MEASURING DEPENDENCE VIA MUTUAL INFORMATION

by

Shan Lu

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
September 2011

Abstract

Considerable research has been done on measuring dependence between random variables. The correlation coefficient [10] is the most widely studied linear measure of dependence. However, the limitation of linearity limits its application. The informational coefficient of correlation [17] is defined in terms of mutual information. It also has some deficiencies, such as it is only normalized to continuous random variables.

Based on the concept of the informational coefficient of correlation, a new dependence measure, which we call the L-measure, is proposed in this work which generalizes Linfoot's measure for both continuous and discrete random variables. To further elucidate its properties, simulated models are used, and estimation algorithms are also discussed. Furthermore, another measure based on the L-measure, which we call the intrinsic L-measure, is studied for the purpose of studying nonlinear dependence. Based on criteria for a dependence measure presented by Renyi [21] and simulation results in this thesis, we believe that the L-measure is satisfactory as a dependence measure.

Acknowledgments

I would like to express my sincere appreciation to my supervisors, Professor Fady Alajaji and Professor Glen Takahara. I have been fortunate to have them as my supervisors during my master studies. I am very thankful to them for their guidance and advices throughout this thesis.

I would also like to thank Queen's University and Kingston, where I spent many memorable moments; and many thanks are to the Department of Mathematics and Statistics for offering me the opportunity to study and research on my favorite topics.

Finally, I would like to thank my family for their love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Dependence Measures

Measuring the dependence between two random variables is a fundamental and interesting problem. It has many applications in different fields, such as statistics, demography, economics, epidemiology and signal processing among other. Obtaining a measure that can sensibly describe the dependence relationship between random variables has received considerable attention in the past, with several dependence measures proposed.

The classical and most popular measure of linear dependence is the *correlation coefficient* [10]. For two random variables, their correlation coefficient is the quotient of their covariance and the product of their standard deviations. It is commonly used in many areas due to its simplicity, low computational cost and ease of estimation.

However, it is well known that correlation is not equivalent to dependence. Two

independent random variables are surely uncorrelated, which means that their correlation coefficient is zero; yet, for uncorrelated random variables, they are not necessarily independent [10].

1.2 Mutual Information as a Dependence Measure

Mutual information is a concept from information theory first introduced by Shannon in the context of digital communication [23]. It describes how much information two random variables share with each other, i.e. the amount of uncertainty about one random variable given knowledge of the other random variable. The mutual information for two random variables is symmetric and always nonnegative. It equals zero if and only if the two random variables are independent. In addition, the mutual information between two continuous random variables equals infinity if there is a functional relationship between these two random variables. These properties provide a possibility for the mutual information to be used as a dependence measure.

In 1957, Linfoot proposed a new dependence measure between two random variables, the *informational coefficient of correlation*, which is a monotone increasing function of mutual information [17]. It successfully preserves the properties of mutual information of being a symmetric nonnegative function and equaling to zero if and only if the arguments are independent. Furthermore, it has attractive properties as a dependence measure for continuous random variables. First, the value of the informational coefficient of correlation always lies between zero and one. This property is a useful standardization when comparing different dependence measures. Second, it is equal to the absolute value of the correlation coefficient when the random variables

are Gaussian. After Linfoot's initial work on this measure, more of its properties were studied and applied by several researchers, namely Granger and Lin [8], and Dionisio and Menezes [6]. Their work will be further discussed in Chapter 2.

1.3 Motivation

Measuring dependence lies at the heart of many statistical problems. Although the correlation coefficient is widely employed, it is not completely satisfactory to measure the dependence between random variables as it provides limited information about their dependence structure [10]. The absence of correlation is equivalent to independence in very rare cases, such as when the random variables are Gaussian distributed. The informational coefficient of correlation introduced by Linfoot [17] successfully addressed some deficiencies of the correlation coefficient. It is able to measure dependence when there exists a nonlinear structure between the random variables, while the correlation coefficient only measures linear dependence between random variables.

The informational coefficient of correlation was originally introduced for continuous random variables. We have found that it has some limitations when applied to discrete random variables. For example, it does not approach one when there is a functional relationship between the discrete random variables. This will be further examined in Chapter 2. Additionally, the estimation of the informational coefficient of correlation is not satisfactorily accurate in the existing literature. Therefore, we wish to define a new dependence measure which extends Linfoot's informational coefficient of correlation to discrete random variables and attempt to improve the accuracy of its estimation. Furthermore, since the correlation coefficient between two random

variables can always be reduced to zero after some linear transformation, the new dependence measure will be ideal if it can offer information about dependence via the minimum value this measure can achieve after linear transformation.

1.4 Contribution

A new dependence measure, which we call the *L-measure*, which extends Linfoot's informational coefficient of correlation, is proposed in this thesis. This measure can be used both for continuous and discrete random variables. Its properties and related theorems are discussed and proved in detail.

A method for estimating the L-measure is presented. Specifically, Gaussian kernel density estimation [24] and Gauss-Legendre quadrature are used. Issues that can create obstacles for estimation accuracy are discussed. Moreover, the histogram estimation method [22] is presented for the purpose of comparison.

A measure based on the L-measure, which we call the *intrinsic L-measure*, is next introduced. This measure is defined by what remains of the L-measure after a linear transformation that minimizes mutual information is applied. Its properties are discussed and it is applied on nonlinear data sets.

Four continuous models and two discrete models are used as examples. Implementations illustrating the advantages of the L-measure as a dependence measure are provided. Two of the continuous models have linear structures and the other two have nonlinear structures. Their intrinsic L-measures are calculated. The L-measure

is also applied on discrete Markov chains. It is observed that the L-measure is able to detect the nonlinear dependence between random variables when the correlation coefficient fails.

1.5 Outline

This thesis is organized as follows. Corresponding literature is reviewed at the beginning of each chapter. Chapter 2 reviews elementary concepts of dependence, mutual information, Linfoot's informational coefficient of correlation and introduces the new definition of the L-measure. The properties of the L-measure are discussed for continuous and discrete random variables. Moreover, four continuous models and two discrete models are generated to illustrate these properties. In Chapter 3, methods of estimation for the L-measure are presented and simulation results for four continuous models and two discrete models are implemented. In addition, the factors that influence the estimation accuracy are discussed. In Chapter 4, the intrinsic L-measure is defined and a numerical method of implementation is presented. The intrinsic L-measure of the four continuous examples are computed and the intrinsic L-measures of two nonlinear data sets are estimated. Chapter 5 summarizes the thesis and presents future directions.

Chapter 2

Definition and Fundamental Properties of the L-measure

In this chapter, related literature and preliminary definitions are first reviewed. Second, the definition of the L-measure is given and its properties are studied. Finally, several examples are presented for the purpose of understanding the L-measure.

2.1 Literature Review

Renyi [21] proposed the following criteria consisting of seven postulates that a measure of dependence should satisfy:

(a) it is defined for any pair of random variables;

(b) it is symmetric;

(c) its value lies between 0 and 1;

(d) it equals 0 if and only if the random variables are independent;

(e) it equals 1 if there is a strict dependence between the random variables;

(f) it is invariant under marginal one-to-one transformations of the random variables;

(g) if the random variables are Gaussian distributed, it equals the absolute value of their correlation coefficient.

The correlation coefficient is the most widely employed measure of linear dependence. It is defined as follows.

Definition 2.1.1. *For any two random variables X and Y, their correlation coefficient $\rho(X,Y)$ is given by*

$$\rho(X,Y) = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \tag{2.1}$$

where μ_X and μ_Y are the means of X and Y, respectively, and σ_X and σ_Y are the standard deviations of X and Y, respectively, provided the expectations exist.

This well known measure, however, fails to satisfy criteria (d) and (f).

The problem of obtaining a measure of dependence between two random variables is connected to that of acquiring a measure of the quantity of information about one contained in the other. Several measures of dependence have been proposed in the literature based on information theory [9, 14, 18, 25].

Joe's relative entropy dependence measure [14] was introduced based on the relative entropy. Relative entropy (or divergence) quantifies the similarity or difference between two different distributions [4]. The relative entropy between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in \chi} p(x) \log \frac{p(x)}{q(x)}.$$

Based on this concept, Joe [14] defined a measure of multivariate dependence for a random vector $(X_1, \ldots, X_m)$ as

$$\delta_{X_1,\ldots,X_m} = \int f_{X_1,\ldots,X_m} \log \left[\frac{f_{X_1,\ldots,X_m}}{\prod_j f_j} \right] d\mu.$$

where $f_{X_1,\ldots,X_m}$ is the joint density of $(X_1, \ldots, X_m)$ and f_j is the marginal density of X_j, $j = 1, \ldots, m$.

Given two random continuous variables X and Y, the parametric centered correntropy [18] is defined as

$$\begin{aligned} \eta_{a,b}(X,Y) &= E_{X,Y}\left[G_\sigma\left(aX + b - Y\right)\right] \\ &= \int\int G_\sigma(ax + b - y)\left[p(x,y) - p(x)p(y)\right] dxdy, \end{aligned}$$

where G_σ is the Gaussian kernel, a and b are real numbers with $a \neq 0$, $p(x,y)$ is the joint density of X and Y and $p(x)$ and $p(y)$ are the marginal densities. Then, the correntropy dependence measure between two continuous random variables is given by [18]

$$\Gamma(X,Y) = \sup_{a,b} |\eta_{a,b}(X,Y)|.$$

Silvey [25] adopted the concept from communication theory that the nature and extent of association between two random variables is captured by the ratio $\phi(x,y)$ of their joint density and the produts of their densities, i.e. $\phi(x,y) = p(x,y)/[p(x)p(y)]$. He introduced a dependence measure defined as

$$\Delta = E\left[d(x)\right],$$

where $d(x) = \int_{[y:\phi(x,y)>1]} \{p(y|x) - p(y)\}dy$. Therefore, Silvey's Δ measure can be written as

$$\Delta = \int\int_{\{(x,y):\phi(x,y)>1\}} [p(x,y) - p(x)p(y)]\, dxdy.$$

Granger al et al. [9] considered a different formula to achieve a dependence measure given by

$$S_\rho = \frac{1}{2}\int\int \left(p^{1/2}(x,y) - [p(x)p(y)]^{1/2}\right)^2 dxdy.$$

None of these four measures, Joe's relative entropy dependence measure, the correntropy dependence measure, Silvey's Δ coefficient, or Granger's measure, satisfy criterion (g); for more details, refer to [14, 18, 25, 9]. Moreover, Joe's measure is also not necessarily symmetric [4]. The correntropy dependence measure and Granger's measure also fail criteria (f) and have a large computational complexity [18].

A relatively new dependence measure, the distance correlation [26], $\Re$, generalizes the classical definition of correlation in two fundamental ways, for all distributions with finite first moments:

- $\Re(X,Y)$ is defined for X and Y in arbitrary dimensions;
- $\Re(X,Y) = 0$ holds if and only if the random vectors X and Y are independent.

The distance covariance between two random vectors of different dimensions, X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$, used to calculate their distance correlation is first presented as

$$\mho^2(X,Y) = \frac{1}{c_p c_q}\int\int \frac{|f(x,y) - f(x)f(y)|^2}{|x|_p^{1+p}|y|_q^{1+q}} dxdy,$$

where $c_d = \frac{\pi^{(d+1)/2}}{\Gamma\{(d+1)/2\}}$, and $\Gamma(\cdot)$ is the complete gamma function. Then, the distance correlation is defined by

$$\Re^2(X,Y) = \begin{cases} \frac{\mho^2(X,Y)}{\sqrt{\mho^2(X,X)\mho^2(Y,Y)}} & \text{if } \mho^2(X,X)\mho^2(Y,Y) > 0; \\ 0 & \text{if } \mho^2(X,X)\mho^2(Y,Y) = 0. \end{cases}$$

Recalling Renyi's criteria, the distance covariance is symmetric, its value always lies between 0 and 1, it reaches 0 if and only if X and Y are independent, and equals to 1 when there exist a vector a, a nonzero real number b and an orthogonal matrix C such that $Y = a + bXC$.

2.2 Preliminary Definitions

Entropy and mutual information are key concepts from information theory which were proposed by Shannon in 1948 [4]. Entropy is a measure of uncertainty in a random variable and mutual information measures how much information one random variable contains about another one. They are defined as follows for the discrete cases and continuous cases, respectively. In this thesis, all the logarithms we use are natural logarithms.

Definition 2.2.1. *The entropy of a discrete random variable X is defined by*

$$H(X) = -\Sigma_x p(x) \log p(x), \tag{2.2}$$

where $p(x)$ is the marginal probability mass function (pmf) of X.

For any two discrete random variables X and Y, their joint entropy is given by

$$H(X,Y) = -\Sigma_{x,y} p(x,y) \log p(x,y), \tag{2.3}$$

where $p(x,y)$ is the joint pmf of X and Y. The conditional entropy of Y given X is

defined as

$$H(Y|X) = \Sigma_x p(x) H(Y|X = x) = -\Sigma_{x,y} p(x, y) \log p(y|x), \tag{2.4}$$

where $H(Y|X = x)$ is the entropy of the conditional distribution of Y given $X = x$ and $p(y|x)$ is the conditional pmf of Y given $X = x$.

Definition 2.2.2. *The differential entropy of a continuous random variable X (admitting a density) is defined by*

$$h(X) = -\int f(x) \log f(x) dx, \tag{2.5}$$

where $f(x)$ is the marginal probability density function (pdf) of X.

For any two continuous random variables X and Y (admitting a joint density), their joint differential entropy is given by

$$h(X, Y) = -\int \int f(x, y) \log f(x, y) dx dy, \tag{2.6}$$

where $f(x, y)$ is the joint pdf of X and Y.

Their conditional entropy is defined by

$$h(Y|X) = -\int \int f(x, y) \log f(y|x) dx dy, \tag{2.7}$$

where $f(y|x)$ is the conditional pdf of Y given $X = x$.

Definition 2.2.3. *For any two discrete random variables X and Y, their mutual information is given by*

$$I(X; Y) = \Sigma_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \tag{2.8}$$

where $p(x, y)$ is the joint pmf of X and Y and $p(x)$ and $p(y)$ are the marginal pmfs of X and Y, respectively.

It follows from the definitions of entropy and mutual information that

$$I(X; Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y). \tag{2.9}$$

Definition 2.2.4. *For any two continuous random variables X and Y (admitting a joint density), their mutual information is given by*

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy, \tag{2.10}$$

where $f(x,y)$ is the joint pdf of X and Y and $f(x)$ and $f(y)$ are the marginal pdfs of X and Y, respectively.

It also follows from the definitions of differential entropy and mutual information that

$$I(X;Y) = h(Y) - h(Y|X) = h(X) - h(X|Y) = h(X) + h(Y) - h(X,Y). \tag{2.11}$$

Mutual information has properties that are desirable for a dependence measure. For example, (1) $I(X;Y) \geq 0$; (2) $I(X;Y) = 0$ if and only if X and Y are independent; (3) $I(X;Y) = +\infty$ if X and Y are continuous and there is a functional relationship between X and Y. Thus, $I(X;Y)$ satisfies Renyi's criteria except (c), (e) and (g). The informational coefficient of correlation, introduced by E. H. Linfoot in 1957 [17] addressed these problems for continuous random variables and is based on the mutual information.

Definition 2.2.5. *For two random variables X and Y, let $I(X;Y)$ denote the mutual information between X and Y. Their informational coefficient of correlation is given by*

$$r(X,Y) = \sqrt{1 - e^{-2I(X;Y)}}. \tag{2.12}$$

The informational coefficient of correlation was introduced for continuous random variables only. When it is applied to discrete random variables, a problems arises in that it does not equal to one when there is a functional relationship between the random variables. The following example illustrates this point.

Example 2.1.1

Let $P(X=-1)=P(X=0)=P(X=1)=\frac{1}{3}$ and $Y=X^2$. Then we get

X	-1	0	1
$p_X(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

,

Y	0	1
$p_Y(y)$	$\frac{1}{3}$	$\frac{2}{3}$

, and

$p_{X,Y}(x,y)$	$X=-1$	$X=0$	$X=1$
$Y=1$	$\frac{1}{3}$	0	$\frac{1}{3}$
$Y=0$	0	$\frac{1}{3}$	0

.

Thus,

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= H(Y) \\ &= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \\ &= 0.6365 \end{aligned}$$

So $r(X,Y)=\sqrt{1-e^{-2I(X;Y)}}=0.8485\neq 1$, while Y is a function of X.

Therefore, a new dependence measure based on the informational coefficient of correlation is necessary to extend its useful properties to discrete random variables.

2.3 Definition and Fundamental Properties

Definition 2.3.1. *For two arbitrary random variables X and Y, with alphabet $\mathscr{X}$ and $\mathscr{Y}$, respectively, let $\mathscr{A}_{X,Y}$ denote the set of all bivariate random vectors (U,V) on*

$\mathscr{X} \times \mathscr{Y}$ with the same marginal distributions as X and Y , and let $I(U;V)$ represent the mutual information between two random variables U and V. Then the L-measure of X and Y is defined as

$$L(X,Y) = \left[1 - \exp\left\{\frac{-2I(X;Y)}{1 - I(X;Y)/\sup_{U,V\in\mathscr{A}_{X,Y}} I(U;V)}\right\}\right]^{1/2}. \tag{2.13}$$

With this definition of the L-measure, its properties are next studied. First note that $L(X,Y)$ is defined for arbitrary random variables.

When X and Y are both discrete, it is clear that

$$\sup_{U,V\in\mathscr{A}_{X,Y}} I(X;Y) = \min\{H(X), H(Y)\}$$

holds when X and Y share a functional relationship. Thus (2.13) can be rewritten as

$$L(X,Y) = \left[1 - \exp\left\{\frac{-2I(X;Y)}{1 - I(X;Y)/\min\{H(X), H(Y)\}}\right\}\right]^{1/2}. \tag{2.14}$$

This yields the following result.

Theorem 2.3.1. *If X and Y are two discrete random variables with finite alphabets, and Y is possibly a function of X, then $L(X,Y) = 1$ if and only if Y is a function of X.*

Proof. If $L(X,Y) = 1$, then the following equation holds

$$\frac{-2I(X;Y)}{1 - I(X;Y)/\sup_{U,V\in\mathscr{A}_{X,Y}} I(U;V)} = -\infty.$$

Since $0 \le I(X;Y) \le \min\{\log|\mathscr{X}|, \log|\mathscr{Y}|\}$, hence $1 - I(X;Y)/\sup_{U,V\in\mathscr{A}_{X,Y}} I(U;V) = 0$, i.e, $I(X;Y) = \sup_{U,V\in\mathscr{A}_{X,Y}} I(U;V)$. Since Y can be a function of X, we have $\sup_{U,V\in\mathscr{A}_{X,Y}} I(U;V) = H(Y)$. Therefore,

$$I(X;Y) = H(Y) \Leftrightarrow H(Y|X) = 0.$$

Hence, Y is a function of X.

Conversely, if Y is a function of X, i.e. $Y = g(X)$, then

$$I(X;Y) = H(Y) - H(Y|X) = H(Y).$$

Therefore $\sup_{U,V \in \mathscr{A}_{X,Y}} I(U;V) = H(Y)$, and thus $L(X,Y) = 1$. □

On the other hand, if X and Y are both continuous random variables, $L(X,Y)$ can be reduced to the informational coefficient of correlation $r(X,Y)$.

First, we note that if Y is a continuous random variable (with a *pdf* $f_Y(y)$), then $I(Y;Y) = +\infty$. This holds by the data processing theorem [4]: $I(Y,Y) \geq I(q_n(Y), q_n(Y))$ for any function $q_n(\cdot)$ indexed by integer $n \geq 1$. Now we can always choose $q_n(\cdot)$ such that $P(q_n(Y) = i) = \frac{1}{n}$ for $i = 1, 2, ...n$. Thus for this choice of $q_n(\cdot)$, we have that $I(q_n(Y), q_n(Y)) = H(q_n(Y)) = \log n$. Thus $I(Y,Y) \geq \log n$, $\forall n \geq 1$. Therefore, $I(Y,Y) \geq \lim_{n \to +\infty} \log n = +\infty$.

With this result, we thus obtain that if X and Y are two continuous random variables (with *pdf*s $f_X(x)$ and $f_Y(y)$), where $Y = g(X)$ for some function $g(\cdot)$, then by the data processing theorem,

$$\begin{aligned} I(X;Y) &= I(X;g(X)) \\ &\geq I(g(X);g(X)) \\ &= I(Y;Y) = +\infty. \end{aligned}$$

Lemma 2.3.2. *Suppose that X is a continuous random variable with cumulative distribution function (cdf) $F_X(x)$. Given another cdf $F(\cdot)$, we can always construct a*

continuous random variable with cdf $F(\cdot)$ such that it can be expressed as a function in terms of X.

Proof. Let $U = F_X(X)$, then $U \sim U(0,1)$; i.e., U is a uniformly distributed random variable over the interval $(0,1)$. Now construct a random variable $Z = F^{-1}(U)$, where $F^{-1}(\cdot)$ is the inverse function of $F(\cdot)$ defined by

$$F^{-1}(y) = \inf_{x \in \mathbb{R}} \{F(x) \geq y\}. \tag{2.15}$$

From (2.15), it is obvious that $F^{-1}(y) \leq x$ if and only if $y \leq F(x)$. Then the cdf of Z can be obtained as follows,

$$F_Z(z) = P(Z \leq z) = P(U \leq F(z)) = F(z).$$

Thus Z is the random variable that has cdf $F(\cdot)$ and is a function of X. □

Based on the above, we can reach the following conclusion.

Theorem 2.3.3. *If X and Y are continuous random variables, $L(X,Y)$ can be simplified as the informational coefficient of correlation $r(X,Y)$.*

Proof. By Lemma 2.3.2, we can choose U and V such that $U = g(V)$, which yields that $\sup_{U,V \in \mathscr{A}_{\mathscr{X},\mathscr{Y}}} I(U;V) = +\infty$ by the previous discussion. Then,

$$\begin{aligned} L(X,Y) &= \left[1 - \exp\left\{\frac{-2I(X;Y)}{1 - I(X;Y)/\sup_{U,V \in \mathscr{A}_{\mathscr{X},\mathscr{Y}}} I(U;V)}\right\}\right]^{1/2} \\ &= \left[1 - \exp\left\{-2I(X;Y)\right\}\right]^{1/2}, \end{aligned}$$

where we use the convention that $\frac{\infty}{\infty} = 1$. □

Furthermore, if X is a continuous random variable and Y is a discrete random variable, we have a lemma as follows.

Lemma 2.3.4. *If discrete random variable Y is a function of continuous random variable X, then $I(X;Y) = H(Y)$.*

Proof. Since Y is a function of X, then $H(Y|X) = 0$. Hence,

$$I(X,Y) = H(Y) - H(Y|X) = H(Y).$$

□

In this case, since X is a continuous random variable and Y is a discrete random variable, Y can always be expressed as a function of X. Hence, $\sup_{U,V\in\mathscr{A}_{X,Y}} I(U;V) = H(Y)$ and thus $L(X,Y)$ can be written as

$$L(X,Y) = \left[1 - \exp\left\{\frac{-2I(X;Y)}{1 - I(X;Y)/H(Y)\}}\right\}\right]^{1/2}. \tag{2.16}$$

Some other properties of the L-measure are listed and discussed as follows.

1. $L(X,Y) = L(Y,X)$ *and* $0 \le L(X,Y) \le 1$*;*

Proof. Since $0 \le I(X;Y)/\sup_{U,V\in\mathscr{A}_{X,Y}} I(U;V) \le 1$, then

$$-\infty \le \frac{-2I(X;Y)}{1 - I(X;Y)/\sup_{U,V\in\mathscr{A}_{X,Y}} I(U;V)} \le 0,$$

thus we have $0 \le L(X,Y) \le 1$. □

2. $L(X,Y)$ *is 0 if and only if* X *and* Y *are independent;*

Since X and Y are independent if and only if $I(X,Y) = 0$, consequently $L(X,Y) = 0$ if and only if X and Y are independent.

3. $L(X,Y) = 1$ *when there is an functional relationship between* X *and* Y*;*

From Theorem 2.3.1, $L(X,Y) = 1$ holds directly from the fact that the supremum value is achieved if there is a functional relationship between the variables.

4. $L(X,Y) = |\rho(X,Y)|$ *when* X, Y *are bivariate normally distributed with correlation coefficient* ρ*;*

Let $(X,Y) \sim N(\mu, K)$, where $K = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$.

$$\begin{aligned} I(X;Y) &= h(X) + h(Y) - h(X,Y) \\ &= \frac{1}{2}\log(2\pi e)\sigma^2 + \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\left(\log\left(2\pi e\right)\right)^2 |K| = -\frac{1}{2}\log(1-\rho^2) \\ L(X,Y) &= \sqrt{1-e^{-2I}} = \sqrt{1-(1-\rho^2)} = |\rho(X,Y)|. \end{aligned}$$

5. *$L(X,Y)$ is invariant under continuous and strictly increasing marginal transformations.*

Proof. Discrete case:

Assume that $g_1(\cdot)$ and $g_2(\cdot)$ are continuous and strictly increasing functions, then we have

$$\begin{aligned} &P(X = x) = P(g_1(X) = g_1(x)), \\ &P(Y = y) = P(g_2(Y) = g_2(y)), \\ &P(X = x, Y = y) = P(g_1(X) = g_1(x), g_2(Y) = g_2(y)). \end{aligned}$$

Thus, $I(X;Y) = I(g_1(X); g_2(Y))$. Hence we have $L(X,Y) = L(g_1(X), g_2(Y))$.

Continuous case:

Assuming that the joint probability density function of (X,Y) is $f_{X,Y}(x,y)$, and

$$U = g_1(X) \Longrightarrow X = g_1^{-1}(U), \tag{2.17}$$

$$V = g_2(Y) \Longrightarrow Y = g_2^{-1}(V). \tag{2.18}$$

Based on (2.17) and (2.18),

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{dg_1^{-1}(u)}{du} & 0 \\ 0 & \frac{dg_2^{-1}(v)}{dv} \end{vmatrix} = \left(\frac{dg_1^{-1}(u)}{du} \cdot \frac{dg_2^{-1}(v)}{dv}\right).$$

Moreover, we can have

$$f_U(u) = f_X\left(g_1^{-1}(u)\right) \cdot \frac{dg_1^{-1}(u)}{du},$$

$$f_V(v) = f_Y\left(g_2^{-1}(v)\right) \cdot \frac{dg_2^{-1}(v)}{dv},$$

$$f_{U,V}(u,v) = f_{X,Y}\left(x(u,v), y(u,v)\right)|J|,$$

hence,

$$\frac{f_{U,V}(u,v)}{f_U(u)f_V(v)} = \frac{f_{X,Y}\left(g_1^{-1}(u), g_2^{-1}(v)\right)}{f_X\left(g_1^{-1}(u)\right) f_Y\left(g_2^{-1}(v)\right)},$$

$$\begin{aligned}
I(U;V) &= \int\int f_{U,V}(u,v) \log\left\{\frac{f_{U,V}(u,v)}{f_U(u)f_V(v)}\right\} dudv \\
&= \int\int f_{X,Y}\left(g_1^{-1}(u), g_2^{-1}(v)\right) \log\left\{\frac{f_{X,Y}\left(g_1^{-1}(u), g_2^{-1}(v)\right)}{f_X\left(g_1^{-1}(u)\right) f_Y\left(g_2^{-1}(v)\right)}\right\} \frac{dg_1^{-1}}{du}\frac{dg_2^{-1}}{dv} dudv \\
&= \int\int f_{X,Y}(x,y) \log\left\{\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right\} dxdy \\
&= I(X;Y).
\end{aligned}$$

So $L(X,Y) = L(U,V)$. □

2.4 Examples

2.4.1 Application to the t-distribution

Let the random vector (X,Y) have the bivariate t distribution [16] with degrees of freedom ν, mean vector $\mu = 0$ and correlation matrix R, where their joint *pdf* is given

by

$$f_{X,Y}(x,y) = \frac{\Gamma\left(\frac{(\nu+2)}{2}\right)}{(\nu\pi)\Gamma(\nu/2)}|R|^{-0.5}\left[1+\frac{1}{\nu}(x,y)^T R^{-1}(x,y)\right]^{-(\nu+2)/2}, \tag{2.19}$$

where $\Gamma(\cdot)$ is the Gamma function defined by

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,\mathrm{d}t\,. \tag{2.20}$$

Guerrero-Cusumano [11] derived the form of the mutual information for the central multivariate t-distribution:

$$I(X;Y) = \Omega - \frac{1}{2}\log|R|,$$

where Ω is given by

$$\Omega = \ln\left[\frac{1}{\pi}\frac{B^2(\frac{1}{2},\frac{\nu}{2})}{B(1,\frac{\nu}{2})}\right] + (\nu+1)\left[\Psi(\frac{1+\nu}{2}) - \Psi(\frac{\nu}{2})\right] - \frac{\nu+2}{\nu}. \tag{2.21}$$

where $B(\cdot)$ is the Beta function defined by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t\,. \tag{2.22}$$

and where $\Psi(\cdot)$ is the Digamma function defined by

$$\psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}. \tag{2.23}$$

Thus,

$$\begin{aligned} L(X,Y) &= \sqrt{1-\exp\{-2I\}} = \sqrt{1-\exp\{-2\Omega\}|R|} \\ &= \sqrt{1-\exp\{-2\Omega\}(1-\rho^2(X,Y))}. \end{aligned} \tag{2.24}$$

Table 2.1 provides values of the L-measure for a range of ν and ρ and Figure 2.1 plots the L-measure in this range. The figure indicates that the dependence increases

as ν decreases and as ρ increases. In fact, the derivative of Ω with respect to ν,

$$\begin{aligned}\Omega' &= \left[\ln\left[\frac{\Gamma(\frac{\nu}{2}+1)}{\Gamma^2(\frac{1}{2}+\frac{\nu}{2})}\right] + (\nu+1)\left[\psi(\frac{1+\nu}{2}) - \psi(\frac{\nu}{2})\right] - \frac{\nu+2}{2}\right]' \\ &= \frac{2}{\nu} - \psi\left(\frac{1+\nu}{2}\right) + (\nu+1)\left[\psi'(\frac{1+\nu}{2}) - \psi'(\frac{\nu}{2})\right] - \frac{1}{2} \qquad (2.25)\end{aligned}$$

Since $\psi'(\frac{1+\nu}{2}) - \psi'(\frac{\nu}{2}) < 0$ and when $\nu \geq 2$, $\psi(\frac{1+\nu}{2}) > 0$, $\frac{2}{\nu} - \frac{1}{2} < 0$, $\Omega' < 0$, thus Ω increases as ν decreases and obviously, $\Omega \to 0$ as $\nu \to +\infty$. So the L-measure increases as ν decreases and $L(X,Y) \to |\rho|$ when $\nu \to +\infty$, which agrees with the fact that the joint t-distribution as $\nu \to +\infty$ is the same as a joint normal distribution with same mean vector and covariance matrix.

Moreover, $L(X,Y) = 1$ when $\rho = 1$ for any value of ν, as there is a linear functional relationship between X and Y.

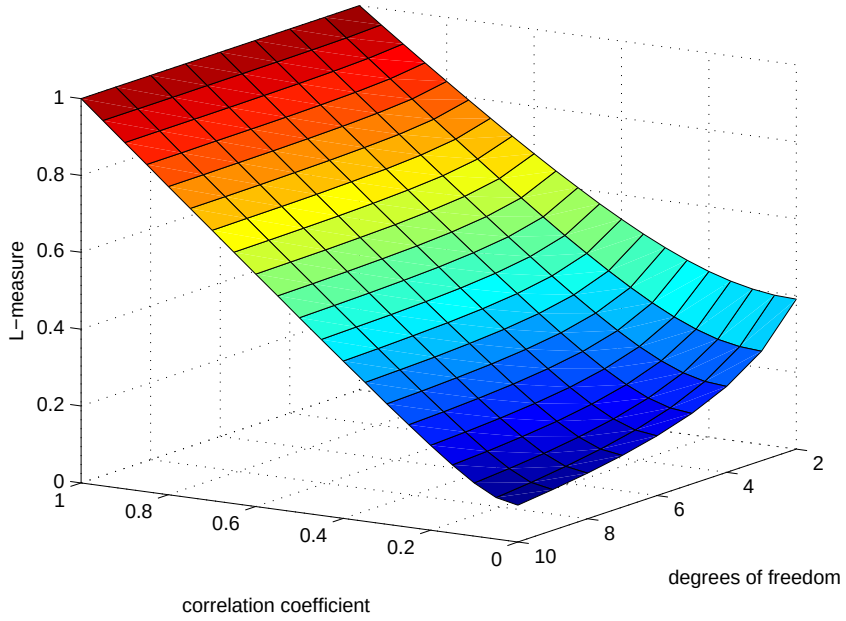
Table 2.1: L-measure of bivariate t-distribution

ρ	$\nu=2$	$\nu=4$	$\nu=6$	$\nu=8$	$\nu=10$
0	0.3904	0.2236	0.1555	0.1189	0.0962
0.2	0.4316	0.2966	0.2514	0.2315	0.2211
0.4	0.5367	0.4494	0.4246	0.4146	0.4096
0.6	0.6764	0.6261	0.6128	0.6075	0.6049
0.8	0.8336	0.8112	0.8054	0.8032	0.8021
1	1.0000	1.0000	1.0000	1.0000	1.0000

In particular, since Ω is always positive, we have

$$L(X,Y) = \sqrt{1 - \exp -2\Omega(1-\rho^2)} \geq \sqrt{1-(1-\rho^2)} = |\rho|.$$

Figure 2.1: L-measure of the t-distribution



2.4.2 Continuous Examples

In this section, two dependence measures, the L-measure and the correlation coefficient, are examined. Four continuous time series are used as examples. They are all strictly stationary and we let Z_t be i.i.d gaussian noise such that $Z_t \sim N(0,1)$.

We first recall the definition of autocorrelation [2].

Definition 2.4.1. *Let Y_t be a stationary time series. The autocorrelation function of Y_t at lag h is*

$$R(h) = R(Y_t, Y_{t+h}) = \rho(Y_t, Y_{t+h}) = \frac{E[(Y_t - \mu_{Y_t})(Y_{t+h} - \mu_{Y_{t+h}})]}{\sigma_{Y_t}\sigma_{Y_{t+h}}}. \tag{2.26}$$

Examples

Model 1: $Y_t = Z_t + 0.8Z_{t-1}^2$

(a) Autocorrelation:

Lag 1:

$$R(Y_t, Y_{t-1}) = \frac{EY_tY_{t-1} - EY_tEY_{t-1}}{\sigma_{Y_t}\sigma_{Y_{t-1}}}.$$
$$= \frac{E[(Z_t + 0.8Z_{t-1}^2)(Z_{t-1} + 0.8Z_{t-2}^2)] - 0.64}{\sigma_{Y_t}\sigma_{Y_{t-1}}} = 0.$$

Lag 2 to 5: Since Y_t and Y_{t-m} are independent, $R(Y_t, Y_{t-m}) = 0$.

(b) L-measure:

Lag 1:

Since $(Z_t, Z_{t-1}, Z_{t-2})^T \sim N(\mathbf{0}, I)$, and $\begin{pmatrix} Z_t \\ Y_t \\ Y_{t-1} \end{pmatrix} = \begin{pmatrix} Z_t \\ Z_t + 0.8Z_{t-1}^2 \\ Z_{t-1} + 0.8Z_{t-2}^2 \end{pmatrix}$. Thus,

$f_{Z_{t-2},Y_t,Y_{t-1}}(z_{t-2}, y_t, y_{t-1}) = f_{Z_{t-2},Z_t,Z_{t-1}}(z_{t-2}, z_{t-1}(y_t, y_{t-1}, z_{t-2}), z_t(y_t, y_{t-1}, z_{t-2})) * |J|$

$$= \frac{1}{(2\pi)^{3/2}} \exp\left\{-\frac{1}{2}\left[z_{t-2}^2 + (y_{t-1}^2 - 0.8z_{t-2}^2)^2 + \left(y_t - 0.8(y_{t-1}^2 - 0.8z_{t-2}^2)^2\right)^2\right]\right\}.$$

Thus $f_{Y_t,Y_{t-1}}(Y_t, Y_{t-1}) = \int f_{e_{t-2},Y_t,Y_{t-1}}(e_{t-2}, Y_t, Y_{t-1})de_{t-2}$. To calculate the integral, we use Gaussian quadrature with 100 nodes and weight at interval$(-15, 15)$ (Note: this method is used in all examples listed in this section; therefore it will not be reintroduced again).

$$h(Y_t, Y_{t-1}) = -E\left(\log\left(f_{Y_t,Y_{t-1}}(y_t, y_{t-1})\right)\right) = 3.3713.$$

Also, since $Z_t \sim N(0,1)$ and $Z_{t-1}^2 \sim \chi^2(1)$ are independent, and $\begin{pmatrix} Z_{t-1}^2 \\ Y_t \end{pmatrix} = \begin{pmatrix} Z_{t-1}^2 \\ Z_t + 0.8Z_{t-1}^2 \end{pmatrix}$. We have

$$\begin{aligned} f_{Z_{t-1}^2,Y_t}(z_{t-1}^2, y_t) &= f_{Z_{t-1}^2,Z_t}\left(z_{t-1}^2, z_t(y_t, z_{t-1}^2)\right) * |J| \\ &= \frac{1}{2\pi}\exp\left\{-\frac{1}{2}\left[z_t^2 + (y_t - 0.8z_t^2)^2\right]\right\}\left(z_t^2\right)^{-\frac{1}{2}}. \end{aligned}$$

So $h(y_t) = h(y_{t-1}) = 1.7540$. Thus,

$$I(y_t; y_{t-1}) = h(y_t) + h(y_{t-1}) - h(y_t, y_{t-1}) = 1.7540 + 1.7540 - 3.3713 = 0.1367,$$

and hence $L(Y_t; Y_{t-1}) = 0.4891$.

Lags 2-5:

Since Y_t and Y_{t-m} are independent, $L(Y_t, Y_{t-m}) = 0$.

To summarize, the results are shown in Table 2.2.

Table 2.2: Autocorrelation and L-measure of Model 1

Lag	R_m	L_m
0	1	1
1	0	0.4891
2	0	0
3	0	0
4	0	0
5	0	0

Model 2: $Y_t = Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2$

(a) Autocorrelation:

Lag 1:

$$R(Y_t, Y_{t-1}) = \frac{EY_tY_{t-1} - EY_tEY_{t-1}}{\sigma_{Y_t}\sigma_{Y_{t-1}}}$$
$$= \frac{E[(Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2)(Z_{t-1} + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2 + 0.8Z_{t-4}^2)] - 0.24^2}{\sigma_{Y_t}\sigma_{Y_{t-1}}}$$

$$= 0.5289.$$

Lag 2: Similarly,

$$R(Y_t, Y_{t-2}) = \frac{EY_tY_{t-2} - EY_tEY_{t-2}}{\sigma_{Y_t}\sigma_{Y_{t-2}}}$$
$$= \frac{E[(Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2)(Z_{t-2} + 0.8Z_{t-3}^2 + 0.8Z_{t-4}^2 + 0.8Z_{t-5}^2)] - 0.24^2}{\sigma_{Y_t}\sigma_{Y_{t-2}}}$$

$$= 0.2645.$$

Lag 3:

$$R(Y_t, Y_{t-3}) = \frac{EY_tY_{t-3} - EY_tEY_{t-3}}{\sigma_{Y_t}\sigma_{Y_{t-3}}}$$
$$= \frac{E[(Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2)(Z_{t-3} + 0.8Z_{t-4}^2 + 0.8Z_{t-5}^2 + 0.8Z_{t-6}^2)] - 0.24^2}{\sigma_{Y_t}\sigma_{Y_{t-3}}}$$

$$= 0.$$

Lags 4, 5: Since Y_t and Y_{t-m} are independent, $R(Y_t, Y_{t-m}) = 0$.

(b) L-measure:

Lag 1:

Since $(Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4})^T \sim N(\mathbf{0}, I)$, and

$$\begin{pmatrix} Z_{t-2} \\ Z_{t-3} \\ Z_{t-4} \\ Y_t \\ Y_{t-1} \end{pmatrix} = \begin{pmatrix} Z_{t-2} \\ Z_{t-3} \\ Z_{t-4} \\ Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-2}^2 \\ Z_{t-1} + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2 + 0.8Z_{t-4}^2 \end{pmatrix},$$

thus,

$$\begin{aligned} & f_{Z_{t-2},Z_{t-3},Z_{t-4},Y_t,Y_{t-1}}(z_{t-2}, z_{t-3}, z_{t-4}, y_t, y_{t-1}) \\ & = f_{Z_{t-2},Z_{t-3},Z_{t-4},Z_t,Z_{t-1}}(z_{t-2}, z_{t-3}, z_{t-4}, z_{t-1}(y_t, y_{t-1}, z_{t-2}, z_{t-3}, z_{t-4}), \\ & \quad z_t(y_t, y_{t-1}, z_{t-2}, z_{t-3}, z_{t-4})) * |J| \\ & = \frac{1}{(2\pi)^{5/2}} \exp\{-\frac{1}{2}[z_{t-2}^2 + z_{t-3}^2 + z_{t-4}^2 \\ & \quad + (y_{t-1} - 0.8z_{t-2}^2 - 0.8z_{t-3}^2 - 0.8z_{t-4}^2)^2 + (y_t - 0.8z_{t-2}^2 - 0.8z_{t-3}^2 \\ & \quad - 0.8(y_{t-1} - 0.8z_{t-2}^2 - 0.8z_{t-3}^2 - 0.8z_{t-4}^2)^2)^2]\}. \end{aligned}$$

Thus,

$$f_{Y_t,Y_{t-1}}(y_t, y_{t-1}) = \int\int\int f_{Z_{t-2},Z_{t-3},Z_{t-4},Y_t,Y_{t-1}}(z_{t-2}, z_{t-3}, z_{t-4}, y_t, y_{t-1}) dz_{t-2} dz_{t-3} dz_{t-4},$$

and $h(Y_t, Y_{t-1}) = -E(\log(f_{Y_t,Y_{t-1}}(y_t, y_{t-1}))) = 4.0094$.

Also, since $Z_t \sim N(0,1)$ and $Z_{t-1}^2 + Z_{t-2}^2 + Z_{t-3}^2 \sim \chi^2(3)$ are independent, and

$$\begin{pmatrix} Z_{t-1}^2 + Z_{t-2}^2 + Z_{t-3}^2 \\ Y_t \end{pmatrix} = \begin{pmatrix} Z_{t-1}^2 + Z_{t-2}^2 + Z_{t-3}^2 \\ Z_t + 0.8Z_{t-1}^2 \end{pmatrix},$$

thus, setting $X = Z_{t-1}^2 + Z_{t-2}^2 + Z_{t-3}^2$, we have

$$\begin{aligned} f_{X,Y_t}(x, y_t) &= f_{X,Z_t}\left(x, z_t(y_t, x)\right) * |J| \\ &= \frac{1}{2\pi} \exp\left\{-\frac{1}{2}[x + (y_t - 0.8x)^2]\right\} (x)^{\frac{1}{2}}. \end{aligned}$$

So $h(y_t) = h(y_{t-1}) = 2.1182$. Therefore,

$$I(Y_t; Y_{t-1}) = h(Y_t) + h(Y_{t-1}) - h(Y_t, Y_{t-1}) = 2.1182 + 2.1182 - 4.0094 = 0.2270$$

and $L(Y_t; Y_{t-1}) = 0.6041$.

Lag 2:

Since $(Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4}, Z_{t-5})^T \sim N(\mathbf{0}, I)$, and

$$\begin{pmatrix} Z_{t-1} \\ Z_{t-3} \\ Z_{t-4} \\ Z_{t-5} \\ Y_t \\ Y_{t-2} \end{pmatrix} = \begin{pmatrix} Z_{t-1} \\ Z_{t-3} \\ Z_{t-4} \\ Z_{t-5} \\ Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-2}^2 \\ Z_{t-2} + 0.8Z_{t-3}^2 + 0.8Z_{t-4}^2 + 0.8Z_{t-5}^2 \end{pmatrix}.$$

Similarly, $f_{Y_t,Y_{t-2}}(y_t, y_{t-2}) = \int\int\int\int f(z_{t-1}, z_{t-3}, z_{t-4}, z_{t-5}, y_t, y_{t-2}) dz_{t-1} dz_{t-3} dz_{t-4} dz_{t-5}$, and $h(Y_t, Y_{t-2}) = -E(\log(f(Y_t, Y_{t-2}))) = 4.1820$.

Thus,

$$I(Y_t; Y_{t-2}) = h(Y_t) + h(Y_{t-2}) - h(Y_t, Y_{t-2}) = 2.1182 + 2.1182 - 4.1820 = 0.0544$$

and hence $L(Y_t; Y_{t-2}) = 0.3211$.

Lag 3:

Since $(Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4}, Z_{t-5}, Z_{t-6})^T \sim N(\mathbf{0}, I)$,and

$$\begin{pmatrix} Z_{t-1} \\ Z_{t-2} \\ Z_{t-4} \\ Z_{t-5} \\ Z_{t-6} \\ Y_t \\ Y_{t-3} \end{pmatrix} = \begin{pmatrix} Z_{t-1} \\ Z_{t-2} \\ Z_{t-4} \\ Z_{t-5} \\ Z_{t-6} \\ Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-2}^2 \\ Z_{t-3} + 0.8Z_{t-4}^2 + 0.8Z_{t-5}^2 + 0.8Z_{t-6}^2 \end{pmatrix}.$$

Similarly,

$f_{y_t,y_{t-3}}(y_t, y_{t-3}) = \int \cdots \int f(z_{t-1}, z_{t-2}, z_{t-4}, z_{t-5}, z_{t-6}, y_t, y_{t-3}) de_{t-1} de_{t-2} de_{t-4} de_{t-5} de_{t-6}$,

and $h(Y_t, Y_{t-3}) = -E(\log(f(Y_t, Y_{t-2}))) = 4.2230$.

Thus, $I(Y_t; Y_{t-2}) = h(Y_t) + h(Y_{t-2}) - h(Y_t, Y_{t-2}) = 2.1182 + 2.1182 - 4.2230 = 0.0134$ and hence $L(Y_t; Y_{t-2}) = 0.1626$.

Lags 4, 5: Since Y_t and Y_{t-m} are independent, $L(Y_t, Y_{t-m}) = 0$.

To summarize, the results are shown in Table 2.3.

Model 3: $Y_t = Z_t + 0.8Z_{t-1}$

(a) Autocorrelation:

Table 2.3: Autocorrelation and L-measure of Model 2

Lag	R_m	L_m
0	1.0000	1
1	0.5289	0.6041
2	0.2645	0.3211
3	0	0.1626
4	0	0
5	0	0

Lag 1:

$$R(Y_t, Y_{t-1}) = \frac{EY_tY_{t-1} - EY_tEY_{t-1}}{\sigma_{Y_t}\sigma_{Y_{t-1}}}$$
$$= \frac{E[(Z_t + 0.8Z_{t-1})(Z_{t-1} + 0.8Z_{t-2})] - 0}{1 + 0.8 * 0.8} = 0.4878.$$

Lags 2 to 5: Since Y_t and Y_{t-m} are independent, $R(Y_t, Y_{t-m}) = 0$.

(b) L-measure:

Lag 1:

Since $(Z_t, Z_{t-1}, Z_{t-2})^T \sim N(\mathbf{0}, I)$,and $\begin{pmatrix} Z_t \\ Y_t \\ Y_{t-1} \end{pmatrix} = \begin{pmatrix} Z_t \\ Z_t + 0.8Z_{t-1} \\ Z_{t-1} + 0.8Z_{t-2} \end{pmatrix}$.

$$f(Y_t, Y_{t-1}) = \int f(z_t, z_{t-1}(y_t, y_{t-1}, z_t), z_{t-2}(y_t, y_{t-1}, z_t)) * |J| dz_t$$
$$= \int \frac{1}{(2\pi)^{3/2}} \exp\left\{-0.5\left(z_t^2 + \frac{(y_t - z_t)^2}{0.64} + \frac{(y_{t-1} - \frac{y_t - z_t}{0.8})^2}{0.64}\right)\right\} * \frac{1}{0.64} dz_t. \tag{2.27}$$

Thus, $I(Y_t; Y_{t-1}) = h(Y_t) + h(Y_{t-1}) - h(Y_t, Y_{t-1}) = 1.6664 + 1.6664 - 3.1970 =$

0.1358, hence $L(Y_t; Y_{t-1}) = 0.4878$.

Lags 2 to 5: Since Y_t and Y_{t-m} are independent, $L(Y_t, Y_{t-m}) = 0$.

To summarize, the result is shown in table in Table 2.4.

Table 2.4: Autocorrelation and L-measure of Model 3

Lag	R_m	L_m
0	1	1
1	0.4878	0.4878
2	0	0
3	0	0
4	0	0
5	0	0

Model 4: $Y_t = Z_t + 0.8Z_{t-1} + 0.8Z_{t-2} + 0.8Z_{t-3}$

(a) Autocorrelation:

Lag 1:

$$R(Y_t, Y_{t-1}) = \frac{EY_tY_{t-1} - EY_tEY_{t-1}}{\sigma_{Y_t}\sigma_{Y_t}}$$
$$= \frac{E[(Z_t + 0.8Z_{t-1} + 0.8Z_{t-2} + 0.8Z_{t-3})(Z_{t-1} + 0.8Z_{t-2} + 0.8Z_{t-3} + 0.8Z_{t-4})] - 0}{1 + 0.64 * 3}$$
$$= \frac{E[0.8Z_{t-1}^2 + 0.64Z_{t-2}^2 + 0.64Z_{t-3}^2]}{1 + 0.64 * 3}$$
$$= 0.7123.$$

Lag 2: Similarly, $R(Y_t, Y_{t-2}) = 0.4932$.

Lag 3: We have $R(Y_t, Y_{t-3}) = 0.2740$.

Lags 4, 5: Since Y_t and Y_{t-m} are independent, $R(Y_t, Y_{t-m}) = 0$.

(b) L-measure:

Lag 1:

Since $(Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4})^T \sim N(\mathbf{0}, I)$, and

$$((Z_t, Z_{t-1}, Z_{t-2}, Y_t, Y_{t-1})^T = A(Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4})^T,$$

then $(Y_t, Y_{t-1}) \sim N(\mathbf{0}, B_1)$, where $B_1 = \begin{pmatrix} 2.92 & 2.08 \\ 2.08 & 2.92 \end{pmatrix}$. Then

$$f_{Y_t, Y_{t-1}}(y_t, y_{t-1}) = \frac{1}{2\pi\sqrt{det(B_1)}} \exp(-0.5(y_t, y_{t-1})B_1^{-1}(y_t, y_{t-1})').$$

Thus, $I(Y_t; Y_{t-1}) = h(Y_t) + h(Y_{t-1}) - h(Y_t, Y_{t-1}) = 1.9547 * 2 - 3.5554 = 0.3540$, hence $L(Y_t, Y_{t-1}) = 0.7123$.

Lag 2: Similarly we have

$(Y_t, Y_{t-2}) \sim N(\mathbf{0}, B_2)$, where $B_2 = \begin{pmatrix} 2.92 & 1.44 \\ 1.44 & 2.92 \end{pmatrix}$.

$I(Y_t; Y_{t-2}) = h(Y_t) + h(Y_{t-2}) - h(Y_t, Y_{t-2}) = 1.9547 * 2 - 3.7701 = 0.1393$; therefore $L(Y_t, Y_{t-2}) = 0.4931$.

Lag 3: Similarly we have

$(Y_t, Y_{t-3}) \sim N(\mathbf{0}, B_3)$, where $B_3 = \begin{pmatrix} 2.92 & 0.8 \\ 0.8 & 2.92 \end{pmatrix}$.

$$I(Y_t; Y_{t-3}) = h(Y_t) + h(Y_{t-3}) - h(Y_t, Y_{t-3}) = 1.9547 * 2 - 3.8704 = 0.0390,$$

so $L(Y_t, Y_{t-3}) = 0.2739$.

Lags 4, 5: Since they are independent with each other, $L(Y_t, Y_{t-m}) = 0$.

To summarize, the result is shown in Table 2.5.

Table 2.5: Autocorrelation and L-measure of Model 4

Lag	R_m	L_m
0	1.0000	1
1	0.7123	0.7123
2	0.4932	0.4932
3	0.2740	0.2740
4	0	0
5	0	0

Model 1 is a nonlinear MA(1) model, so Y_t and Y_{t-1} are dependent and Y_t and Y_{t-m} are independent when $m > 1$. Table 2.2 shows that the correlation coefficient of Y_t and Y_{t-1} is 0 while the L-measure of them is 0.47.

Similarly, Model 2 is a nonlinear MA(3) model, so Y_t and Y_{t-m} with lags 1, 2, 3 are dependent, and their dependence tends to decrease when the lag increases. However, in the Table 2.3, the correlation coefficient is 0, when lag is 3, while the L-measure has a significant value.

From the results of Models 1 and 2, we can draw the conclusion that the L-measure captures well the dependence structure for nonlinear MA models while the correlation coefficient fails.

Models 3 and 4 are two linear MA models and we find that the value of the correlation coefficients and the L-measures are exactly the same for any number of lags, which supports property 4 of the L-measure as well. The dependence between Y_t and Y_{t-m} agrees with the correlation coefficient since they are all bivariate Gaussian

distributed for any number of lags.

2.4.3 Discrete Examples

In this section, we examine the L-measure and correlation coefficient for two discrete Markov chains.

Model 5: Stationary Two-state First-order Markov Chain

Consider a two-state stationary Markov chain with probability transition matrix $A = \begin{pmatrix} 0.5 & 0.5 \\ 0.6 & 0.4 \end{pmatrix}$ and stationary distribution $P(X_n = 0) = \frac{6}{11}$ and $P(X_n = 1) = \frac{5}{11}$.

(a) Autocorrelation:

Lag 1:

$$R(X_n, X_{n-1}) = \frac{EX_nX_{n-1} - EX_nEX_{n-1}}{\sqrt{\sigma_{X_n}\sigma_{X_{n-1}}}}$$
$$= \frac{\frac{2}{11} - \frac{25}{121}}{\frac{30}{121}} = -0.1.$$

Lag 2:

$$R(X_n, X_{n-2}) = \frac{EX_nX_{n-2} - EX_nEX_{n-2}}{\sqrt{\sigma_{X_n}\sigma_{X_{n-2}}}}$$
$$= \frac{\frac{23}{110} - \frac{25}{121}}{\frac{30}{121}} = 0.01.$$

Lag 3:

$$R(X_n, X_{n-3}) = \frac{EX_nX_{n-3} - EX_nEX_{n-3}}{\sqrt{\sigma_{X_n}\sigma_{X_{n-3}}}}$$
$$= \frac{0.2064 - \frac{25}{121}}{\frac{30}{121}} = -0.0009.$$

(b) L-measure:

Lag 1:

$$H(X_n) = -\sum p_{X_n}(x_n) \log p_{X_n}(x_n) = 0.6890;$$
$$H(X_{n-1}, X_n) = -\sum p_{X_{n-1},X_n}(x_{n-1}, x_n) \log p_{X_{n-1},X_n}(x_{n-1}, x_n) = 1.3730.$$

Thus,

$$I(X_n; X_{n-1}) = H(X_n) + H(X_{n-1}) - H(X_n, X_{n-1}) = 0.6890 * 2 - 1.3730 = 0.0050,$$

and since it is stationary, we have

$$L(X_n, X_{n-1}) = \left[1 - \exp\left\{\frac{-2I(X_n; X_{n-1})}{1 - I(X_n; X_{n-1})/H(X_n)}\right\}\right]^{1/2} = 0.1001.$$

Lag 2:

$$H(X_n) = -\sum p_{X_n}(x_n) \log p_{X_n}(x_n) = 0.6890;$$
$$H(X_{n-2}, X_n) = -\sum p_{X_{n-2},X_n}(x_{n-2}, x_n) \log p_{X_{n-2},X_n}(x_{n-2}, x_n) = 1.3780.$$

Thus,

$$I(X_n; X_{n-2}) = H(X_n) + H(X_{n-2}) - H(X_n, X_{n-2}) = 0.6890 * 2 - 1.3780 = 0.00005,$$

hence $L(X_n, X_{n-2}) = \left[1 - \exp\left\{\frac{-2I(X_n;X_{n-2})}{1-I(X_n;X_{n-2})/H(X_n)}\right\}\right]^{1/2} = 0.0100.$

Lag 3:

$$H(X_n) = -\sum p_{X_n}(x_n) \log p_{X_n}(x_n) = 0.6890;$$
$$H(X_{n-3}, X_n) = -\sum p_{X_{n-3},X_n}(x_{n-3}, x_n) \log p_{X_{n-3},X_n}(x_{n-3}, x_n) = 1.3780.$$

Thus,

$$I(X_n; X_{n-3}) = H(X_n) + H(X_{n-3}) - H(X_n, X_{n-3}) = 0.6890 * 2 - 1.3780 = 0.0000005,$$

hence $L(X_n, X_{n-3}) = \left[1 - \exp\left\{\frac{-2I(X_n;X_{n-3})}{1-I(X_n;X_{n-3})/H(X_n)}\right\}\right]^{1/2} = 0.0010.$

The results are summarized in Table 3.6.

Table 2.6: Autocorrelation and L-measure of Model 5

Lag	R_m	L_m
0	1	1
1	-0.1	0.1001
2	0.01	0.0100
3	-0.0009	0.0010

Model 6: Stationary Two-state Second-order Markov Chain

Consider a two-state stationary Markov second-order chain with probability transition $P(0|00) = P(1|11) = 0.8$, $P(0|10) = P(0|01) = 0.5$ and stationary distribution $P(X_n = 0) = \frac{1}{2}$ and $P(X_n = 1) = \frac{1}{2}$.

(a) Autocorrelation:

Lag 1:

$$R(X_n, X_{n-1}) = \frac{EX_nX_{n-1} - EX_nEX_{n-1}}{\sqrt{\sigma_{X_n}\sigma_{X_{n-1}}}}$$
$$= \frac{\frac{5}{14} - \frac{1}{4}}{\frac{1}{4}} = 0.4286.$$

Lag 2:

$$R(X_n, X_{n-2}) = \frac{EX_nX_{n-2} - EX_nEX_{n-2}}{\sqrt{\sigma_{X_n}\sigma_{X_{n-2}}}}$$
$$= \frac{\frac{5}{14} - \frac{1}{4}}{\frac{1}{4}} = 0.4286.$$

Lag 3:

$$R(X_n, X_{n-3}) = \frac{EX_nX_{n-3} - EX_nEX_{n-3}}{\sqrt{\sigma_{X_n}\sigma_{X_{n-3}}}}$$
$$= \frac{0.3143 - \frac{1}{4}}{\frac{1}{4}} = 0.2571.$$

(b) L-measure:

Lag 1:

$$H(X_n) = -\sum p_{X_n}(x_n) \log p_{X_n}(x_n) = 0.6931;$$
$$H(X_{n-1}, X_n) = -\sum p_{X_{n-1},X_n}(x_{n-1}, x_n) \log p_{X_{n-1},X_n}(x_{n-1}, x_n) = 1.2914.$$

Thus,

$$I(X_n; X_{n-1}) = H(X_n) + H(X_{n-1}) - H(X_n, X_{n-1}) = 0.6931 * 2 - 1.2914 = 0.0949,$$

and since it is stationary, we have

$$L(X_n, X_{n-1}) = \left[1 - \exp\left\{\frac{-2I(X_n; X_{n-1})}{1 - I(X_n; X_{n-1})/H(X_n)}\right\}\right]^{1/2} = 0.4443.$$

Lag 2:

$$H(X_n) = -\sum p_{X_n}(x_n) \log p_{X_n}(x_n) = 0.6931;$$
$$H(X_{n-2}, X_n) = -\sum p_{X_{n-2},X_n}(x_{n-2}, x_n) \log p_{X_{n-2},X_n}(x_{n-2}, x_n) = 1.2914.$$

Thus,

$$I(X_n; X_{n-2}) = H(X_n) + H(X_{n-2}) - H(X_n, X_{n-2}) = 0.6931 * 2 - 1.2914 = 0.0949,$$

hence $L(X_n, X_{n-2}) = \left[1 - \exp\left\{\frac{-2I(X_n;X_{n-2})}{1-I(X_n;X_{n-2})/H(X_n)}\right\}\right]^{1/2} = 0.4443.$

Lag 3:

$$H(X_n) = -\sum p_{X_n}(x_n) \log p_{X_n}(x_n) = 0.6931;$$
$$H(X_{n-3}, X_n) = -\sum p_{X_{n-3},X_n}(x_{n-3}, x_n) \log p_{X_{n-3},X_n}(x_{n-3}, x_n) = 1.3529.$$

Thus,

$$I(X_n; X_{n-3}) = H(X_n) + H(X_{n-3}) - H(X_n, X_{n-3}) = 0.6931 * 2 - 1.3529 = 0.0000005,$$

hence $L(X_n, X_{n-3}) = \left[1 - \exp\left\{\frac{-2I(X_n;X_{n-3})}{1-I(X_n;X_{n-3})/H(X_n)}\right\}\right]^{1/2} = 0.2603.$

The results are summarized in Table 2.7.

Table 2.7: Autocorrelation and L-measure of Model 6

Lag	R_m	L_m
0	1	1
1	0.4286	0.4443
2	0.4286	0.4443
3	0.2571	0.2603

Model 5 is stationary two-state first-order Markov chain, so X_t and X_{t-1} are dependent and X_t and X_{t-m} are less dependent when $m > 1$. Table 2.6 shows that the correlation coefficient and the L-measure of X_t and X_{t-m} both decrease when the lag increases.

Model 6 is a stationary two-state second-order Markov chain, so X_t and X_{t-m} are dependent in the first two lags and also have dependence structure with higher lags. The correlation coefficient and the L-measure in Table 2.7 both illustrate this dependence.

Chapter 3

Estimation

This chapter discusses the method we used for estimating the L-measure from time series data. Examples of both the continuous time series and discrete ones, described in Chapter 2, are utilized in this chapter to illustrate the estimation method.

In this chapter, it is assumed that we estimate the pdf $f_X(x)$ of X_t for continuous random variables or pmf $p_X(x)$ of X_t for discrete random variables given a sample $X_1, X_2, ..., X_n$. The symbol $\widehat{\cdot}$ is used to denote the estimation of the function being estimated. MATLAB is used for all the implementations.

3.1 Literature Review

Before proceeding to our approach we first briefly review some existing methods from the literature for estimating the mutual information and entropy. The two main

methods, histogram-based estimators and Kernel-based estimators, to estimate entropy or mutual information in literature, will be discussed.

Histogram-based estimators are widely used for their simplicity of implementation. Moddemeijer [19] and Tambakis [28] utilized the histogram method with equidistant cells. From the sequence of pairs of observations, a bivariate histogram is constructed with the cells having identical bandwidths. The probability of a point is approximated by the ratio of the number of pairs in the bin containing the point to the product of the number of pairs and the area of the bin. The the marginal *pdf*s, $f_X(x)$ and $f_Y(y)$, are estimated by taking the marginal of the estimate of the joint *pdf*. On the other hand, Darbellay [5] and Dionisio [6] utilized a histogram method with equiprobable cells. The bivariate histogram is constructed by dividing each edge into cells with approximately the same number of points. Then the mutual information is estimated by

$$\widehat{I}(X,Y) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \widehat{f}_{X,Y}(t_i,s_j)\log\left\{\frac{\widehat{f}_{X,Y}(t_i,s_j)}{\widehat{f}_X(t_i)\widehat{f}_Y(s_j)}\right\}, \tag{3.1}$$

where t_i and s_j are the points selected from the domain of $f_{X,Y}(x,y)$.

Ahmad and Lin [1] and Granger [8] proposed estimating the mutual information using a kernel estimate. They first estimate the joint *pdf* $f_{X,Y}(x,y)$ by a kernel estimate, then estimate the marginal *pdf*s, $f_X(x)$ and $f_Y(y)$, from the estimate of the joint *pdf*. Then they evaluate the marginal entropy and the joint entropy as follows:

$$\widehat{h}(X) = \frac{1}{N}\sum_{i=1}^{N}\log\left\{\widehat{f}_X(t_i)\right\}, \tag{3.2}$$

where $t_i, i = 1, ..., N$ are the points selected from the domain of $f_X(x)$,

$$\widehat{h}(Y) = \frac{1}{N}\sum_{j=1}^{N}\log\left\{\widehat{f}_Y(s_j)\right\}, \tag{3.3}$$

where $s_j, j = 1, ..., N$ are the points selected from the domain of $f_Y(y)$,

$$\widehat{h}(X,Y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\left\{\widehat{f}_{X,Y}(t_i,s_j)\right\}, \tag{3.4}$$

where t_i and s_j are the points selected from the domain of $f_{X,Y}(x,y)$.

3.2 Continuous Case

In this section, our methods for estimating the L-measure and correlation coefficient for continuous time series are described. Then, the four continuous time series from Chapter 2 are generated and the L-measure and the correlation coefficient are estimated based on these series. Possible reasons resulting in estimation bias are discussed at the end of the section.

3.2.1 Method of Estimation of the L-measure

Let $\{X_t\}$ denote a strictly stationary time series. Our goal is to estimate the lag m L-measure, $L(X_k, X_{k+m})$ which for a stationary time series depends only on m, and so we denote this by $L(m)$. The following equation holds:

$$I(X_k; X_{k+m}) = 2h(X_k) - h(X_k, X_{k+m}), \tag{3.5}$$

since $h(X_k) = h(X_{k+m})$.

To estimate the marginal density $f_X(x)$ of X_t, univariate Gaussian kernel density estimation [24] is used:

$$\widehat{f}_X(x) = \frac{1}{n\sigma}\sum_{i=1}^{n} K\left(\frac{x - X_i}{\sigma}\right) \tag{3.6}$$

where $K(\cdot)$ is the Gaussian kernel defined as $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-0.5x^2)$, n is the sample size of the series and σ is the smoothing parameter also called the bandwidth, which we select using Silverman's Rule [24],

$$\sigma = 0.9An^{-1/5} \tag{3.7}$$

where $A = \min\{\text{standard deviation of } \{X_t\}, \text{data interquartile range}/1.34\}$, where the data interquartile range is defined as the difference between the third and first quartiles.

Recall that the differential entropy $h(X_t)$ is defined as an integral in (2.5). To estimate the integration in $h(X_k)$, Gauss-Legendre quadrature [20] is applied:

$$\widehat{h}(X_k) = \sum_{i=1}^{N} w_i \widehat{f}(t_i) \log\left\{\widehat{f}(t_i)\right\}, \tag{3.8}$$

where $t_i, i = 1, ..., N$ are the evaluation points, $w_i, i = 1, ..., N$ are the weights for these points in the sum and N is the number of points in the approximation, which influences the accuracy of the estimation.

For the estimation of $f_m(x, y)$, the joint *pdf* of X_k and X_{k+m}, we apply a bivariate kernel density estimation method:

$$\widehat{f}_m(x, y) = \frac{1}{(n-m)\sigma^2} \sum_{i=1}^{n-m} K\left(\frac{(x, y)^T - (X_i, X_{i+m})^T}{\sigma}\right), \tag{3.9}$$

where $K(\cdot, \cdot)$ is the Gaussian kernel defined as $K(x, y) = \frac{1}{2\pi} \exp\{-0.5(x^2 + y^2)\}$ and the bandwidth is selected by Silverman's Rule again such that $\sigma = 0.96N^{-1/6}$ [24].

Recall that the joint differential entropy $h(X_k, X_{k+m})$ is defined as a double integral (2.6). To estimate the double integral in $h(X_k, X_{k+m})$, Gauss-Legendre quadrature is used again:

$$\widehat{h}(X_k, X_{k+m}) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} w_{ij}\widehat{f}_m(t_i, s_j)\log\left\{\widehat{f}_m(t_i, s_j)\right\} \tag{3.10}$$

where t_i and s_j are the evaluation points; w_{ij} is the product of w_i and w_j, the two weights corresponding to these two points, and N_1 and N_2 are the numbers of the evaluation points chosen.

By now the estimation for mutual information, $I(X_k, X_{k+m})$, is available and is given by

$$\widehat{I}(X_k; X_{k+m}) = 2\widehat{h}(X_k) - \widehat{h}(X_k, X_{k+m}). \tag{3.11}$$

Then, the estimate of the lag m L-measure, $L(m)$, is given by

$$\widehat{L}(m) = \widehat{L}(X_k, X_{k+m}) = \sqrt{1 - \exp\{-2\widehat{I}(X_k, X_{k+m})\}}. \tag{3.12}$$

We use the usual estimate of the autocorrelation function [2] in which the lag m correlation between X_k and X_{k+m}, $R(m)$, is given by

$$\widehat{R}(m) = \frac{1}{(n-m)\widehat{\sigma}_X^2}\sum_{k=1}^{n-m}(x_k - \widehat{\mu_X})(x_{k+m} - \widehat{\mu_X}), \tag{3.13}$$

where $\widehat{\mu_X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean and $\widehat{\sigma_X}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu_X})^2$ is the sample variance.

3.2.2 Simulation of continuous examples

In this subsection, the four continuous examples from Chapter 2 are generated, and estimates of $L(m)$ and $R(m)$ are computed. In the Gauss-Legendre quadrature method in (3.10) for $h(X_k)$, N is selected as 200, on the interval $(\min\{X_k\} - 3\sigma, \max\{X_k\} + 3\sigma)$; and for $h(X_k, X_{k+m})$, N_1 and N_2 both are defined as 50, and domain from $(\min\{X_k\} - 3\sigma, \min\{X_{k+m}\} - 3\sigma)$ to $(\max\{X_k\} + 3\sigma, \max\{X_{k+m}\} + 3\sigma)$.

Model 1: $Y_t = Z_t + 0.8Z_{t-1}^2$

The simulation with sample size 20,000 and replication number 10 is shown in Table 3.1.

Table 3.1: Simulation results for Model 1

Lag	$\widehat{R_m}$	$\widehat{L_m}$	R_m	L_m
0	1.0000	0.9827	1	1
1	-0.0003	0.4626	0	0.4891
2	-0.0004	0.0161	0	0
3	-0.0007	0.0157	0	0
4	-0.0005	0.0164	0	0
5	-0.0025	0.0160	0	0

Model 2: $Y_t = Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2$

The simulation with sample size 20,000 and replication number 10 is shown in Table 3.2.

Model 3: $Y_t = Z_t + 0.8Z_{t-1}$

The simulation with sample size 20,000 and replication number 10 is shown in table

Table 3.2: Simulation results for Model 2

Lag	$\widehat{R_m}$	$\widehat{L_m}$	R_m	L_m
0	1.0000	0.9915	1.0000	1
1	0.5329	0.5912	0.5289	0.6041
2	0.2678	0.3563	0.2645	0.3211
3	0.0022	0.1839	0	0.1626
4	0.0002	0.0218	0	0
5	0.0006	0.0199	0	0

in Table 3.3.

Table 3.3: Simulation results for Model 3

Lag	$\widehat{R_m}$	$\widehat{L_m}$	R_m	L_m
0	1.0000	0.9796	1	1
1	0.48894	0.4782	0.4878	0.4878
2	-0.0006	0.0166	0	0
3	-0.0056	0.0163	0	0
4	-0.0068	0.0153	0	0
5	-0.0034	0.0149	0	0

Model 4: $Y_t = Z_t + 0.8Z_{t-1} + 0.8Z_{t-2} + 0.8Z_{t-3}$

The simulation with sample size 20,000 and replication number 10 is shown in Table 3.4.

3.2.3 Bias Analysis

In this section, the histogram estimation method is used to compare with the kernel density estimation method that was used in this thesis. Furthermore, reasons that

Table 3.4: Simulation results for Model 4

Lag	$\widehat{R_m}$	$\widehat{L_m}$	R_m	L_m
0	1.0000	0.9884	1.0000	1
1	0.7122	0.7035	0.7123	0.7123
2	0.4941	0.4890	0.4932	0.4932
3	0.2750	0.2747	0.2740	0.2740
4	0.0002	0.0223	0	0
5	-0.0001	0.0225	0	0

may cause numerical bias are discussed.

Histogram Estimation

The histogram is one of the most classic and widely used density estimators thanks to its simplicity. In this method, the estimation of the density function of variable X, $\widehat{f}_X(x)$, is calculated as the ratio of the number of observations that fall into the bin containing x to the total number of observations:

$$\widehat{f}_X(x) = \frac{1}{nh}(\text{the number of } X_i \text{ in same bin as } x), \tag{3.14}$$

where n is the number of all the observations and h is the bandwidth that we choose by the Freedman-Diaconis' Rule [29],

$$h = 2 * \frac{\text{data interquartile range}}{n^{1/3}}. \tag{3.15}$$

Similarly to the univariate case, a bivariate histogram method for estimating $f_m(x, y)$, the joint *pdf* of X_k and X_{k+m}, is defined by

$$\widehat{f}_m(x, y) = \frac{1}{nh_1h_2}(\text{no. of } X_i, Y_j \text{ in same bin as } x), \tag{3.16}$$

where h_1 and h_2 are the bandwidths of X and Y, respectively, and n is the number of all the observations. Now the histogram estimator is applied to estimate the L-measure for different lags in time series.

Let $\{X_t\}$ denote a strictly stationary time series. We already know that

$$\widehat{I}(X_k; X_{k+m}) = 2\widehat{h}(X_k) - \widehat{h}(X_k, X_{k+m}). \tag{3.17}$$

To estimate the integration in $h(X_k)$, the histogram estimator is substituted, $\widehat{f}_X(\cdot)$, into Gauss-Legendre quadrature,

$$\widehat{h}(X_k) = \sum_{i=1}^{N} w_i \widehat{f}(t_i) \log\left\{\widehat{f}(t_i)\right\}, \tag{3.18}$$

where $t_i, i = 1, ..., N$ are the evaluation points, $w_i, i = 1, ..., N$ are the weights for these points in the sum and N is the number of points in the approximation, which influences the accuracy of the estimation.

To estimate the double integral in $h(X_k, X_{k+m})$, again, the bivariate histogram estimator is taken and plugged, $\widehat{f}_m(\cdot, \cdot)$, into Gauss-Legendre quadrature,

$$\widehat{h}(X_k, X_{k+m}) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} w_{ij} \widehat{f}_m(t_i, s_j) \log\left\{\widehat{f}_m(t_i, s_j)\right\}, \tag{3.19}$$

where t_i and s_j are the evaluation points; w_{ij} is the product of w_i and w_j, the two weights corresponding to these two points, and N_1 and N_2 are the numbers of the evaluation points chosen. Then the L-measure is calculated as 3.17.

To evaluate the accuracy of both estimators, they are both applied to estimate the L-measure of the first three lags of i.i.d $N(0, 1)$ distributed time series with same sample size 1000 and 5000. The estimated mean $\widehat{\mu}$ and standard deviation $\widehat{\sigma}$ are

shown in the following tables.

Table 3.5: $\widehat{\mu}$ and $\widehat{\sigma}$ for both estimators with sample size 1000

Lag	mean of hist	std of hist	mean of kernel	std of kernel
1	0.1901	0.1590	0.1036	0.0073
2	0.1744	0.1663	0.1011	0.0190
3	0.1795	0.1859	0.0900	0.0159

Table 3.6: $\widehat{\mu}$ and $\widehat{\sigma}$ for both estimators with sample size 5000

Lag	mean of hist	std of hist	mean of kernel	std of kernel
1	0.1082	0.1273	0.0531	0.0100
2	0.1379	0.1329	0.0536	0.0109
3	0.1089	0.1292	0.0509	0.0078

From the tables above, we note that the bias of the histogram estimation method is bigger than the bias of the kernel density estimation method; this is due to the discontinuity of histograms and the lack of information about tails of distribution which is not appropriate for the integral.

Moreover, the standard deviation of the histogram estimation method is also larger than that of the kernel method, because the selections of an origin and the bandwidth have important effects on the result and are highly depended on the data. It should be better to choose an origin and bandwidth for different contexts individually.

Based on these two reasons, the kernel density estimation method instead of histogram estimation method is selected as the estimation method in this thesis.

Bias

In this paragraph, several reasons that may lead to bias of the kernel density estimation method are presented.

From the examples above, we have seen that the bias of the L-measure with sample size 20,000 is around 0.02 when the value of L-measure approaches zero, and decreases when the value is bigger. Sample size is a very important element that affects the accuracy. To see how it influences the bias, several Gaussian distributed i.i.d. series with sample size 1000, 5000, 10000, 20000, 50000 are generated and their L-measures are calculated. The simulation with replication number 10 for different sample sizes is shown in Table 3.8.

Table 3.7: L-measure of i.i.d. series with different sample sizes

Lag	1000	5000	10000	20000	50000
1	0.1036	0.0531	0.0392	0.0211	0.0124
2	0.1011	0.0536	0.0374	0.0204	0.0132
3	0.0900	0.0509	0.0360	0.0207	0.0140

The bias of the L-measure for the sample of size 1000 is around 0.10 and can decrease to 0.015 for the one of the size of 20,000. Thus these values in this form can be regarded as criteria to examine whether there is a dependence between random variables for a given sample size.

Besides sample size, there also exist some other reasons leading to the inaccuracies. First, the interval in the Gauss-Legendre quadrature was selected based on the

assumption that the interval on the domain that is out of this interval is small enough to ignore; this also result in a bias. Second, the particular form of the L-measure can also exaggerate the bias in the estimation of mutual information. Last, the methods of generating the data sets and defining the bandwidth may also bring some variances.

3.3 Discrete Case

In this section, the method of estimation of the L-measure and the correlation coefficient for discrete time series are described. Then, the two discrete Markov chains from Chapter 2 are generated and the two measures are estimated.

3.3.1 Method of Estimation

Let $\{Y_t\}$ denote a strictly stationary discrete time series. Our goal is to estimate the lag m L-measure, $L(Y_k, Y_{k+m})$ which for a stationary time series depends only on m, and so we denote this by $L(m)$. The following equation holds

$$I(Y_k; Y_{k+m}) = 2H(Y_k) - H(Y_k, Y_{k+m}), \tag{3.20}$$

since $H(Y_k) = H(Y_{k+m})$.

Suppose that there are a set of possible states $a_1, a_2..., a_r$ for $\{Y_t\}$, to estimate the marginal pmf $p_Y(y)$. The ratio of the number of observations of a given point to the number of all observations is considered:

$$\widehat{P}(y = a_i) = \frac{n(y = a_i)}{n}, i = 1, ...r, \tag{3.21}$$

where $n(y = a_i)$ is the number of observations with value a_i and n is the sample size of the series.

Then the estimate of $H(Y_k)$ is given by

$$\widehat{H}(Y_k) = \sum_{i=1}^{r} \widehat{P}(Y = a_i) \log \left\{ \widehat{P}(Y = a_i) \right\}. \tag{3.22}$$

Similarly, the joint pmf of Y_k and Y_{k+m} can be estimated as

$$\widehat{P}(Y_k = a_i, Y_{k+m} = a_j) = \frac{n(Y_k = a_i, Y_{k+m} = a_j)}{n - m}, \tag{3.23}$$

where $i, j = 1, ...r$, $n(y_k = a_i, y_{k+m} = a_j)$ is the number of observation pairs that $Y_k = a_i$ and $Y_{k+m} = a_j$.

Thus the estimate of $H(Y_k, Y_{k+m})$ is given by

$$\widehat{H}(Y_k, Y_{k+m}) = \sum_{i=1}^{r} \sum_{j=1}^{r} \widehat{P}(Y_k = a_i, Y_{k+m} = a_j) \log \left\{ \widehat{P}(Y_k = i, Y_{k+m} = j) \right\}. \tag{3.24}$$

By now, the estimation for mutual information, $I(Y_k, Y_{k+m})$ is given by

$$\widehat{I}(Y_k; Y_{k+m}) = 2\widehat{H}(Y_k) - \widehat{H}(Y_k, Y_{k+m}). \tag{3.25}$$

Then, the estimate of the lag m L-measure, $L(m)$, is given by

$$\widehat{L}(m) = \widehat{L}(Y_k, Y_{k+m}) = \sqrt{1 - \exp\left\{ \frac{-2\widehat{I}(Y_k; Y_{k+m})}{1 - \widehat{I}(Y_k; Y_{k+m})/\widehat{H}(Y_k)} \right\}}. \tag{3.26}$$

On the other hand, the estimate of the autocorrelation function at lag m, $R(m)$,

as for the continuous time series, is given by

$$\widehat{R}(m) = \frac{1}{(n-m)\widehat{\sigma}_Y^2} \sum_{k=1}^{n-m} (Y_k - \widehat{\mu_Y})(Y_{k+m} - \widehat{\mu_Y}). \tag{3.27}$$

where $\widehat{\mu_Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the sample mean and $\widehat{\sigma_Y}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{\mu_Y})^2$ is the sample variance.

3.3.2 Simulation of Discrete Examples

In this subsection the two discrete Markov Chains from Chapter 2 are generated and estimates of $L(m)$ and $R(m)$ are computed.

Model 5: Stationary two-state first-order Markov chain

The simulation with sample size 20,000 and replication number 10 is shown in Table 3.8.

Table 3.8: Simulation results for Model 5

Lag	$\widehat{R_m}$	$\widehat{L_m}$	R_m	L_m
0	1.0000	1.0000	1	1
1	-0.1013	0.1015	-0.1	0.1001
2	0.0095	0.0105	0.01	0.0100
3	-0.0029	0.0061	-0.0009	0.0010

Model 6: Stationary two-state second-order Markov chain

The simulation with sample size 20,000 and replication number 10 is shown in Table 3.9.

The bias of L-measure for discrete cases is about 0.01 with sample size 1,000 and is

Table 3.9: Simulation results for Model 6

Lag	$\widehat{R_m}$	$\widehat{L_m}$	R_m	L_m
0	1.0000	1.0000	1	1
1	0.4302	0.4461	0.4286	0.4443
2	0.4278	0.4434	0.4286	0.4443
3	0.2572	0.2606	0.2571	0.2603

almost able to reach 0.005 with sample size increases more than 10000.

Chapter 4

Intrinsic L-measure

For any two random variables, the correlation coefficient between them can always be reduced to zero after an appropriate linear transformation. In this chapter, first, the properties of the L-measure under such transformations are discussed and the intrinsic L-measure is defined based on these properties. Second, the intrinsic L-measure of the four examples is calculated and the intrinsic L-measure of two nonlinear data sets is numerically estimated.

4.1 Preliminary Discussion

If X, Y is a pair of random variables with covariance matrix

$$K_{XY} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix},$$

there exists more than one linear transformation that can make the transformed variables uncorrelated.

For example, since K_{XY} is symmetric, we may write $K_{XY} = BDB^T$, where D is diagonal and B is orthonormal (i.e., $BB^T = B^TB = I$). Then setting

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = B^T \begin{pmatrix} X \\ Y \end{pmatrix},$$

the covariance matrix of $(Z_1, Z_2)^T$ is $K_{Z_1Z_2} = B^TK_{XY}B = B^TBDB^TB = D$, and so Z_1 and Z_2 are uncorrelated.

Independent component analysis (ICA) was first proposed by Juttena and Herault [15]. It describes how to separate the components to minimize their dependency by linear transformation. This technique is a powerful tool for data analysis for its ability of not only the decorrelating the random variables but also minimizing high-order statistical moments.

The mutual information, was first proposed as the dependence measure in ICA by Common [3] in 1994. Minimization of the mutual information is considered as a criterion for ICA. Since the L-measure is a monotone increasing function of the mutual information, we next consider the properties of the L-measure under a linear transformation.

4.2 Definition

Definition 4.2.1. *Let (X, Y) be a pair of continuous random variables. The intrinsic L-measure between X and Y is defined as $\min_A L(\widetilde{X}, \widetilde{Y})$, where*

$$\begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix} = A \begin{pmatrix} X \\ Y \end{pmatrix},$$

and the minimum is taken over all nonsingular 2×2 matrices A.

This section discusses how to find out the matrix which can make the transformed variables obtain the minimum value of their L-measure.

The L-measure is a monotonic increasing function of mutual information; therefore it is equivalent to look for the minimum value of mutual information instead of the one of the L-measure. Based on Property 5 of the L-measure that the L-measure is invariant under continuous and strictly increasing transformations for the marginals, the L-measure is then invariant under the multiplication with a diagonal matrix.

$$I(aX; dY) = h(aX) + h(dY) - h\left(\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right)$$

$$= h(X) + \log a + h(Y) + \log d - h(X,Y) - \log(a * d)$$

$$= I(X;Y). \tag{4.1}$$

Assume that

$$A_u = \begin{pmatrix} \frac{1}{\sqrt{a^2+b^2}} & 0 \\ 0 & \frac{1}{\sqrt{c^2+d^2}} \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \frac{a}{\sqrt{a^2+b^2}} & \frac{b}{\sqrt{a^2+b^2}} \\ \frac{c}{\sqrt{c^2+d^2}} & \frac{d}{\sqrt{c^2+d^2}} \end{pmatrix},$$

thus, the mutual information after transformation matrix A is the same as after transformation matrix A_u which has unit row vectors.

Therefore, we can set the range of a and c from -1 to 1, respectively. Furthermore, $b = \sqrt{1-a^2}$ and $d = \sqrt{1-c^2}$ are given.

A special class of random variables we need to mention here are bivariate Gaussian distributed random variables. As we know, for Gaussian distributed random variables, independence is equivalent to zero correlation. Thus the minimum value of the L-measure of transformed random variables is 0 if and only if the transformation also makes them uncorrelated. In this case we can say the intrinsic L-measure between them is zero.

Obviously, the intrinsic L-measure between one variable and itself is 0. Choose $A = \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}$, then $\widetilde{X} = X$ and $\widetilde{Y} = 0$, thus $I(\widetilde{X}; \widetilde{Y}) = H(\widetilde{Y}) - H(\widetilde{Y}|\widetilde{X}) = 0$.

The intrinsic L-measure of purely linearly dependent random variables also equals to 0. Assume that $Y = aX + Z$, where a is constant and Z is a random variable that is independent of X. Choose $A = \begin{pmatrix} a & -1 \\ 1 & 0 \end{pmatrix}$, then $\widetilde{X} = -Z$ and $\widetilde{Y} = X$, thus $I(\widetilde{X}; \widetilde{Y}) = 0$.

4.3 Examples

Model 1: $Y_t = Z_t + 0.8Z_{t-1}^2$

Since $(Z_t, Z_{t-1}, Z_{t-2})^T \sim N(\mathbf{0}, I)$, and $\begin{pmatrix} Z_t \\ Y_t \\ Y_{t-1} \end{pmatrix} = \begin{pmatrix} Z_t \\ Z_t + 0.8Z_{t-1}^2 \\ Z_{t-1} + 0.8Z_{t-2}^2 \end{pmatrix}$,

$$f_{Z_{t-2},Y_t,Y_{t-1}}(z_{t-2}, y_t, y_{t-1}) = f_{Z_{t-2},Z_t,Z_{t-1}}(z_{t-2}, z_{t-1}(y_t, y_{t-1}, z_{t-2}), z_t(y_t, y_{t-1}, z_{t-2})) * |J|$$
$$= \frac{1}{(2\pi)^{3/2}} \exp\left\{ -\frac{1}{2}\left[z_{t-2}^2 + (y_{t-1}^2 - 0.8z_{t-2}^2)^2 + \left(y_t - 0.8(y_{t-1}^2 - 0.8z_{t-2}^2)^2\right)^2 \right] \right\}.$$

Thus $f_{Y_t,Y_{t-1}}(Y_t, Y_{t-1}) = \int f_{e_{t-2},Y_t,Y_{t-1}}(e_{t-2}, Y_t, Y_{t-1}) de_{t-2}$.

Then based on the assumption that $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, we have

$$\begin{cases} \widetilde{Y}_t = aY_t + bY_{t-1} \\ \widetilde{Y}_{t-1} = cY_t + dY_{t-1}. \end{cases} \tag{4.2}$$

Therefore,

$$f_{\widetilde{Y}_t,\widetilde{Y}_{t-1}}(\widetilde{y}_t, \widetilde{y}_{t-1}) = F_{Y_t,Y_{t-1}}(y_t(\widetilde{y}_t, \widetilde{y}_{t-1}), y_{t-1}(\widetilde{y}_t, \widetilde{y}_{t-1}))|A^{-1}|$$

$$= \int f_{Z_{t-2},Y_t,Y_{t-1}}(z_{t-2}, y_t(\widetilde{y}_t, \widetilde{y}_{t-1}), y_{t-1}(\widetilde{y}_t, \widetilde{y}_{t-1})) * |A^{-1}| dz_{t-2}. \tag{4.3}$$

Thus,

$$\begin{cases} f_{\widetilde{Y}_t}(\widetilde{y}_t) = \int f_{\widetilde{Y}_t,\widetilde{Y}_{t-1}}(\widetilde{y}_t, \widetilde{y}_{t-1}) d\widetilde{y}_{t-1} \\ f_{\widetilde{Y}_{t-1}}(\widetilde{y}_{t-1}) = \int f_{\widetilde{Y}_t,\widetilde{Y}_{t-1}}(\widetilde{y}_t, \widetilde{y}_{t-1}) d\widetilde{y}_t. \end{cases}$$

To calculate the integral $h(\widetilde{Y}_t)$, here we chose Gaussian quadrature with 200 nodes and weight at interval $(-15(a+b)/2, 15(a+b)/2)$; and for $h(\widetilde{Y}_{t-1})$, Gaussian quadrature with the same number of nodes but weight at interval $(-15(c+d)/2, 15(c+d)/2)$ instead.

From Chapter 2, we know that $I(Y_t; Y_{t-1}) = h(Y_t) + h(Y_{t-1}) - h(Y_t, Y_{t-1}) = 1.7540 + 1.7540 - 3.3713 = 0.1367$, and $L(Y_t, Y_{t-1}) = 0.4891$.

Now we find (by numerical search) that when $A = \begin{pmatrix} -0.95 & 0.3122 \\ 0.35 & 0.9367 \end{pmatrix}$, $I(\widetilde{Y}_t, \widetilde{Y}_{t-1})$ would achieve its minimum value and thus $h(\widetilde{Y}_t) = 1.7609$, $h(\widetilde{Y}_{t-1}) = 1.6734$ and $h(\widetilde{Y}_t, \widetilde{Y}_{t-1}) = h(Y_t, Y_{t-1}) + \log(\det|A|) = 3.3713 + \log(\det|A|) = 3.3705$.

Thus $I(\widetilde{Y}_t; \widetilde{Y}_{t-1}) = 0.0638$ and $L(\widetilde{Y}_t, \widetilde{Y}_{t-1}) = 0.3461$.

Model 2: $Y_t = Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2$

Lag 1:

Since $(Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4})^T \sim N(\mathbf{0}, I)$, and

$$\begin{pmatrix} Z_{t-2} \\ Z_{t-3} \\ Z_{t-4} \\ Y_t \\ Y_{t-1} \end{pmatrix} = \begin{pmatrix} Z_{t-2} \\ Z_{t-3} \\ Z_{t-4} \\ Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-2}^2 \\ Z_{t-1} + 0.8Z_{t-2}^2 + 0.8Z_{t-3}^2 + 0.8Z_{t-4}^2 \end{pmatrix},$$

thus,

$$f_{Y_t, Y_{t-1}}(y_t, y_{t-1}) = \int\int\int f_{Z_{t-2}, Z_{t-3}, Z_{t-4}, Y_t, Y_{t-1}}(z_{t-2}, z_{t-3}, z_{t-4}, y_t, y_{t-1}) dz_{t-2} dz_{t-3} dz_{t-4}.$$

Then we assume that $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Therefore,

$$f_{\widetilde{Y}_t,\widetilde{Y}_{t-1}}(\widetilde{y}_t,\widetilde{y}_{t-1}) = f_{Y_t,Y_{t-1}}(y_t(\widetilde{y}_t,\widetilde{y}_{t-1}),y_{t-1}(\widetilde{y}_t,\widetilde{y}_{t-1}))|A^{-1}|$$

$$= \int\int\int f_{Z_{t-2},Z_{t-3},Z_{t-4},Y_t,Y_{t-1}}(z_{t-2},z_{t-3},z_{t-4},y_t(\widetilde{y}_t,\widetilde{y}_{t-1}),y_{t-1}(\widetilde{y}_t,\widetilde{y}_{t-1}))|A^{-1}|dz_{t-2}dz_{t-3}dz_{t-4}. \tag{4.4}$$

Thus,

$$\begin{cases} f_{\widetilde{Y}_t}(\widetilde{y}_t) = \int f_{\widetilde{Y}_t,\widetilde{Y}_{t-1}}(\widetilde{y}_t,\widetilde{y}_{t-1})d\widetilde{y}_{t-1} \\ f_{\widetilde{Y}_{t-1}}(\widetilde{y}_{t-1}) = \int f_{\widetilde{Y}_t,\widetilde{Y}_{t-1}}(\widetilde{y}_t,\widetilde{y}_{t-1})d\widetilde{y}_t. \end{cases} \tag{4.5}$$

To calculate the integral $h(\widetilde{Y}_t)$, here we chose Gaussian quadrature with 150 nodes and weight at interval $(-15(a+b)/2, 15(a+b)/2)$; and for $h(\widetilde{Y}_{t-1})$, Gaussian quadrature with the same number of nodes but weight at interval $(-15(c+d)/2, 15(c+d)/2)$ instead.

From Chapter 2 we have $I(Y_t;Y_{t-1}) = h(Y_t) + h(Y_{t-1}) - h(Y_t,Y_{t-1}) = 2.1182 + 2.1182 - 4.0094 = 0.2270$ and $L(Y_t,Y_{t-1}) = 0.6041$.

Now we find $A = \begin{pmatrix} -0.8000 & 0.6000 \\ 0.2700 & 0.9629 \end{pmatrix}$ can make $I(\widetilde{Y}_t,\widetilde{Y}_{t-1})$ achieve the minimum value and thus $h(\widetilde{Y}_t) = 1.7971$, $h(\widetilde{Y}_{t-1}) = 2.1929$ and $h(\widetilde{Y}_t,\widetilde{Y}_{t-1}) = h(Y_t,Y_{t-1}) + \log(\det|A|) = 4.0094 + \log(\det|A|) = 3.9393$.

Thus $I(\widetilde{Y}_t;\widetilde{Y}_{t-1}) = 0.0507$ and $L(\widetilde{Y}_t,\widetilde{Y}_{t-1}) = 0.3105$.

Lag 2:

Since $(Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4}, Z_{t-5})^T \sim N(\mathbf{0}, I)$, and

$$\begin{pmatrix} Z_{t-1} \\ Z_{t-3} \\ Z_{t-4} \\ Z_{t-5} \\ Y_t \\ Y_{t-2} \end{pmatrix} = \begin{pmatrix} Z_{t-1} \\ Z_{t-3} \\ Z_{t-4} \\ Z_{t-5} \\ Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-2}^2 \\ Z_{t-2} + 0.8Z_{t-3}^2 + 0.8Z_{t-4}^2 + 0.8Z_{t-5}^2 \end{pmatrix}.$$

Similarly,

$$f_{Y_t,Y_{t-2}}(y_t, y_{t-2}) = \int\int\int\int f(z_{t-1}, z_{t-3}, z_{t-4}, z_{t-5}, y_t, y_{t-2}) dz_{t-1} dz_{t-3} dz_{t-4} dz_{t-5}.$$

Therefore,

$$f_{\widetilde{Y}_t,\widetilde{Y}_{t-2}}(\widetilde{y}_t, \widetilde{y}_{t-2}) = f_{Y_t,Y_{t-2}}(y_t(\widetilde{y}_t, \widetilde{y}_{t-2}), y_{t-2}(\widetilde{y}_t, \widetilde{y}_{t-2}))|A^{-1}|$$

$$= \int\int\int\int f_{Z_{t-1},Z_{t-3},Z_{t-4},Z_{t-5},Y_t,Y_{t-2}} \left(z_{t-1}, z_{t-3}, z_{t-4}, z_{t-5}, y_t\left(\widetilde{y}_t, \widetilde{y}_{t-2}\right), y_{t-2}(\widetilde{y}_t, \widetilde{y}_{t-2})\right)$$

$$* |A^{-1}| dz_{t-1} dz_{t-3} dz_{t-4} dz_{t-5}. \tag{4.6}$$

$$\tag{4.7}$$

To calculate the integral $h(\widetilde{Y}_t)$, here we chose Gaussian quadrature with 60 nodes and weight at interval $(-15(a+b)/2, 15(a+b)/2)$; and for $h(\widetilde{Y}_{t-2})$, Gaussian quadrature with the same number of nodes but weight at interval $(-15(c+d)/2, 15(c+d)/2)$

instead.

From Chapter 2 we have

$$I(Y_t; Y_{t-2}) = h(Y_t) + h(Y_{t-2}) - h(Y_t, Y_{t-2}) = 2.1182 + 2.1182 - 4.1820 = 0.0544,$$

and $L(Y_t; Y_{t-2}) = 0.3211$.

Now we find $A = I$ can make $I(\widetilde{Y}_t, \widetilde{Y}_{t-2})$ achieve the minimum value and thus $I(\widetilde{Y}_t; \widetilde{Y}_{t-2}) = I(Y_t; Y_{t-2}) = 0.0544$ and $L(\widetilde{Y}_t, \widetilde{Y}_{t-2}) = L(Y_t; Y_{t-2}) = 0.3211$.

Lag 3:

Since $(Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4}, Z_{t-5}, Z_{t-6})^T \sim N(\mathbf{0}, I)$,and

$$\begin{pmatrix} Z_{t-1} \\ Z_{t-2} \\ Z_{t-4} \\ Z_{t-5} \\ Z_{t-6} \\ Y_t \\ Y_{t-3} \end{pmatrix} = \begin{pmatrix} Z_{t-1} \\ Z_{t-2} \\ Z_{t-4} \\ Z_{t-5} \\ Z_{t-6} \\ Z_t + 0.8Z_{t-1}^2 + 0.8Z_{t-2}^2 + 0.8Z_{t-2}^2 \\ Z_{t-3} + 0.8Z_{t-4}^2 + 0.8Z_{t-5}^2 + 0.8Z_{t-6}^2 \end{pmatrix}.$$

Similarly,

$$f_{y_t, y_{t-3}}(y_t, y_{t-3}) = \int\int\int\int\int f(z_{t-1}, z_{t-2}, z_{t-4}, z_{t-5}, z_{t-6}, y_t, y_{t-3}) de_{t-1} de_{t-2} de_{t-4} de_{t-5} de_{t-6}.$$

Therefore,

$$f_{\widetilde{Y}_t,\widetilde{Y}_{t-3}}(\widetilde{y}_t,\widetilde{y}_{t-3}) = f_{Y_t,Y_{t-3}}(y_t(\widetilde{y}_t,\widetilde{y}_{t-3}),y_{t-3}(\widetilde{y}_t,\widetilde{y}_{t-3}))|A^{-1}|$$
$$= \int\cdots\int f_{Z_{t-1},Z_{t-2},Z_{t-4},Z_{t-5},Z_{t-6},Y_t,Y_{t-2}}(z_{t-1},z_{t-2},z_{t-4},z_{t-5},z_{t-6},$$
$$y_t(\widetilde{y}_t,\widetilde{y}_{t-3}),y_{t-3}(\widetilde{y}_t,\widetilde{y}_{t-3}))|A^{-1}|dz_{t-1}dz_{t-2}dz_{t-4}dz_{t-5}dz_{t-6}. \quad (4.8)$$

To calculate the integral $h(\widetilde{Y}_t)$, here we chose Gaussian quadrature with 40 nodes and weight at interval $(-15(a+b)/2, 15(a+b)/2)$; and for $h(\widetilde{Y}_{t-3})$, Gaussian quadrature with the same number of nodes but weight at interval $(-15(c+d)/2, 15(c+d)/2)$ instead.

From Chapter 2 we have $I(Y_t;Y_{t-2}) = h(Y_t) + h(Y_{t-2}) - h(Y_t,Y_{t-2}) = 2.1182 + 2.1182 - 4.2230 = 0.0134$ and $L(Y_t;Y_{t-2}) = 0.1626$.

Now we find $A = \begin{pmatrix} 0.9900 & 0.1411 \\ -0.1600 & 0.9871 \end{pmatrix}$ can make $I(\widetilde{Y}_t,\widetilde{Y}_{t-2})$ achieve the minimum value and thus $h(\widetilde{Y}_t) = 2.1159$, $h(\widetilde{Y}_{t-3}) = 2.1192$ and $h(\widetilde{Y}_t,\widetilde{Y}_{t-3}) = h(Y_t,Y_{t-3}) + \log(\det|A|) = 4.2230 + \log(\det|A|) = 4.2228$.

Thus $I(\widetilde{Y}_t;\widetilde{Y}_{t-3}) = 0.0123$ and $L(\widetilde{Y}_t,\widetilde{Y}_{t-3}) = 0.1559$.

Model 3 and Model 4:

Since for any lag m for both models, (Y_t,Y_{t-m}) is bivariate Gaussian distributed, their intrinsic L-measure are all zero.

4.3.1 Application to Data Sets

In this section, the Santa Fe data set A [7] and the Lorenz data [27] are presented and their L-measure and intrinsic L-measure are estimated.

The Santa Fe data set A [7] is a univariate time record including 1,000 points from a physics laboratory experiment. It is stationary and is approximately described by three coupled nonlinear ordinary differential equations. Figure 4.1 shows the Santa Fe data Set A and Figure 4.2 shows the estimated L-measure and the estimated intrinsic L-measure for lags from 0 to 150.

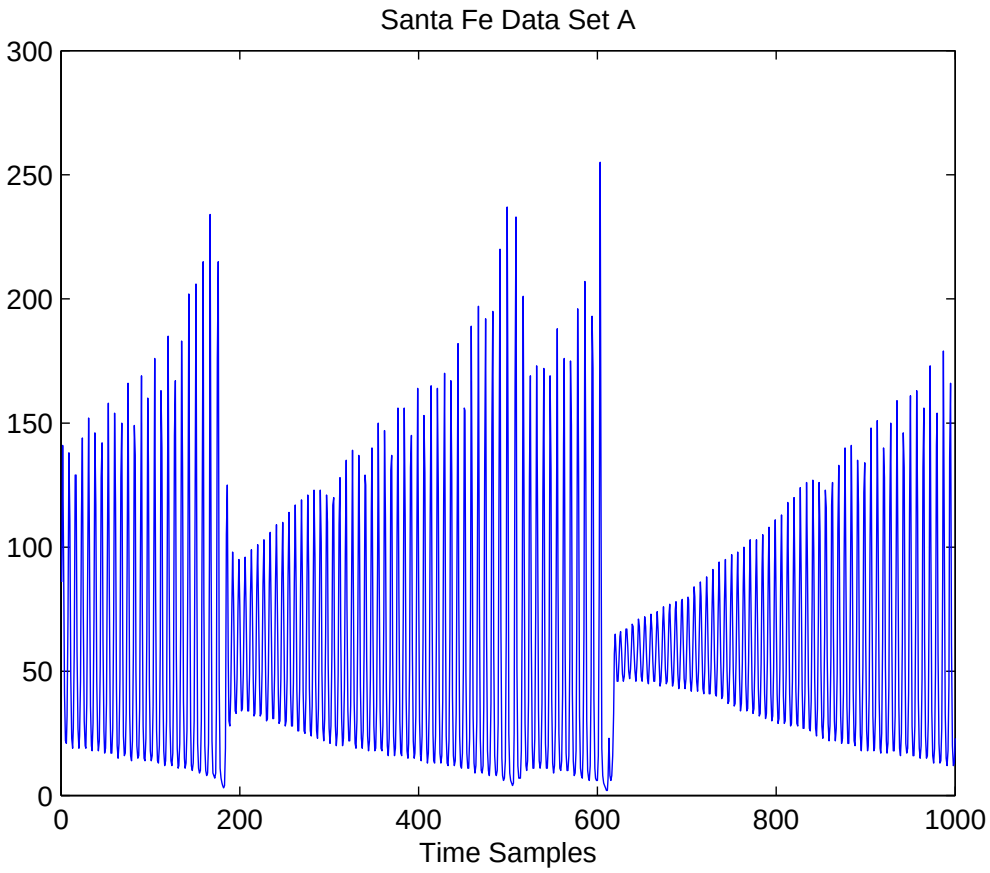
The Lorenz data [27] (of length 5000 points) is the y-component of the data generated from a chaotic Lorenz system by equations

$$\begin{cases} x' = \sigma(y - x) \\ y' = x(\rho - z) - y) \\ z' = xy - \beta z \end{cases}, \tag{4.9}$$

where $\rho = 28$, $\beta = 8/3$ and $(x_0, y_0, z_0) = (0, 0, 0)$.

These two data sets were tested by comparing of their correntropy and the correntropy of their surrogate data by Gunduz [12]. The results indicate that they have nonlinear structures. From Figure 4.2 we can see that the trend of the L-measure and of the intrinsic L-measure are almost the same. The intrinsic L-measure is decreasing when the L-measure decreases and is increasing when the L-measure increases. The value of the intrinsic L-measure is very close to the value of the L-measure and stabilizes around 0.2 after lag 100. This illustrates that it is hard to reduce the value of the L-measure through linear transformation after lag 100. It is not clear that

Figure 4.1: Santa Fe data set A



dependence still exists for high lags due to the 0.1 estimation bias. Figure 4.4 shows a similar result. The value of the intrinsic L-measure is close to the L-measure and becomes almost identical after lag 50.

Figure 4.2: L-measure and intrinsic L-measure of Santa Fe data set A

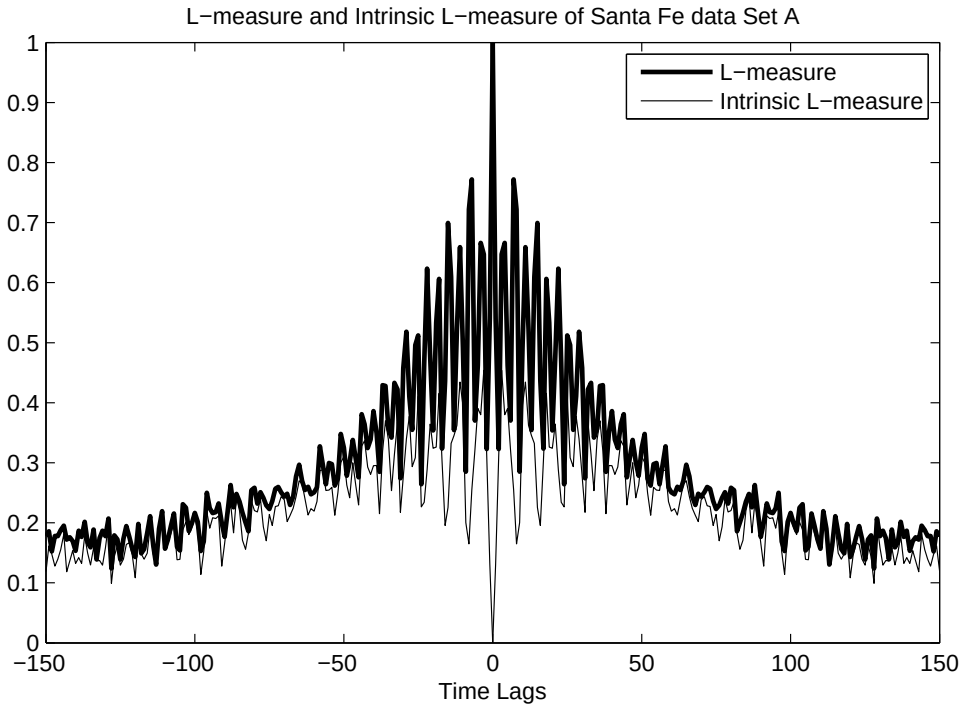


Figure 4.3: Lorenz data

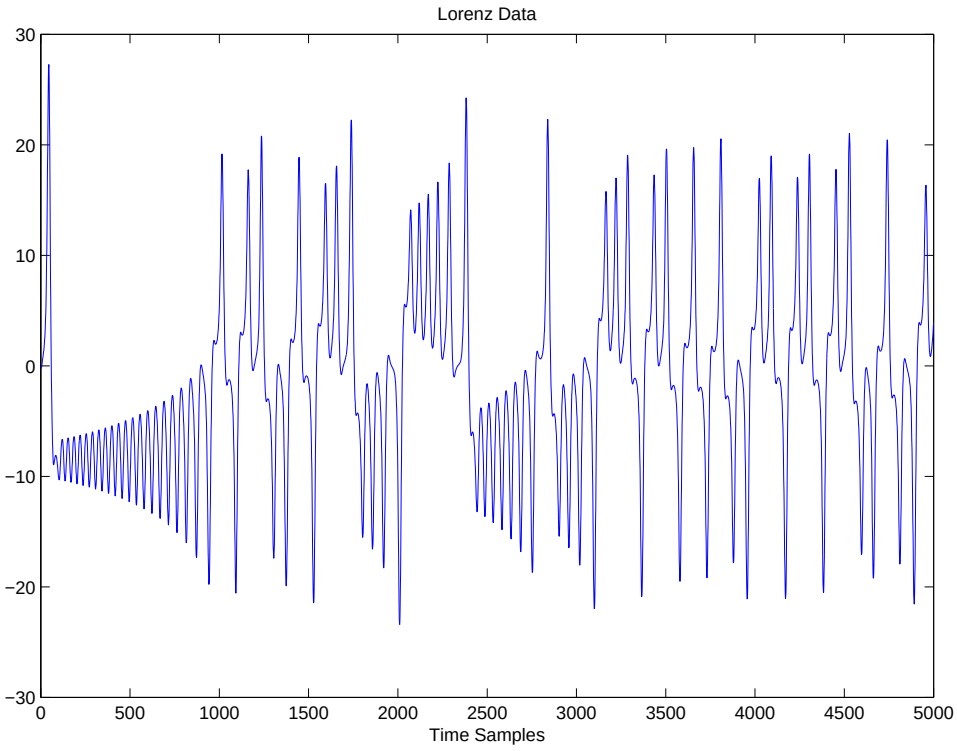
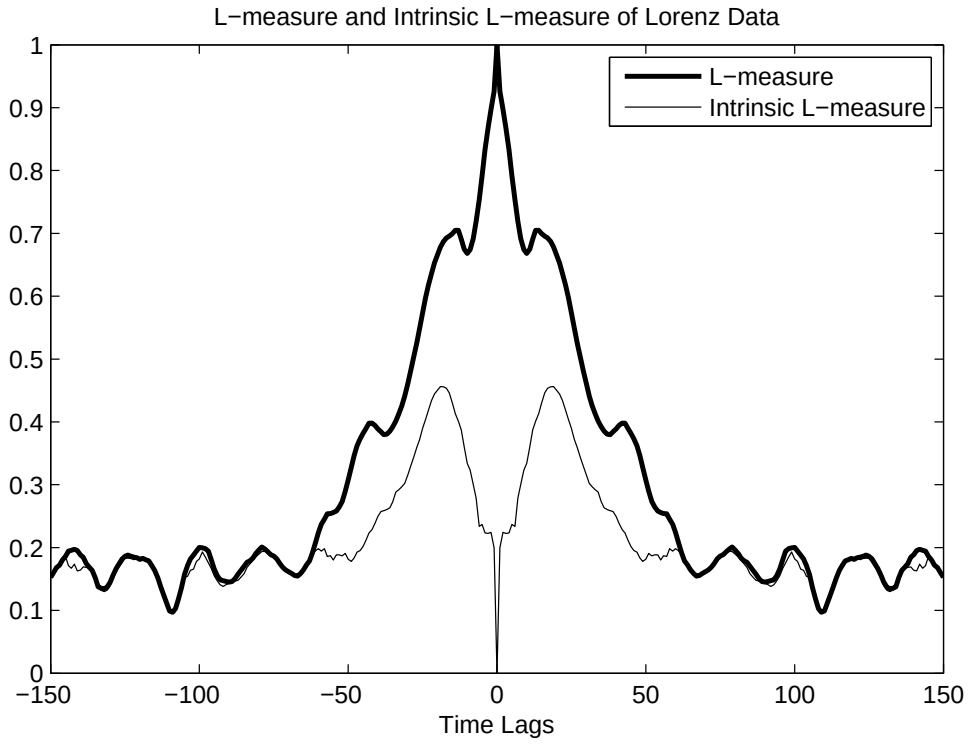


Figure 4.4: L-measure and intrinsic L-measure of Lorenz data



Chapter 5

Conclusion

5.1 Summary

Measuring dependence is a central and interesting research topic in statistics and is also a challenge because there is no known standard constructive dependence measure. In this thesis, the L-measure was introduced as a new measure of dependence. The L-measure is defined based on the concept of informational coefficient of correlation introduced by Linfoot [17]. It expands the application range from continuous random variables to arbitrary ones. All of the good properties of the informational coefficient of correlation for continuous cases are also inherited by the L-measure and were proved in detail. For example, the L-measure is symmetric, its value lies between 0 to 1 and equals to 0 if and only if its arguments are independent. Corresponding properties for discrete cases were also discussed and proved.

To estimate the L-measure for continuous cases, Gaussian kernel density estimation and Gauss-Legendre quadrature were combined and applied. The performances

of estimations of continuous examples are satisfactory. Compared to the histogram estimation method and some other methods from previous studies in the literature, the proposed method in this thesis has more accurate results. To calculate the L-measure for discrete cases, the ratio of the number of occurrences of a event to the total number of occurrences of all events has been considered as the probability of this event. Results of estimation of discrete examples are quite accurate.

Since all the bivariate random variables can be uncorrelated after some linear transformation, the intrinsic L-measure was defined to search for intrinsic dependence between two continuous random variables. Some interesting properties of the intrinsic L-measure were introduced. The intrinsic L-measure for bivariate Gaussian distributed random variables was shown to be zero. It reaches zero between a random variable and itself or between two random variables if they have a pure linear functional relationship. Two nonlinear data sets were used and the implementation of their L-measures and intrinsic L-measures show that the L-measure and the intrinsic L-measure can offer more information than the traditional correlation coefficient.

In conclusion, the L-measure is satisfactory as a new dependence measure since not only it can describe the dependence for random variables with linear structure, but it is also sensitive for those with nonlinear structure. It also applies to both discrete cases and continuous cases.

5.2 Further Research

The L-measure is defined as a new dependence measure based on the informational coefficient of correlation. This thesis offered one step for extending the range of application from continuous cases to arbitrary cases. Further research can start from several aspects as follows.

First, the dependence measure between two random variables can expand to a measure among multiple random variables. It is indeed known that mutual information is a special case of Kullback-Leiber number for measuring dependence between multiple random variables [13]. The Kullback-Leiber number is defined as the expectation of the logarithm of the ratio of the joint *pdf* and the product of marginal *pdf*s. Therefore, it would be very interesting if the dependence measure can consider the dependence between more than two variables.

Second, the accuracy of estimation for continuous cases can be improved. In this thesis, the Gaussian kernel density estimation was implemented. Alternative kernels may yield better results.

Third, regarding the estimation accuracy, the bandwidth is also a key aspect that influences the accuracy. Besides Silverman's Rule, alternative ways to choose the bandwidth can also be further studied.

Fourth, more properties of the intrinsic L-measure can be discovered. We found that the intrinsic L-measure is zero when there is a pure linear relationship between

the random variables, but the relationship between the intrinsic L-measure and non-linear dependence is still an open question.

Finally, because the results of the intrinsic L-measure are all numerical, there is ample room for future analytical investigations.

Bibliography

[1] I. Ahmadm and P. Lin. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Transactions on Information Theory*, 22(3):372–375, 1976.

[2] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2002.

[3] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 2006.

[5] G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

[6] A. Dionisio, R. Menezes, and D. A. Mendes. Mutual informaion: a measure of dependency for nonlinear time series. *Physica A*, 344(1-2):326–329, 2004.

[7] N. A. Gershenfeld and A. S. Weigend. *The Future of Time Series*. Xerox Corp., Palo Alto Research Center, 1993.

[8] C. Granger and J. Lin. Using the mutual information coefficient to identify lags in nonlinear models. *Journal of Time Series Analysis*, 15(4):371–384, 1994.

[9] C. W. Granger, E. Maasoumi, and J. Racine. A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*, 25(5):649–669, 2004.

[10] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[11] J. L. Guerrero-Cusumano. Measures of dependence for the multivariate t distribution with applications to the stock market. *Communications in Statistics-Theory and Methods*, 27(12):2985–3006, 1998.

[12] A. Gunduz and J. C. Principe. Correntropy as a novel measure for nonlinearity tests. *Signal Processing*, 89(1):14–23, 2009.

[13] S. Ihara. *Information Theory for Continuous Systems*. World Scientific, 1993.

[14] H Joe. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):157–164, 1989.

[15] C. Juttena and J. Herault. Analytic structure of the lorenz system. *Physical Review A*, 24(1):1–10, 1981.

[16] S. Kotz and S. Nadarajah. *Multivariate t Distributions and Their Applications*. Cambridge University Press, 2004.

[17] E. H. Linfoot. An informational measure of correlation. *Information and Control*, 1(1):85–89, 1957.

[18] W. Liu, P. Pokharel, J. Xu, and S. Seth. Correntropy for random variables: Properties and applications in statistical inference. In *Information Theoretic Learning Renyi's Entropy and Kernel Perspectives*, chapter 10, pages 385–413. Springer New York, 2010.

[19] R. Moddemeijer. A nonparametric estimation of the entropy for absolutely continuous distributions. *Signal Processing*, 75(1):51–63, 1999.

[20] W. H. Press. *Numerical Recipes: the Art of Scientific Computing*. Cambridge University Press, 1986.

[21] A. Renyi. On measures of dependence. *Acta Mathematica Hungarica*, 10(3-4):441–451, 1959.

[22] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand J Statist*, 9(2):65–78, 1982.

[23] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(1):379–423,623–656, 1948.

[24] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[25] S. D. Silvey. On a measure of association. *The Annals of Mathematical Statistics*, 35(3):1157–1166, 1964.

[26] G. J. Szekely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[27] M. Tabor and J. Weiss. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(4):2157–2167, 1991.

[28] D. N. Tambakis. On the informational content of asset prices. *Parallel Applications in Statistics and Economics Conference*, May 2000.

[29] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Springer, 2002.