INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(4), 475-491, August 2017.

Asymptotically Unbiased Estimation of A Nonsymmetric Dependence Measure Applied to Sensor Data Analytics and Financial Time Series

A. Caţaron, R. Andonie, Y. Chueh

Angel Caţaron
Department of Electronics and Computers
Transilvania University of Braşov, Romania
cataron@unitbv.ro

Razvan Andonie
1. Department of Computer Science
Central Washington University, USA
2. Department of Electronics and Computers
Transilvania University of Braşov, Romania
*Corresponding author: andonie@cwu.edu

Yvonne Chueh
Department of Mathematics
Central Washington University, USA
chueh@cwu.edu

Abstract: A fundamental concept frequently applied to statistical machine learning is the detection of dependencies between unknown random variables found from data samples. In previous work, we have introduced a nonparametric unilateral dependence measure based on Onicescu's information energy and a kNN method for estimating this measure from an available sample set of discrete or continuous variables. This paper provides the formal proofs which show that the estimator is asymptotically unbiased and has asymptotic zero variance when the sample size increases. It implies that the estimator has good statistical qualities. We investigate the performance of the estimator for data analysis applications in sensor data analysis and financial time series.
Keywords: machine learning, sensor data analytics, financial time series, statistical inference, information energy, nonsymmetric dependence measure, big data analytics.

1 Introduction

Statistical machine learning is based on the strong assumption that we use a *representative* training set of samples to infer a model. In this case, we select a random sample of the population, perform a statistical analysis on this sample, and use these results as an estimation to the desired statistical characteristics of the population as a whole. The accuracy of the estimation depends on the representativeness of the data sample. We gauge the representativeness of a sample by how well its statistical characteristics reflect the probabilistic characteristics of the entire population. Many standard techniques may be used to select a representative sample set [16]. However, if we do not use expert knowledge, selecting the most representative training set from a given dataset was proved to be computationally difficult (NP-hard) [10].

The problem is actually more complex, since in most applications the complete dataset is either unknown or too large to be analyzed. Therefore, we have to rely on a more or less representative training set. For example, a common statistical machine learning problem is to estimate information theory measures (such as entropy) from available training sets. This can be

reduced to the construction of an estimate of a density function from the observed data [21]. We refer here only to nonparametric estimation, where less rigid assumptions will be made about the distribution of the observed data. Although it will be assumed that the distribution has the probability density f, the data will be allowed to speak for themselves in determining the estimate of f more than would be the case if f were constrained to fall in a given parametric family.

The estimation of information theory measures has an important application area - the detection of dependency relationships between unknown random variables represented by data samples. There are two information theory strategies one can adopt when studying the relationship between two random variables: the first is to measure their interdependence thought as a mutual attribute and the second is to measure how much one system depends on the other. In the first case we have symmetric (bilateral) measures of dependence, whereas in the second one we have nonsymmetric (unilateral) measures. The literature review summarizes these two strategies as follows.

Strategy I. Several symmetric dependence measure were proposed (see [20]). Among them, the Shannon entropy based mutual information (MI), $MI(X,Y) = H(X) + H(Y) - H(X,Y) = MI(Y,X)$, which measures the information interdependence between two random variables X and Y. Estimating MI is known to be a non-trivial task [3]. Naïve estimations (which attempt to construct a histogram where every point is the center of a sampling interval) are plagued with both systematic (bias) and statistical errors. Other more sophisticated estimation methods were proposed [19], [23], [14]: kernel density estimators, B-spline estimators, k-th nearest neighbor (kNN) estimators, and wavelet density estimators. An ideal estimator does not exist, instead the choice of the estimator depends on the structure of data to be analyzed. It is not possible to design an estimator that minimizes both the bias and the variance to arbitrarily small values. The existing studies have shown that there is always a delicate trade off between the two types of errors [3].

Strategy II. Several continuous entropy-like nonsymmetric dependence measures have been proposed [15]. But information measures can also refer to certainty (not only to uncertainty, like Shannon's entropy), and probability can be considered as a measure of certainty. For instance, Onicescu's information energy (IE) was interpreted by several authors as a measure of expected commonness, a measure of average certainty, or as a measure of concentration. The second strategy can be illustrated by a unilateral dependence measure defined for discrete random variables by Andonie *et al.* [2]: $o(X,Y) = IE(X|Y) - IE(X)$, where $IE(X) = \sum_{k=1}^{n} p_k^2$ is Onicescu's information energy of a discrete random variable X with probabilities p_k, and $IE(X|Y)$ is the conditional information energy between two variables, as defined in [18].

For a continuous random variable X with probability density function $f(x)$, the IE is [11,18]:

$$IE(X) = E(f(X)) = \int_{-\infty}^{+\infty} f(x)f(x)dx = \int_{-\infty}^{+\infty} f^2(x)dx \tag{1}$$

where X is a random variable with probability density function $f(x)$. In other words, $IE(X)$ is the expectation of the values of a density function f. For two continuous random variables X and Y, with their joint probability density function $f(x,y)$, we introduced the continuous version of the $o(X,Y)$ measure in [1,8]:

$$o(X,Y) = IE(X|Y) - IE(X) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)f(x|y)dy\,dx - \int_{-\infty}^{+\infty} f^2(x)dx \tag{2}$$

The discrete and the continuous $o(X,Y)$ measure the "additional" average certainty (or information) of X occurring under the condition that Y has already or simultaneously occurred

(or is certain) "over" the average certainty of X when the certainty (or information) of Y is not available. Thus, $o(X,Y)$ can be regarded as an indicator of the unilateral dependence of X upon Y. As a unilateral measure, $o(X,Y)$ gives a different insight than the MI into the dependencies between pairs of random variables.

The $o(X,Y)$ measure can easily be computed if the data sample is extracted from known parametric probability distributions. When the underlying distribution of data sample is unknown, the $o(X,Y)$ has to be estimated. More formally, we have to estimate $o(X,Y)$ from a random sample x_1, x_2, ..., x_n of n d-dimensional observations from a distribution with the unknown probability density $f(x)$. This problem is even more difficult if the number of available points is small.

Contributions. In previous work [4,5], we introduced a kNN method for estimating IE and $o(X,Y)$ from data samples of discrete or continuous variables. In the present paper, we give a more detailed description of this method. For the first time, we provide the formal proofs showing that the estimator is asymptotically unbiased and has asymptotic zero variance when the sample size increases. This implies that the estimator has good statistical qualities. The potential application of our results in statistical machine learning for multivariate data drives the motivation for the present work and we refer to two real world applications. The first one is a study of the unilateral interactions between temperature sensor data in a refrigerator room. The second one is an analysis of the dynamics of interaction between assets and liabilities evidenced by the historical event that matches the time series formed by the unilateral dependence measures observed from the financial time series reported by Kodak and Apple. This evolution is captured by our unilateral dependence measure over time.

The rest of the paper is structured as follows. In Section 2 we summarize our previous work, we review the kNN estimation method for the continuous $o(X,Y)$ measure, and we derive our IE estimator in a d-dimensional space. In Section 3 we analyze and prove the asymptotic unbiased behavior of this IE estimator. In Section 4 we derive the $o(X,Y)$ estimator for the purpose of measuring the unilateral dependence relation between a pair of random variables. Section 5 discusses applications and Section 6 contains our final remarks.

2 Background

Fitting multi-dimensional data to joint density functions is challenging. We aim to summarize the kNN estimation technique which can be used for this problem. We will refer to our previous results on IE and $o(X,Y)$ kNN estimation.

2.1 kNN estimation of the unilateral dependence measure in an unidimensional space

Our goal is to approximate IE and $o(X,Y)$ from the available dataset, for random variables X and Y with unknown densities, using the kNN method. The kNN estimators represent an attempt to adapt the amount of smoothing to the "local" density of data. The degree of smoothing is controlled by an integer k, chosen to be considerably smaller than the sample size. We define the distance $R_{i,k,n}$ between two points on the line as $|x_i - x_k|$ out of n available points, and for each x_i we define the distances $R_{i,1,n} \leq R_{i,2,n} \leq \ldots \leq R_{i,n,n}$, arranged in ascending order, from x_i to the points of the sample.

The kNN density estimate $f(x_i)$ is defined by [21]:

$$\hat{f}(x_i) = \frac{k}{2nR_{i,k,n}}$$

The kNN was used for non-parametrical estimate of the entropy based on the k-th nearest neighbor distance between n points in a sample, where k is a fixed parameter and $k \leq n-1$. Based on the first nearest neighbor distances, Leonenko *et al.* [13] introduced an asymptotic unbiased and consistent estimator H_n of the entropy $H(f)$ in a multidimensional space. When the sample points are very close one to each other, small fluctuations in their distances produce high fluctuations of H_n. In order to overcome this problem, Singh *et al.* [22] defined an entropy estimator based on the k-th nearest neighbor distances. A kNN estimator of the Kullback-Leibler divergence was introduced by Wang *et al.* [24]. Faivishevsky *et al.* used a mean of several kNN estimators, corresponding to different values of k, to increase the smoothness of the estimation [9].

The kNN method works well if the value of k is optimally chosen, and it outperforms histogram methods [23]. However, there is no model selection method for determining the number of nearest neighbors k, and this is a known limitation.

2.2 kNN estimation of the IE and the $o(X,Y)$ measure

In [5], we introduced a kNN IE estimator, and we stated in [4], without mathematical proofs, that this estimator is consistent. By definition [12], a statistic whose mathematical expectation is equal to a parameter is called *unbiased* estimator of that parameter. Otherwise, the statistic is said to be *biased*. A statistical estimator that converges asymptotically to a parameter is called *consistent* estimator of that parameter. In the following, we will review this result, introducing also the basic mathematical notations used in Section 3.

The IE is the average of $f(x)$, therefore we have to estimate $f(x)$. The n observations from our samples have the same probability $\frac{1}{n}$. A convenient estimator of the IE is:

$$\hat{IE}_k^{(n)}(X) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_i). \tag{3}$$

The mass probability at value x_i determined by its k nearest surrounding points and standardized by the sphere volume $V_k(x_i)$ these k nearest observations occupy measures the density of probability at value x_i. The k-th nearest neighbor estimate is defined by [4], [21]:

$$\hat{f}(x_i) = \frac{\frac{k}{n}}{V_k(x_i)}, i = 1...n \tag{4}$$

where $V_k(x_i)$ is the volume of the d-dimensional sphere centered in x_i, with the radius $R_{i,k,n}$ measuring the Euclidean distance from x_i to the k-th nearest data point. The estimate (4) can be re-written as:

$$\hat{f}(x_i) = \frac{k}{nV_d R_{i,k,n}^d}, i = 1...n. \tag{5}$$

given $V_k(x_i) = V_d R_{i,k,n}^d$ where V_d is the radius of the unit sphere in d-dimensions, that is $V_1 = 2$, $V_2 = \pi$, $V_3 = \frac{4\pi}{3}$, etc.

By substituting $\hat{f}(x_i)$ in (3), we finally obtain the following IE approximation:

$$\hat{IE}_k^{(n)}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{k}{nV_d R_{i,k,n}^d}. \tag{6}$$

In [6], we showed how to estimate the $o(X,Y)$ dependence measure from an available sample set of discrete or continuous variables and we experimentally compared the results of the kNN estimator with the naïve histogram estimator.

3 Asymptotic behavior of the IE approximator

Consistency of an estimator means that as the sample size gets large the estimate gets closer and closer to the true value of the parameter. Unbiasedness is a statement about the expected value of the sampling distribution of the estimator. The ideal situation, of course, is to have an unbiased consistent estimator. This may be very difficult to achieve. Yet unbiasedness is not essential, and just a little bias is permitted as long as the estimator is consistent. Therefore, an asymptotically unbiased consistent estimator may be acceptable.

In [4], we stated (without proofs) that the IE approximator is asymptotically unbiased and consistent. In the this section, we provide the formal proofs for this statement.

Lemma 1. *Let us consider the identically distributed random variables* $T_1^{(n)}$, $T_2^{(n)}$, ..., $T_n^{(n)}$ *defined by*

$$T_i^{(n)} = \frac{k}{nV_d R_{i,k,n}^d}. \tag{7}$$

For a given x*, the random variable* T_x *has the following probability density function:*

$$h_{T_x}(y) = \frac{f(x)\left[\frac{k}{y}f(x)\right]^k}{y^2(k-1)!}e^{-\frac{k}{y}f(x)}, 0 < y < \infty.$$

Proof: See Appendix 6. □

Theorem 2 (*Asymptotically Unbiasedness*)**.** *The information energy estimator* $\hat{IE}_k^{(n)}(f)$ *is asymptotically unbiased.*

Proof:

We will find the asymptotic value of the average:

$$\lim_{n\to\infty} E\left[\hat{IE}_k^{(n)}(f)\right].$$

From Lemma 1, for a given x, the random variable T_x has the pdf:

$$h_{T_x}(y) = \frac{f(x)\left[\frac{k}{y}f(x)\right]^k}{y^2(k-1)!}e^{-\frac{k}{y}f(x)}, 0 < y < \infty.$$

The conditional expectation of $T_1^{(n)}|X_1 = x$, given the fixed center $X_1 = x$ is:

$$\begin{aligned} lim_{n\to\infty}E\left[T_1^{(n)}|X_1 = x\right] &= \int_0^\infty yh(y)dy \\ &= \int_o^\infty \frac{f(x)\left(kf(x)\right)^k}{(k-1)!}\cdot y^{-1-k}e^{-\frac{kf(x)}{y}}dy \\ &= \frac{f(x)\left(kf(x)\right)^k}{(k-1)!}\int_0^\infty y^{-1-k}e^{-\frac{kf(x)}{y}}dy. \end{aligned}$$

We make the following change of variable: $u = \frac{kf(x)}{y}$, with $y \in (0, +\infty)$. The corresponding $u \in (+\infty, 0)$ is decreasing. We have:

$$du = -\frac{1}{y^2}kf(x)dy$$

and

$$y^{-1-k} = \left(\frac{u}{kf(x)}\right)^{k+1}.$$

We obtain:

$$\int_0^\infty yh(y)dy = \frac{f(x)\,(kf(x))^k}{(k-1)!}\int_\infty^0 -u^{k+1}e^{-u}du \cdot \frac{1}{(kf(x))^{k+1}}\left(\frac{kf(x)}{u}\right)^2 \cdot \frac{1}{kf(x)}.$$

The negative sign before u^{k+1} allows us to change the limits of the integral:

$$\int_0^\infty yh(y)dy = \frac{f(x)\,(kf(x))^k}{(k-1)!}\int_0^\infty u^{k-1}e^{-u}du \cdot \frac{1}{(kf(x))^k}$$
$$= \frac{f(x)}{(k-1)!}\int_0^\infty u^{k-1}e^{-u}du = \frac{f(x)}{(k-1)!}\Gamma(k) = f(x).$$

We take the complete expectation of the above, then we have:

$$lim_{n\to\infty}E\left[E\left[T_1^{(n)}|X_1 = x\right]\right] = lim_{n\to\infty}E\left[T_1^{(n)}\right] = E[f(x)].$$

Let us define $\hat{G}^{(n)}(f) = \frac{1}{n}\sum_{i=1}^n T_i^{(n)}$, the unbiased estimator of $f(x)$ (like in [22]). We obtain:

$$E\left[\hat{G}^{(n)}(f)\right] = \frac{1}{n}\sum_{i=1}^n E\left[T_i^{(n)}\right] = \frac{1}{n}\left[E\left[f(x)\right] + \cdots + E\left[f(x)\right]\right] = E\left[f(x)\right].$$

$$lim_{n\to\infty}E\left[\hat{G}^{(n)}(f)\right] = E\left[f(x)\right] = \int f(x)\cdot f(x)dx = IE(f).$$

This proves that our IE estimator is asymptotically unbiased. □

Next, we aim to prove that it has asymptotic zero variance.

Lemma 3. *Given the identically distributed random variables $T_1^{(n)}$, $T_2^{(n)}$, ..., $T_n^{(n)}$ defined by (7), we have:*

$$\lim_{n\to\infty}\frac{1}{n}E\left[T_1^{(n)^2}\right] - \lim_{n\to\infty}\frac{1}{n}E\left[T_1^{(n)}\right]^2 = 0.$$

Proof: See Appendix 6. □

Lemma 4. *Given the identically distributed random variables $T_1^{(n)}$, $T_2^{(n)}$, ..., $T_n^{(n)}$ defined by (7), the following relation is true:*

$$\lim_{n\to\infty}\frac{n(n-1)}{n^2}\left(E\left[T_1^{(n)}T_2^{(n)}\right] - E\left[T_1^{(n)}\right]E\left[T_2^{(n)}\right]\right) = 0.$$

Proof: See Appendix 6. □

Theorem 5 (*Asymptotic Zero Variance*)**.**

$$\lim_{n\to\infty} Var\left[\hat{G}_k^{(n)}(f)\right] = 0.$$

Proof: $T_1^{(n)}, T_2^{(n)}, \ldots, T_n^{(n)}$ are identically distributed.

$$Var\left[\hat{G}_k^{(n)}(f)\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} Var\left[T_i^{(n)}\right] + \sum_{\substack{i\neq j \\ i<j}} 2Cov\left(T_i^{(n)}, T_j^{(n)}\right)\right]$$
$$= \frac{1}{n} Var\left[T_1^{(n)} + \frac{n(n-1)}{n^2}\right] Cov\left(T_1^{(n)}, T_2^{(n)}\right).$$

$$\lim_{n\to\infty} Var\left[\hat{G}_k^{(n)}(f)\right] = \lim_{n\to\infty} \frac{1}{n} E\left[T_1^{(n)^2}\right] - \lim_{n\to\infty} \frac{1}{n} E\left[T_1^{(n)}\right]^2$$
$$+ \lim_{n\to\infty} \frac{n(n-1)}{n^2}\left(E\left[T_1^{(n)} T_2^{(n)}\right] - E\left[T_1^{(n)}\right] E\left[T_2^{(n)}\right]\right).$$

From Lemma 3 and Lemma 4, we obtain:

$$\lim_{n\to\infty} Var\left[\hat{G}_k^{(n)}(f)\right] = 0.$$

□

Our main result is synthesized in:

Theorem 6 (*Consistency*)**.** *The information energy estimator* $\hat{IE}_k^{(n)}(f)$ *is consistent.*

Proof: This results from Theorems 2 and 5, and the following property: An asymptotically unbiased estimator with asymptotic zero variance is consistent [17]. □

4 The kNN $o(X,Y)$ estimator

Our goal is to infer $o(X,Y)$ from the random samples x_1, x_2, ..., x_n. We will use the results from Section 2.2 to deduct the kNN estimator of $o(X,Y)$.

First, we substitute $\widehat{IE}_k^{(n)}(X)$ from eq. (6) in eq. (2):

$$\hat{o}(X,Y) = \widehat{IE}_k^{(n)}(X|Y) - \widehat{IE}_k^{(n)}(X) \tag{8}$$

where:

$$\widehat{IE}_k^{(n)}(X|Y) = \sum_{j=1}^{m} \hat{f}(y_j)\widehat{IE}_k^{(n)}(X|y_j) \tag{9}$$

and

$$\widehat{IE}_k^{(n)}(X) = \frac{1}{n}\sum_{i=1}^{n} \frac{k_1}{nV_{d_1(X)}R_i^{d_1}}, \tag{10}$$

is an adaptation of eq. (6).

We can write:

$$\widehat{IE}_k^{(n)}(X|y_j) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_i|y_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}(x_i,y_j)}{\hat{f}(y_j)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}(x_i,y_j)}{\hat{f}(y_j)}, \tag{11}$$

where

$$\hat{f}(x_i|y_j) = \frac{\hat{f}(x_i,y_j)}{\hat{f}(y_j)}. \tag{12}$$

From (9) and (11) we can write:

$$\widehat{IE}_k^{(n)}(X|Y) = \sum_{j=1}^{m}\hat{f}(y_j)\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}(x_i,y_j)}{\hat{f}(y_j)} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}\hat{f}(x_i,y_j).$$

The estimate of the joint probability density can be written as:

$$\hat{f}(x_i,y_j) = \frac{k_2}{pV_{d_2(X,Y)}R_{i,j}^{d_2}},$$

where p is the number of pairs (x_i, y_j), then re-write the eq. (9) as:

$$\widehat{IE}_k^{(n)}(X|Y) = \frac{k_2}{npV_{d_2(X,Y)}}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{1}{R_{i,j}^{d_2}}.$$

R_i is the Euclidean distance between the reference point x_i and its k_1^{th} nearest neighbor, when the points are drawn from the one-dimensional probability distribution $f(x)$: $R_i = \|x_i - x_{i,k_1}\|$. Similarly, R_j is the Euclidean distance between the reference point y_j and its k_1^{th} nearest neighbor, when the points are drawn from the one-dimensional probability distribution $f(Y)$: $R_j = \|y_j - y_{j,k_1}\|$. Then, R_{ij} is the Euclidean distance between the reference point (x_i, y_j) and its k_2^{th} nearest neighbor, when the points are drawn from the joint probability distribution $f(X,Y)$: $R_{ij} = \sqrt{(x_{ij} - x_{ij,k_2})^2 + (y_{ij} - y_{ij,k_2})^2}$.

The estimate of $o(X,Y)$ is:

$$\hat{o}(X,Y) = \frac{k_2}{npV_{d_2(X,Y)}}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{1}{R_{i,j}^{d_2}} - \frac{k_1}{n^2V_{d_1(X)}}\sum_{i=1}^{n}\frac{1}{R_i^{d_1}}. \tag{13}$$

Although we do not have a general method to set the nearest neighbor parameter, Silverman [21] suggested that k should be proportional to

$$n^{4/(d+4)}. \tag{14}$$

In our case, the optimal values of k_1 and k_2 may not be equal, because the these two parameters refer to different samples. The R code for computing the estimate of $o(X,Y)$ is publicly available on GitHub[1].

5 Applications

There are tremendous opportunities for applications using $o(X,Y)$ estimation to analyze potential dependencies between data represented by samples. We illustrate with an example with numerical implementations and two applications using real-world data, all extracted from our previous work and summarized here.

[1] https://github.com/cataron/information-energy

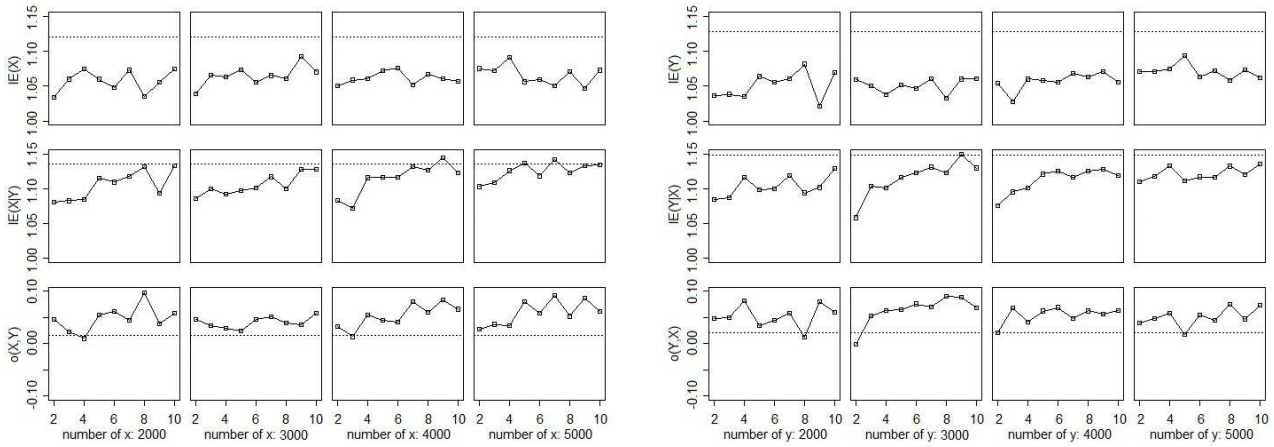

Figure 1: The kNN estimated values, where k was determined with formula (14). The theoretical values $IE(X) = 1.12$, $IE(X|Y) = 1.1351$, $o(X,Y) = 0.0151$, $IE(Y) = 1.128$, $IE(Y|X) = 1.14787$, $o(Y,X) = 0.01987$ are marked with dashed lines. The estimates converge towards their theoretical values for an increasing number of samples, under various k values (the horizontal axis): 2, 4, 6, 8, 10.

5.1 A specified joint probability distribution

Let us illustrate with the following joint probability density function from [6]:

$$f_{X,Y}(x,y) = \frac{6}{5}\left(x+y^2\right), x \in [0,1], y \in [0,1], \tag{15}$$

which has the marginal probability density functions

$$f_X(x) = \int_0^1 \frac{6}{5}\left(x+y^2\right) dy = \frac{6}{5}\left(x+\frac{1}{3}\right) \tag{16}$$

and

$$f_Y(y) = \int_0^1 \frac{6}{5}\left(x+y^2\right) dx = \frac{6}{5}\left(\frac{1}{2}+y^2\right). \tag{17}$$

The conditional probability density function is:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x+y^2}{\frac{1}{2}+y^2},$$

and the theoretical value of $o(X,Y)$ is:

$$o(X,Y) = IE(X|Y) - IE(X)$$

$$IE(X) = \int_0^1 f^2(x)dx = 1.12$$

$$IE(X|Y) = \int_0^1 \int_0^1 f_{X,Y}(x,y) f_{X|Y}(x|y) dxdy = 1.1351$$

and

$$o(X,Y) = 1.1351 - 1.12 = 0.0151.$$

Similarly, we obtain $IE(Y) = 1.128$, $IE(Y|X) = 1.14787$, and $o(Y,X) = 0.01987$. We observe that $o(X,Y) < o(Y,X)$. The theoretical values of $IE(X)$, $IE(Y)$, $IE(X|Y)$, $IE(Y|X)$, $o(X,Y)$, and (Y,X) are represented in Fig. 1 with dashed lines. Their kNN estimates are represented with continuous lines, for an increasing number of data samples.

5.2 Sensor data

When studying the interaction between two random variable, why is a unilateral dependence measure useful, and why do we not simply use the well-known MI? Let us consider two sets of experiments, characterized respectively by random variables X and Y. The experiments run simultaneously and interact probabilistically. Our question is which variable influences probabilistically more the other one. Thus, X can be viewed as $X|Y$ and Y can be viewed as $Y|X$. While the correlation quantifies linear dependence and MI describes the degree of interdependence between two random variables, $o(X,Y)$ helps us understand which random variable, X or Y, has a higher influence on the other one. If both X and $X|Y$ are available, we can estimate $IE(X|Y)$ as well as $o(X,Y)$, and similarly for Y and $Y|X$.

When the data is acquired from real world experiments or from simulators, we need to store series of values for X and Y featuring the two phenomena running independently, as well as measurements of values generated by the two phenomena running simultaneously, in order to capture $X|Y$ and $Y|X$. Moreover, the precision of the $IE(X|Y)$ and $o(X,Y)$ estimators increase when more values of $X|Y$ are available for each value of Y.

In our application, which can be found in [6], we study the evolution of the temperature measured at the surface of two packages, X and Y, introduced in a refrigerated room. When one package at the outside temperature is placed in the room, it is getting cold under the influence of the air conditioning devices. The situation changes when a second package is placed in the room because it will change the ambient conditions and the temperature may decrease slower. Emulating the real world circumstances, we are able to measure the temperatures with various experimental setups. When we only have one package in the room, we measure the time series of temperatures as samples of an unconditional random variables. When we have both packages in the room, we measure simultaneously the time series of temperatures as conditional random values.

According to these experiments, we can assess if package X has a stronger influence on package Y, or vice versa. This might lead to the decision to remove the sensor which is more influenced by the other.

5.3 Financial time series

Studying the relationship between stocks is a very interesting business case for finance. This is just an example which can be extended to all kind of financial parameters of companies. In [7] we have presented an application using the public available data offered by Kodak and Apple for the period 1999–2014. We analyzed the unilateral dependence between the Kodak and Apple series in order to understand how they influence one to each other. We aim to describe in the following our general approach for detecting unilateral dependencies between time series variables, without specifically referring to the Kodak-Apple case.

The information energy and the unilateral dependence measure both are the result of two random variables. The estimation becomes more precise when the number of observations is high. When we analyze time series, we consider the series as a set of n observations $x_i, i = 1...n$ of the random variable X. Our goal is to find synchronous relations between two time series which have been considered as two samples of two random variables X and Y. A point x_i from the first series was paired with a point y_i from the second one producing a joint observation, with

$i = 1, \ldots, n$. Therefore, from two time series with n points each we obtain n pairs (x_i, y_i). The unilateral dependence $o(X, Y)$ between X and Y can be estimated with formula (13) based on X and (X, Y), while the unilateral dependence $o(Y, X)$ can be estimated with the same formula based on Y and (Y, X). We note that the set of pairs (X, Y) is equal with the set of pairs (Y, X) since for each observation x_i exactly one observation y_i is used.

It is interesting to study the evolution of the unilateral dependence in time. The values $o(X, Y)$ and $o(Y, X)$ can be estimated at the moment of time t_m using the history of the m observations of x_i and y_i between the initial moments t_0 and t_m. A new set of two observations x_{m+1} and y_{m+1} allow us to re-estimate $o(X, Y)$ and $o(Y, X)$.

Our model captures not only the unilateral dependencies between two simultaneous time series, but also how these dependencies evolve in time, which can be valuable and insightful for forecast or investment in financial research or applications.

6 Conclusion

We introduced a non-parametric asymptotically unbiased and consistent estimator of the IE, and the unilateral dependency measure $o(X, Y)$ between random variables X and Y. We estimated $o(X, Y)$ from available data samples using the kNN technique. In our applications, we showed how the nonsymmetric dependence measure can provide information that cannot be expressed by a symmetric measure. The examples illustrated in this paper are all one-dimensional random variables, for the purpose of simplicity. The kNN estimation can be also applied on multi-dimensional variables. For instance, the $o(X, Y)$ measure can be used to analyze the unilateral dependence between bi-variate time-series data acquired from various fields.

Because of the mathematical properties proved here, the $o(X, Y)$ estimator works the best for large data sets, so it is suitable for big data analytics. Our method can be applied to both continuous and discrete variable spaces, meaning that we can use it both in classification and regression problems.

Bibliography

[1] Andonie R., Caţaron A. (2004), An informational energy LVQ approach for feature ranking, *European Symposium on Artificial Neural Networks 2004, pages In d-side publications*, 471–476, 2004.

[2] Andonie R. (1986), Interacting systems and informational energy, *Foundation of Control Engineering*, 11, 53–59, 1986.

[3] Bonachela J.A., Hinrichsen H., Miguel A. Munoz M.A. (2008), Entropy estimates of small data sets, *MATH.THEOR.*, 41(20), 1-20, 2008.

[4] Caţaron A., Andonie R., Chueh Y. (2013), Asymptotically unbiased estimator of the informational energy with kNN, *International Journal of Computers Communications & Control*, 8(5), 689–698, 2013.

[5] Caţaron A., Andonie R. (2012), How to infer the informational energy from small datasets, *Optimization of Electrical and Electronic Equipment (OPTIM), 2012 13th International Conference on*, 1065 –1070, 2012.

[6] Caţaron A., Andonie R., Chueh Y. (2014), kNN estimation of the unilateral dependency measure between random variables, *2014 IEEE Symposium on Computational Intelligence and Data Mining, (CIDM 2014)*, Orlando, FL, USA, 471–478, 2014.

[7] Caţaron A., Andonie R., Chueh Y. (2015), Financial data analysis using the informational energy unilateral dependency measure, *Proceedings of the International Joint Conference on Neural Networks, (IJCNN 2015)*, Killarney, Ireland, 1-8, 2015.

[8] Chueh Y., Caţaron A., Andonie R. (2016), Mortality rate modeling of joint lives and survivor insurance contracts tested by a novel unilateral dependence measure, *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, Athens, Greece, December 6-9, 2016, 1–8, 2016.

[9] Faivishevsky L., Goldberger J. (2008), ICA based on a smooth estimation of the differential entropy, *NIPS*, 1-8, 2008.

[10] Gamez J.E., Modave F., Kosheleva O. (2008), Selecting the most representative sample is NP-hard: Need for expert (fuzzy) knowledge, *Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence). IEEE International Conference on*, 1069–1074, 2008.

[11] Guiaşu S. (1977), *Information theory with applications*, McGraw Hill, New York, 1977.

[12] Hogg R.V., McKean J., Allen T. Craig A.T. (2006), *Introduction To Mathematical Statistics, 6/E*, Pearson Education, 2006.

[13] Kozachenko L. F., Leonenko N. N. (1987), Sample estimate of the entropy of a random vector, *Probl. Peredachi Inf.*, 23(2), 9–16, 1987.

[14] Kraskov A., Stögbauer H., Grassberger P. (2004), Estimating mutual information, *Phys. Rev. E*, 69, 1–16, 2004.

[15] Li H. (2015), On nonsymmetric nonparametric measures of dependence, arXiv:1502.03850, 2015.

[16] Lohr H. (1999), *Sampling: Design and Analysis*, Duxbury Press, 1999.

[17] Miller M., Miller M. (2003), *John E. Freund's mathematical statistics with applications*, Pearson/Prentice Hall, Upper Saddle River, New Jersey, 7th edition, 2003.

[18] Onicescu O. (1966), Theorie de l'information. Energie informationelle, *C. R. Acad. Sci. Paris, Ser. A–B*, 263, 841–842, 1966.

[19] Paninski L. (2003), Estimation of entropy and mutual information, *Neural Comput.*, 15, 1191–1253, 2003.

[20] Schweizer B., Wolff E. F. (1981), On nonparametric measures of dependence for random variables, *Ann. Statist.*, 9:879–885, 1981.

[21] Silverman B.W. (1986), *Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*, Chapman and Hall/CRC, 1986.

[22] Singh H., Misra N., Hnizdo V., Fedorowicz A., Demchuk E. (2003), Nearest neightboor estimates of entropy, *American Journal of Mathematical and Management Sciences*, 23, 301–321, 2003.

[23] Walters-Williams J., Li Y. (2009), Estimation of mutual information: A survey, *Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology*, Springer-Verlag, Berlin, Heidelberg, 389–396, 2009.

[24] Wang Q., Kulkarni S. R., Verdu S. (2006), A nearest-neighbor approach to estimating divergence between continuous random vectors, *Proc. of the IEEE International Symposium on Information Theory*, Seattle, WA, 242-246, 2006.

Appendix

Proof of Lemma 1

The random variables $T_1^{(n)}$, $T_2^{(n)}$, ..., $T_n^{(n)}$ are identically distributed, therefore:

$$E\left[\hat{IE}_k^{(n)}(f)\right] = E\left[T_1^{(n)}\right].$$

Let us denote $\rho_{r,n} = \left(\frac{k}{nV_d r}\right)^{\frac{1}{d}}$, with $\rho_{r,n} \to 0$ when $n \to \infty$. For a real number r we have:

$$P\left[T_1^{(n)} > r | X_1 = x\right] = P\left[\rho_{r,n} > R_{1,k,n} | X_1 = x\right], \tag{18}$$

because:

$$T_1^{(n)} > r \Leftrightarrow \frac{k}{nV_d R_{1,k,n}^d} > r \Rightarrow \frac{k}{nV_d r} > R_{1,k,n}^d \Rightarrow \left(\frac{k}{nV_d r}\right)^{\frac{1}{d}} > R_{1,k,n}$$

We write the probability $P\left[T_1^{(n)} > r | X_1 = x\right]$ as a binomial distribution, and approximate it with the Poisson probability function.

The binomial distribution is given by $f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$, and the binomial distribution of the cumulative distribution function is $P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} f(x)$. The probability expressed by the Poisson formula is $P(x) \approx \frac{e^{-np}(np)^x}{x!}$, where p is the successful probability out of n samples.

Using the Poisson approximation for the binomial distribution by letting the sample size $n \to \infty$ we get:

$$\begin{aligned} &P[\rho_{r,n} > R_{1,k,n} | X_1 = x] = 1 - P[R_{1,k,n} \geq \rho_{r,n}] \\ &= 1 - \sum_{i=1}^{k} \binom{n-1}{i} \left[P(S_{\rho_{r,n}})(1 - P(S_{\rho_{r,n}}))\right]^{n-i-1} = P[T_x > r]. \end{aligned}$$

We note that:

$$\begin{aligned} P(S_{\rho_{r,n}}) &= \int_{S_{\rho_{r,n}}} f(t)dt, \\ \rho_{r,n} &= \left(\frac{k}{nV_d r}\right)^{\frac{1}{d}}, \\ V_\rho = V_d \rho^d &= V_d \frac{k}{nV_d r} = \frac{k}{nr} \Rightarrow n = \frac{k}{rV_\rho}. \end{aligned}$$

Let $n \to \infty$, then:

$$\lim_{n\to\infty} nP(S_{\rho_{r,n}}) = \frac{k}{r} \lim_{n\to\infty} \frac{P(S_\rho)}{V_\rho} = \frac{k}{r} f(x).$$

Thus, from the Poisson approximation we obtain:

$$P[T_x > r] = 1 - \sum_{i=0}^{k} \frac{\left(\frac{k}{r} f(x)\right)^i}{i!} e^{-\frac{k}{r} f(x)}.$$

To calculate the pdf for random variable T_x, we differentiate its cumulative density function with respect to r:

$$\begin{aligned}
&\frac{d}{dr} P[T_x \le r] = \frac{d}{dr} \left[1 - P[T_x > r]\right] = \frac{d}{dr} \left[1 - \left(1 - \sum_{i=0}^{k} \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} e^{-\frac{k}{r} f(x)}\right)\right] \\
&= \frac{d}{dr} \sum_{i=0}^{k} \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} e^{-\frac{k}{r} f(x)} = \frac{d}{dr} \left[e^{-\frac{k}{r} f(x)} + \sum_{i=1} k \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} e^{-\frac{k}{r} f(x)}\right] \\
&= \frac{kf(x)}{r^2} e^{-\frac{k}{r} f(x)} + \sum_{i=1}^{k} \frac{\left[\frac{k}{r} f(x)\right]^{i-1}}{(i-1)!} \left(-\frac{kf(x)}{r^2}\right) e^{-\frac{k}{r} f(x)} + \sum_{i=1}^{k} \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} \cdot \frac{kf(x)}{r^2} e^{-\frac{k}{r} f(x)} \\
&= \frac{kf(x)}{r^2} e^{-\frac{k}{r} f(x)} - \sum_{i=1}^{k} \frac{kf(x)}{r^2} \cdot \frac{\left[\frac{k}{r} f(x)\right]^{i-1}}{(i-1)!} e^{-\frac{k}{r} f(x)} + \sum_{i=1}^{k} \frac{kf(x)}{r^2} \cdot \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} e^{-\frac{k}{r} f(x)} \\
&= \frac{kf(x)}{r^2} e^{-\frac{k}{r} f(x)} - \frac{kf(x)}{r^2} e^{-\frac{k}{r} f(x)} + \frac{kf(x)}{r^2} \cdot \frac{\left[\frac{k}{r} f(x)\right]^k}{k!} e^{-\frac{k}{r} f(x)} = \frac{e^{-\frac{k}{r} f(x)} f(x) \left[\frac{k}{r} f(x)\right]^k}{r^2 (k-1)!}.
\end{aligned}$$

Then, for a given x, the random variable T_x has the pdf:

$$h_{T_x}(y) = \frac{f(x) \left[\frac{k}{y} f(x)\right]^k}{y^2 (k-1)!} e^{-\frac{k}{y} f(x)}, 0 < y < \infty.$$

Proof of Lemma 3

We have:

$$\begin{aligned}
\lim_{n\to\infty} E\left[T_1^{(n)^2} \,\middle|\, X_1 = x\right] &= \int_0^\infty y^2 h(y) dy \\
&= \int_0^\infty \frac{f(x) \left(kf(x)\right)^k}{(k-1)!} y^{-k} e^{-\frac{kf(x)}{y}} dy \\
&= \frac{f(x) \left(kf(x)\right)^k}{(k-1)!} \int_0^\infty y^{-k} e^{-\frac{kf(x)}{y}} dy.
\end{aligned}$$

We make the following substitution: $u = \frac{kf(x)}{y}$, with $y \in (0, +\infty)$, $u \in (+\infty, 0)$ decreasing, and $y = \frac{kf(x)}{u}$. Then $du = -\frac{1}{y^2} kf(x) dy$ and $y^{-k} = \left(\frac{u}{kf(x)}\right)^k$.

We obtain:

$$\begin{aligned}\lim_{n\to\infty} E\left[\left(T_1^{(n)}\right)^2 |X_1 = x\right] &= \frac{f(x)\left(kf(x)\right)^k}{(k-1)!}\int_{\infty}^{0} -\left(\frac{u}{kf(x)}\right)^k e^{-\frac{kf(x)}{y}} y^2 \frac{1}{kf(x)} du \\ &= \frac{f(x)\left(kf(x)\right)^k}{(k-1)!}\int_{\infty}^{0} -u^{k-2}e^{-u}\left(\frac{1}{kf(x)}\right)^{k-1} du \\ &= \frac{f(x)\left(kf(x)\right)^k}{(k-1)!}\int_{0}^{\infty} -u^{k-2}e^{-u} du \\ &= \frac{kf(x)^2}{(k-1)!}\Gamma(k-1) = \frac{k}{k-1}f(x)^2.\end{aligned}$$

$$\begin{aligned}\lim_{n\to\infty} E\left[\left(T_1^{(n)}\right)^2\right] &= \lim_{n\to\infty} E\left[E\left[\left(T_1^{(n)}\right)^2 |X_1 = x\right]\right] \\ &= \lim_{n\to\infty} E\left[\frac{k}{k-1}f(x)^2\right] \\ &= \frac{k}{k-1}E\left[f(x)^2\right] \\ &= \frac{k}{k-1}\left[Var\left[f(x)\right] + \left(E\left[f(x)\right]\right)^2\right] \\ &= \frac{k}{k-1}\left[Var\left[f(x)\right] + IE(f)^2\right].\end{aligned}$$

$$\begin{aligned}\lim_{n\to\infty} Var\left[T_1^{(n)}\right] &= \lim_{n\to\infty} E\left[\left(T_1^{(n)}\right)^2\right] - \lim_{n\to\infty}\left(E\left[T_1^{(n)}\right]\right)^2 \\ &= \frac{k}{k-1}Var\left[f(x)\right] + \frac{k}{k-1}IE(f)^2 - IE(f)^2 \\ &= \frac{k}{k-1}Var\left[f(x)\right] + \frac{1}{k-1}IE(f)^2.\end{aligned}$$

$$\lim_{n\to\infty}\frac{1}{n}Var\left[T_1^{(n)}\right] = \lim_{n\to\infty}\frac{1}{n}\left[\frac{k}{k-1}Var\left[f(x)\right] + \frac{1}{k-1}IE(f)^2\right] = 0.$$

Proof of Lemma 4

The relation above can be proved if the limiting covariance between $T_1^{(n)}$ and $T_2^{(n)}$ is zero. This can be achieved by showing limiting probability distributions of $T_1^{(n)}$ and $T_2^{(n)}$ are independent.

Given the real number r, $s > 0$:

$$\begin{aligned}
&P\left[T_1^{(n)} > r, T_1^{(n)} > s | X_1 = x, X_2 = y\right] \\
&= P\left[R_{1,k,n} < \rho_{r,n}, R_{2,k,n} < \rho_{s,n} | X_1 = x, X_2 = y\right] \\
&= P\left[\text{at least } k \text{ of } X_3, X_4, \ldots, X_n \in S_{\rho_{r,n;x}} \text{ and} \right. \\
&\left. \text{at least } k \text{ of } X_3, X_4, \ldots, X_n \in S_{\rho_{s,n;y}}\right].
\end{aligned}$$

Note: For $x \neq y$, $\rho_{r,n} \to 0$ and $\rho_{s,n} \to 0$ as $n \to +\infty$. For large enough n, we have $S_{\rho_{r,n;x}} \cap S_{\rho_{s,n;y}} = \Phi$.

We obtain:

$$\begin{aligned}
&P\left[T_1^{(n)} > r, T_1^{(n)} > s | X_1 = x, X_2 = y\right] \\
&= \sum_{\substack{k \le i,j \\ i+j \le n-2}} \frac{(n-2)!}{i!j!(n-2-i-j)!} \left[P(S_{\rho_{r,n;x}})\right]^i \left[P(S_{\rho_{s,n;x}})\right]^j \cdot \\
&\left[1 - P(S_{\rho_{r,n;x}}) - P(S_{\rho_{s,n;y}})\right]^{n-2-i-j}
\end{aligned}$$

where $P(S_{\rho_{r,n;x}}) = \int_{S_{\rho_{r,n;x}}} f(t)dt$.

Recall:

$$\lim_{n\to\infty} nP(S_{\rho_{r,n;x}}) = \frac{k}{r} f(x) = \lim_{n\to\infty} \frac{k}{r} \cdot \frac{P(S_{\rho_{r,n;x}})}{V_\rho}.$$

Then, we have:

$$\begin{aligned}
&P\left[T_1^{(n)} > r, T_1^{(n)} > s | X_1 = x, X_2 = y\right] \\
&= \sum_{\substack{k \le i,j \\ i+j \le n-2}} \frac{(n-2)!}{i!j!(n-2-i-j)!} \left(\frac{k}{nr}\right)^i \left(\frac{k}{ns}\right)^j \cdot \\
&\left(\frac{P(S_{\rho_{r,n;x}})}{V_{\rho_{r,n}}}\right)^i \left(\frac{P(S_{\rho_{s,n;x}})}{V_{\rho_{s,n}}}\right)^j \cdot \\
&\left[1 - \frac{1}{n}\left(\frac{k}{r} \cdot \frac{P(S_{\rho_{r,n;x}})}{V_{\rho_{r,n}}} + \frac{k}{s}\frac{P(S_{\rho_{s,n;x}})}{V_{\rho_{s,n}}}\right)\right]^{n-2-i-j}
\end{aligned}$$

Since:

$$\begin{aligned}
&\frac{(n-2)!}{(n-2-i-j)!n^i n^j} \to 1 \text{ as } n \to \infty \\
&\left[1 - \frac{1}{n}\left(\frac{k}{r} \cdot \frac{P(S_{\rho_r})}{V_{\rho_r}} + \frac{k}{s}\frac{P(S_{\rho_s})}{V_{\rho_s}}\right)\right]^{n-2-i-j} \to e^{-\left(\frac{k}{r}f(x) + \frac{k}{s}f(y)\right)} \text{ as } n \to \infty \\
&\lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n = e^{-1}.
\end{aligned}$$

we obtain:

$$\begin{aligned}
&\lim_{n\to\infty} P\left[T_1^{(n)} > r, T_2^{(n)} > s | X_1 = x, X_2 = y\right] \\
&= \sum_{i=k}^{n-2-k} \sum_{j=k}^{n-2-k} \frac{k^i}{i! r^i} \left[f(x)\right]^i \cdot \frac{k^j}{j! s^j} \left[f(y)\right]^j \cdot e^{-\left(\frac{k}{r} f(x) + \frac{k}{s} f(y)\right)} \\
&= \left[\sum_{i=k}^{n-2-k} \frac{k^i}{i! r^i} \left[f(x)\right]^i e^{-\frac{k}{r} f(x)}\right] \cdot \left[\sum_{j=k}^{n-2-k} \frac{k^j}{j! s^j} \left[f(y)\right]^j e^{-\frac{k}{s} f(y)}\right] \\
&= P[T_x > r] \cdot P[T_y > s],
\end{aligned}$$

where for given z, the random variable T_z has the *pdf*:

$$h_{T_x} = \frac{f(z)\left(\frac{k}{y} f(z)\right)^k}{(k-1)! y^2} e^{-\frac{k f(z)}{y}} = \frac{k^k f(z)^{k+1}}{(k-1)! y^{k+2}} e^{-\frac{k f(z)}{y}}.$$

Therefore, by the theorem of independent variables:

$$\begin{aligned}
&\lim_{n\to\infty} E[T_1^{(n)} T_2^{(n)}] = \lim_{n\to\infty} \left[E[T_1^{(n)}] \cdot E[T_2^{(n)}]\right] \\
&\Rightarrow \lim_{n\to\infty} Cov[T_1^{(n)}, T_2^{(n)}] = 0.
\end{aligned}$$