$. 
Figure 10: Mean-squared errors of the MI estimates obtained via nearest-neighbor methods of Kraskov et al. (2004) and our Voronoi diagram based approaches from data sampled from the functional form described in Table 1 with an amount of noise added to $Y$ which is specified by the coefficients of determination ($R^2$ values). We have run SVORMI under two choices for the number of local-averaging updates $T$ (1 and $\log_2(n)$) and the $k$-NN method with $k = 3$ (the default in the parmigene $R$ package) and $k = \log_2(n)$.
Figure 11: Mean-squared errors of the MI estimates obtained via nearest-neighbor methods of Kraskov et al. (2004) and our Voronoi diagram based approaches from data sampled from the functional form described in Table 1 with an amount of noise added to $Y$ which is specified by the coefficients of determination ($R^2$ values). We have run SVORMI under two choices for the number of local-averaging updates $T$ ($1$ and $\log_2(n)$) and the $k$-NN method with $k = 3$ (the default in the parmigene $R$ package) and $k = \log_2(n)$.
Figure 12: Mean-squared errors of the MI estimates obtained via nearest-neighbor methods of Kraskov et al. (2004) and our Voronoi diagram based approaches from data sampled from the functional form described in Table 1 with an amount of noise added to $Y$ which is specified by the coefficients of determination ($R^2$ values). We have run SVORMI under two choices for the number of local-averaging updates $T$ (1 and $\log_2(n)$) and the $k$-NN method with $k = 3$ (the default in the parmigene $R$ package) and $k = \log_2(n)$.
**Fact 1** If $Y = g(X)$ for monotone function $g$, then $I(X,Y) = +\infty$ for “nice” joint distributions (which are smooth and supported on a set of positive measure).

**Proof**

\[
I(X,Y) = \mathbb{E}_{x \sim F_X} \left[ D_{KL}(F_Y \mid \mid F_X) \right] \quad (D_{KL} \text{ denotes the Kullback-Leibler divergence})
\]

\[
= \mathbb{E}_{x \sim F_X} \left[ \int_{-\infty}^{\infty} \log \left( \frac{\delta_x(g^{-1}(y))}{f_X(g^{-1}(y))} \right) f_X(g^{-1}(y)) \left| \frac{d}{dy}(g^{-1}(y)) \right| \, dy \right] \quad (\delta_x(\cdot) \text{ is the Dirac delta function})
\]

and for each value of $x$:

\[
\int_{-\infty}^{\infty} \log \left( \frac{\delta_x(g^{-1}(y))}{f_X(g^{-1}(y))} \right) f_X(g^{-1}(y)) \left| \frac{d}{dy}(g^{-1}(y)) \right| \, dy = +\infty
\]

since $\delta_x(g^{-1}(y)) \equiv 0$ on a set of positive measure where $f_X(g^{-1}(y)) > 0$

Note that the continuous formulation differs significantly from the discrete-setting in which the mutual information of $X$ with itself is simply its entropy.