Information Dependency and Its Applications

Inaugural dissertation

zur

Erlangung des akademischen Grades eines

Doctor rerum naturalium (Dr. rer. nat.)

an der Mathematisch-Naturwissenschaftlichen Fakultät

der

Ernst-Moritz-Arndt-Universität Greifswald

vorgelegt von

Le Trung Kien

geboren am 12.07.1982

in T.T. Hue, Vietnam

Greifswald, November 28th, 2012
Dekan: Prof. Dr. Klaus Fesser

1. Gutachter: Prof. Dr. Christoph Bandt
2. Gutachter: Prof. Dr. Friedrich Liese

Tag der Promotion: November 28th, 2012
Dedicated to my parents,
Le Dinh Phuc
Bui Thi Tan
my wife,
Nguyen Huyen Anh
and my brothers,
Le Anh Vu
Le Trung Hieu
Abstract

Independence is a basic concept of probability theory and statistics. In a lot of fields of sciences, dependency of different variables is gained lots of attention from scientists. A measure, named information dependency, is proposed to express the dependency of a group of random variables. This measure is defined as the Kullback-Leibler divergence of a joint distribution with respect to a product-marginal distribution of these random variables. In the bivariate case, this measure is known as mutual information of two random variables. Thus, the measure information dependency has a strong relationship with the Information Theory.

The thesis aims to give a thorough study of the information dependency from both mathematical and practical viewpoints. Concretely, we would like to research three following problems:

i. Proving that the information dependency is a useful tool to express the dependency of a group of random variables by comparing it with other measures of dependency.

ii. Studying the methods to estimate the information dependency based on the samples of a group of random variables.

iii. Investigating how the Independent Component Analysis problem, an interesting problem in statistics, can be solved using information dependency.
# Contents

<table>
<thead>
<tr>
<th>Contents</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>Notations</td>
<td>xi</td>
</tr>
</tbody>
</table>

## 1 Introduction

## 2 Measures of Dependency, Information Dependency

### 2.1 Measures of Dependency

#### 2.1.1 Copulas

#### 2.1.2 Rényi Axioms

#### 2.1.3 Bivariate measures of dependency

#### 2.1.4 Multivariate measures of dependency

### 2.2 Information Dependency

#### 2.2.1 Information

#### 2.2.2 Divergence

#### 2.2.3 Information Dependency

### 2.3 Normalized Information Dependency

### 2.4 Examples

#### 2.4.1 Linear convolution of Discrete distribution and Gaussian

#### 2.4.2 Linear convolution of Diagonal Uniform distribution and Gaussian

#### 2.4.3 Linear convolution of Circular Uniform distribution and Gaussian

## 3 Information Dependency Estimation

### 3.1 Estimations

#### 3.1.1 Data-dependent partition estimations

#### 3.1.2 $k$-Nearest neighbor estimations

#### 3.1.3 Gaussian kernel estimations
## CONTENTS

3.2 Numerical Experiments ........................................... 74  
   3.2.1 Bivariate Experiments ....................................... 74  
   3.2.2 Multivariate Experiments .................................... 84  
3.3 Strong consistency of Data dependent partitions ................. 89  
   3.3.1 Simultaneous approximation .................................. 90  
   3.3.2 Vapnik-Chervonenkis Theory and its improvement .......... 95  
   3.3.3 Strongly consistent partition rules ......................... 106  

4 Independent Component Analysis  .................................. 115  
   4.1 The problem .................................................... 115  
      4.1.1 The properties of ICA .................................... 116  
      4.1.2 Contrast functions and Information dependency .......... 120  
   4.2 The Simplest case of ICA ....................................... 123  
   4.3 Blind Audio Sources Separation ................................. 132  

5 Conclusions  .................................................................. 139  

Appendix  ......................................................................... 141  

Bibliography ...................................................................... 149
List of Figures

2.1 Examples of Linear convolution of Discrete distribution and Gaussian corresponding with parameters $r = 0, 0.2, 0.5$ and 1. .......................... 41
2.2 Linear convolution of Discrete distribution and Gaussian with parameter $r = 0.1$. 45
2.3 Linear convolution of Discrete distribution and Gaussian with parameter $r = 2$. 46
2.4 Examples of Linear convolution of Diagonal Uniform distribution and Gaussian corresponding with parameters $r = 0, 0.2, 0.5$ and 1. 48
2.5 Examples of Linear convolution of Circular Uniform distribution and Gaussian corresponding with parameters $r = 0, 0.2, 0.5$ and 1. 50

3.1 Illustration of the Data-dependent partitions for given 64 samples in $[0, 1]^2$. The grey rectangles in Darbellay-Vajda sub-figure are the local independent cells. ........................ 61
3.2 Illustration of two different Kraskov’s k-nearest neighbor information dependency estimation. First Estimation estimates the information dependency based on the hyper-cubics. Second Estimation estimates the information dependency based on the hyper-rectangles ......................................................... 68
3.3 Illustration of the models to estimate the information dependency based on the $\alpha$-divergence. The blue star-points are given 64-samples, the red diamond-points and square-points are independent samples generated by this sample. ........................ 68
3.4 Illustration of the Gaussian-Kernel density estimation ......................................................... 72
3.5 Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Discrete distribution and Gaussian with parameter $r = 0.1$. .......................... 76
3.6 Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Diagonal Uniform distribution and Gaussian with parameter $r = 0.2$. .......................... 77
3.7 Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Diagonal Uniform distribution and Gaussian with parameter $r = 0.5$. .......................... 78
3.8 Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Circular Uniform distribution and Gaussian with parameter $r = 0.2$. 

3.9 Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Circular Uniform distribution and Gaussian with parameter $r = 0.5$.

3.10 Two examples of the Gamma-Exponential distribution corresponding with parameters (1, 0.5, 1) and (1, 1, 1), respectively. The small diagrams are the zoom in of the big diagrams at the original coordinate (0,0).

3.11 Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the Gamma-Exponential distribution with parameter (1, 0.5, 1).

3.12 Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the Gamma-Exponential distribution with parameter (1, 1, 1).

3.13 Mean and standard deviation of 100 independent estimations of the information dependency. Estimations are done for the Seven-variate Gaussian distribution.

3.14 Mean and standard deviation of 100 independent estimations of the information dependency. Estimations are done for the Four-variate Exponential distribution with parameter (2, 2, 2, 2).

3.15 Mean and standard deviation of 100 independent estimations of the information dependency. Estimations are done for the Five-variate Logistics distribution with mean parameter (1, 1, 2, 2, 4) and variance parameter (2, 2, 6, 6, 9).

4.1 Interpreting a geometric view of the proof of Lemma 4.5.

4.2 Three simulations the unique local solution of the Simplest case of ICA. The star in each sub-figure are the positions of the local solutions. We use the operator (4.23) to simulate the values of Darbella-Vajda information dependency estimation.

4.3 The visual samples of original sound-signals (original) and observed sound-signals (mixture) in 5 seconds of three experiments.


1 Differences between the density functions $f_1$, $f$ and their lower and upper bounds. The case $0 < r < 0.15$.

2 Differences between the marginal density $f_1$ and their bounds. The cases $r \geq 2$.

3 Differences between the joint density $f$ and their bounds. The cases $r \geq 2$. 

1 Differences between the density functions $f_1$, $f$ and their lower and upper bounds. The case $0 < r < 0.15$. 

2 Differences between the marginal density $f_1$ and their bounds. The cases $r \geq 2$.

3 Differences between the joint density $f$ and their bounds. The cases $r \geq 2$. 

viii
List of Tables

2.1 Relationship between Rényi axioms and bivariate measures of dependency. The symbol (x) means that this measure has this axiom respectively. .............................. 24
2.2 Different measures of dependency for the example of the discrete distribution. ........................................... 41
2.3 Different measures of dependency for the example of figure 2.2. ........................................................... 45
2.4 Different measures of dependency for the example of figure 2.3. ........................................................... 46
2.5 Different measures of dependency for the example of figure 2.4. ........................................................... 48
2.6 The dependency and estimating dependency values of the examples in figure 2.5. ........................................ 50

3.1 Means, standard deviations and 90% confidence intervals for figure 3.5 corresponding with the sample sizes 10,000 and 30,000. ......................................................... 76
3.2 Means, standard deviations and 90% confidence intervals for figure 3.6 corresponding with the sample sizes 10,000 and 30,000. ................................................. 77
3.3 Means, standard deviations and 90% confidence intervals for figure 3.7 corresponding with the sample sizes 10,000 and 30,000. ................................................. 78
3.4 Means, standard deviations and 90% confidence intervals for figure 3.8 corresponding with the sample sizes 10,000 and 30,000. ................................................. 79
3.5 Means, standard deviations and 90% confidence intervals for figure 3.9 corresponding with the sample sizes 10,000 and 30,000. ......................................... 80
3.6 Means, standard deviations and 90% confidence intervals for figure 3.11 corresponding with the sample sizes 10,000 and 30,000. ........................................... 82
3.7 Means, standard deviations and 90% confidence intervals for figure 3.12 corresponding with the sample sizes 10,000 and 30,000. ........................................... 83
3.8 Means, standard deviations and 90% confidence intervals for figure 3.13 corresponding with the sample sizes 10,000 and 50,000. ........................................... 86
3.9 Means, standard deviations and 90% confidence intervals for figure 3.14 corresponding with the sample sizes 10,000 and 50,000. ........................................... 87
3.10 Means, standard deviations and 90% confidence intervals for figure 3.15 corresponding with the sample sizes 10,000 and 50,000. ................................. 88
Notations

\( \mathbb{R} \) space of real numbers and \( \pm \infty \)
\( \mathbb{R}^{m\times n} \) space of sequences of \( n \) \( m \)-dimensional real vectors
\( I \) \([0, 1]\)

\((\Omega, \mathcal{B}, \mathbb{P})\) probability space
\((\Omega^m, \mathcal{B}^m)\) product of \( m \) measurable spaces \((\Omega, \mathcal{B})\)

\( A, A_i \) event of \( \sigma \)-field \( \mathcal{B} \) or \( \mathcal{B}^m \)
\( X, X_i \) random variable
\( x, x_i \) realization of random variable
\( X = (X_1, \ldots, X_m) \) random vector
\( x = (x_1, \ldots, x_m) \) realization of random vector
\( X_1^n \) sequence of \( n \) random vectors \( X_1, X_2, \ldots, X_n \)
\( x_1^n \) realize of sequence of random vector \( X_1^n \)

\( \mu, \nu, \mu_1, \ldots, \mu_m \) joint, product marginal, and marginal distributions
\( \mu_n, \nu_n, \mu_{n,1}, \ldots, \mu_{n,m} \) empirical joint, product marginal, and marginal distributions
\( F, G, F_1, \ldots, F_m \) joint, product marginal, and marginal cumulative distribution functions (cdf)

\( F^{-1} \) pseudo-inverse of cumulative distribution function \( F \)
\( f, g, f_1, \ldots, f_m \) joint, product marginal, and marginal density functions
\( C \) copula
\( \Pi_m \) independent copula
\( M_m, W_2 \) upper and lower Fréchet-Hoeffding bound copulas

\( \mathbb{E} \) expectation operator
\( Var \) variance operator
\( cov \) covariance operator
\( \sigma(X) \) standard deviation of \( X \)
\( \rho(X) = f'(X)/f(X) \) score function of a random variable \( X \)
## NOTATIONS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{F}(X)$</td>
<td>Fisher’s information of a random variable $X$</td>
</tr>
<tr>
<td>$\text{mmsr}(X)$</td>
<td>minimum mean square error in estimating $X$ with condition of appearing $Y$</td>
</tr>
<tr>
<td>$\delta(X_1, \ldots, X_m)$</td>
<td>measure of dependency</td>
</tr>
<tr>
<td>$H$</td>
<td>Shannon entropy, general entropy</td>
</tr>
<tr>
<td>$h$</td>
<td>differential entropy</td>
</tr>
<tr>
<td>$\mathcal{D}, \mathcal{D}<em>\varphi, \mathcal{D}</em>\alpha$</td>
<td>divergence, $\varphi$-divergence, $\alpha$-divergence</td>
</tr>
<tr>
<td>$\mathcal{I}$</td>
<td>information-dependency</td>
</tr>
<tr>
<td>$\hat{\mathcal{I}}$</td>
<td>estimation of information-dependency</td>
</tr>
<tr>
<td>$\mathcal{I}^*$</td>
<td>normalized information-dependency</td>
</tr>
<tr>
<td>$\pi$</td>
<td>countable and measurable partition</td>
</tr>
<tr>
<td>$</td>
<td>\pi</td>
</tr>
<tr>
<td>$\sigma(\pi)$</td>
<td>$\sigma$-field generated by partition $\pi$</td>
</tr>
<tr>
<td>$\sigma(\mathcal{F})$</td>
<td>$\sigma$-field generated by a collection of events $\mathcal{F}$</td>
</tr>
<tr>
<td>$\mathcal{P}, \Omega$</td>
<td>collection of finite measurable partitions</td>
</tr>
<tr>
<td>${\pi_n}_{n \in \mathbb{N}}$</td>
<td>partition rule</td>
</tr>
<tr>
<td>$S_A^*(n)$</td>
<td>$n$-th shatter coefficient of a collection of sets $\mathcal{A}$</td>
</tr>
<tr>
<td>$\mathfrak{M}(\mathcal{P})$</td>
<td>maximal cell count of a collection of partitions $\mathcal{P}$</td>
</tr>
<tr>
<td>$\Delta_{\mathcal{P}}^*(n)$</td>
<td>$n$-th growth function of a collection of partitions $\mathcal{P}$</td>
</tr>
<tr>
<td>$</td>
<td>a</td>
</tr>
<tr>
<td>$\lfloor a \rfloor$</td>
<td>integer part of real number $a$</td>
</tr>
<tr>
<td>$|a|$</td>
<td>Euclidean length of a vector $a$</td>
</tr>
<tr>
<td>$\det(A)$</td>
<td>determinant of a matrix $A$</td>
</tr>
<tr>
<td>$A^T$</td>
<td>transpose of a matrix or a vector $A$</td>
</tr>
<tr>
<td>$A^H$</td>
<td>transpose and conjugate of a complex matrix or vector $A$</td>
</tr>
<tr>
<td>$\mathbf{1}$</td>
<td>identity matrix</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>diagonal matrix</td>
</tr>
<tr>
<td>$P$</td>
<td>permutated matrix or permutation</td>
</tr>
<tr>
<td>$\mathbf{1}_A$</td>
<td>identity function of a set $A$</td>
</tr>
<tr>
<td>$|A|$</td>
<td>diameter of a set $A$</td>
</tr>
<tr>
<td>$S_x(r)$</td>
<td>hypersphere with a center $x$ and a radius $r$</td>
</tr>
<tr>
<td>$*$</td>
<td>linear convolution operator</td>
</tr>
</tbody>
</table>
### NOTATIONS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{X}(f,t)$</td>
<td>discrete Fourier transform of signals $X(t)$</td>
</tr>
<tr>
<td>$\hat{W}_f$</td>
<td>discrete Fourier transform of a sequence ${W_k}_{k \in \mathbb{Z}}$</td>
</tr>
<tr>
<td>$(a_n) \approx (b_n)$</td>
<td>there exists $C &gt; 0$ such that $\lim_{n \to \infty} \frac{a_n}{b_n} = C$</td>
</tr>
<tr>
<td>$(a_n) \geq (b_n)$</td>
<td>there exists $C &gt; 0$ and $k \in \mathbb{N}$ such that $Ca_n \geq b_n$, $\forall n \geq k$</td>
</tr>
<tr>
<td>$\rho(x,y)$</td>
<td>Euclidean distance</td>
</tr>
<tr>
<td>$\eta(x,y)$</td>
<td>maximum distance</td>
</tr>
<tr>
<td>$\phi_r$</td>
<td>Gaussian density function with zeros mean and variance $r^2$</td>
</tr>
<tr>
<td>$\Phi_r$</td>
<td>Gaussian distribution function with zeros mean and variance $r^2$</td>
</tr>
<tr>
<td>$\Gamma(z)$</td>
<td>Gamma function</td>
</tr>
<tr>
<td>$\psi(z)$</td>
<td>digamma function</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Studying a way to express the dependency of a group of random variables is always an interesting work in probability theory and statistics. This is a very difficult problem since we have not had a comprehensive concept of dependency. Fortunately, there are a unique definition of independence in probability theory and various methods to check independence in an empirical setting. Thus, most mathematicians agree to express the dependency based on measuring of the deviation from the independence.

In the thesis, we would like to propose a measure, named the information dependency, as a way to express the dependency of a group of random variables. As the meaning of its name, the information dependency evaluate the dependency from the information concept which is studied carefully in Information Theory. This measure is defined as the Kullback-Leibler divergence of a joint distribution of these random variables with respect to their product-marginal distribution. In the bivariate case, this measure is known the mutual information of two random variables, however it is different from the multivariate mutual information when the number of random variable is larger than three.

We study the information dependency and its applications in the both theoretical and practical fields. Concretely, we find it interested in answering the following questions:

1. How does the information dependency express the dependency?

2. How do we estimate the information dependency from samples of a group of random variables?

3. How is the information dependency used to solve the Independent Component Analysis problem in the theoretical and practical fields?

Chapter 2, Measures of Dependency and Information Dependency, provide an answer to the first question. To do that, firstly we introduce some theoretical results about the
1. INTRODUCTION

Fréchet class, the Copula, and the Rényi axioms, which seem the basic criteria to express the dependency. In the situation that we do not have any exact definition of the dependency, many mathematicians and statisticians use the Rényi axioms [Rényi, 1959] as a criterion to evaluate the quality of the measures of dependency. We also agree with this idea, and use the Rényi axioms to compare the information dependency with other measures of dependency such as Pearson’s correlation, monotone correlation, maximal correlation, Spearman’s ρ, Kendall’s τ, Blomquist’s β, Gini’s γ, Höpfding’s Δ, and \(L_p\)-distances measures.

The definitions and the properties of the measures of dependency will be collected in the second section of this chapter. We recognize that several measures of dependency also express the dependency in various ways of the difference between the joint distribution and the product-marginal distribution of a group of random variables. The Rényi axioms are also checked for these measures. Considering the bivariate case, only two measures of dependency, the normalized information dependency and the maximal correlation, fulfill all the axioms of the Rényi axiom 1 (see in table 2.1); and only the normalized information dependency fulfills all the axioms of the Rényi axiom 2. Note that, the maximal correlation cannot be applied for the multivariate case. For this reason, we decide to use the normalized information dependency to evaluate the dependency of a group of random variables.

The definitions and the properties of the information dependency and the normalized information dependency are presented in the third section of Chapter 2. The content of this section is based on the well-known books [Cover and Thomas, 2006; Gray, 2010]. Some important notions in Information Theory such as the Shannon entropy, the differential entropy, the general entropy, and the Kullback-Leibler divergence are also introduced.

Chapter 3, Information Dependency Estimations, would like to answer the second question. Most of the non-parametric estimations of the density function, entropy and Kullback-Leibler divergence come from the three following classes of methods data-dependent partition estimations, \(k\)-nearest neighbor estimations, and Gaussian kernel estimations. Since the information dependency can be presented as the formulae of these above notions, it will be determined by various methods from all three classed. In the following this idea, we briefly describe how these estimators work.

The idea of the class of data-dependent partition estimations comes from the definition of the information dependency which contains a supremum over all finite and measurable partitions of the sample space. The problem is to find the best partition. It is clear that one partition for all samples such as the Bin-structure partition rule will not give accurate estimations. In 1970, Gessaman introduced the idea of data-dependent partition and
proposed the *Gessaman’s partition rule* [Gessaman, 1970]. Some other data-dependent partitions such as *Tree-quantization partition rule* and *Darbellay-Vajda partition rule* are introduced in [Breiman et al., 1984; Devroye et al., 1996] and [Darbellay and Vajda, 1999] respectively. In this thesis, we use these data-dependent partition rules to study the estimation of the information dependency.

The $k$-nearest neighbor estimations and the Gaussian kernel estimations use the results in the estimations of the densities and the differential entropies to apply for estimating the information dependency. Both estimators are based on the Euclidean distances between each pair of sample points. The $k$-nearest neighbor estimations uses the Euclidean distance from each sample point to its $k$-th closest point to estimate the density at this point (see in [Goria et al., 2005; Kozachenko and Leonenko, 1987; Loftsgaarden and Quesenberry, 1965]), while the Gaussian kernel uses the Euclidean distances from each sample point to all other sample points and the standard Gaussian function to estimate that one (see in [Silverman, 1986; Wand and Jones, 1995]). With the $k$-nearest neighbor estimations, there are two different ways to apply the results of the estimations of the density and the differential entropy to estimate the information dependency. The first way is proposed by Kraskov et al. [2004], and the second one is proposed by Póczos and Schneider [2011]. With the Gaussian kernel estimations, we are interested in the adaptive Gaussian kernel density estimation proposed by Botev et al. [2010]. Not convenient as the data-dependent partition estimations, the $k$-nearest neighbor estimations and the Gaussian kernel estimations can not work without some assumptions of each marginal random variables such as their density functions should be continuous and bounded.

All these information dependency estimations are evaluated in the section 3.2 based on seven bivariate examples and three multivariate examples. They are the *bivariate linear convolution of discrete distribution and Gaussian with parameter* $r = 0.1$ (see in equations (2.64) and (2.65), figure 2.2, table 2.3), the *bivariate linear convolution of diagonal uniform distribution and Gaussian with parameters* $r = 0.2$ and $r = 0.5$ (see in equations (2.71) and (2.72), figure 2.4, table 2.5), the *bivariate linear convolution of circular uniform distribution and Gaussian with parameters* $r = 0.2$ and $r = 0.5$ (see in equations (2.75) and (2.76), figure 2.5, table 2.6), the *bivariate Gamma-Exponential distribution with parameters* $(\theta_1, \theta_2, \theta_3) = (1, 0.5, 1)$ and $(\theta_1, \theta_2, \theta_3) = (1, 1, 1)$ (see in equations (3.33) and (3.34), figure 3.10), the *seven-dependent variate Gaussian distribution* (see in equations (3.37) and (3.38)), the *four-variate Exponential distribution with parameter* $(2, 2, 2, 2)$ (see in equations (3.43) and (3.44)), and the *five-variate Logistic distribution with the mean parameter* $(1, 1, 2, 2, 4)$ and the variance parameter $(2, 2, 6, 6, 9)$ (see in equations (3.40) and (3.41)). Fortunately, we can determine or approximate the information dependency values of these examples, hence we can give the evaluation of these estimations. They are
simulated from figure 3.5 to figure 3.14 and from table 3.1 to table 3.10 respectively. From these simulations, we realize that the Darbellay-Vajda partition estimator is the best estimator in the low-dimensional cases (i.e. the number of random variables is smaller than 3), and the Kraskov’s $k$-nearest neighbor estimator is the best one in the high-dimensional cases (i.e. the number of random variables is larger than 4).

Estimation of information dependency is a difficult problem. Even the simplest statistical properties could not be verified for many estimators. It is not known whether these estimators are unbiased for fixed simple size $n$. For the Darbellay-Vajda partition estimator, it is even not known whether it is consistent, that is asymptotically unbiased for $n \to \infty$. The consistency of the Gaussian kernel estimator, the Kraskov’s and Póczos’s $k$-nearest neighbor estimations have not been solved yet.

In the next section of Chapter 3, we would like to study the strong consistency of the class of data-dependent partition estimators. Specially the conditions of parameters of Darbellay-Vajda partition estimator which confirm that this estimator is strongly consistent are given out. The data-dependent partition estimations work as follows: Let $X = (X_1, \ldots, X_m)$ be a group of $m$ random variables with a joint distribution $\mu$, and a product-marginal distribution $\nu$. The information dependency of the random vector $X$ is defined as follows:

$$ I(X) = \sup_{\pi \in P} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)}, $$

where $P$ is a collection of all finite and measurable partitions of the space. Given any samples $X^n = \{X_1, \ldots, X_n\}$ of a random vector $X$, the data-dependent partition estimations try to find an optimal partition $\pi_n(X^n)$ such that

$$ \hat{J}(X^n) = \sum_{A \in \pi_n(X^n)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \text{ is a strongly consistent estimator for } J(X), $$

where $\mu_n$, $\nu_n$ are the empirical joint and product-marginal distributions generated by the samples. The strong consistency of the Gessaman’s partition estimation and the Tree-quantization partition estimation are proved by Silva and Narayan [2010a,b]. The skeleton of their proofs relate to the simultaneous approximation concept (see in subsection 3.3.1) and the Vapnik-Chervonenkis theory (see in subsection 3.3.2). The simultaneous approximation concept helps us to find out the conditions of the partition $\pi_n(X^n)$ such that

$$ \sum_{A \in \pi_n(X^n)} \mu(A) \frac{\mu(A)}{\nu(A)} \to J(X) = \sup_{\pi \in P} \sum_{A \in \pi} \mu(A) \frac{\mu(A)}{\nu(A)} \quad \mathbb{P}\text{-a.s.,} $$
where $\mu$, $\nu$ are the joint distribution and the product-marginal distribution of $X$; and the Vapnik-Chervonenkis theory helps us to find the conditions such that

$$\sum_{A \in \pi_n(X^*_1)} \mu_n(A) \mu(A) \rightarrow \sum_{A \in \pi_n(X^*_1)} \mu(A) \mu(A) \nu(A) \quad \text{P-a.s.}$$

The strong consistency of the Darbellay-Vajda partition estimation is proved in the subsection 3.3.3. This is the main result of our thesis. The interest of this result is that it replaces some conditions in the cases of the Gessman’s partition and the Tree-quantization partition estimations which are very difficult to check by some simpler conditions. These conditions are given in Theorem 3.6 and Proposition 3.3.

Chapter 4, Independent Component Analysis, would like to study the third question. Since 1990, the Independent Component Analysis (ICA) [Comon and Jutten, 2010; Hyvärinen and Oja, 2001] has gained a lot of attention of statisticians because it has many applications in practice. This optimization problem tries to find out an optimal linear transformation such that it transforms a random vector to a new random vector of which its components are as independent as possible. Hence, the quality of the measure of dependency is very important to this problem. Comon [1994] showed that the information dependency was a good tool to solve the Independent Component Analysis. Moreover, in [Kien, 2011], we proved that the simplest case of ICA is perfectly solved by the information dependency. The simplest case of ICA is stated shortly as follows: Assuming that $Y = (Y_1, Y_2)$, a two-dimensional random vector, is an unknown linear transformation of another two-dimensional random vector $(X, N)$, where $X$ is a non-Gaussian, $N$ is a Gaussian, and $X, N$ are independent. The simplest case of ICA would like to recover the unknown random vector $(X, N)$ based on the known-random vector $Y$. Kien [2011] proved that with the information dependency tool, the simplest case of ICA is solvable, i.e. the random vector $(X, N)$ is the unique local solution of these optimal problem. This result will be presented carefully in section 4.2.

To see how the information dependency works in the practical fields, we focus to study the signification of the Darbellay-Vajda information dependency estimations in the simplest case of ICA problem and the blind audio sources separation problem. We present three examples of the simplest case of ICA and use the Darbellay-Vajda partition rule to estimation all possible linear transformations of a random vector $Y$ for each example. The results, given in figure 4.2, show that this estimation can exhibit the meaning of the result in [Kien, 2011], i.e. the simplest case of ICA has a unique local solution. The blind audio sources separation is one of the popular applications of ICA. Using the ICA model, it would like to recover the original audio sources from the mixture audio
signals which are the recording of these original audio sources together (see Chapter 19 in [Comon and Jutten, 2010]). Clearly, we need to assume that the original audio sources are nearly mutual independent and the mixture audio signals are very dependent. Thus, we measure the Darbellay-Vajda information dependency estimations for some original speech sources and mixture speech signals. It shows that all of these estimations measuring for the original sources are nearly zero and that ones measuring for mixture signals are very high. From these numerical experiments, we believe that the ICA model and the Darbellay-Vajda partition method can solve the *blind speech sources separation*. This will be worked out in the future.
Chapter 2

Measures of Dependency and Information Dependency

2.1 Measures of Dependency

Dependency of a group of random variables is one of the most widely studied topics in probability theory and statistics. It has attracted scientists' attention for over a century since independence in the data is often considered a desired property. Pearson [1904] and Spearman [1904] seem to be the first scientists who give the ways to measure the dependency in the statistical methodology. However, the concepts of dependency really appeared around two hundred and fifty years ago. In 1763, T. Bayes introduced the notion of independence as follows (see in [Mari and Kots, 2001, Ch. 2, p. 7])

"Events are independent when the happening of any one of them does neither increase nor abate the probability of the rest."

In 1767, R. Price, a close friend of T. Bayes, stated in a footnote of his volume *Four Dissertations* a similar definition (see in [Mari and Kots, 2001, Ch. 2, p. 7])

"Two events are independent when the happening of one of them has no influence on the other."

Possibly, the earliest mathematical description of independence was presented by P.S. Laplace in his book *Théorie Analytiques des Probabilités* (1812). He stated that the sequence of simple events \( \{ A_i \}_{i=1}^{m} \) was independent if the probability of all events’ occurrence equaled the product of the probability of each event’s occurrence, i.e. \( P(A_1, \ldots, A_m) = \prod_{i=1}^{m} P(A_i) \). After more than one hundred years, the notion of independence had been extended for random variables, and accepted by most scientists. The extension is presented in the following definition. Let \( X_1, \ldots, X_m \) be random variables on a probability space \( (\Omega, \mathcal{B}, P) \).
Definition 2.1. \( X_1, \ldots, X_m \) are called independent random variables if \( \forall A_i \in \mathcal{B}, i = 1, \ldots, m, \)

\[
\mathbb{P}(X_1 \in A_1, \ldots, X_m \in A_m) = \prod_{i=1}^{m} \mathbb{P}(X_i \in A_i).
\] (2.1)

Let \( \mu \) be a joint distribution of \( (X_1, \ldots, X_m) \), \( \mu_i \) be marginal distributions of \( X_i, i = 1, \ldots, m, \) and \( F \) be a joint cumulative distribution function (cdf) of \( (X_1, \ldots, X_m) \), \( F_i \) be marginal cumulative distribution functions of \( X_i, i = 1, \ldots, m. \) Then the independence of \( X_1, \ldots, X_m \) can be presented as follows:

\[
\mu(A_1 \times \cdots \times A_m) = \prod_{i=1}^{m} \mu_i(A_i), \quad \forall A_1, \ldots, A_m \in \mathcal{B},
\] (2.2)

or

\[
F(x_1, \ldots, x_m) = \prod_{i=1}^{m} F_i(x_i), \quad \forall x_1, \ldots, x_m \in \mathbb{R}.
\] (2.3)

Nowadays, the concept of independence and dependency pervades our life. It is obviously not only a mathematical concept but also a concept of nature life.

Let a function \( G(x_1, \ldots, x_m) = \prod_{i=1}^{m} F_i(x_i), \forall x_1, \ldots, x_m \in \mathbb{R} \) be a product of marginal (product-marginal) cumulative distribution function of \( (X_1, \ldots, X_m) \). The formula (2.3) shows us that the independence, a special case of dependency, occurs if and only if the joint cdf equals to the product-marginal cdf. Generating this idea for the other cases of dependency, most mathematicians and statisticians agree that

\textit{The dependency of a group of random variables can be expressed by the difference between its joint cdf} \( F \) \textit{and its product-marginal cdf} \( G \).

Thus, most well-known measures of dependency such as Spearman’s \( \rho \), Kendall’s \( \tau \), Hoeffding’s measure, \( L_p \)-distances measure and information dependency are the presentations of the difference between a joint cdf \( F \) and a product-marginal cdf \( G \).

Studying the multivariate cumulative distribution functions is very difficult, especially in the high dimensional cases. For example, giving a function or a class of functions, in some cases, we find it impossible to check whether they are cumulative distribution functions or not. It is said that the history of studying the structure of multivariate cdfs was started by Fréchet. In 1951, he studied the class of cdfs \( \Gamma(F_1, F_2) \) named \textit{Fréchet class} of \( F_1 \) and \( F_2 \), which was described as follows: Given the univariate cdfs \( F_1 \) and \( F_2 \), the class \( \Gamma(F_1, F_2) \) is a collection of all bivariate cdfs of which marginal cdfs are \( F_1 \) and \( F_2 \). It is clear that the class \( \Gamma(F_1, F_2) \) always contains the bivariate cdf \( G(x_1, x_2) = F_1(x_1)F_2(x_2), \forall x_1, x_2 \in \mathbb{R} \); however, it is impossible to describe all bivariate cdfs of \( \Gamma(F_1, F_2) \). The Fréchet class can be generated for \( m \)-univariate cdfs \( F_1, \ldots, F_m \),
and denoted by $\Gamma(F_1, \ldots, F_m)$. When $F_1, \ldots, F_m$ are uniform distributions on $[0, 1]$, the class $\Gamma(F_1, \ldots, F_m)$ is known an $m$-dimensional Copula. Since 1990, Fréchet class and Copulas had been paid lots of attention by mathematician and statistician. Thus, there have been six big international conferences devoted to research the above fields such as the “Symposium on Distributions with Given Marginals (Fréchet Classes)” in Rome in 1990, the conference on “Distributions with Fixed Marginals, Doubly Stochastic Measures, and Markov Operators” in Seattle in 1993, the conference on “Distributions with Given Marginals and Moment Problems” in Prague in 1996, the conference on “Distributions with Given Marginals and Statistical Modelling” in Barcelona in 2000, the conference on “Dependence Modelling: Statistical Theory and Applications in Finance and Insurance” in Québec in 2004 (see in [Nelsen, 2006]), and the conference on “Copula Theory and Its Applications” in Warsaw in 2009 [Jaworski et al., 2010]. With the purpose of answering the question “how Fréchet class and Copulas have significance in probability theory and statistics”, scientists proved that they were tools of studying scale-free measures of dependency.

2.1.1 Copulas

In this subsection, we give a short introduction of copulas and its relationship with the dependency.

**Definition 2.2** (Jaworski et al. [2010]). For every $m \geq 2$, an $m$-dimensional copula is an $m$-variate distribution function on $\mathbb{I}^m$ of which univariate marginals are uniformly distributed on $\mathbb{I}$.

The class of all $m$-dimensional copulas will be denoted by $\mathcal{C}_m$. Since copulas are multivariate cdfs, as a consequence of the properties of multivariate cdf, they can be determined in the following equivalent way.

**Theorem 2.1.** A function $C : \mathbb{I}^m \to \mathbb{I}$ is a copula if and only if the following properties hold:

i. for every $i \in \{1, \ldots, m\}$, $C(u) = u_i$ when all the components of $u$ are equal to 1 with the exception of the $i$-th one that is equal to $u_i \in \mathbb{I}$,

ii. $C$ is isotonic, i.e. $C(u) \leq C(v)$ for all $u = (u_1, \ldots, u_m), v = (v_1, \ldots, v_m) \in \mathbb{R}^m$, $u_i \leq v_i \forall i = 1, \ldots, m$,

iii. $C$ is $m$-increasing.

The relationship between Fréchet class and copulas is reflected in a following theorem [Schweizer and Sklar, 1974; Sklar, 1959].
Theorem 2.2 (Sklar’s Theorem). Let $F$ be an $m$-dimensional cdf with univariate margins $F_1, \ldots, F_m$. Let $A_i$ denote the range of $F_i$, $A_i = F_i(\mathbb{R})$, $i = 1, \ldots, m$. Then there exists an $m$-dimensional copula $C$ such that for all $(x_1, \ldots, x_m) \in \mathbb{R}^m$, 

$$F(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m)).$$ \hfill (2.4)

Such a copula $C$ is uniquely determined on $A_1 \times \cdots \times A_m$ and, hence, it is unique when $F_1, \ldots, F_m$ are all continuous. Conversely, if $C$ is a copula and $F_1, \ldots, F_m$ are univariate margins, the function $F$ defined by (2.4) is a joint cdf with margins $F_1, \ldots, F_m$.

The proofs of Sklar’s Theorem for the bivariate case appeared in [Carley and Taylor, 2002; Schweizer and Sklar, 1974], and for the multivariate cases appeared in [Rüschendorf, 2009]. When univariate margins $F_1, \ldots, F_m$ are continuous, the relationship between joint cdf $F$ and copula $C$ is clear based on the transformations $T = (F_1^{-1}, \ldots, F_m^{-1})$. When $F_1, \ldots, F_m$ are discontinuous, Ferguson [1967] introduced a distribution transform notion: Let $X$ be a random variable with a cumulative distribution function $F$, and let $F(x, \lambda) = \mathbb{P}(X < x) + \lambda \mathbb{P}(X = x)$ for any real number $\lambda \in (0, 1)$. Let $V$ be uniformly distributed on $I$ and independent of $X$, the distributional transform of $X$ is defined by

$$U = F(X, V) = F(X) + V[F(X) - F(X^-)].$$ \hfill (2.5)

The distributional transform $U$ has interesting properties.

Lemma 2.1 (Rüschendorf [2009]). Let $U$ be the distributional transform of $X$, then $U$ is uniformly distributed on $I$ and

$$X = F^{-1}(U) \quad \text{almost surely},$$ \hfill (2.6)

where $F^{-1}$ denoted the pseudo-inverse of $F$ given by $F^{-1}(u) = \inf\{t : F(t) \geq u\}$.

Proof. For any $0 < \alpha < 1$, let $q_{\alpha}(X) = \sup\{x : \mathbb{P}(X \leq x) < \alpha\}$ the lower $\alpha$-quantile of $X$. We know $F(X, V) \leq \alpha$ if and only if $(X, V) \in \{(x, \lambda) : \mathbb{P}(X < x) + \lambda \mathbb{P}(X = x) \leq \alpha\}$. If $\beta = \mathbb{P}(X = q_{\alpha}(X)) > 0$, with $q = \mathbb{P}(X < q_{\alpha}(X))$, we have

$$\mathbb{P}(U \leq \alpha) = \mathbb{P}(F(X, V) \leq \alpha) = q + \mathbb{P}(X = q_{\alpha}(X), q + V \beta \leq \alpha)$$

$$= q + \beta \mathbb{P}\left(V \leq \frac{\alpha - q}{\beta}\right) = q + \beta \frac{\alpha - q}{\beta} = \alpha.$$

If $\beta = 0$, $\mathbb{P}(U \leq \alpha) = \mathbb{P}(X \leq q_{\alpha}(X)) = \alpha$. Thus, $U$ is uniformly distributed on $I$.

By definition of $U$, $F(X^-) \leq U \leq F(X)$. It infers that $F^{-1}(U) = X$ almost surely since $\forall u \in (F(x^-), F(x)], F^{-1}(u) = x$. \qed
2.1. MEASURES OF DEPENDENCY

**Proof of Sklar’s Theorem [Rüschendorf, 2009].** Let $X_1, \ldots, X_m$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the joint cdf $F$ and the marginal cdfs $F_1, \ldots, F_m$. Let $V$ be independent of $X = (X_1, \ldots, X_m)$ and uniformly distributed on $[0,1]$. Considering the distributional transforms $U_i = F_i(X_i, V)$, by Lemma 2.1, $U$ will be uniformly distributed on $[0,1]$ and $X_i = F_i^{-1}(U_i)$ almost surely, $i = 1, \ldots, m$. Thus we define $C$ the joint cdf of $U = (U_1, \ldots, U_m)$, and have

$$F(x_1, \ldots, x_m) = \mathbb{P}[F_1^{-1}(U_i) \leq x_i, i = 1, \ldots, m] = \mathbb{P}[U_i \leq F_i(x_i), i = 1, \ldots, m] = C(F_1(x_1), \ldots, F_m(x_m)),$$

i.e. $C$ is a copula of $F$.

The converse is a matter of straightforward verification of joint cumulative distribution function. □

Sklar’s Theorem allowed us to define a copula of a random vector as follows: Let $X = (X_1, \ldots, X_m)$ be random vector with joint cdf $F$ and marginal cdfs $F_1, \ldots, F_m$. The new random vector

$$C(X_1, \ldots, X_m) = F(F_1^{-1}(X_1), \ldots, F_m^{-1}(X_m)),$$

(2.7)

where $F_i^{-1}$ denote the pseudo-inverse of $F_i$ given by $F_i^{-1}(v) = \inf\{t : F_i(t) \geq v\}$, is called a copula of $X$.

The next four following propositions show the significance of copulas in the dependency. These propositions are described and proven carefully in [Jaworski et al., 2010; Joe, 1997; Mari and Kots, 2001; Nelsen, 2006].

**Proposition 2.1.** Let $X = (X_1, \ldots, X_m)$ be a random vector with a continuous joint cdf $F$ and a copula $C$. Let $f_1, \ldots, f_m$ be strictly increasing functions from $\mathbb{R}$ to $\mathbb{R}$. Then $C$ is also the copula of the random vector $(f_1(X_1), \ldots, f_m(X_m))$.

**Proposition 2.2.** Let $(X_1, \ldots, X_m)$ be a random vector with a continuous joint cdf $F$. Then the copula of $(X_1, \ldots, X_m)$ is $\Pi_m$ if and only if $X_1, \ldots, X_m$ are independent, where $\Pi_m(u_1, \ldots, u_m) = u_1 \cdots u_m$, $\forall (u_1, \ldots, u_m) \in \mathbb{I}^m$.

**Proposition 2.3.** Let $(X_1, \ldots, X_m)$ be a random vector with a continuous joint cdf $F$. The the copula of $(X_1, \ldots, X_m)$ is $M_m$ if and only if there exists a random variable $Y$ and increasing functions $f_1, \ldots, f_m$ such that $X_i = f_i(Y)$ almost surely $i = 1, \ldots, m$, where $M_m(u_1, \ldots, u_m) = \min\{u_1, \ldots, u_m\}$, $\forall (u_1, \ldots, u_m) \in \mathbb{I}^m$. $M_m$ is called the upper Fréchet-Hoeffding bound of $m$ dimensional copulas.

**Proposition 2.4.** Let $(X_1, X_2)$ be a random vector with continuous joint cdf $F$. Then $(X_1, X_2)$ has the copula $W_2$ if and only if for some strictly decreasing function $f$, $X_2 = $
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

\[ f(X_1) \text{ almost surely, where } W_2(u_1, u_2) = \max\{u_1 + u_2 - 1, 0\}, \forall (u_1, u_2) \in \mathbb{I}^2. \]  
\[ W_2 \text{ is called the lower Fréchet-Hoeffding bound of 2 dimensional copulas.} \]

If the dependency of a group of random variables is invariant under increasing and continuous functions of each variable, Proposition 2.1 describes that the copulas have the same property. Proposition 2.2 shows that the product copula \( \Pi_m \) characterizes independent random variables when the distribution functions are continuous while copulas \( M_m \) and \( W_2 \) characterizing strictly dependency are showed in next two propositions. Note that Proposition 2.4 cannot be extended to dimension \( m \) with \( m \geq 3 \) because it is impossible for each one of the \( m \) random variables \( X_1, \ldots, X_m \) to be an almost surely decreasing function of the remaining ones.

2.1.2 Rényi Axioms

In most fields of application of statistics, we often encounter the problem of estimating the strength of dependency of \( X_1, \ldots, X_m \) by a numerical value, which is denoted \( \delta(X_1, \ldots, X_m) \). This value is non-negative, equaling zero when \( X_1, \ldots, X_m \) are mutual independent, and the bigger value it receives, the more dependent \( X_1, \ldots, X_m \) are. Moreover, since dependency is a stochastic quantity, it should be invariant under deterministic transformations. In this spirit, Rényi [1959] proposed a set of seven axioms for measures of dependency of two random variables.

**Rényi Axiom 1.** A measure \( \delta \) is called a measure of dependency of two random variables if it satisfies seven following conditions

(A) \( \delta(X, Y) \) is defined for any pair of random variables \( (X, Y) \), neither of them being constant with probability 1.

(B) \( \delta(X, Y) = \delta(Y, X) \).

(C) \( 0 \leq \delta(X, Y) \leq 1 \).

(D) \( \delta(X, Y) = 0 \) if and only if \( X \) and \( Y \) are independent.

(E) \( \delta(X, Y) = 1 \) if and only if there is a strict dependence between \( X \) and \( Y \), i.e. either \( X = g(Y) \) or \( Y = f(X) \) almost surely, where \( f, g \) are Borel-measurable functions.

(F) For every Borel-measurable injective functions \( f \) and \( g \) on \( \mathbb{R} \), \( \delta(f(X), g(Y)) = \delta(X, Y) \).

(G) If the joint distribution of \( X \) and \( Y \) is normal, then \( \delta(X, Y) \) is a strictly increasing function of \( |\rho(X, Y)| \), where \( \rho(X, Y) \) the correlation coefficient of \( X \) and \( Y \).
The generalization of Rényi axiom for \( m \) random variables is introduced by Micheas and Zografos [2006].

**Rényi Axiom 2.** A measure \( \delta \) is called a measure of dependency of \( m \) random variables if it satisfies seven following conditions

(A) \( \delta(X_1, \ldots, X_m) \) is defined for any \( m \) random variables \( X_1, \ldots, X_m \), neither of them being constant with probability 1.

(B) For any permutation \( \sigma = (i_1, \ldots, i_m) \) of the indices \( (1, \ldots, m) \), we always have \( \delta(X_{i_1}, \ldots, X_{i_m}) = \delta(X_1, \ldots, X_m) \).

(C) \( 0 \leq \delta(X_1, \ldots, X_m) \leq 1 \).

(D) \( \delta(X_1, \ldots, X_m) = 0 \) if and only if \( X_1, \ldots, X_m \) are mutual independent.

(E) \( \delta(X_1, \ldots, X_m) = 1 \) if and only if there exist \( m \) Borel-measurable functions \( f_1, \ldots, f_m \) such that \( X_i = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m) \) almost surely, \( i = 1, \ldots, m \).

(F) For every Borel-measurable injective transformation \( T : \mathbb{R}^m \to \mathbb{R}^m \), \( \delta(T(X_1, \ldots, X_m)) = \delta(X_1, \ldots, X_m) \).

(G) If the joint distribution of \( X \) and \( Y \) is normal, then \( \delta(X, Y) \) is a strictly increasing function of \( |\rho(X, Y)| \), where \( \rho(X, Y) \) the correlation coefficient of \( X \) and \( Y \).

In the situation that there isn’t an exact way to express the dependency of groups of random variables, and there are a lot of proposed measures of dependency, Rényi axiom can work as a basic criterion to decide a measure of dependency is good or bad. Some researchers Bell [1962]; Kimberdof and Sampson [1978]; Schmid et al. [2010]; Schweizer and Wolff [1981] have discussed that the axiom (F) of Rényi axioms is an austere condition, sometimes unnecessary in practicality. They suggested to change this condition to a simpler condition

\( (F') \) For every strictly monotone and continuous functions \( f_1, \ldots, f_m \), \( \delta(f_1(X_1), \ldots, f_m(X_m)) = \delta(X_1, \ldots, X_m) \).

For the continuous case, the axiom (F’) is a sufficient condition to show that copulas can replace Fréchet class to study the measures of dependency. Thus, for any group of random variables \( X_1, \ldots, X_m \) with a continuous joint cdf \( F \) and continuous marginal cdfs \( F_1, \ldots, F_m \), we know that \( C(X_1, \ldots, X_m) = F(F_1^{-1}(X_1), \ldots, F_m^{-1}(X_m)) \), where \( C \) is a copula of \( (X_1, \ldots, X_m) \). Since \( F_1^{-1}, \ldots, F_m^{-1} \) are strictly increasing and continuous functions, axiom (F’) shows that \( \delta(X_1, \ldots, X_m) = \delta(F_1^{-1}(X_1), \ldots, F_m^{-1}(X_m)) \). Hence, we can use \( C \), the joint cdf of \( (F_1^{-1}(X_1), \ldots, F_m^{-1}(X_m)) \), to measure the dependency of group of random variables \( (X_1, \ldots, X_m) \).
Schweizer and Wolff [1981] also suggested to add a continuity axiom in Rényi axioms to constructed the measures of dependency. The continuity axiom is stated as follows:

\((H)\) Let \((X_1, \ldots, X_m)\) and \(\{(X_1^{(n)}, \ldots, X_m^{(n)})\}_{n \in \mathbb{N}}\) be a group and a sequence of groups of random variables with a joint cdf \(F\) and joint cdfs \(\{F_n\}_{n \in \mathbb{N}}\) respectively. If the sequence \(\{F_n\}\) converges weakly to \(F\), then \(\lim_{n \to \infty} \delta(X_1^{(n)}, \ldots, X_m^{(n)}) = \delta(X_1, \ldots, X_m)\).

In our works, we are interested in the original Rényi axioms with seven axioms, and based on these conditions to evaluate some well-known measures of dependency. Some popular measures of dependency will be introduced and investigated in the next subsections. We divide these measures into two classes named bivariate dependence measures and multivariate dependence measures because the multivariate case cannot be considered as an extension of bivariate case. For example, a concondant property, which is exploited by some measures such as Spearman’s \(\rho\), Kendall’s \(\tau\), Blomqvist’s \(\beta\), etc., is very difficult to extend from the bivariate case to the multivariate case.

### 2.1.3 Bivariate measures of dependency

In this subsection, we would like to introduce the well-known bivariate measures of dependency such as Pearson’s correlation, monotone correlation, maximal correlation, Spearman’s \(\rho\), Kendall’s \(\tau\), Blomqvist’s \(\beta\), Gini’s \(\gamma\), Hölding’s \(\Delta\), \(L_p\)-distance measure and Information-based measures, and study how they satisfy the conditions in Rényi axiom 1.

**Pearson’s correlation:** The most familiar measure of dependency between two random variables is the Pearson’s correlation. Pearson [1904] defined the correlation by dividing the covariance of \(X\) and \(Y\) by the product of their standard deviations.

\[
\rho_P(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)} = \frac{\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]}{\sigma(X)\sigma(Y)} = \frac{1}{\sigma(X)\sigma(Y)} \int_{\mathbb{R}^2} [F(x, y) - F_1(x)F_2(y)] dxdy. \tag{2.8}
\]

Pearson’s correlation is defined only if the standard deviations of \(X\) and \(Y\) are finite and nonzero. This quantity is symmetric, i.e. \(\rho_P(X, Y) = \rho_P(Y, X)\). Pearson’s correlation value is smaller than 1 and bigger than -1. It receives value 1 when \(X, Y\) are strictly positive linear, and -1 when \(X, Y\) are strictly negative linear. If \(X\) and \(Y\) are independent, Pearson’s correlation is 0, but the converse is not true. However, in a special case when \(X\) and \(Y\) are jointly normal distributions, the modulus of Pearson’s correlation is a perfect measure of dependency.
Monotone correlation: Kimeldorf and Sampson [1978] extended the Pearson’s correlation by a new measure named *monotone correlation* given following:

\[
\rho^+(X, Y) = \sup \rho_P[f(X), g(Y)],
\]

(2.9)

where the supremum is taken over all monotone functions \(f, g\) with \(0 < \sigma(f(X)), \sigma(g(X)) < \infty\).

Maximal correlation: And Rényi [1959] extended the Pearson’s correlation by another measure named *maximal correlation* given following:

\[
\rho^*(X, Y) = \sup \rho_P[f(X), g(Y)],
\]

(2.10)

where the supremum is taken over all Borel-measurable functions \(f, g\) for which \(0 < \sigma(f(X)), \sigma(g(X)) < \infty\).

*Monotone correlation* and *maximal correlation* are defined when the standard deviations of \(X\) and \(Y\) are nonzero. The maximal correlation \(\rho^*\) is a generalization of Pearson’s correlation which is sufficient to confirm that this measure satisfies axiom (F) in Rényi axiom 1. Obviously, \(\rho^*\) has axioms (A), (B), (C) and (E). For axiom (D), if \(X\) and \(Y\) are independent, \(f(X)\) and \(g(Y)\) are independent for all Borel-measurable functions \(f, g\). Contrarily, \(\forall u, v \in \mathbb{R}\), we choose Borel-measurable functions \(f, g\) with \(f(x) = 1\) if \(x < u\) and 0 if \(x \geq u\), and \(g(y) = 1\) if \(y < v\) and 0 if \(y \geq v\). Since \(\rho^*(X, Y) = 0\), we have

\[
\rho^*(f(X), g(Y)) = 0 \Rightarrow \mathbb{P}(X < u, Y < v) = \mathbb{P}(X < u)\mathbb{P}(Y < v).
\]

It infers that \(X\) and \(Y\) are independent. The satisfaction of axiom (G) of \(\rho^*(X, Y)\) has been shown in [Gebelein, 1941]. Thus, maximal correlation is one of a few of measures of dependency which have all axioms.

The monotone correlation \(\rho^+\) is a generalization of Pearson’s correlation which is sufficient to make this measure satisfies the axiom (F’). Being similar to the maximal correlation, the monotone correlation also has all axioms (A), (B), (C), (D), (E) and (G). The monotone correlation is interested in due to a *monotone dependent* notion. Random variables \(X\) and \(Y\) are called *monotone dependent* if there exists a monotone function \(f\) such that \(\mathbb{P}(Y = f(X)) = 1\). We know that, if a sequence \(\{(X_n, Y_n)\}\) of pairs of independent random variables converges in distribution to a pair \((X, Y)\) of random variables, \(X\) and \(Y\) must be independent. However, this judgment is not true if we replace the hypothesis “independent” with “strictly dependent”. Kimeldorf and Sampson [1978] proved if a sequence \(\{(X_n, Y_n)\}\) of pairs of “monotone dependent” random variables converges in distribution to a pair \((X, Y)\) of random variables, \(X\) and \(Y\) are monotone
dependent.

Although the monotone and maximal correlations satisfy nearly all axioms in Rényi axiom 1, they are not much significant in practice. It is because we cannot determine these quantities without knowledge of the exact joint cdf $F$ and univariate margins $F_1, F_2$. However, in practice, the joint and marginal cdfs are often unknown.

**Concordance:** There are many bivariate measures of dependency such as Spearman’s $\rho$, Kendall’s $\tau$, Blomquist’s $\beta$, Gini’s $\gamma$ and Höfling’s $\Delta$, which are reflected as measures of concordance since the concordant property is a simple condition to recognize the dependency of two random variables. Informally, random variables $X$ and $Y$ are concordant when large values of $X$ go with large values of $Y$. More precisely, let $(x_i, y_i)$ and $(x_j, y_j)$ be observations of $(X, Y)$, we say that $(x_i, y_i)$ and $(x_j, y_j)$ are concordant if $(x_i - x_j)(y_i - y_j) \geq 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$. When $(x_i, y_i)$ and $(x_j, y_j)$ are concordant, we say that $(X, Y)$ is concordance, and when $(x_i, y_i)$ and $(x_j, y_j)$ are discordant, it is not $(X, Y)$ which is in concordance [Joe, 1997; Nelsen, 2006; Scarsini, 1984].

**Spearman’s $\rho$** Spearman’s $\rho$ firstly studied by Spearman [1904] is also one of the best well-known measures of dependency between two random variables. If we denote $F_1(X)$ and $F_2(Y)$ ranking order variables of $X$ and $Y$, respectively, then the Spearman’s $\rho$ of $X$ and $Y$, denoted $\rho_S(X, Y)$, is the Pearson’s $\rho$ of their ranking order variables,

$$
\rho_S(X, Y) = \rho_P(F_1(X), F_2(Y)) = \frac{\text{cov}(F_1(X), F_2(Y))}{\sigma(F_1(X))\sigma(F_2(Y))}.
$$

(2.11)

Note that, for any random variable $X$, if $F(X)$ is a ranking order variable of $X$, it is also a cdf of $X$. Nelsen [2006] (Ch. 5, p. 167) gave another definition of Spearman’s $\rho$ based on the measure of concordance as follows: Let $(X_1, Y_1), (X_2, Y_2)$ and $(X_3, Y_3)$ be three independent and identity distributions of $(X, Y)$, then

$$
\rho_S(X, Y) = 3\left[\mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) \geq 0] - \mathbb{P}[(X_1 - X_2)(Y_1 - Y_3) < 0]\right].
$$

(2.12)

The proof of the following result should infer that the definitions (2.11) and (2.12) are similar.

**Proposition 2.5.** Let $X, Y$ be continuous random variables with a joint cdf $F$ and marginal cdfs $F_1, F_2$ respectively. Denote $C$ is a copula of $(X, Y)$. Then

$$
\rho_S(X, Y) = 12 \int_{\mathbb{R}^2} \left[ F(x, y) - F_1(x)F_2(y) \right] dF_1(x)dF_2(y),
$$

$$
= 12 \int_{\mathbb{R}^2} [C(u, v) - uv] dudv.
$$

(2.13)
2.1. MEASURES OF DEPENDENCY

Proof. (2.11) infers (2.13), since $\mathbb{E}[F_1(X)] = \frac{1}{2}$, $\sigma^2[F_1(X)] = \frac{1}{4}$, and

$$\text{cov}[F_1(X), F_2(Y)] = \mathbb{E}[(F_1(X) - 1/2)(F_2(Y) - 1/2)] = \mathbb{E}[F_1(X)F_2(Y)] - 1/4$$

$$= \int_{\mathbb{R}^2} [F(x, y) - F_1(x)F_2(y)] dF_1(x)dF_2(y).$$

To prove (2.12) inferring (2.13), we see that

$$\mathbb{P}(X_1 > X_2, Y_1 > Y_3) = \int_{\mathbb{R}^2} \mathbb{P}(X_2 < x, Y_3 < y)dF(x, y)$$

$$= \int_{\mathbb{R}^2} \mathbb{P}(X_2 < x)\mathbb{P}(Y_3 < y)dF(x, y) \quad (X_2, Y_3 \text{ independent}),$$

$$= \int_{\mathbb{R}^2} F_1(x)F_2(y)dF(x, y).$$

Doing similarly with other quantities, we have

$$\rho_S(X, Y) = 3 \int_{\mathbb{R}^2} [4F_1(x)F_2(y) + 1 - 2F_1(x) - 2F_2(y)]dF(x, y)$$

$$= 12 \int_{\mathbb{R}^2} F_1(x)F_2(y)dF(x, y) - 3$$

$$= 12 \int_{\mathbb{R}^2} [F(x, y) - F_1(x)F_2(y)]dF_1(x)dF_2(y).$$

The formula of copula is a consequence of the relationship between the copula and the distribution function.

Kendall’s $\tau$: Let $(X, Y)$ be a pair of random variables. Let $(X_1, Y_1)$ and $(X_2, Y_2)$ be independent and identically distributed with $(X, Y)$. Kendall’s $\tau$ measure of $(X, Y)$ is defined as the probability of the concordance minus the probability of the discordance of $(X_1, Y_1)$ and $(X_2, Y_2)$ (see in [Kendall, 1938]).

$$\tau(X, Y) = \mathbb{P}[(X_1 - Y_1)(X_2 - Y_2) \geq 0] - \mathbb{P}[(X_1 - Y_1)(X_2 - Y_2) < 0]. \quad (2.14)$$

The formulae as functions of the distribution functions or the copula of Kendall’s $\tau(X, Y)$ are given in the following proposition.

**Proposition 2.6.** Let $X, Y$ be continuous random variables with a joint cdf $F$ and univariate margins $F_1, F_2$, respectively. Denote $C$ is a copula of $(X, Y)$. Then

$$\tau(X, Y) = 4 \int_{\mathbb{R}^2} [F(x, y) - F_1(x)F_2(y)]dF(x, y),$$

$$= 4 \int_{I^2} [C(u, v) - uv]dC(u, v). \quad (2.15)$$
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

Proof. The proposition is proven since

\[
\mathbb{P}(X_1 < X_2, Y_1 < Y_2) = \int_{\mathbb{R}^2} \mathbb{P}(X_1 < x, X_2 < y) dF(x, y)
\]
\[
= \int_{\mathbb{R}^2} F(x, y) dF(x, y)
\]
\[
= \int_{\mathbb{R}^2} C(u, v) dC(u, v).
\]

So,

\[
\tau(X, Y) = \mathbb{P}(X_1 < X_2, Y_1 < Y_2) + \mathbb{P}(X_1 > X_2, Y_1 > Y_2)
\]
\[
- \mathbb{P}(X_1 < X_2, Y_1 > Y_2) - \mathbb{P}(X_1 > X_2, Y_1 < Y_2)
\]
\[
= 4 \int_{\mathbb{R}^2} \left[ F(x, y) - F_1(x)F_2(y) \right] dF(x, y)
\]
\[
= 4 \int_{\mathbb{R}^2} \left[ C(u, v) - uv \right] dC(u, v).
\]

Proposition 2.5 and 2.6 show us that Spearman’s \( \rho \) and Kendall’s \( \tau \) are the functions of the difference between the joint cdf \( F(x, y) \) and the product-marginal cdf \( F_1(x)F_2(y) \), or the difference between the copula \( C(u, v) \) and the independent copula \( \Pi_2(u, v) = uv \). The difference in Spearman’s \( \rho \) is computed on the distribution of the product-marginal cdf, i.e. \( dF_1(x)dF_2(y) \) while that in Kendall’s \( \tau \) is computed on the distribution of the joint cdf, i.e. \( dF(x, y) \). The explanation comes from the difference between the formulas given in (2.12) and (2.14).

Obviously, both of the modulus of these measures have axioms (A), (B) and (C) of Rényi axiom 1. Since Spearman’s \( \rho \) and Kendall’s \( \tau \) can be written in only the term of copula \( C(u, v) \), they are invariant with respect to strictly increasing functions. Unfortunately, Spearman’s \( \rho \) and Kendall’s \( \tau \) do not satisfy axioms (D), (E) and (F) [Schweizer and Wolff, 1981]. Kruskal [1958] proved that if \( (X, Y) \) is a bivariate normal, the Spearman’s \( \rho_S(X, Y) \) is given by the formula,

\[
\rho_S(X, Y) = \frac{2}{\pi} \arcsin \left( \frac{\rho_P(X, Y)}{2} \right).
\]  
(2.16)

It means that \( |\rho_S(X, Y)| \) is a strictly increasing function of \( |\rho_P(X, Y)| \). The modulus of Spearman’s \( \rho \) satisfies axiom (G).

**Blomqvist’s \( \beta \)**: Blomqvist [1950] proposed a measure of dependency \( (X, Y) \) by using the concordant property of these variables with their population medians \( \bar{x} \) and \( \bar{y} \). This
measure is called Blomqvist’s $\beta$ or the medial correlation coefficient, and is given by

$$
\beta(X, Y) = \mathbb{P}[(X - \bar{x})(Y - \bar{y}) \geq 0] - \mathbb{P}[(X - \bar{x})(Y - \bar{y}) < 0].
$$

(2.17)

If $(X, Y)$ is continuous with a joint cdf $F$ and marginal cdfs $F_1$ and $F_2$ respectively, and a copula $C$,

$$
\beta(X, Y) = 2\mathbb{P}[(X - \bar{x})(Y - \bar{y}) > 0] - 1 = 2\left(\mathbb{P}(X < \bar{x}, Y < \bar{y}) + \mathbb{P}(X > \bar{x}, Y > \bar{y})\right) - 1
$$

$$
= 4F(\bar{x}, \bar{y}) - 1 = 4C(1/2, 1/2) - 1.
$$

Thus, Blomqvist’s $\beta$ can be interpreted as a normalized difference between the joint cdf $F$ and product-marginal cdf $F_1F_2$ at $(\bar{x}, \bar{y})$, or the copula $C$ and the independent copula $\Pi$ at $(1/2, 1/2)$. The measure $|\beta(X, Y)|$ satisfies axioms (A), (B), (C) and (F’), however it doesn’t have other axioms.

Gini’s $\gamma$: In 1914, Corrado Gini introduced a measure of dependency $\gamma$ which he called the indice di coagraduazione semplice. This measure is described as follows [Gini, 1914]:

Let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be $n$-observations of continuous random variables $X$ and $Y$, respectively. Let $\{p_i\}_{i=1}^n$ and $\{q_i\}_{i=1}^n$ be ranking order of $\{x_i\}_{i=1}^n$, and $\{y_i\}_{i=1}^n$ respectively. Then

$$
\gamma(X, Y) = \frac{1}{n^2/2} \left( \sum_{i=1}^n |p_i + q_i - n - 1| - \sum_{i=1}^n |p_i - q_i| \right),
$$

(2.18)

where $\lfloor u \rfloor$ denotes the integer part of $u$. If $(X, Y)$ has a copula $C$, $\gamma(X, Y)$ is given by [Nelsen, 2006; Schmid et al., 2010]

$$
\gamma(X, Y) = 2 \int_{I^2} \left( |u + v - 1| - |u - v| \right) dC(u, v)
$$

$$
= 4 \int_{I^2} \left[ M_2(u, v) + W_2(u, v) \right] dC(u, v) - 2.
$$

(2.19)

To see the significance of Gini’s $\gamma$, let us present the relationship between the formulae of Gini’s $\gamma$ and Spearman’s $\rho$. From second part of the proof of Proposition 2.5, we have

$$
\rho_S(X, Y) = 3 \int_{I^2} \left[ 4uv + 1 - 2u - 2v \right] dC(u, v)
$$

$$
= 3 \int_{I^2} \left[ (u + v - 1)^2 - (u - v)^2 \right] dC(u, v).
$$

If $X$ and $Y$ are independent then $\gamma(X, Y) = 0$, and if $X$ and $Y$ are strictly dependent then $|\gamma(X, Y)| = 1$. The both of converses are not true. The modulus $|\gamma(X, Y)|$ has axioms (A), (B) and (C), and doesn’t have other axioms. Gini’s $\gamma(X, Y)$ is also written
in only the term of copula, then it satisfies axiom (F').

**Hoeffding’s Δ:** The Kendall’s τ measure evaluates the dependency of two samples \( \{x_i\}_{i=1}^n \) and \( \{y_i\}_{i=1}^n \) by considering the concordant property of each pair points \((x_i, y_i)\) and \((x_j, y_j)\), \(i, j = 1, \ldots, n, i \neq j\). More generally, Hoeffding [1948] evaluated the dependency of these two samples by considering the “related position” of group of any four points \((x_{i1}, y_{i1}), (x_{i3}, y_{i3}), (x_{i4}, y_{i4})\) \(\) and \((x_{i5}, y_{i5})\) corresponding with the point \((x_{i1}, y_{i1}), i_u = 1, \ldots, n, i_u \neq i_v\), \(u, v = 1, \ldots, 5, u \neq v\). The “related position” notion depends on the ranking order of samples. Specifically, the Hoeffding’s measure \(D\) is defined as follows:

Let \( \{x_i\}_{i=1}^n \) and \( \{y_i\}_{i=1}^n \) be \(n\)-observations of continuous random variables \(X\) and \(Y\), respectively. Denote \(\text{sign}(u) = 1\) if \(u \geq 0\), and \(0\) if \(u < 0\) be a sign function. Let us denote

\[
\psi(x_1, x_2, x_3) = \text{sign}(x_1 - x_2) - \text{sign}(x_1 - x_3),
\]

\[
\phi(x_1, y_1; \ldots; x_5, y_5) = \frac{1}{4} \psi(x_1, x_2, x_3) \psi(y_1, y_2, y_3) \psi(x_1, x_4, x_5) \psi(y_1, y_4, y_5).
\]

Then

\[
D = D_n = \frac{1}{n(n-1) \cdots (n-4)} \sum \phi(x_{i1}, y_{i1}; \ldots; x_{i5}, y_{i5}),
\]

(2.20)

where \(\sum\) denotes summation over all \(i_u\) such that

\[
i_u = 1, \ldots, n; \quad i_u \neq i_v\] if \(u \neq v\), \(u, v = 1, \ldots, 5\).

The formula (2.20) is presented by the distribution function \(F\) of \((X, Y)\) as follows:

\[
\Delta(X, Y) = \int_{\mathbb{R}^2} \cdots \int_{\mathbb{R}^2} \phi(x_1, y_1; \ldots; x_5, y_5) dF(x_1, y_1) \cdots dF(x_5, y_5).
\]

(2.21)

The simple presentation of \(\Delta(X, Y)\) is given by the following result.

**Proposition 2.7.** If \(F, F_1\) and \(F_2\) are continuous and \(C\) is a copula of \((X, Y)\),

\[
\Delta(X, Y) = \int_{\mathbb{R}^2} \left[ F(x, y) - F_1(x)F_2(y) \right]^2 dF(x, y)
\]

\[
= \int_{\mathbb{R}^2} \left[ C(u, v) - uv \right]^2 dC(u, v).
\]

(2.22)

**Proof.** It is not difficult to see that \(\psi(x_1, x_2, x_3) \psi(y_1, y_2, y_3) = 1\) if and only if \((x_1 - x_2)(y_1 - y_2) > 0, (x_1 - x_3)(y_1 - y_3) > 0\) and \((x_1 - x_2)(x_1 - x_3) < 0, \) and \(-1\) if and only if \((x_1 - x_2)(y_1 - y_2) < 0, (x_1 - x_3)(y_1 - y_3) < 0\) and \((x_1 - x_2)(x_1 - x_3) < 0\). So,

\[
\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \frac{1}{2} \psi(x_1, x_2, x_3) \psi(y_1, y_2, y_3) dF(x_2, y_2) dF(x_3, y_3) =
\]

\[
P(X_2 < x_1, Y_2 < y_1)P(X_3 > x_1, Y_3 > y_1) - P(X_2 < x_1, Y_2 > y_1)P(X_3 > x_1, Y_3 < y_1),
\]
2.1. MEASURES OF DEPENDENCY

where \((X_2, Y_2)\) and \((X_3, Y_3)\) are independent and identically distributed with \((X, Y)\). We have

\[
\begin{align*}
\mathbb{P}(X_2 < x, Y_2 < y) & \mathbb{P}(X_3 > x, Y_3 > y) \\
& = F(x_1, y_1) \left[1 - F_1(x_1) - F_2(y_1) + F(x_1, y_1)\right] \\
\mathbb{P}(X_2 < x, Y_2 > y) & \mathbb{P}(X_3 > x, Y_3 < y) \\
& = \left[F_1(x_1) - F(x_1, y_1)\right] \left[F_2(y_1) - F(x_1, y_1)\right].
\end{align*}
\]

Hence,

\[
\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \frac{1}{2} \psi(x_1, x_2, x_3) \psi(y_1, y_2, y_3) dF_{1}(x_2, y_2) dF_{3}(x_3, y_3) = F(x_1, y_1) - F_1(x_1) F_2(y_1)
\]

\[
\Rightarrow \quad \Delta(X, Y) = \int_{\mathbb{R}^2} \left[F(x_1, y_1) - F_1(x_1) F_2(y_1)\right] \\
\left[\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \frac{1}{2} \psi(x_1, x_4, x_5) \psi(y_1, y_4, y_5) dF_{4}(x_4, y_4) dF_{5}(x_5, y_5)\right] dF_{x_1, y_1} \\
= \int_{\mathbb{R}^2} \left[F(x, y) - F_1(x) F_2(y)\right]^2 dF(x, y).
\]

The formula of copula is a consequence of formula of distribution functions.
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

i.e. \( dF_1(x) dF_2(y) \), or independent copula, i.e. \( dudv \). These measures are given by

\[
\rho_{L_1}(X, Y) = 12 \int_{\mathbb{R}^2} |F(x, y) - F_1(x) F_2(y)| dF_1(x) dF_2(y),
\]

\[
= 12 \int_{\mathbb{R}^2} |C(u, v) - uv| dudv. \tag{2.23}
\]

\[
\rho_{L_2}(X, Y) = \left( 90 \int_{\mathbb{R}^2} [F(x, y) - F_1(x) F_2(y)]^2 dF_1(x) dF_2(y) \right)^{\frac{1}{2}},
\]

\[
= \left( 90 \int_{\mathbb{R}^2} [C(u, v) - uv]^2 dudv \right)^{\frac{1}{2}}. \tag{2.24}
\]

\[
\rho_{L_\infty}(X, Y) = 4 \sup_{x, y \in \mathbb{R}} |F(x, y) - F_1(x) F_2(y)|,
\]

\[
= 4 \sup_{u, v \in [0, 1]} |C(u, v) - uv|. \tag{2.25}
\]

**Theorem 2.3** (Schweizer and Wolff [1981]). Let \( X \) and \( Y \) be continuous random variables with a copula \( C \). Then these measures \( \rho_{L_1}(X, Y) \) and \( \rho_{L_2}(X, Y) \) given by (2.23) and (2.24) respectively, satisfy the Rényi axiom 1’ (i.e. replacing axiom (F) by axiom (F’)) and condition (H).

**Proof.** It is clear from (2.23) that \( \rho_{L_1}(X, Y) \) has axioms (A) and (B). Since

\[
\int_{\mathbb{R}^2} |C(u, v) - uv| dudv \leq \frac{1}{12}, \tag{2.26}
\]

\( \rho_{L_1}(X, Y) \) also satisfies axiom (C). Axiom (D) follows from Proposition 2.2. Since the equality holds in (2.26) if and only if, \( C = M_2 \) or \( C = W_2 \), axiom (E) follows from Proposition 2.3 and Proposition 2.4. Let \( f \) and \( g \) be strictly increasing functions, Proposition 2.1 infers that \( \rho_{L_1}(X, Y) = \rho_{L_1}(f(X), g(Y)) \). Denote \( \tilde{C} \) be copula of \( (f(X), g(Y)) \).

If \( f \) is strictly increasing and \( g \) is strictly decreasing, it is not difficult to see that \( |\tilde{C} - \Pi_2|(u, v) = |\Pi_2 - C|(1 - u, v) \), and if \( f \) and \( g \) are strictly decreasing, \( |\tilde{C} - \Pi_2|(u, v) = |C - \Pi_2|(1 - u, 1 - v) \). In both cases we always have \( \rho_{L_1}(f(X), g(Y)) = \rho_{L_1}(X, Y) \). Thus, \( L_1 \)-distance measure satisfies axiom (F’). The satisfaction of axiom (G) of \( \rho_{L_1}(X, Y) \) is referred in [Slepain, 1962]. Finally, if \( F_n \xrightarrow{w} F \), it follows from (2.7) that the corresponding copulas \( C_n \) converge pointwise to \( C \). By the Lipschitz condition of copula [Nelsen, 1996, p. 11], any family of copulas is equicontinuous, whence the convergence is uniform. Thus, \( \rho_{L_1}(X_n, Y_n) \rightarrow \rho_{L_1}(X, Y) \) as \( n \) tends to infinity.

The proof of satisfaction of \( \rho_{L_2}(X, Y) \) is analogous. However, the explicit formula of strictly increasing function in axiom (G) remains to be determined. \( \square \)
For the $\rho_{L_\infty}$ measure, the axiom (E) is not satisfied because there exist copulas $C$, distinct from $W_2$ and $M_2$, for which $\sup_{u,v \in [1]} |C(u, v) - uv| = 1/4$. Others axioms in Rényi axiom 1' are satisfied by $\rho_{L_\infty}$.

**Information-based measures:** If $X$ and $Y$ have a joint density function $f$ and marginal density functions $f_1$ and $f_2$, the difference between the joint cdf $F$ and the product-marginal cdf $F_1 F_2$ of $(X, Y)$ can be considered based on the difference between the joint density function $f$ and the product-marginal density function $f_1 f_2$. In the Information Theory, we often use a divergence measure to evaluate the difference between two density functions.

The first divergence, named information divergence, Kullback-Leibler divergence or relative entropy, was introduced by Kullback and Leibler [1951]. For any two density functions $f$ and $g$ determined on a probability space $(\Omega, \mathcal{B}, P)$, the information divergence of $f$ and $g$ is given as follows:

$$\mathcal{D}(f \| g) = \int_\Omega f(x) \log \frac{f(x)}{g(x)} dP(x). \quad (2.27)$$

Imagine that $\log \frac{f(x)}{g(x)}$ is the difference of $f$ and $g$ at point $x$, thus the quantity $\mathcal{D}(f \| g)$ is an average of difference of $f$ and $g$ corresponding distribution $f$. This quantity is nonnegative and equals zero if and only if $f = g$ almost surely. Csiszár [1967], Vajda [1972] and Liese and Vajda [1987] generalized the information divergence by $\varphi$-divergence which was defined as follows: Let $\varphi$ be a real and continuous convex function on $[0, +\infty)$ that satisfies the following conditions:

$$0 \varphi\left(\frac{0}{0}\right) = 0, \quad \text{and} \quad 0 \varphi\left(\frac{t}{0}\right) = t \lim_{u \to +\infty} \frac{\varphi(u)}{u}, \quad \forall t \in [0, +\infty). \quad (2.28)$$

Then the quantity

$$\mathcal{D}_\varphi(f \| g) = \int_\Omega g(x) \varphi\left(\frac{f(x)}{g(x)}\right) dP(x) \quad (2.29)$$

called $\varphi$-divergence of $f$ with respect to $g$, is used to measure the difference between the two densities $f$ and $g$.

When $f$ and $g$ are joint and product-marginal density functions of $(X, Y)$ respectively, $\mathcal{D}(f \| g)$ and $\mathcal{D}_\varphi(f \| g)$ are the measures of dependency of $X$ and $Y$. In this case, information divergence $\mathcal{D}(f \| g)$ is called an information dependency of $X$ and $Y$, and denoted $I(X, Y)$. Precisely, if $X$ and $Y$ have a joint density function $f$ and marginal density functions $f_1, f_2$, then the information dependency and the $\varphi$-divergence are given by

$$I(X, Y) = \int_{\Omega^2} f(x, y) \log \frac{f(x, y)}{f_1(x)f_2(y)} dP(x) dP(y), \quad (2.30)$$
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

<table>
<thead>
<tr>
<th>Bivariate measures of dependency</th>
<th>Rényi Axioms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abs. of Pearson’s correlation</td>
<td>(A) (B) (C)</td>
</tr>
<tr>
<td>Monotone correlation</td>
<td>(D) (E) (F)</td>
</tr>
<tr>
<td>Maximal correlation</td>
<td>(F’) (G)</td>
</tr>
<tr>
<td>Abs. of Spearman’s ρ</td>
<td></td>
</tr>
<tr>
<td>Abs. of Kendall’s τ</td>
<td></td>
</tr>
<tr>
<td>Abs. of Blomqvist’s β</td>
<td></td>
</tr>
<tr>
<td>Abs. of Gini’s γ</td>
<td></td>
</tr>
<tr>
<td>Hoeffding’s 30∆</td>
<td></td>
</tr>
<tr>
<td>$L_1$&amp;$L_2$-distance measures</td>
<td></td>
</tr>
<tr>
<td>$L_\infty$-distance measure</td>
<td></td>
</tr>
<tr>
<td>Normalized Information dependency</td>
<td></td>
</tr>
</tbody>
</table>

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
</table>

Table 2.1: Relationship between Rényi axioms and bivariate measures of dependency. The symbol (x) means that this measure has this axiom respectively.

and

$$\mathcal{D}_\varphi(X,Y) = \int_{\Omega^2} f_1(x)f_2(y)\varphi\left(\frac{f(x,y)}{f_1(x)f_2(y)}\right)d\mathbb{P}(x)d\mathbb{P}(y). \quad (2.31)$$

The definition in (2.30) shows that the information dependency of two random variables $X$ and $Y$ is the mutual information of them (see the definition of the mutual information in [Cover and Thomas, 2006; Shannon, 1948]). However, in the next subsection we will see that their definitions in the multivariate case are different.

Unfortunately, the information dependency of $X$ and $Y$ is not bounded by one. Hence, to make the information dependency as a measure of dependency, we normalize it by

$$J^*(X,Y) = \frac{J(X,Y)}{H(X) + H(Y) - \max\{H(X),H(Y)\}}. \quad (2.32)$$

where $H$ is the entropy operator which will be introduced in the next section. The normalization of $\varphi$-divergence has some difficulties which depend on the function $\varphi$. Some interesting properties of the $\varphi$-divergence can be found in [Liese et al., 2006; Liese and
2.1. MEASURES OF DEPENDENCY

Vajda, 1987. For example, they gave the following interesting property: Assuming that \( \varphi \) is standardized by the condition \( \varphi(1) = 0 \). We denote \( \varphi^*(t) = t\varphi(1/t) \) and \( \varphi^*(0) = \lim_{t\downarrow 0} \varphi^*(t) \). Then

\[
0 \leq D_{\varphi}(X,Y) \leq \varphi(0) + \varphi^*(0) \]

where let \( \mu \) and \( \nu \) are joint and product-marginal distributions of \((X,Y)\) respectively, \( D_{\varphi}(X,Y) = 0 \) if and only if \( \mu = \nu \) and \( D_{\varphi}(X,Y) = \varphi(0) + \varphi^*(0) \) if \( \mu \) and \( \nu \) are singular. If \( \varphi(0) + \varphi^*(0) < \infty \), we can replaced the last word “if” by “if and only if”. Note that, when \( \varphi(u) = u \log u, \ \forall u \in [0, +\infty) \), \( D_{\varphi}(X,Y) \equiv \mathcal{I}(X,Y) \).

The information dependency and the \( \varphi \)-divergence using as the measures of dependency are paid attention by researchers such as [Ali and Silvey, 1966; Bell, 1962; Csiszár, 1967; Guerrero-Cusumano, 1996; Joe, 1989; Kent, 1983; Linfoot, 1957; Tjostheim, 1996; Vajda, 1989]. The interest in information-based measures is that the normalization of information dependency satisfies Rényi axiom 1. This statement will be discussed and proven carefully in section 2.3.

Finally, the satisfaction of bivariate measures of dependency with the Rényi axioms 1 and 1’ is summarized in Table 2.1

2.1.4 Multivariate measures of dependency

In the last subsection, we have surveyed some well-known bivariate measures of dependency. These measures evaluate the dependency of two random variables based on three views as follows the correlation (i.e. Pearson’s correlation, monotone correlation, maximal correlation, Spearman’s \( \rho \)), the concordance (i.e. Spearman’s \( \rho \), Kendall’s \( \tau \), Blomqvist’s \( \beta \), Hoeffding’s \( \Delta \)), and the difference between a joint cdf and a product-marginal cdf (i.e. Hoeffding’s \( \Delta \), \( L_p \)-distance measures, Information-based measures). Among three views, only the third one maintains the significance of dependent property when applying for the multivariate case. For example, in the multivariate case, we can not give a scale real value to evaluate the difference of the covariance matrices of the different random vector \( X \). In this spirit, to generalize the bivariate measures of dependency, we are interested in the view of the difference between the joint cdf and the product-marginal cdf. The article [Schmid et al., 2010] is an excellent reference of the multivariate extensions of measures of dependency.

Let \( X_1, \ldots, X_m \) be random variables on a probability space \((\Omega, \mathcal{B}, \mathbb{P})\) with cumulative distribution functions \( F_1, \ldots, F_m \) respectively. Random vector \( X = (X_1, \ldots, X_m) \) has a joint cdf \( F \) and a copula \( C \). In the view of the difference between the joint cdf and the product-marginal cdf, the multivariate extensions of Spearman’s \( \rho \), Kendall’s \( \tau \), Hoeffding’s \( \Delta \), and \( L_p \)-distance measures, and their references are straightforward. The multivariate extensions of Information-based measures are given and discussed in the next
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

section. Denoting \( x = (x_1, \ldots, x_m) \in \mathbb{R}^m \), \( u = (u_1, \ldots, u_m) \in \mathbb{I}^m \).

**Spearman’s \( \rho \):** [Joe, 1990; Nelsen, 1996; Schmid and Schmidt, 2007]

\[
\rho_S(X) = \frac{(m + 1)2^m}{2^m - (m + 1)} \int_{\mathbb{R}^m} \left[ F(x) - \prod_{i=1}^{m} F_i(x_i) \right] dF_1(x_1) \ldots dF_m(x_m),
\]

\[
= \frac{(m + 1)2^m}{2^m - (m + 1)} \int_{\mathbb{I}^m} \left[ C(u) - \prod_{i=1}^{m} u_i \right] du_1 \ldots du_m. \tag{2.33}
\]

**Kendall’s \( \tau \):** [Joe, 1990; Nelsen, 1996; Schmid et al., 2010]

\[
\tau(X) = \frac{1}{2^m - 1} \left\{ 2^m \int_{\mathbb{R}^m} F(x) dF(x) - 1 \right\} = \frac{1}{2^m - 1} \left\{ 2^m \int_{\mathbb{I}^m} C(u) dC(u) - 1 \right\}. \tag{2.34}
\]

**Hoeffding’s measure:**

\[
\Delta(X) = h(m) \int_{\mathbb{R}^m} \left[ F(x) - \prod_{i=1}^{m} F_i(x_i) \right]^2 dF(x) = h(m) \int_{\mathbb{I}^m} \left[ C(u) - \prod_{i=1}^{m} u_i \right]^2 dC(u), \tag{2.35}
\]

where

\[
h(m) = \left( \int_{\mathbb{I}^m} \left[ M_m(u) - \prod_{i=1}^{m} u_i \right]^2 dM_m(u) \right)^{-1},
\]

with \( M_m \) is the upper Fréchet-Hoeffding bounded copula.

**\( L_p \)-distance measures:** [Fernández-Fernández and González-Barrios, 2004; Gaißer et al., 2009; Wolff, 1980]

\[
\rho_{L_1}(X) = h_1(m) \int_{\mathbb{R}^m} \left| F(x) - \prod_{i=1}^{m} F_i(x_i) \right| dF_1(x_1) \ldots dF_m(x_m),
\]

\[
= h_1(m) \int_{\mathbb{I}^m} \left| C(u) - \prod_{i=1}^{m} u_i \right| du_1 \ldots du_m, \tag{2.36}
\]

where

\[
h_1(m) = \left( \frac{1}{m + 1} - \frac{1}{2^m} \right)^{-1}.
\]

\[
\rho_{L_2}(X) = h_2(m) \left( \int_{\mathbb{R}^m} \left[ F(x) - \prod_{i=1}^{m} F_i(x_i) \right]^2 dF_1(x_1) \ldots dF_m(x_m) \right)^{\frac{1}{2}},
\]

\[
= h_2(m) \left( \int_{\mathbb{I}^m} \left[ C(u) - \prod_{i=1}^{m} u_i \right]^2 du_1 \ldots du_m \right)^{\frac{1}{2}}, \tag{2.37}
\]

26
where

\[
h_2(m) = \left( \frac{2}{(m+1)(m+2)} - \frac{1}{2^m} \prod_{i=0}^{m} \left( i + \frac{1}{2} \right) + \frac{1}{3^m} \right)^{-1}.
\]

\[
\rho_{L_{\infty}}(X) = h_\infty(m) \sup_{x \in \mathbb{R}^m} |F(x) - \prod_{i=1}^{m} F_i(x_i)| = h_\infty(m) \sup_{u \in \mathbb{R}^m} |C(u) - \prod_{i=1}^{m} u_i|, \quad (2.38)
\]

Almost the references of these multivariate measures proved that these measures satisfied axioms (A) and (B) of Rényi axiom 2. The Hoeffding’s measure, \( L_p \)-distance measures and the modulus of Spearman’s \( \rho \), Kendall’s \( \tau \) also have axiom (C). The Spearman’s \( \rho_S(X) \) and Kendall’s \( \tau(X) \) equal zero when components of \( X \) are mutual independent, and one when components of \( X \) are strictly increasing dependent. The converse are not true. Since the presentation of Hoeffding’s measure and \( L_p \)-distance measures are the distance between copula \( C \) and independent copula \( \Pi \), these measures satisfy axiom (D) and invariant with respect to strictly increasing functions.

Joe [1990]; Taylor [2007] gave an multivariate extension of concordance, and used the statement that measure of dependency was increasing with respect to increasing of concordance to find the other multivariate extensions of Spearman’s \( \rho \), Kendall’s \( \tau \) and Blomqvist’s \( \beta \). The multivariate extension of concordance is defined as follows: Let \( X = (X_1, \ldots, X_m) \) be random vector with joint cdf \( F \). Let us denote \( \bar{F}(x) = \bar{F}(x_1, \ldots, x_m) = \mathbb{P}(X_1 > x_1, \ldots, X_m > x_m) \) be a survival cdf of \( X \).

**Definition 2.3.** Let \( X, Y \) be random vectors with cdfs \( F, G \) and survival cdfs \( \bar{F}, \bar{G} \), respectively. Then, \( Y \) is more concordant than \( X \) (written \( X \prec_c Y \)) if \( \forall x \in \mathbb{R}^m, \)

\[
F(x) \leq G(x) \quad \text{and} \quad \bar{F}(x) \leq \bar{G}(x). \quad (2.39)
\]

In the bivariate case, we know that \((X_1, X_2)\) which \( X_1 \) and \( X_2 \) are strictly decreasing dependent is the least concordance and \((Y_1, Y_2)\) which \( Y_1 \) and \( Y_2 \) are strictly increasing dependent is the most concordance. However, in the multivariate, we only know the most concordance since the strictly decreasing dependent does not exist when random vector has more than two components. Thus, the multivariate extension of concordance is interested in distinguishing the increasing dependent random vector with other dependent random vectors. In our experience, we are often interested in determining the independence and strict dependence of random vectors, not only the strictly increasing dependence. In this intent, we prefer to use the difference between the joint cdf and the product-marginal cdf to constructing multivariate extension of measures of dependency more than the concordant property.
2.2 Information Dependency

In this section, we would like to study the information dependency concept and answer the question why the information dependency, a special case of the information-based measures, can work as a measure of dependency. The information dependency is a concept of Information Theory, therefore the question why the information dependency can express a dependency of a group of random variables will be answered by investigating an imprint of information theory on probability theory and statistics.

Although researches on information theory was started in the field of engineering science, it is now a great branch of probability theory and statistics. The essential mathematical and statistical nature of information theory were introduced and developed by three famous scientists, Fisher [1956], Shannon [1948], and Wiener [1956]. The greatest mathematician of probability theory and statistics, A.N. Kolmogorov, gave his personal idea about information theory that

“Information theory must precede probability theory and not be based on it”
[Kolmogorov, 1983].

According to A.N. Kolmogorov, the role of information theory is greater than probability theory. In this sense, information theory will be an excellent tool to solve some problems in probability theory and statistics, including a problem of measuring the dependency.

The fundamental notions of information theory such as Shannon entropy, divergence (or relative entropy, information divergence, Kullback-Leibler divergence), and information dependency will be introduced in this section.

2.2.1 Information

In 1948, C.E. Shannon defined a quantity, named entropy, which gives a numerical measure of how far from deterministic a random variable is. This definition is described as follows [Shannon, 1948]: Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space.

**Definition 2.4 (Unit of Information).** For any event $A \in \mathcal{B}$ with probability $\mathbb{P}(A)$, the “information” $I(A)$ gained by knowing that $A$ has occurred is given by

$$I(A) = - \log \mathbb{P}(A).$$

We shall often use logarithms to the base 2 instead of natural logarithms. The natural logarithms are usually more convenient for mathematical view while the base 2 logarithms provide more intuitive descriptions. The intuitive idea of Definition 2.4 is that the rarer an event $A$ is, the more information we gain if we know it has occurred. For example, let $A$ be an event “the sun rises this morning”, and $B$ be an event “the sun does not rise this
morning”. It is clear that when each event occurs, the case of B has so much information compared with the case of A.

**Definition 2.5 (Shannon entropy).** Let X be a discrete random variable taking values in the finite set \( \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R} \) with probabilities \( \{p_1, p_2, \ldots, p_n\} \), respectively. The Shannon entropy of X is defined as the expected amount of information gained on learning the value of X:

\[
H(X) = \mathbb{E}[I(X = x_i)] = - \sum_{i=1}^{n} p_i \log p_i.
\]

We define \( 0 \log 0 \) to be 0 in the above definition. From the definition of information of events, the Shannon entropy \( H(X) \) measures the amount of uncertainty of random variable \( X \); \( H(X) = 0 \) when \( X \) is deterministic, and \( H(X) \) is the highest when \( X \) is purely random.

Rényi [1960] explained in his works why we don’t choose any other decreasing functions of \( P(A) \) but the function \(- \log P(A)\) to define the information of the event \( A \). His works is summarized as follows: Let us call \( P = \{p_1, p_2, \ldots, p_n\} \), a sequence of nonnegative numbers with \( 0 < \sum_{i=1}^{n} p_i \leq 1 \), be a finite discrete generalized probability distribution, and \( \Delta \) denotes the set of all finite discrete generalized probability distributions \( P \). For any generalized distribution \( P = \{p_1, \ldots, p_n\} \in \Delta \), we denote \( W(P) = \sum_{i=1}^{n} p_i \). For any \( P = \{p_1, \ldots, p_n\} \in \Delta \) and \( Q = \{q_1, \ldots, q_m\} \in \Delta \), let us denote

\[
P \ast Q = \{p_iq_j, \ i = 1, n, \ j = 1, m\} \quad \text{and} \quad P \cup Q = \{p_1, \ldots, p_n, q_1, \ldots, q_m\}.
\]

Note that, \( P \ast Q \in \Delta \) and when \( W(P) + W(Q) \leq 1 \), \( P \cup Q \in \Delta \).

**Rényi Axiom 3.** The uncertainty measure or the expected amount of information of any finite discrete generalized probability distribution \( P \), \( H(P) \), should be satisfied the following conditions

(A) \( H(P) \) is a symmetric function of the elements of \( P \).

(B) \( H(\{p\}) \), \( 0 < p \leq 1 \) is a continuous function of \( p \).

(C) \( H(\{1/2\}) = 1 \).

(D) For \( P, Q \in \Delta \), we have \( H(P \ast Q) = H(P) + H(Q) \).

(E) If \( P, Q \in \Delta \), and \( W(P) + W(Q) \leq 1 \), then

\[
H(P \cup Q) = \frac{W(P)H(P) + W(Q)H(Q)}{W(P) + W(Q)}.
\]
Theorem 2.4 ([Rényi, 1960]). If $H(\mathcal{P})$ is defined for all $\mathcal{P} \in \Delta$, and satisfied the Rényi axiom 3, then

$$H(\mathcal{P}) = \frac{-\sum_{i=1}^{n} p_i \log p_i}{\sum_{i=1}^{n} p_i}, \quad (2.42)$$

with $\mathcal{P} = \{p_1, \ldots, p_n\}$.

Proof. Let us put $h(p) = H(\{p\})$ for all $0 < p \leq 1$. Axiom (D) shows that $h(pq) = h(p) + h(q)$ for all $0 < p, q \leq 1$. By axiom (B), $h(p)$ is continuous for $0 < p \leq 1$ and by axiom (C), $h(1/2) = 1$, we have, $\forall p \in (0, 1]$

$$h(p) = H(\{p\}) = -\log p.$$

Now let us consider $\mathcal{P} = \{p_1, p_2, \ldots, p_n\} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_n$, where $\mathcal{P}_i = \{p_i\}, i = 1, \ldots, n$, the theorem is proven by axiom (E). \qed

The definition of Shannon entropy (Definition 2.5) can be easily extended to random variables taking countably many values. The significance of Shannon entropy can be also used for random vectors, groups of random variables. This definition is given by:

**Definition 2.6 (Joint entropy).** Given a random vector $X = (X_1, \ldots, X_m)$ taking values in a countable set $\{x_i\}_{i=1}^{\infty}$, i.e. $x_i = (x_{i1}, \ldots, x_{im})$, with probabilities $\{p_i\}_{i=1}^{\infty}$, respectively. Joint entropy of $X$ is defined as follows:

$$H(X) = -\sum_{i=1}^{\infty} p_i \log p_i. \quad (2.43)$$

Next, we consider a formula of the entropy for continuous random variables. This entropy is often called a differential entropy [Cover and Thomas, 2006, Ch. 08, p. 243].

**Definition 2.7 (Differential entropy).** Let $X$ be a continuous random variable with a density function $f(x)$. The differential entropy of $X$ is defined as

$$h(X) = -\int_{\mathbb{R}} f(x) \log f(x)dx. \quad (2.44)$$

Let $X = (X_1, \ldots, X_m)$ be a continuous random vector with a density $f(x_1, \ldots, x_m)$. The joint differential entropy of $X$ is defined as

$$h(X) = -\int_{\mathbb{R}^m} f(x_1, \ldots, x_m) \log f(x_1, \ldots, x_m)dx_1 \ldots dx_m. \quad (2.45)$$

We also define $0 \log 0$ to be 0 in the above definition. If the density function $f$ is continuous, the relationship between the differential entropy and the Shannon entropy
is given by the following explanation. Consider a random variable \( X \) with a continuous density function \( f \). For any positive number \( \delta \), by the mean value theorem, there exists a value \( x_i \) in interval \((i\delta, (i+1)\delta)\), \( \forall i \in \mathbb{Z} \), such that

\[
\delta f(x_i) = \int_{i\delta}^{(i+1)\delta} f(x)dx.
\]

Consider the quantized random variable \( X^\delta \) taking values in countable set \( \{x_i, i \in \mathbb{Z}\} \) with \( X^\delta(\omega) = x_i \) if \( i\delta \leq X(\omega) < (i+1)\delta \), \( \forall \omega \in \Omega \). Clearly, \( \mathbb{P}(X^\delta = x_i) = \delta f(x_i) \). Then the Shannon entropy of \( X^\delta \) is given as

\[
H(X^\delta) = -\sum_{i \in \mathbb{Z}} \delta f(x_i) \log f(x_i) - \log \delta.
\]

So by Riemann integrability,

\[
\lim_{\delta \to 0} [H(X^\delta) + \log \delta] = -\int_{\mathbb{R}} f(x) \log f(x)dx = h(X).
\]

Naturally, we often estimate the continuous random variable \( X \) by a quantized random variable \( X^\delta \) with enough small positive number \( \delta \). However, the above explanation shows that the smaller \( \delta \) becomes, the more the Shannon entropy of \( X^\delta \) differs from the differential entropy of \( X \). Moreover, unlike the Shannon entropy, the differential entropy can be negative and not invariant under continuous coordinate transformations. Thus, in general the differential entropy is not a good measure of uncertainty or information.

We can recover the disadvantage of the differential entropy by introducing a new notion, divergence. Given a measurable space \((\Omega, \mathcal{B})\) and two probability measures \( \mu \) and \( \nu \). For any event \( A \in \mathcal{B} \), let \( H_1 \) (or \( H_2 \)) be the hypothesis that \( A \) was selected from the population of which probability measure is \( \mu \) (or \( \nu \)).

**Definition 2.8 (Unit of discriminative information).** For any event \( A \in \mathcal{B} \), the “discriminative information” of \( A \) for discriminating between \( H_1 \) and \( H_2 \) is given by

\[
\mathcal{D}(A) = \log \frac{\mu(A)}{\nu(A)}.
\]

(2.46)

**Definition 2.9.** For two discrete random vectors taking values in \( \{x_1, \ldots, x_n\} \) with probabilities \( \mathbf{p} = (p_1, \ldots, p_n) \) and \( \mathbf{q} = (q_1, \ldots, q_n) \) respectively, the divergence of \( \mathbf{p} \) with respect to \( \mathbf{q} \) is defined by

\[
\mathcal{D}(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{n} p_i \log \left( \frac{p_i}{q_i} \right).
\]

(2.47)

In the case of a continuous random vector with density functions \( f \) and \( g \), define the
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

The divergence of $f$ with respect to $g$ can be defined as:

$$
\mathcal{D}(f\|g) = \int_{\mathbb{R}^m} f(x) \log \left( \frac{f(x)}{g(x)} \right) dx.
$$

In the above definition, we define $0 \log \frac{0}{0}$ to be 0. The divergence is also known as the relative entropy, Kullback-Leibler divergence, and Kullback-Leibler discrimination. It is a nonsymmetric measure. From the perspectives of information theory, divergence is the average information for discrimination between two hypotheses $H_1$ and $H_2$ introduced above [Kullback and Leibler, 1951, p. 80]. Divergence is the most important quantity in the information theory since other quantities Shannon entropy, differential entropy, and information dependency can be expressed in its terms.

Note that, all the above definitions of concepts in information theory are defined for the discrete and continuous variables. In the next subsection, we would like to introduce the definition and properties of divergence in the general setting. The general definitions of entropy and information dependency are also introduced. This work is a summary of [Gray, 2010, Ch. 7].

2.2.2 Divergence

Given a measurable space $(\Omega, \mathcal{B})$ and two probability measures $\mu$ and $\nu$. Let us denote $\mathcal{P}$ be the collection of all finite and measurable partitions of $\Omega$.

**Definition 2.10 (Divergence).** The relative entropy or divergence of $\mu$ with respect to $\nu$ is defined as follows

$$
\mathcal{D}(\mu\|\nu) = \sup_{\pi \in \mathcal{P}} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)}.
$$

The divergence is the supremum of the relative entropies or divergences obtained in a collection of all countable alphabet codings of the sample space. A simpler formula to determine the divergence is given by the following result. Let us denote $\mathfrak{F}$ be a generating field of $\mathcal{B}$, i.e. $\mathcal{B} = \sigma(\mathfrak{F})$, and $\Omega$ is a collection of all finite measurable partitions of $\Omega$ which its cells belong to $\mathfrak{F}$.

**Proposition 2.8 (Dobrushin’s Theorem [Gray, 2010], p. 175).** Let $\mu$ and $\nu$ be two probability measures on $(\Omega, \mathcal{B})$,

$$
\mathcal{D}(\mu\|\nu) = \sup_{\pi \in \mathcal{Q}} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)}.
$$

Since the union of a sequence of asymptotically generating sequence of sub-$\sigma$-fields is a generating field, the following corollary is inferred immediately from Dobrushin’s Theorem.
2.2. INFORMATION DEPENDENCY

**Corollary 2.1** (Gray [2010], Corollary 7.3). Suppose that $\mu$ and $\nu$ are probability measures on a measurable space $(\Omega, \mathcal{B})$ and $\mathcal{F}_n$ is an asymptotically generating sequence of sub-$\sigma$-fields and let $\mu^{(n)}$ and $\nu^{(n)}$ denote the restrictions of $\mu$ and $\nu$ to $\mathcal{F}_n$. Then

$$D(\mu^{(n)} \parallel \nu^{(n)}) \uparrow D(\mu \parallel \nu).$$

(2.51)

The divergence has some basic properties given by the following result.

**Proposition 2.9** (Gray [2010], Lemma 7.1 and Lemma 7.4). For any two probability measures $\mu$ and $\nu$,

1. $D(\mu \parallel \nu) \geq 0$, with equality if and only if $\mu = \nu$.

2. If $\mu$ is not absolutely continuous with respect to $\nu$, $D(\mu \parallel \nu) = \infty$.

3. If $\mu$ is absolutely continuous with respect to $\nu$ (denote $\mu \ll \nu$), the Radon-Nikodym derivative $f = d\mu/d\nu$ exists and

$$D(\mu \parallel \nu) = \int_{\Omega} \log f(\omega) d\mu(\omega) = \int_{\Omega} f(\omega) \log f(\omega) d\nu(\omega) < \infty.$$  

(2.52)

**Proof.** For the first property, given any partition $\pi \in \mathcal{P}$, we show that

$$\sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)} \geq 0$$

with equality if, and only if $\mu(A) = \nu(A), \forall A \in \pi$. Thus, since $\log x \leq x - 1$, $\forall x > 0$, with equality if and only if $x = 1$,

$$\sum_{A \in \pi} \mu(A) \log \frac{\nu(A)}{\mu(A)} \leq \sum_{A \in \pi} \mu(A) \left( \frac{\nu(A)}{\mu(A)} - 1 \right) = \sum_{A \in \pi} \nu(A) - \sum_{A \in \pi} \mu(A) = 0$$

with equality if and only if $\mu(A) = \nu(A), \forall A \in \pi$. From the Jensen’s inequality, since log is a concave function, we have

$$\sum_{A \in \pi} \mu(A) \log \frac{\nu(A)}{\mu(A)} \leq \log \left( \sum_{A \in \pi} \mu(A) \frac{\nu(A)}{\mu(A)} \right) = 0$$

with equality if and only if $\mu(A) = \nu(A), \forall A \in \pi$.

Since $D(\mu \parallel \nu)$ is the supremum over all such partitions, it is also nonnegative. It can be 0 only if $\mu$ and $\nu$ assign the same probabilities to all cells in all partitions and hence the divergence is 0 only if the measures are identical.

Secondly, if $\mu$ is not absolutely continuous with respect to $\nu$, there is a measurable set $A$ such that $\nu(A) = 0$ and $\mu(A) > 0$. The relative entropy for partition $\pi = \{A, A^c\}$
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

is infinity, and hence, the divergence is also infinity.

Finally, assume that \( \mu \ll \nu \) and let \( f = d\mu/d\nu \). Suppose that \( A \) is a set with \( \nu(A) > 0 \) and consider the following conditional cumulative distribution function, \( \forall x \in \mathbb{R} \nabla x \cap A \nabla = \frac{\nu(\{f < x\})}{\nu(A)}. \)

We have

\[
E_\nu(f|A) = \int_0^\infty x dF_A(x) = \int_A f(\omega) d\nu(\omega) = \frac{\mu(A)}{\nu(A)},
\]

\[
\int_0^\infty x \log xdF_A(x) = \frac{1}{\nu(A)} \int_A f(\omega) \log f(\omega) d\nu(\omega).
\]

Applying Jensen’s inequality to the convex function \( x \log x \) yields the inequality

\[
\frac{1}{\nu(A)} \int_A \log f(\omega) d\mu(\omega) = \frac{1}{\nu(A)} \int_A f(\omega) \log f(\omega) d\nu(\omega) = \int_0^\infty x \log xdF_A(x)
\]

\[
\geq \left( \int_0^\infty xdF_A(x) \right) \log \left( \int_0^\infty xdF_A(x) \right) = \frac{\mu(A)}{\nu(A)} \log \frac{\mu(A)}{\nu(A)}.
\]

Therefore, for any set \( A \) with \( \nu(A) > 0 \) we have

\[
\int_A \log f(\omega) d\mu(\omega) \geq \mu(A) \log \frac{\mu(A)}{\nu(A)}.
\]

Let \( \pi = \{A_1, \ldots, A_n\} \) be a finite partition, note that \( \mu(A) > 0 \) infers \( \nu(A) > 0 \), we have

\[
\int_\Omega \log f(\omega) d\mu(\omega) \geq \sum_{i: \mu(A_i) > 0} \int_{A_i} \log f(\omega) d\mu(\omega) \geq \sum_{i=1}^n \mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)}.
\]

This proves that

\[
\mathcal{D}(\mu||\nu) \leq \int_\Omega \log f(\omega) d\mu(\omega).
\]

To obtain the converse inequality, let \( q_n : \mathbb{R} \to \mathbb{R} \) as follows:

\[
q_n(x) = \begin{cases} 
  n & n \leq x \\
  (k-1)2^{-n} & (k-1)2^{-n} \leq x < k2^{-n}, \ k = 1, 2, \ldots, n2^n \\
  -k2^{-n} & -k2^{-n} \leq x < -(k-1)2^{-n}, \ k = 1, 2, \ldots, n2^n \\
  -n & x < -n.
\end{cases}
\]

We know that

\[
\int_\Omega \log f(\omega) d\mu(\omega) = \lim_{n \to \infty} \int_\Omega q_n[\log f(\omega)] d\mu(\omega).
\]

For fixed \( n \) the quantizer \( q_n[\log f(\omega)] \) induces a partition \( \pi \) of \( \Omega \) with \( 2n2^n + 1 \) cells. Denote
the cells \( \{ \omega : -n < \log f(\omega) < n \} \) be \( A_i, \ i = 1, \ldots, 2n^{2^n} - 1 \). It is clear that \( \forall \omega_1, \omega_2 \in A_i, i = 1, \ldots, 2n^{2^n} - 1, \ | \log f(\omega_1) - \log f(\omega_2)| \leq 2^{-(n-1)} \). So, \( \forall i = 1, \ldots, 2n^{2^n} - 1 \)

\[
\mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)} \geq \int_{A_i} \log f(\omega) d\mu(\omega) - 2^{-(n-1)} \mu(A_i).
\]

Since \( \mu(\log f \geq n) \) and \( \mu(\log f \leq -n) \) tend to 0 as \( n \to \infty \) and since \( x \log x \to 0 \) as \( x \to 0 \), given any small positive number \( \epsilon \) we can choose \( n \) large enough such that

\[
\sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)} \geq \sum_{i=1}^{2n^{2^n}-1} \mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)} - \epsilon
\]

\[
\geq \sum_{i=1}^{2n^{2^n}-1} \int_{A_i} \log f(\omega) d\mu(\omega) - 2^{-(n-1)} - \epsilon
\]

\[
\geq \int_{\{\omega : |\log f(\omega)| \leq n\}} \log f(\omega) d\mu(\omega) - 2^{-(n-1)} - \epsilon.
\]

The above inequalities are true for arbitrarily large \( n \) and arbitrarily small \( \epsilon \), then

\[
\mathcal{D}(\mu\|\nu) \geq \int_{\Omega} \log f(\omega) d\mu(\omega),
\]

completing the proof of the proposition.

If \( \mu \) and \( \nu \) are discrete measures with corresponding probability measure function’s \( p \) and \( q \), the Radon-Nikodym derivative is simply \( d\mu/d\nu(\omega) = p(\omega)/q(\omega) \) and the proposition gives the known formula for the discrete case. If \( \mu \) and \( \nu \) are the distributions of some continuous random vectors with corresponding density functions \( f \) and \( g \), then

\[
\mathcal{D}(\mu\|\nu) = \int_{\mathbb{R}^m} f(x) \log \frac{f(x)}{g(x)} dx.
\]

(2.53)

2.2.3 Information Dependency

Let \( X \) and \( Y \) be random variables on a probability space \( (\Omega, \mathcal{B}, \mathbb{P}) \) with corresponding distributions \( \mu_1 \) and \( \mu_2 \). The random vector \((X, Y)\) has a joint distribution \( \mu \) and a product-marginal distribution \( \nu = \mu_1 \times \mu_2 \).

**Definition 2.11 (Information Dependency).** The information dependency of \( X \) and \( Y \) is defined as the divergence between its joint distribution and its product-marginal distribution,

\[
J(X, Y) = \mathcal{D}(\mu\|\nu) = \sup_{\pi \in \mathcal{P}^2} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{[\mu_1 \times \mu_2](A)},
\]

(2.54)

where \( \mathcal{P}^2 \) is a collection of all finite and measurable partitions of \( \Omega^2 \).
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

From the definition of the divergence, the information dependency of $X$ and $Y$ is the average information to discriminate between two hypotheses $H_1$: “$X$ and $Y$ are not mutually independent” and $H_2$: “$X$ and $Y$ are mutually independent”. Intuitively, information dependency measures the information to get one variable from the other, i.e. if two variables are independent, the information to get one variable from the other is zero, and if these variables are strictly dependent, the information to get one variable from the other is the highest value.

The information dependency can be extended for a group of more than two random variables as follows: Let $X_1, \ldots, X_m$ be $m$ random variables on a probability space $(\Omega, \mathcal{B}, P)$ with a joint distribution $\mu$ and marginal distributions $\mu_1, \ldots, \mu_m$. Let us denote $\nu = \mu_1 \times \cdots \times \mu_m$ be a product marginal distribution of $(X_1, \ldots, X_m)$.

**Definition 2.12 (Multivariate Information Dependency).** The information dependency of $X_1, \ldots, X_m$ is defined as the divergence between their joint distribution and their product marginal distribution,

$$I(X_1, \ldots, X_m) = D(\mu || \nu) = \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\mu_1 \times \cdots \times \mu_m}(A),$$

where $\mathcal{P}(m)$ is a collection of all finite and measurable partition of $\Omega^m$.

As a consequence of the presentation of the divergence, we also have the formula of information dependency in some simple cases: If $X_i$ is a discrete random variable receiving $n_i$ values $\{x_{i1}, \ldots, x_{in_i}\}$, $i = 1, \ldots, m$, the information dependency of $(X_1, \ldots, X_m)$ is given by simple formula:

$$I(X_1, \ldots, X_m) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} \mathbb{P}(X_1 = x_{i1}^{(1)}, \ldots, X_m = x_{im}^{(m)}) \times \log \frac{\mathbb{P}(X_1 = x_{i1}^{(1)}, \ldots, X_m = x_{im}^{(m)})}{\prod_{j=1}^{m} \mathbb{P}(X_j = x_{ij}^{(j)})}.$$ (2.56)

If $X_1, \ldots, X_m$ are continuous random variables with a joint density function $f$ and $m$ marginal density functions $f_1, \ldots, f_m$, their information dependency is given by the following formula of integral:

$$I(X_1, \ldots, X_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_m) \log \frac{f(x_1, \ldots, x_m)}{\prod_{i=1}^{m} f_i(x_i)} dx_1 \cdots dx_m.$$ (2.57)

The generalization of Shannon entropy (Definition 2.5) is given by:

**Definition 2.13 (Entropy).** For any random variable $X$ with distribution $\mu$, the entropy
2.3. NORMALIZED INFORMATION DEPENDENCY

of \( X \) is the information dependency of \( X \) and itself,

\[
H(X) = J(X, X) = \sup_{\pi \in \mathcal{P}} \sum_{A \in \pi} -\mu(A) \log \mu(A),
\]

(2.58)

where \( \mathcal{P} \) is a collection of all finite and measurable partition of \( \Omega \).

Definition 2.13 is also used with the purpose of defining the entropy of random vector \( X = (X_1, \ldots, X_m) \), i.e. \( H(X) = J(X, X) \). Note that, let \( X \) be a group of \( m \) random variables \( X \), it is not difficult to show that \( H(X) = \frac{1}{m-1}J(X) \), the generalization of formula (2.58).

### 2.3 Normalized Information Dependency

In this section, we would like to study how the information dependency works as a measure of dependency. Concretely, we study how the information dependency measure satisfies the Rényi Axioms. Let us consider two important results which seem to be fundamental reasons to consider the information dependency as a measure of dependency. Given \( m \geq 2 \) random variables \( X_1, \ldots, X_m \) on a probability space \( (\Omega, \mathcal{B}, \mathbb{P}) \) with a joint distribution \( \mu \) and \( m \) marginal distributions \( \mu_1, \ldots, \mu_m \).

**Proposition 2.10.**

\[
J(X_1, \ldots, X_m) \geq 0,
\]

(2.59)

with equality if and only if \( X_1, \ldots, X_m \) are mutual independent.

This result is clearly a consequence of the first statement of Proposition 2.9.

**Proposition 2.11.**

\[
J(X_1, \ldots, X_m) \leq \sum_{i=1}^{m} H(X_i) - \max\{H(X_1), \ldots, H(X_m)\},
\]

(2.60)

with equality if and only if \( X_1, \ldots, X_m \) are strictly dependent.

**Proof.** For each \( n \in \mathbb{N} \), let \( \pi_n \) be a partition of \( \Omega^m \) with \((2n^2 + 1)^m\) cells given as follows:

\[
\pi_n = \{ A_{i_1, \ldots, i_m} : 1 \leq i_j \leq 2n^2 + 1, \ j = 1, \ldots, m \},
\]

where

\[
A_{i_1, \ldots, i_m} = \prod_{j=1}^{m} \{ \omega : t_{ij-1} < X_j(\omega) \leq t_{ij} \},
\]

with

\[
t_k = \begin{cases} 
-\infty & k = 1 \\
\frac{k-1}{2n} - n & 2 \leq k \leq 2n^2 \\
+\infty & k = 2n^2 + 1.
\end{cases}
\]
Clearly, $\sigma(\pi_n)$, the $\sigma$-field generating by the partition $\pi_n$, is an asymptotically generating sequence of sub-$\sigma$-fields. Hence, Corollary 2.1 shows that

$$\sum_{A \in \pi_n} \mu(A) \log \frac{\mu(A)}{\prod_{i=1}^{m} \mu_i(A_i)} \uparrow I(X_1, \ldots, X_m),$$

and

$$\sum_{A \in \pi_n} \mu(A) \log \frac{\mu(A)}{\prod_{i=1}^{m} \mu_i(A_i)} = \sum_{A \in \pi_n} \mu(A) \log \mu(A) - \sum_{i=1}^{m} \mu_i(A_i) \log \mu_i(A_i)$$

$$\to \sum_{i=1}^{m} H(X_i) - H(X_1, \ldots, X_m),$$

where $A = A_1 \times \cdots \times A_m$. Thus, we have

$$I(X_1, \ldots, X_m) = \sum_{i=1}^{m} H(X_i) - H(X_1, \ldots, X_m).$$

We also have $H(X_1, \ldots, X_m) \geq H(X_i)$, $i = 1, \ldots, m$ with equality if and only if there exist $m$ Borel-measurable function $f_1, \ldots, f_m$ such that $X_j = f_j(X_i)$ almost surely, $j = 1, \ldots, m$. Therefore, the inequality (2.60) is true. The equality occurs if and only if for some $1 \leq i \leq m$, $X_j$ is a Borel-measurable function of $X_i$ for all $j \neq i$, i.e. $X_1, \ldots, X_m$ are strictly dependent.

Based on the two above propositions, we would like to consider a presentation of the normalized information dependency as a measure of dependency

$$J^*(X_1, \ldots, X_m) = \frac{J(X_1, \ldots, X_m)}{\sum_{i=1}^{m} H(X_i) - \max\{H(X_1), \ldots, H(X_m)\}},$$

with the assumption that $0 \div 0 = 0$. Note that, when $J(X_1, \ldots, X_m) = \infty$, the normalized information dependency is computed as follows:

$$J^*(X_1, \ldots, X_m) = \lim_{n \to \infty} \frac{J_{\pi_n}(X_1, \ldots, X_m)}{\sum_{i=1}^{m} \mu_{\pi_n}(X_i) - \max\{H_{\pi_n}(X_1), \ldots, H_{\pi_n}(X_m)\}},$$

when the above limitation exists, where $J_{\pi_n}$ and $H_{\pi_n}$ are the information dependency and entropy quantities computed only based on the partition $\pi_n$ introduced above.

The formulae (2.61) and (2.62) infer that the normalized information dependency $J^*(X_1, \ldots, X_m)$ is defined for any $m$ random variables $X_1, \ldots, X_m$ (axiom (A)). The normalized information dependency is also invariant with any permutation of the order of random variables (axiom (B)). Proposition 2.10 and Proposition 2.11 confirm that $J^*(X_1, \ldots, X_m)$ has axioms (C), (D) and (E). The satisfaction of axiom (F) is given by
2.3. NORMALIZED INFORMATION DEPENDENCY

following result

**Proposition 2.12.** Given \( m \) random variables \( X_1, \ldots, X_m \) on a probability space \( (\Omega, \mathcal{B}, \mathbb{P}) \). Let \( Y_i = f_i(X_i) \), \( i = 1, \ldots, m \), be new random variables, for some Borel-measurable functions \( f_i \). Then

\[
I(X_1, \ldots, X_m) \geq I(Y_1, \ldots, Y_m),
\]

(2.63)

with equality if and only if all of the functions \( f_i \) are one-to-one.

**Proof.** Ali and Silvey [1966] showed that: Let \( \mu \) and \( \nu \) be any two probability measures on \( (\Omega_1, \mathcal{B}_1) \), and \( \phi \) be the generalized Radon-Nikodym derivative of \( \nu \) with respect to \( \mu \). Given any continuous convex function \( C \), and for any measurable transformation \( f \) from \( (\Omega_1, \mathcal{B}_1) \) onto another measurable space \( (\Omega_2, \mathcal{B}_2) \), we have the following inequality

\[
\int_{\Omega_1} C[\phi(\omega_1)] d\mu(\omega_1) \geq \int_{\Omega_2} C[\phi(\omega_2)] d[\mu f^{-1}](\omega_2),
\]

where \([\mu f^{-1}](A) = \mu[f^{-1}(A)] \), \( \forall A \in \mathcal{B}_1 \). There is equality here for all \( C \) if and only if \( f \) is sufficient for \( (\mu, \nu) \) (see more detail in [Ali and Silvey, 1966]).

Considering in our case with \( \mu \) as joint distribution, \( \nu \) as product-marginal distribution of \( (X_1, \ldots, X_m) \), \( C \) as a function \( u \log u \) on \((0, +\infty)\) and transformation \( f = (f_1, \ldots, f_m) \) from \( (\Omega^m, \mathcal{B}^m) \) onto \((\Omega^m, \mathcal{B}^m) \), we have

\[
I(X_1, \ldots, X_m) = \int_{\Omega^m} \log \frac{d\mu(\omega)}{d\nu(\omega)} d\mu(\omega)
= \int_{\Omega^m} C\left(\frac{d\mu(\omega)}{d\nu(\omega)}\right) d\nu(\omega)
\geq \int_{\Omega^m} C\left(\frac{d[\mu f^{-1}](\omega)}{d[\nu f^{-1}](\omega)}\right) d[\nu f^{-1}](\omega)
= \int_{\Omega^m} \log \frac{d[\mu f^{-1}](\omega)}{d[\nu f^{-1}](\omega)} d[\mu f^{-1}](\omega)
= I(Y_1, \ldots, Y_m).
\]

The equality occurs if and only if \( f_i \) are one-to-one Borel-measurable functions, \( i = 1, \ldots, m \).

Let \((X, Y)\) be a bivariate normal with correlation coefficient \( r(X,Y) \), the information dependency \( I(X, Y) \) is \(-1/2 \log[1 - 2r^2(X, Y)]\). Since this quantity is a strictly increasing function of \( |r(X,Y)| \) and \( H(X) + H(Y) - \max\{H(X), H(Y)\} \) is constant with respect to \( r(X,Y) \), measure \( I^* \) has axiom (G). In brief, the normalized information dependency measure \( I^* \) satisfies all of seven postulates of Rényi Axiom 2.
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

2.4 Examples

In this section, we construct three simple examples of classes of bivariate random vectors: linear convolution of discrete distribution and Gaussian, linear convolution of diagonal uniform distribution and Gaussian, and linear convolution of circular uniform distribution and Gaussian. These examples are called simple because we can determine the exact formulae of joint distribution and joint density functions of their bivariate random vectors. However, excepting the first example which we can compute the exact formulae of \( \rho_{L_1} \) and \( \rho_{L_\infty} \) measures, all other examples are impossible for us to determine the formulae of the \( \rho_{L_p} \) measures or the information dependency measures.

Although we cannot find the exact formulae of some measures of dependency such as \( L_p \)-distance and the normalized information dependency, we can use some numerical methods to estimate the dependency values getting from these measures which have significance in evaluating and comparing these measures. To estimate the normalized information dependency \( J^*_n(X,Y) \), we do not only estimate the information dependency \( J(X,Y) \) but also estimate the univariate entropies \( H(X), H(Y) \). Estimating the entropy is not mentioned in our works, so we give the simple way to bound the entropy of a random variable \( X \) as follows: For each \( k = 1, \ldots, 20 \), let \( k \times 10^3 \) independent samples \( x_1, \ldots, x_{k \times 10^3} \) of \( X \). Let \( a_k = \min\{x_1, \ldots, x_{k \times 10^3}\} \), \( b_k = \max\{x_1, \ldots, x_{k \times 10^3}\} \), and \( \pi_k \) be a partition of \([a_k, b_k]\) which has 100 equal length interval-cells. \( \hat{H}_k(X) \) is an empirical entropy of these samples computed based on the partition \( \pi_k \). Finally, we estimate the entropy of \( X \) as follows:

\[
\min_{1 \leq k \leq 20} \hat{H}_k(X) \lesssim H(X) \lesssim \max_{1 \leq k \leq 20} \hat{H}_k(X).
\]

2.4.1 Linear convolution of Discrete distribution and Gaussian

Let \( U \) be a discrete bivariate random vector which receives five values corresponding with probabilities \( P[U = (0,0)] = P[U = (-1,-1)] = P[U = (-1,1)] = P[U = (1,-1)] = P[U = (1,1)] = \frac{1}{5} \). Given a nonnegative real number \( r \), let \( N_r \) be a bivariate Gaussian random vector with a mean vector \((0,0)\) and a covariance matrix \( r^2 I \), where \( I \) is the \((2 \times 2)\)-identify matrix. Random vectors \( U \) and \( N_r \) are independent. In this example, we would like to express the dependency value getting from some measures of a bivariate random vector

\[
(X, Y) = U + N_r.
\]

Figure 2.1 illustrates the values of 2000 independent samples of \((X,Y)\) corresponding \( r = 0, 0.2, 0.5 \) and 1. When \( r = 0 \), \((X,Y)\) is a discrete random vector and has same distribution with \( U \). The dependency values of \((X,Y)\) getting from some bivariate measures are given in Table 2.2. When \( r > 0 \), \((X,Y)\) is a continuous random vector with a
2.4. EXAMPLES

Figure 2.1: Examples of Linear convolution of Discrete distribution and Gaussian corresponding with parameters $r = 0, 0.2, 0.5$ and $1$.

<table>
<thead>
<tr>
<th>Measuring Dependency Values ($r = 0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_P(X,Y) = \rho_S(X,Y) = \tau(X,Y) = 0$</td>
</tr>
<tr>
<td>$\beta(X,Y) = \frac{3}{5} = 0.6$ ; $30\Delta(X,Y) = \frac{36}{27} = 0.0576$</td>
</tr>
<tr>
<td>$\rho_{L_1}(X,Y) = \frac{12}{54} = 0.1728$ ; $\rho_{L_2}(X,Y) = \frac{9\sqrt{10}}{53} = 0.2277$ ; $\rho_{L_\infty}(X,Y) = \frac{4}{23} = 0.16$</td>
</tr>
<tr>
<td>$\rho^<em>(X,Y) = 1$ ; $J^</em>(X,Y) = \frac{\log 5 - 0.8 \log 4}{\log 3 - 0.8 \log 2} = 0.4744$</td>
</tr>
</tbody>
</table>

Table 2.2: Different measures of dependency for the example of the discrete distribution.
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

joint distribution function, marginal distribution functions, a joint density function and marginal density functions given as follows: Denote

\[
\phi_r(x) = \frac{1}{\sqrt{2\pi r}} e^{-\frac{x^2}{2r}} \quad \text{and} \quad \Phi_r(x) = \int_{-\infty}^{x} \phi_r(u) du, \quad \forall x \in \mathbb{R},
\]

The joint distribution function and the marginal distribution functions:

\[
F(x, y) = \frac{1}{5} \left[ \Phi_r(x) \Phi_r(y) + \Phi_r(x-1) \Phi_r(y-1) + \Phi_r(x-1) \Phi_r(y+1) + \Phi_r(x+1) \Phi_r(y-1) + \Phi_r(x+1) \Phi_r(y+1) \right],
\]

\[
F_1(x) = \frac{1}{5} \left[ \Phi_r(x) + 2\Phi_r(x-1) + 2\Phi_r(x+1) \right],
\]

\[
F_2(y) = \frac{1}{5} \left[ \Phi_r(y) + 2\Phi_r(y-1) + 2\Phi_r(y+1) \right], \quad \forall x, y \in \mathbb{R}.
\]

The joint density function and the marginal density functions:

\[
f(x, y) = \frac{1}{5} \left[ \phi_r(x) \phi_r(y) + \phi_r(x-1) \phi_r(y-1) + \phi_r(x-1) \phi_r(y+1) + \phi_r(x+1) \phi_r(y-1) + \phi_r(x+1) \phi_r(y+1) \right],
\]

\[
f_1(x) = \frac{1}{5} \left[ \phi_r(x) + 2\phi_r(x-1) + 2\phi_r(x+1) \right],
\]

\[
f_2(y) = \frac{1}{5} \left[ \phi_r(y) + 2\phi_r(y-1) + 2\phi_r(y+1) \right], \quad \forall x, y \in \mathbb{R}.
\]

Using the following formula of integral

\[
\int_{-\infty}^{\infty} \Phi_r(a + bx) \phi_r(x) dx = \Phi_1 \left( \frac{a}{\sqrt{1 + r^2 b^2}} \right),
\]

the Pearson's correlation, the Spearman's \( \rho \), the Kendall's \( \tau \) and the Blomqvist's \( \beta \) of \((X, Y)\) are computed and equal zero. The \( L_1 \)-distance measure is computed as follows:

\[
\rho_{L_1}(X, Y) = 12 \int_{\mathbb{R}^2} |F(x, y) - F_1(x) F_2(y)| f_1(x) f_2(y) dxdy
\]

\[
= \frac{12}{5^2} \left[ \int_{-\infty}^{\infty} \left| \Phi_r(x - 1) + \Phi_r(x + 1) - 2\Phi_r(x) \right| \left( \phi_r(x) + 2\phi_r(x - 1) + 2\phi_r(x + 1) \right) dx \right]^2
\]

\[
= \frac{24}{5^2} \left[ \int_{0}^{\infty} \left( 2\Phi_r(x) - \Phi_r(x - 1) - \Phi_r(x + 1) \right) \left( \phi_r(x) + 2\phi_r(x - 1) + 2\phi_r(x + 1) \right) dx \right]^2.
\]

Using some techniques such as \( \Phi_r(x) + \Phi_r(-x) = 1 \), and

\[
\int \Phi_r(x) \phi_r(x - a) dx = \Phi_r(x) \Phi_r(x - a) - \int \Phi_r(x - a) \phi_r(x) dx,
\]
2.4. EXAMPLES

we infer
\[\rho_{L_1}(X,Y) = \frac{6}{125} \left(2\Phi_r(1) + 1 - 4 \int_0^1 \Phi_r(x+1)\phi_r(x)dx\right).\] (2.66)

The \(L_\infty\)-distance measure is given by
\[\rho_{L_\infty}(X,Y) = \frac{4}{25} \left(\sup_{x \in \mathbb{R}}|\Phi_r(x-1) + \Phi_r(x+1) - 2\Phi_r(x)|\right)^2\] (2.67)
\[= \frac{4}{25}(\Phi_r(x^* - 1) + \Phi_r(x^* + 1) - 2\Phi_r(x^*))^2,\]
where \(x^*\) is a solution of the following equations
\[\frac{d}{dx}[\Phi_r(x-1) + \Phi_r(x+1) - 2\Phi_r(x)] = \phi_r(x-1) + \phi_r(x+1) - 2\phi_r(x)\]
\[= e^{-\frac{1}{2\pi r}(x-1)^2} + e^{-\frac{1}{2\pi r}(x+1)^2} - 2e^{-\frac{1}{2\pi r}x^2}\]
\[= 0, \quad x \in (0, \infty)\] (2.68)

These equations have a unique solution
\[x^* = r^2 \log\left(e^{\frac{1}{2\pi r}} + e^{-\frac{1}{2\pi r}} - 1\right).\]

Computing exactly the information dependency in this case is impossible. However, in some special cases of \(r\), we can give the significant estimation of the information dependency. For example, when \(r\) is small \((0 < r \leq 0.15)\), we construct bounds of the information dependency as follows:
\[(2C_1 - D_1) + \left(\frac{2}{(1+\epsilon)^2}C_2 - D_2\right) \leq I(X,Y) \leq (2C_1 - D_1) + \left(2C_2 - \frac{1}{(1+\epsilon)^2}D_2\right),\] (2.69)
where
\[C_1 = \log 5 - \frac{2}{5} \log 2\left[1 + 2\Phi_r\left(-\frac{3}{2}\right) + \Phi_r\left(\frac{1}{2}\right)\right] + \log(\sqrt{2\pi r})\]
\[C_2 = \frac{1}{10r^2}\left[5r^2 + 12\Phi_r\left(-\frac{3}{2}\right) + 6\Phi_r\left(-\frac{1}{2}\right) - 8r^2\phi_r\left(\frac{3}{2}\right) - 12r^2\phi_r\left(\frac{1}{2}\right)\right],\]
\[D_1 = \log 5 + 2 \log(\sqrt{2\pi r})\]
\[D_2 = \frac{1}{5r^2}\left(4(r^2 + 4)\Phi_r\left(-\frac{3}{2}\right) + 5r^2\Phi_r\left(\frac{1}{2}\right) + 2\Phi_r\left(-\frac{1}{2}\right)\Phi_r\left(\frac{1}{2}\right)(r^2 + 2) + r^2\phi_r^2\left(-\frac{1}{2}\right)\right.\]
\[\left.- 8\Phi_r\left(-\frac{1}{2}\right)r^2\phi_r(0) - \left[\frac{9}{2} - 5\Phi_r\left(-\frac{1}{2}\right)\right]r^2\phi_r\left(\frac{1}{2}\right) - 10r^2\phi_r\left(\frac{3}{2}\right)\right].\]

The way of constructing the bounds of the information dependency and choosing the parameter \(\epsilon\) are explained carefully in Appendix.

We would like to illustrate an example with \(r = 0.1\) and choose \(\epsilon = 5.10^{-3}\) to estimate
the information dependency. The bounds of the information dependency are

\[ 0.4835 \leq I(X,Y) \leq 0.4973. \]

The bounds of the entropies are simply given by \( 3.7734 \leq H(X), H(Y) \leq 3.7965 \). The visual samples of this example are given in figure 2.2, and the measuring dependency values are given in table 2.3. When \( r \) is large, i.e. \( (r \geq 2) \), the bounds of the information dependency are given by simple formula:

\[
\frac{5r^2 + 2}{5} \left( \frac{1}{(r + \alpha_2)^2} - \frac{1}{(r + \beta_1)^2} \right) \leq I(X,Y) \leq \frac{5r^2 + 2}{5} \left( \frac{1}{(r + \alpha_1)^2} - \frac{1}{(r + \beta_2)^2} \right),
\]

(2.70)

where the parameters \( \alpha_1, \alpha_2, \beta_1 \) and \( \beta_2 \) depend on \( r \) and \( 0 < \alpha_1 < \alpha_2, 0 < \beta_1 < \beta_2 \). The formula of bounds and the way to choose \( \alpha_1, \alpha_2, \beta_1, \beta_2 \) are also presented carefully in Appendix. For the case in which \( r \) is large, we would like to illustrate an example with \( r = 2 \) and choose \( \alpha_1 = 0.16, \alpha_2 = 0.23, \beta_1 = 0.16, \) and \( \beta_2 = 0.23 \) (see in Appendix). The bounds of the information dependency and the entropies are

\[ 0 \leq I(X,Y) \leq 0.0556, \quad 3.9533 \leq H(X), H(Y) \leq 4.1163. \]

The visual samples of this example are given in figure 2.3, and the measuring dependency values are given in table 2.4.

### 2.4.2 Linear convolution of Diagonal Uniform distribution and Gaussian

Let \( U \) be a bivariate random vector which has a uniform distribution on the diagonal of the square \([-1, 1]^2\). Random vector \( U \) has a joint distribution function and marginal distribution functions:

\[
F(x,y) = \begin{cases} 
\frac{1}{2}(\min(x,y) + 1) & x,y > -1, x > 1 \text{ or } y > 1 \\
\frac{1}{4}(x + y + \min(x,y) + 1) & x,y \leq 1, x + y > 0 \\
\frac{1}{4}(\min(x,y) + 1) & x,y > -1, x + y \leq 0 \\
0 & \text{otherwise},
\end{cases}
\]

\[
F_1(x) = \begin{cases} 
1 & 1 < x \\
\frac{1}{2}(x + 1) & -1 < x \leq 1 \\
0 & x \leq -1
\end{cases}, \quad F_2(y) = \begin{cases} 
1 & 1 < y \\
\frac{1}{2}(y + 1) & -1 < y \leq 1 \\
0 & y \leq -1.
\end{cases}
\]
2.4. EXAMPLES

Figure 2.2: Linear convolution of Discrete distribution and Gaussian with parameter $r = 0.1$.

Measuring Dependency Values ($r = 0.1$)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_P(X, Y)$</td>
<td>$\rho_S(X, Y)$</td>
</tr>
<tr>
<td>$\rho_{L_1}(X, Y)$</td>
<td>$\approx 0.048$</td>
</tr>
<tr>
<td>$0.1274 \lesssim J^*(X, Y) \lesssim 0.1472$</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Different measures of dependency for the example of figure 2.2.
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

Figure 2.3: Linear convolution of Discrete distribution and Gaussian with parameter $r = 2$.

<table>
<thead>
<tr>
<th>Measuring Dependency Values ($r = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_P(X, Y) = \rho_S(X, Y) = \tau(X, Y) = \beta(X, Y) = 0$</td>
</tr>
<tr>
<td>$\rho_{L_1}(X, Y) \approx 0.0861$ ; $\rho_{L_\infty}(X, Y) \approx 4.29 \times 10^{-4}$</td>
</tr>
<tr>
<td>$0 \leq I^*(X, Y) \lesssim 0.0141$</td>
</tr>
</tbody>
</table>

Table 2.4: Different measures of dependency for the example of figure 2.3.
Let $\mathbf{N}_r$ be a bivariate Gaussian random vector with a mean vector $(0, 0)$ and a covariance matrix $r^2 \mathbf{I}$. Random vectors $\mathbf{U}$ and $\mathbf{N}_r$ are independent. Now we consider a new random vector 

$$(X, Y) = \mathbf{U} + \mathbf{N}_r$$

and express the dependency of $X$ and $Y$. Figure 2.4 illustrates 2000 independent samples of $(X, Y)$ corresponding $r = 0, 0.2, 0.5$ and 1. When $r > 0$, using an integral formula

$$\int \Phi_r(a + bx) dx = \frac{1}{b} (a + bx) \Phi_r(a + bx) + \frac{r^2}{b} \phi_r(a + bx) + C,$$

a joint distribution function and marginal distribution functions of $(X, Y)$ are computed as follows: $\forall x, y \in \mathbb{R},$

$$F(x, y) = \frac{1}{4} \int_{-1}^{1} \left[ \Phi_r(x - t) \Phi_r(y - t) + \Phi_r(x - t) \Phi_r(y + t) \right] dt$$

$$F_1(x) = \frac{1}{2} \int_{-1}^{1} \Phi_r(x - t) dt$$

$$= \frac{1}{2} \left[ (x + 1) \Phi_r(x + 1) - (x - 1) \Phi_r(x - 1) \right] + \frac{r^2}{2} \left[ \phi_r(x + 1) - \phi_r(x - 1) \right] \quad (2.71)$$

$$F_2(y) = \frac{1}{2} \int_{-1}^{1} \Phi_r(y - t) dt$$

$$= \frac{1}{2} \left[ (y + 1) \Phi_r(y + 1) - (y - 1) \Phi_r(y - 1) \right] + \frac{r^2}{2} \left[ \phi_r(y + 1) - \phi_r(y - 1) \right].$$

Using an integral formula

$$\int \phi_1(x) \phi_1(a + bx) dx = \frac{1}{\sqrt{1 + b^2}} \phi_1 \left( \frac{a}{\sqrt{1 + b^2}} \right) \Phi_1 \left( \sqrt{1 + b^2}x + \frac{ab}{\sqrt{1 + b^2}} \right) + C,$$

a joint density function and marginal density functions of $(X, Y)$ are computed as follows: $\forall x, y \in \mathbb{R}$

$$f(x, y) = \frac{1}{4} \int_{-1}^{1} \left[ \phi_r(x - t) \phi_r(y - t) + \phi_r(x - t) \phi_r(y + t) \right] dt$$

$$= \frac{1}{4 \sqrt{2}} \left[ \phi_r \left( \frac{y - x}{\sqrt{2}} \right) \left[ \Phi_r \left( \frac{y + x + 2}{\sqrt{2}} \right) - \Phi_r \left( \frac{y - x - 2}{\sqrt{2}} \right) \right] \right.$$

$$+ \phi_r \left( \frac{y + x}{\sqrt{2}} \right) \left[ \Phi_r \left( \frac{x - y + 2}{\sqrt{2}} \right) - \Phi_r \left( \frac{x - y - 2}{\sqrt{2}} \right) \right] \right) \quad (2.72)$$

$$f_1(x) = \frac{1}{2} \int_{-1}^{1} \phi_r(x - t) dt = \frac{1}{2} [\Phi_r(x + 1) - \Phi_r(x - 1)]$$

$$f_2(y) = \frac{1}{2} \int_{-1}^{1} \phi_r(y - t) dt = \frac{1}{2} [\Phi_r(y + 1) - \Phi_r(y - 1)].$$
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

Figure 2.4: Examples of Linear convolution of Diagonal Uniform distribution and Gaussian corresponding with parameters \( r = 0, 0.2, 0.5 \) and 1.

<table>
<thead>
<tr>
<th>( r )</th>
<th>Measuring Dependency Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 0 )</td>
<td>( \rho_P(X,Y) = \rho_S(X,Y) = \tau(X,Y) = \beta(X,Y) = 0 )</td>
</tr>
<tr>
<td></td>
<td>( \rho_{L_1}(X,Y) = \rho_{L_2}(X,Y) = \rho_{L_\infty}(X,Y) = \frac{1}{4} )</td>
</tr>
<tr>
<td></td>
<td>( 30\Delta(X,Y) = 0.005 ); ( \rho^<em>(X,Y) = \beta^</em>(X,Y) = 1 )</td>
</tr>
<tr>
<td>( r = 0.2 )</td>
<td>( \rho_{L_1}(X,Y) \approx 0.1702 ); ( \rho_{L_2}(X,Y) \approx 0.1651 ); ( \rho_{L_\infty}(X,Y) \approx 0.1393 )</td>
</tr>
<tr>
<td></td>
<td>( \beta(X,Y) \approx 0.2929 ); ( 4.377 \leq H(X) \leq 4.399 ) ( \Rightarrow ) ( 0.0666 \leq \beta^*(X,Y) \leq 0.0669 )</td>
</tr>
<tr>
<td>( r = 0.5 )</td>
<td>( \rho_{L_1}(X,Y) \approx 0.0429 ); ( \rho_{L_2}(X,Y) \approx 0.0409 ); ( \rho_{L_\infty}(X,Y) \approx 0.0347 )</td>
</tr>
<tr>
<td></td>
<td>( \beta(X,Y) \approx 0.0147 ); ( 4.104 \leq H(X) \leq 4.239 ) ( \Rightarrow ) ( 0.0035 \leq \beta^*(X,Y) \leq 0.0036 )</td>
</tr>
<tr>
<td>( r = 1 )</td>
<td>( \rho_{L_1}(X,Y) \approx 0.0064 ); ( \rho_{L_2}(X,Y) \approx 0.0063 ); ( \rho_{L_\infty}(X,Y) \approx 0.0066 )</td>
</tr>
<tr>
<td></td>
<td>( \beta(X,Y) \approx -0.0022 ); ( 4.026 \leq H(X) \leq 4.129 ) ( \Rightarrow ) ( \beta^*(X,Y) \approx 0 )</td>
</tr>
</tbody>
</table>

Table 2.5: Different measures of dependency for the example of figure 2.4.
The coefficients of multiple $\frac{1}{4}$ and $\frac{1}{2}$ appear to confirm that the integrals of $f(x, y)$ and $f_1(x), f_2(y)$ equal one. In this case, the Pearson’s correlation, the Spearman’s $\rho$, the Kendall’s $\tau$ and the Blomqvist’s $\beta$ are also zero. Finding exact formulae of the $L_p$-distances and the information dependency measures is impossible. Since with $r \leq 1$, $f(x, y) \approx 0$ for all $x, y \in \mathbb{R}$, $|x|, |y| > 5$, the $L_p$-distance measures and the information dependency measure can be estimated based on the following formulae, for example

$$
\rho_{L_1}(X, Y) \approx \lim_{n \to \infty} 12 \sum_{k_1=0}^{10n} \sum_{k_2=0}^{10n} \left| F\left(\frac{k_1}{n} - 5, \frac{k_2}{n} - 5\right) - F_1\left(\frac{k_1}{n} - 5\right) F_2\left(\frac{k_2}{n} - 5\right) \right|
\times \left[ F_1\left(\frac{k_1+1}{n} - 5\right) - F_1\left(\frac{k_1}{n} - 5\right) \right] \left[ F_2\left(\frac{k_2+1}{n} - 5\right) - F_2\left(\frac{k_2}{n} - 5\right) \right],
$$

(2.73)

and

$$
\begin{align*}
&h(X) \approx - \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{10n} f_1\left(\frac{k_1}{n} - 5\right) \log f_1\left(\frac{k_1}{n} - 5\right) \\
&h(X, Y) \approx - \lim_{n \to \infty} \frac{1}{n^2} \sum_{k_1=0}^{10n} \sum_{k_2=0}^{10n} \left[ f\left(\frac{k_1}{n} - 5, \frac{k_2}{n} - 5\right) \right] \log f\left(\frac{k_1}{n} - 5, \frac{k_2}{n} - 5\right) \\
&I(X, Y) \approx 2h(X) - h(X, Y).
\end{align*}
$$

(2.74)

The estimations of the $L_p$-distances and the normalized information dependency measures are given in the table 2.5 corresponding the cases $r = 0.2$, $0.5$ and $1$. The values of dependency getting from the measures of dependency for the case $r = 0$ are also given in this table.

### 2.4.3 Linear convolution of Circular Uniform distribution and Gaussian

Let $\mathbf{U}$ be a bivariate random vector which has a uniform distribution on the circle $S_2(0, 1) = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$. The random vector $\mathbf{U}$ has a joint distribution function and marginal distribution functions:

$$
F(x, y) = \begin{cases} 
\frac{3}{4} - \frac{1}{2\pi} (\arccos x + \arccos y) & x^2 + y^2 \leq 1 \\
1 - \frac{1}{2} (\arccos x + \arccos y) & x^2 + y^2 > 1, \ x, y > 0 \\
1 - \frac{1}{2} \arccos x & x^2 + y^2 > 1, \ x \leq 0, y > 0 \\
1 - \frac{1}{2} \arccos y & x^2 + y^2 > 1, \ x > 0, y \leq 0 \\
0 & x^2 + y^2 > 1, \ x, y \leq 0 
\end{cases}
$$

$$
F_1(x) = 1 - \frac{1}{\pi} \arccos x, \quad F_2(y) = 1 - \frac{1}{\pi} \arccos y,
$$
2. MEASURES OF DEPENDENCY, INFORMATION DEPENDENCY

Figure 2.5: Examples of Linear convolution of Circular Uniform distribution and Gaussian corresponding with parameters $r = 0$, $0.2$, $0.5$ and $1$.

\[ \forall r = 0, 0.2, 0.5 \text{ and } 1. \]

Table 2.6: The dependency and estimating dependency values of the examples in figure 2.5.
where we denote \( \arccos x = \pi, \forall x \geq 1 \) and \( \arccos x = 0, \forall x \leq -1 \). Let \( N_r \) be a bivariate Gaussian random vector independent with \( U \), and let

\[
(X, Y) = U + N_r.
\]

Figure 2.5 illustrates 2000 independent samples of \( (X, Y) \) corresponding \( r = 0, 0.2, 0.5 \) and 1. The dependent values of \( (X, Y) \) expressed by the measures of dependency for the case \( r = 0 \) are collected in table 2.6. When \( r > 0 \), a joint distribution function and marginal distribution functions of \( (X, Y) \) are given by:

\[
F(x, y) = \frac{1}{2\pi} \int_{0}^{2\pi} \Phi_r(x - \sin \theta)\Phi_r(y - \cos \theta) d\theta
\]

\[
F_1(x) = \frac{1}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1 - t^2}} \Phi_r(x - t) dt
\]

\[
F_2(y) = \frac{1}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1 - t^2}} \Phi_r(y - t) dt, \quad \forall x, y \in \mathbb{R},
\]

and a following joint density function and marginal density functions are given by:

\[
f(x, y) = \frac{1}{2\pi} \int_{0}^{2\pi} \phi_r(x - \sin \theta)\phi_r(y - \cos \theta) d\theta
\]

\[
f_1(x) = \frac{1}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1 - t^2}} \phi_r(x - t) dt
\]

\[
f_2(y) = \frac{1}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1 - t^2}} \phi_r(y - t) dt, \quad \forall x, y \in \mathbb{R}.
\]

Similarly to the linear convolution of diagonal uniform distribution and Gaussian example, we can not determine the exact formulae of \( L_p \)-distance measures and the information dependency measure. Therefore, we use the numerical estimations (2.73) to estimate \( L_p \)-distance measures, and (2.74) for the information dependency measure. The estimations of the cases \( r = 0.2, 0.5, \) and 1 are also presented in table 2.6. The Pearson’s correlation, the Spearman’s \( \rho \), the Kendall’s \( \tau \) and the Blomqvist’s \( \beta \) of the linear convolution of circular uniform distribution and Gaussian random vectors equal zero for all cases of \( r \).

The expressions of dependency in the second and third examples show us that some traditional measures of dependency such as the Pearson’s correlation, the Spearman’s \( \rho \), the Kendall’s \( \tau \) and the Blomqvist’s \( \beta \) work not well. In the cases \( r = 0 \) of these examples, in which the random variables \( X \) and \( Y \) are strictly dependent, the \( L_p \)-distances measures also work not well since they express very small dependent values, \( \frac{1}{4} \). The normalized information dependency will express the value 1 for these cases because it satisfies the axiom (E) of Rényi Axioms. The \( L_p \)-distances and the normalized information dependency
measures educe the degree of dependency in the cases $r > 0$ in all the three examples with the meaning that the larger $r$ gets, the smaller their expressions of dependency are. The Hoeffding’s $\Delta$ and the maximal correlation measures are very difficult to determine not only the theoretical values but also the approximating values from the samples. In conclusion, to finish this chapter, we give the personal idea that to measure the dependency of a group of random variables we would like to use the normalized information dependency measure. The reason is that: firstly, this measure satisfies the Rényi axioms; secondly, it can work well in the multivariate case; and finally, its expressions of dependent values of a random vector can be approximated exactly based on the samples of this random vector. The third reason will be studied carefully in the next chapter.
Chapter 3

Information Dependency Estimation

Chapter 2 presents how the information dependency is used for measuring the dependency of a group of random variables from the theoretical perspective. In this chapter, we would like to prove that the information dependency not only works in the theoretical aspect, but also can work in the practical aspect. Concretely, we would like to study ways to find out the theoretical value of the information dependency of a group of random variables based on its samples.

The problem of estimating the information dependency is stated as follows: Given $m$ random variables $X_1, \ldots, X_m$ on a probability space $(\Omega, \mathcal{B}, P)$. A random vector $X = (X_1, \ldots, X_m)$ has a joint distribution $\mu$ and marginal distributions $\mu_1, \ldots, \mu_m$. Denoting $\nu = \mu_1 \times \cdots \times \mu_m$ be a product-marginal distribution of $X$, the random vector $X$ has the information dependency

$$I(X) = \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)},$$

where $\mathcal{P}(m)$ is a collection of all finite and measurable partitions of $\Omega^m$. For each $n \in \mathbb{N}$, let $X^n = \{X_1, X_2, \ldots, X_n\}$ be $n$ samples of $X$, i.e. $X_1, X_2, \ldots, X_n$ are $n$ independent and identically distributed (i.i.d.) random vectors with $X_j \sim X$, $j = 1, \ldots, n$. We construct a function $\hat{I}$ of $X^n$ such that it is an estimator for the information dependency $I(X)$.

**Definition 3.1** ([Shao, 1999], p.102). Let $\hat{I}(X^n)$ be an estimator for $I(X)$.

(i) $\hat{I}(X^n)$ is said “unbiased” if $E[\hat{I}(X^n)] = I(X)$.

(ii) For any two unbiased estimators $\hat{I}^{(1)}$ and $\hat{I}^{(2)}$, $\hat{I}^{(1)}$ is said “better” than $\hat{I}^{(2)}$ if

$$\text{Var}[\hat{I}^{(1)}(X^n)] \leq \text{Var}[\hat{I}^{(2)}(X^n)].$$
(iii) \( \hat{I}(X^n_1) \) is said “weak consistency” if for all \( \epsilon > 0 \),
\[
\lim_{n \to \infty} P(|\hat{I}(X^n_1) - I(X)| > \epsilon) = 0.
\]

(iv) \( \hat{I}(X^n_1) \) is said “strong consistency” if
\[
\lim_{n \to \infty} \hat{I}(X^n_1) = I(X) \quad \text{P-a.s.}
\]

(v) Given a constant \( p > 0 \), \( \hat{I}(X^n_1) \) is said “\( L_p \)-consistency” if
\[
\lim_{n \to \infty} E|\hat{I}(X^n_1) - I(X)|^p = 0.
\]

From the notions in Definition 3.1, we wish to find the “best and consistent” estimator for \( I(X) \). Section 3.1 should present some popular information dependency estimators such as the Gessaman’s partition estimator [Gessaman, 1970], the Tree-quantization partition estimator [Breiman et al., 1984; Devroye et al., 1996], the Darbellay-Vajda partition estimator [Darbellay and Vajda, 1999], the \( k \)-Nearest neighbor estimators [Kraskov et al., 2004; Pöczos and Schneider, 2011], and the Gaussian kernel estimator [Botev et al., 2010; Moon et al., 1995; Steuer et al., 2002]. In the next section, we will give some numerical experiments of these estimators and discuss their implements based on the results of these experiments. The practical results in section 3.2 show us that the Darbellay-Vajda partition estimator is the best estimator in the low dimension (\( m \leq 3 \)), and the \( k \)-Nearest neighbor estimators are the best estimators in the high dimension (\( m \geq 4 \)). The Darbellay-Vajda partition estimator belongs to a data-dependent partition estimator class. The final section of this chapter would like to study the conditions to confirm that some data-dependent partition estimations such as the Gessaman’s partition estimator, the Tree-quantization partition estimator, and specially the Darbellay-Vajda partition estimator are strongly consistent. There are the main results of our works.

Until now, among the nonparametric methods to estimate the information dependency, there have been only some data-dependent partition estimators which are consistent for the information dependency. Finding out the conditions of random vector \( X \) or the parameters of estimators such that the \( k \)-nearest neighbor estimators and the Gaussian kernel estimator are consistent for the information dependency is still an unsolved problem. We will discuss this problem in section 3.1.
3.1 Estimations

The definition of the information dependency

\[ I(X) = \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)}, \tag{3.1} \]

where \( \mathcal{P}(m) \) is a collection of all finite and measurable partitions of \( \Omega^m \), suggests us the natural way to estimate the information dependency: Based on the samples \( X_1^n \) of the random vector \( X \), we would like to find an optimal partition \( \pi_n^* \in \mathcal{P}(m) \) such that

\[ \hat{I}(X^n_1) = \sum_{A \in \pi_n^*} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \text{ is a consistent for } I(X), \tag{3.2} \]

where \( \mu_n \) and \( \nu_n \) which are introduced in next subsection are empirical distributions of \( \mu, \nu \). Since the estimator \( \hat{I}(X^n_1) \) is a function of \( X^n_1 \), the optimal partition \( \pi_n^* \) should be a function of \( X^n_1 \). Therefore, this method of estimation is often called a data-dependent partition estimation.

When \( X \) is a continuous random vector, i.e. \( X \) has a joint density function \( f \) and a product-marginal density function \( g \), the Proposition 2.9 and Definition 2.7 give other formulae of the information dependency as follows:

\[ I(X) = \int_{\mathbb{R}^m} f(x) \log \frac{f(x)}{g(x)} dx = \sum_{j=1}^m h(X_j) - h(X), \tag{3.3} \]

where \( X_1, \ldots, X_m \) are \( m \) elements of \( X \) and \( h \) is the differential entropy operator. If these density functions \( f \) and \( g \) have some conditions such as continuous and bounded, etc., there are some useful nonparametric methods to estimate density functions and differential entropies such as the k-nearest neighbor estimations and the Gaussian kernel estimations. Thus, the formula (3.3) infers that with some conditions of the density functions \( f \) and \( g \), the k-nearest neighbor method and the Gaussian kernel method can be proposed to estimate the information dependency.

Now, we would like to present how the data-dependent partition, the k-nearest neighbor, and the Gaussian kernel estimation methods are used for estimating the information dependency. After, we will present some numerical experiments of these estimators corresponding with some examples presented in the section 2.4, Examples.

3.1.1 Data-dependent partition estimations

Let \((\Omega, \mathcal{B})\) be a measurable space, and denote \((\Omega^m, \mathcal{B}^m)\) be a product measurable space of \( m \) spaces \((\Omega, \mathcal{B})\). A set \( A \) is said a measurable rectangle (or simply rectangle) in \( \mathcal{B}^m \) if
3. INFORMATION DEPENDENCY ESTIMATION

there exist unique \( m \) measurable sets \( A_1, \ldots, A_m \in \mathcal{B} \) such that \( A = A_1 \times \cdots \times A_m \).

**Definition 3.2.** A collection of sets \( \pi = \{A_1, \ldots, A_r\} \) is said a partition of \( \Omega^m \) if (i) \( A_1, \ldots, A_r \in \mathcal{B}^m \), (ii) \( A_i \cap A_j = \emptyset \) for all \( i \neq j \), and (iii) \( \bigcup_{i=1}^r A_i = \Omega^m \). The partition \( \pi \) is said rectangle partition if \( A_1, \ldots, A_r \) are rectangles in \( \mathcal{B}^m \).

Let \( \mathcal{P}(m) \) be a collection of all finite partitions of \( \Omega^m \), and denote \( \mathbb{R}^{m \times n} \) be a space of all \( n \) points in \( \mathbb{R}^m \), \( n \in \mathbb{N} \).

**Definition 3.3.** For each \( n \in \mathbb{N} \), a mapping \( \pi_n \) from \( \mathbb{R}^{m \times n} \) to \( \mathcal{P}(m) \) is said a (data-dependent) partition rule of order \( n \), i.e. for any collection of \( n \) points in \( \mathbb{R}^m \), \( x^n = \{x_1, \ldots, x_n\} \), \( \pi_n(x^n) \) is a partition of \( \Omega^m \). The partition \( \pi_n(x^n) \) is said an \( x^n \)-dependent partition. A sequence \( \{\pi_n\}_{n \in \mathbb{N}} \), which each \( \pi_n \) is a partition rule of order \( n \), is said a (data-dependent) partition rule. If for all \( x^n \), \( \pi_n(x^n) \) is a rectangle partition, then \( \{\pi_n\}_{n \in \mathbb{N}} \) is said a rectangle partition rule.

Let \( X = (X_1, \ldots, X_m) \) be a random vector on \((\Omega^m, \mathcal{B}^m)\) with a joint distribution \( \mu \) and \( m \) marginal distributions \( \mu_1, \ldots, \mu_m \). Let \( X^n_i = \{X_1, \ldots, X_n\} \) be \( n \) samples of \( X \), i.e. \( X_i = (X_{i1}, \ldots, X_{im}) \) is the random vector with \( X_i \sim X \), \( i = 1, \ldots, n \). An empirical joint distribution \( \mu_n \) of samples \( X^n_i \) is defined as follows: \( \forall A \in \mathcal{B}^m \)

\[
\mu_n(A) = \frac{1}{n} \sum_{i=1}^n 1\{x_i \in A\},
\]

where \( 1_B = 1 \) if \( B \) occurs, and \( 0 \) otherwise. **Empirical marginal distributions** and an empirical product-marginal distribution of \( X^n_i \) are defined as follows: \( \forall A_j \in \mathcal{B}, j = 1, \ldots, m \)

\[
\mu_{n,j} = \frac{1}{n} \sum_{i=1}^n 1\{x_{ij} \in A_j\}, \quad \text{and} \quad \nu_n(A) = \prod_{j=1}^m \mu_{n,j}(A_j),
\]

where \( A = A_1 \times \cdots \times A_m \). The **Law of Large Number** shows us that when \( n \) tends to infinity, these empirical distributions converge to the corresponding theoretical distribution, i.e., \( \forall A, B \in \mathcal{B}^m, B \) is a rectangle, \( \forall C \in \mathcal{B}, j = 1, \ldots, m, \)

\[
\lim_{n \to \infty} \mu_n(A) = \mu(A), \quad \lim_{n \to \infty} \nu_n(B) = \nu(B) \quad \text{and} \quad \lim_{n \to \infty} \mu_{n,j}(C) = \mu_j(C).
\]

Note that, the results \( \mu_n(A) \to \mu(A), \nu_n(A) \to \nu(A) \) for all rectangle \( A \) as \( n \) tends to infinity can not conclude that for all rectangle partition \( \pi \in \mathcal{P}(m) \),

\[
\sum_{A \in \pi} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \to I(X), \quad \mathbb{P}\text{-a.s.}
\]

Thus, since the empirical product-marginal distribution \( \nu_n \) is only defined on the space of all rectangle sets, the motivation of the data-dependent partition estimation method
3.1. ESTIMATIONS

is to find a rectangle partition rule \( \{ \pi_n \}_{n \in \mathbb{N}} \) such that

\[
\hat{I}(X^n_1) = \sum_{A \in \pi_n(X^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)}
\]
is a strong consistency for \( I(X) \), \( \hat{I}(X) \) is a strong consistency for \( I(X) \), \( (3.6) \)

or

\[
\lim_{n \to \infty} \sum_{A \in \pi_n(X^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} = I(X) \quad \mathbb{P}-a.s., \quad (3.7)
\]

where \( \mu_n, \nu_n \) are the empirical joint and product-marginal distributions of \( X^n_1 \). To study how to construct a rectangle partition rule such that \( (3.7) \) occurs, we would like to introduce four popular partition rules as follows the bin-structure partition rule, the Gessaman’s partition rule, the tree-quantization partition rule, and the Darbellay-Vajda partition rule, and find the conditions of these partition rules to make their estimators are strongly consistent.

**Bin-structure partition rule:** The bin-structure partition rule is a classical partition method since this rule is very simple and useful to estimate some quantities such as density functions, entropies, etc. However, until now, we have not found any result to prove that the bin-structure partition rule is a consistency for estimating the information dependency yet. Let \( X_1, \ldots, X_n \) be a sequence of \( n \) i.i.d. random vectors, and \( r_n \) be a positive integer number. The bin-structure partition rule, \( \{ \pi^n_n \}_{n \in \mathbb{N}} \), construct a \( X^n_1 \)-dependent partition which has \( r^n_n \) cells as follows:

**Step 1.** For each coordinate \( j \in \{ 1, \ldots, m \} \), let \( X_{1,j}, \ldots, X_{n,j} \) be projections of \( X_1, \ldots, X_n \) into the coordinate \( j \), and

\[
a_j = \min_{1 \leq i \leq n} X_{i,j}, \quad b_j = \max_{1 \leq i \leq n} X_{i,j}.
\]

We partition the \( j \)-th coordinate line by \( r_n \) intervals

\[
\{(A^{(k)}_j)^{r_n-1} : k = 0, \ldots, r_n-1\} = \left\{(-\infty, a_j], (a_j, a_j + \frac{1}{r_n-2}(b_j - a_j)], \ldots, (a_j + \frac{r_n-3}{r_n-2}(b_j - a_j), b_j], (b_j, +\infty)\right\}.
\]

**Step 2.** For all sequence \( (k_1, \ldots, k_m) \) with \( k_a \in \{ 1, \ldots, r_n \} \), let \( B_{(k_1, \ldots, k_m)} = A^{(k_1)}_1 \times \cdots \times A^{(k_m)}_m \), a rectangle on \( \mathbb{R}^m \).

**Step 3.** The partition \( \pi^n_n(X^n_1) \) is a collection of all rectangles \( \{ B_{(k_1, \ldots, k_m)} : k_1, \ldots, k_m \in \{1, \ldots, r_n\} \} \).
An example of the bin-structure partition rule for 64 samples on the space $\mathbb{R}^2$ with $r_n = 8$ is given in the figure 3.1.

**Tree-structure partition:** Given an integer number $\kappa \geq 2$, a graph $T$ which has no circle, only one node with degree $\kappa$ (root node), and the remaining nodes with degree $\kappa + 1$ (internal nodes) or degree 1 (terminal nodes) is called a $\kappa$-order tree [Devroye et al., 1996, Ch. 20]. Let $\text{depth}(t)$ denote the depth of $t \in T$, the number of edges that connect $t$ with the root of $T$, and $\mathcal{L}(T)$ be the collection of all terminal nodes of $T$. If there is an edge connecting $u$ with $v$, $u, v \in T$ and $\text{depth}(u) = \text{depth}(v) - 1$, $v$ is called posterity of $u$.

**Definition 3.4.** Given an integer number $\kappa \geq 2$, a $\kappa$-order tree $T$ is said a $\kappa$-order tree of partition of $\mathbb{R}^m$ if it satisfies

i. its root is $\mathbb{R}^m$,

ii. its terminal nodes are rectangles on $\mathbb{R}^m$,

iii. any its internal node or its root is a rectangle which is a union of all its posterity rectangles,

iv. and for any node $u, v \in T$, $u \neq v$, if $\text{depth}(u) = \text{depth}(v)$ then $u, v$ are disjointed.

In this case, the collection of internal nodes of $T$, $\mathcal{L}(T)$, is called a $\kappa$-order tree-structure partition of $\mathbb{R}^m$.

**Definition 3.5.** A partition rule of order $n$, $\pi_n$, is said a ($\kappa$-order) tree-structure partition rule of order $n$, if for all $n$ points $\mathbf{x}_n^1 \in \mathbb{R}^{m \cdot n}$, $\pi_n(\mathbf{x}_n^1)$ are $\kappa$-order tree-structure partitions of $\mathbb{R}^m$. $\{\pi_n\}_{n \in \mathbb{N}}$ is called a ($\kappa$-order) tree-structure partition rule if for any $n \in \mathbb{N}$, $\pi_n$ is a ($\kappa$-order) tree-structure partition rule of order $n$.

**Gessaman’s partition rule:** The Gessaman’s partition rule is introduced and described in [Gessaman, 1970; Lugosi and Nobel, 1996]. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be $n$ i.i.d. random vectors, and $k_n$ be a positive integer number. Given an order of the coordinates $(j_1, \ldots, j_m)$, a permutation of $(1, \ldots, m)$. The Gessaman’s partition rule, $\{\pi_n^0\}_{n \in \mathbb{N}}$, constructs a $\mathbf{X}_n^0$-dependent partition such that each cell of this partition contains nearly $k_n$ samples of $\mathbf{X}_n^0$ as follows: Let $r_n = \lfloor (n/k_n)^{1/m} \rfloor$, $\{\pi_n^0\}_{n \in \mathbb{N}}$ be an $r_n$-order tree-structure partition rule which its $r_n$-order tree of partition corresponding to $\mathbf{X}_n^0$ is constructed by the following steps:

Step 1. Let the root of tree, denoted $t_{(0,0)}$, be the space $\mathbb{R}^m$. 

58
Step 2. Choosing the first coordinate $j_1$, let $X_{1,j_1}, \ldots, X_{n,j_1}$ be projections of $X_1, \ldots, X_n$ into $j_1$-th coordinate, and let $\{X_{j_1}^{(1)}, \ldots, X_{j_1}^{(n)}\}$ be a permutation of $\{X_{1,j_1}, \ldots, X_{n,j_1}\}$ such that $X_{j_1}^{(1)} \leq \cdots \leq X_{j_1}^{(n)}$.

Step 3. Dividing the $j_1$-th-coordinate line into $r_n$ intervals

$$\{A_u\}_{u=0}^{r_n-1} = \left\{(-\infty, X_{j_1}^{(s_n)}], (X_{j_1}^{(s_n)}, X_{j_1}^{(2s_n)}], \ldots, (X_{j_1}^{(|r_n-1| s_n)}), +\infty\right\}$$

where $s_n = |t_{(0,0)}|/r_n$ with $|t|$ be the number of samples in the rectangle $t$. Partitioning the space $\mathbb{R}^m$ by $r_n$ parts $\{t_{(1,u)}\}_{u=0}^{r_n-1}$ where $t_{(1,u)} = A_u \times \mathbb{R}^{m-1}$, $u = 0, \ldots, r_n-1$. These new rectangles are chosen as posterity nodes of $t_{(0,0)}$.

Step 4. Choosing the second $j_2$-th coordinate and continue approach in Steps 2 and 3 for new rectangles $t_{(1,0)}, \ldots, t_{(1,r_n-1)}$ to make their posterities $\{t_{(2,u)}\}_{u=0}^{r_n^2-1}$. Noting that, when working with the rectangle $t_{(1,u)}$, we only consider the samples contained in this rectangle, and replace $s_n = |t_{(1,u)}|/r_n$, $u = 0, \ldots, r_n - 1$.

Step 5. Iterating this approach until the last coordinate, we get an $r_n$-order tree partition of $\mathbb{R}^m$. The collection of all terminal nodes of this tree is said the $X^n_{1}$-dependent Gessaman’s partition.

An example of the Gessaman’s partition rule for 64 samples on the space $\mathbb{R}^2$, choosing the order coordinate $(1,2)$ and $k_n = 1$ is given in the figure 3.1. The conditions to make the Gessaman’s partition rule which is a strong consistency for the information dependency are given in the following result.

**Proposition 3.1.** Let $\{\pi^n_n\}_{n \in \mathbb{N}}$ be the Gessaman’s partition rule associated with a sequence positive integer numbers $\{k_n\}_{n \in \mathbb{N}}$. Let $X$ be a random vector, and let $X_1, X_2, \ldots$ be a sequence of i.i.d. random vectors with $X_n \sim X, \forall n \in \mathbb{N}$. If there exists a real number $\tau \in \left(\frac{1}{2}, 1\right)$ such that $(k_n) \approx n^{\frac{1}{1+\tau}}$, then

$$\widehat{J}_G(X^n_1) = \sum_{A \in \pi^n_1(X^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \text{ is a strong consistency for } J(X). \quad (3.8)$$

The proof of the Proposition 3.1 is given in section 3.3.

**Tree-quantization partition rule:** The tree-quantization partition rule is introduced and described in [Breiman et al., 1984; Devroye et al., 1996; Nobel, 2002]. This is a binary (2-order) tree-structure partition rule. Let $X_1, \ldots, X_n$ be i.i.d. random vectors, and $k_n$ be a positive integer number. Given an order of the coordinates $(j_1, \ldots, j_m)$, a permutation of $(1, \ldots, m)$. The tree-quantization partition rule, $\{\pi^n_1\}_{n \in \mathbb{N}}$, constructs a
3. INFORMATION DEPENDENCY ESTIMATION

\( \mathbf{X}_1^n \)-dependent partition such that each cell of this partition contains nearly \( k_n \) samples. The construction of its binary tree of partition is described in the following steps:

Step 1. Let the root of tree, denoted \( t_{(0,0)} \), be the space \( \mathbb{R}^m \).

Step 2. Choosing the first coordinate \( j_1 \), let \( X_{1,j_1}, \ldots, X_{n,j_1} \) be projections of \( \mathbf{X}_1, \ldots, \mathbf{X}_n \) into the \( j_1 \)-th coordinate. Let \( \{X_{j_1}^{(1)}, \ldots, X_{j_1}^{(n)}\} \) be a permutation of \( \{X_{1,j_1}, \ldots, X_{n,j_1}\} \) such that \( X_{j_1}^{(1)} \leq \cdots \leq X_{j_1}^{(n)} \).

Step 3. Using a hyper-plane

\[
H_{j_1}(\mathbf{X}_1^n) = \{ \mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m : x_{j_1} \leq X_{j_1}^{(\lfloor n/2 \rfloor)} \}
\]

to divide the space \( \mathbb{R}^m \) into two statistically equivalent parts with respect to \( j_1 \)-th coordinate, denoted by \( t_{(1,0)} \) and \( t_{(1,1)} \). The new rectangles \( t_{(1,0)} \) and \( t_{(1,1)} \) are chosen as posterity nodes of \( t_{(0,0)} \).

Step 4. Choosing the next coordinate, \( j_2 \)-th coordinate, and continue approach in Steps 2 and 3 for rectangles \( t_{(1,0)} \) and \( t_{(1,1)} \) to make their posterities \( t_{(2,0)}, t_{(2,1)}, t_{(2,2)} \) and \( t_{(2,3)} \). Note that, when working on the rectangle \( t \), we only consider the samples contained in this rectangle.

Step 5. After \( k \) steps, we will have a binary tree with the root node \( t_{(0,0)} \), the internal nodes \( \{t_{(u,v)} : 0 < u < k, 0 \leq v \leq 2^j - 1 \} \), and the terminal nodes \( \{t_{(k,v)} : 0 \leq v \leq 2^k - 1 \} \). The iteration is stopped when we guarantee that a minimum number of samples in the terminal nodes \( t_{(k,v)} \) is larger than \( k_n \).

An example of the tree-quantization partition rule for 64 samples on the space \( \mathbb{R}^2 \), choosing the order coordinate \( (1, 2) \) and \( k_n = 1 \) is given in the figure 3.1. The conditions to confirm that the tree-quantization partition rule is a strong consistency for the information dependency are given in the following proposition.

**Proposition 3.2.** Let \( \{\pi_n^T\}_{n \in \mathbb{N}} \) be the tree-quantization partition rule associated with a sequence integer number \( \{k_n\}_{n \in \mathbb{N}} \). Let \( \mathbf{X} \) be a random vector, and let \( \mathbf{X}_1, \mathbf{X}_2, \ldots \) be a sequence of i.i.d. random vectors with \( \mathbf{X}_n \sim \mathbf{X} \), \( \forall n \in \mathbb{N} \). If there exists a real number \( \tau \in (\frac{1}{3}, 1) \) such that \( (k_n) \approx (n^{\frac{1+\tau}{2}}) \), then

\[
\hat{I}_\tau(\mathbf{X}_n^T) = \sum_{A \in \pi_n^T(\mathbf{X}_n^T)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \quad \text{is a strong consistency for} \quad I(\mathbf{X}). \tag{3.9}
\]

The proof of this proposition is given in section 3.3.
Figure 3.1: Illustration of the Data-dependent partitions for given 64 samples in $[0,1]^2$. The grey rectangles in Darbellay-Vajda sub-figure are the local independent cells.
3. INFORMATION DEPENDENCY ESTIMATION

**Darbelay-Vajda partition rule:** The Darbelay-Vajda partition rule was introduced by Darbelay and Vajda [1999]. Let $X_1, \ldots, X_n$ be $n$ i.i.d. random vectors on $\mathbb{R}^m$, $h_n$ be a positive integer number and $\epsilon_n$ be a positive real number. Before describing the Darbelay-Vajda partition rule of $X^n_t$, we introduce the two following definitions. Let $A = \times_{j=1}^m (a_j,b_j)$, $a_j < b_j$, be a rectangle on $\mathbb{R}^m$. For each coordinate $j$, assuming that the rectangle $[a_j,b_j) \times \mathbb{R}^{-1}$ contains $n_j$ samples $X_{i_1}, \ldots, X_{i_{n_j}}$. Finding $X_j^{(1)}, \ldots, X_j^{(n_j)}$, a permutation of $X_{i_1}, \ldots, X_{i_{n_j}}$, such that $X_j^{(1)} \leq \cdots \leq X_j^{(n_j)}$. Given an integer number $2 \leq d \leq h_n$, we partition the interval $[a_j,b_j)$ by $d$ intervals

$$B_j^{(1)} = [a_j, X^{[n_j/d]}), B_j^{(2)} = [X^{[n_j/d)}, X^{(2[n_j/d])}), \ldots, B_j^{(d)} = [X^{((d-1)n_j/d)}), b_j)$$

and the rectangle $A$ by a collection of $d^m$ rectangles $\{B_{(u_1,\ldots,u_m)} : u_v \in \{1,\ldots,d\}, v = 1,\ldots,m\}$, which

$$B_{(u_1,u_2,\ldots,u_m)} = B_1^{(u_1)} \times B_2^{(u_2)} \times \cdots \times B_m^{(u_m)}.$$

This partition is said a $d$-marginal statistically equivalent partition of $A$.

A rectangle $A$ is said local independent associated with an integer number $h_n > 2$ and real positive number $\epsilon_n$ if for any $2 \leq d \leq h_n$, and for any $B = \{B_{(u_1,\ldots,u_m)}\}$ is $d$-marginal statistically equivalent partitions of $A$,

$$\left| \sum_{B \in \mathcal{B}} \mu_n(B) \log \frac{\mu_n(B)}{\nu_n(B)} - \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \right| < \epsilon_n.$$

Given $n$ sample points $X^n_t$, the Darbelay-Vajda partition rule $\{T^n_t\}_{n \in \mathbb{N}}$ of $X^n_t$ is an $2^m$-order tree-structure partition. This tree depends on two parameters $h_n \geq 2$ and $\epsilon_n > 0$, and it is constructed as follows:

**Step 1.** Let the root of tree, denoted $t_{(0,0)}$, be the space $\mathbb{R}^m$.

**Step 2.** For any terminal node $t_{(u,v)}$ of the tree, if $t_{(u,v)}$ is local independent or depth $t_{(u,v)} = h_n$, it does not continue partitioning. In other cases, $t_{(u,v)}$ is partitioned by its $2$-marginal statistically equivalent partition. The cells of the $2$-marginal statistically equivalent partition of $t_{(u,v)}$ are chosen as posterities of $t_{(u,v)}$.

**Step 3.** We iterate the Step 2 for the new terminal nodes. This iteration will stop when all the terminal nodes of the tree are local independent or have the depth $h_n$.

The Darbelay-Vajda partition $T^n_t(X^n_t)$ is a collection of all terminal nodes of the above tree. An example of the Darbelay-Vajda partition rule for 64 samples on the space $\mathbb{R}^2$, choosing $h_n = 4$ and $\epsilon_n = 10^{-3}$ is given in the figure 3.1. The conditions to confirm that
the Darbellay-Vajda partition rule is a strong consistency for the information dependency are given in the following result.

**Proposition 3.3.** Let \( \{\pi_n^{DV}\}_{n \in \mathbb{N}} \) be the Darbellay-Vajda partition rule associated with a sequence integer \( \{h_n\}_{n \in \mathbb{N}} \) and a sequence positive real number \( \{\epsilon_n\}_{n \in \mathbb{N}} \). Let \( X \) be a random vector, and let \( X_1, X_2, \ldots \) be a sequence of i.i.d. random vectors with \( X_n \sim X, \forall n \in \mathbb{N} \). If \( (\epsilon_n) \approx \left( n^{-\frac{1}{2}} \right) \) and there exists real numbers \( \tau \in \left( \frac{1}{3}, 1 \right) \) such that \( (2^{m h_n}) \approx \left( n^{\frac{1}{2} - \tau} \right) \), then

\[
\hat{I}_{DV}^{\pi_n}(X^n) = \sum_{A \in \pi_n^{DV}(X^n)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)}
\]

is a strong consistency for \( I(X) \). \( (3.10) \)

The proof of this proposition is given in section 3.3.

### 3.1.2 k-Nearest neighbor estimations

Let \( \rho(x, y) \) represent the \( m \)-dimensional Euclidean distance on \( \mathbb{R}^m \)-space

\[
\rho(x, y) = \left[ \sum_{j=1}^{m} (x_j - y_j)^2 \right]^{\frac{1}{2}},
\]

where \( x = (x_1, \ldots, x_m) \), \( y = (y_1, \ldots, y_m) \in \mathbb{R}^m \). For each point \( x \), an \( m \)-dimensional hypersphere with a center \( x \) and a radius \( r \), \( S_x(r) = \{ y \in \mathbb{R}^m : \rho(x, y) < r \} \) has a volume, which is denoted \( |S_x(r)| \),

\[
|S_x(r)| = c_m r^m \quad \text{where} \quad c_m = \frac{\pi^\frac{m}{2}}{\Gamma(\frac{m}{2} + 1)}.
\]

and \( \Gamma \) is the gamma function,

\[
\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \quad z \in \mathbb{C} \quad \text{with its real part} \ \Re(z) > 0,
\]

\[
\Gamma(z + 1) = z\Gamma(z), \quad \Gamma(1) = 1 \quad \Rightarrow \quad \Gamma(k) = (k-1)! \quad \forall k \in \mathbb{N}, \ k \geq 1.
\] \( (3.11) \)

**Estimation of the density function:** Let \( X \) be a continuous random vector with density function \( f \), Lebesgue [1910] gave the following result as a useful tool to estimate the density function \( f \).

**Theorem 3.1** (Lebesgue [1910]). If \( f \in L_1(\mathbb{R}^m) \), then for almost surely \( x \in \mathbb{R}^m \) and any sequence of hyperspheres \( S_x(r_k) \) of radius \( r_k \to 0 \),

\[
\lim_{k \to \infty} \frac{1}{|S_x(r_k)|} \int_{S_x(r_k)} f(y) dy = f(x).
\] \( (3.12) \)
3. INFORMATION DEPENDENCY ESTIMATION

Note that, the expression “for almost surely $x$” means that the corresponding assertion holds with respect to a probability measure generated by the density function $f$. From the result of Theorem 3.1, Loftsgaarden and Quesenberry [1965] proposed an interested estimator of density function as follows: Let $\{k_n\}_{n \in \mathbb{N}}$ be a non-decreasing of positive integer numbers such that

$$\lim_{n \to \infty} k_n = \infty \quad \text{and} \quad \lim_{n \to \infty} \frac{k_n}{n} = 0. \quad (3.13)$$

For each $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be a sequence of $n$ i.i.d. random vectors with $X_i \sim X$, $i = 1, \ldots, n$. Then the density function $f$ at the point $x \in \mathbb{R}^m$ is estimated by the following formula:

$$\hat{f}_n(x) = \frac{k_n - 1}{n} \frac{1}{|S_x(r_{k_n})|} = \frac{k_n - 1}{nc_m r_{k_n}^m}, \quad (3.14)$$

where $r_{k_n}$ is the Euclidean distance from $x$ to the $k_n$-th closest $X_i$ to $x$ ($X_i \neq x$).

**Proposition 3.4** (Loftsgaarden and Quesenberry [1965]). If $f$ is continuous then the density estimator $\hat{f}_n(x)$ is a weakly consistent for $f(x)$.

**Proof.** Let $U_{k_n} = \mathbb{P}[S_x(r_{k_n})]$, some results in [Wald, 1943] inferred that the probability density function of $U_{k_n}$ is given by

$$\frac{\Gamma(n + 1)}{\Gamma(k_n)\Gamma(n - k_n + 1)} U_{k_n}^{k_n - 1} (1 - U_{k_n})^{n - k_n} dU_{k_n}, \quad (3.15)$$

a beta density with parameters $k_n$ and $(n - k_n + 1)$.

Using the Chebyshev’s inequality, the equation (3.15) yields $U_{k_n} \to 0$ in probability. However, this happens only if $|S_x(r_{k_n})| \to 0$ in probability which in turn occurs only if $r_{k_n} \to 0$ in probability. From Theorem 3.1, $r_{k_n} \to 0$ in probability infers that

$$\frac{U_{k_n}}{|S_x(r_{k_n})|} \to f(x) \quad \text{in probability.}$$

In other words, continuing apply the Chebyshev’s inequality we have

$$\frac{n}{k_n - 1} U_{k_n} \to 1 \quad \text{in probability.}$$

Thus, $\hat{f}_n(x) \to f(x)$ in probability. \hfill $\square$

Once $n$ and $k_n$ are given, the estimator $\hat{f}_n(x)$ only depends on the value $r_{k_n}$, the distance from $x$ to the $k_n$-th nearest points in $\{X_1, \ldots, X_n\}$ to $x$. Therefore, this estimator is said the $k_n$-nearest neighbor ($k_n$-NN) estimator.
3.1. ESTIMATIONS

**Estimation of the differential entropy:** Based on the samples $X^n_1$, the differential entropy $h(X)$ is often estimated by the formula:

$$
\hat{h}(X^n_1) = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{f}(X_i),
$$

(3.16)

where $\hat{f}(X_i)$ is a density estimation of $f$ at $X_i$. We can use the estimation of density function in the formula (3.14) with the condition in (3.13) to estimate the differential entropy. However, since the estimation of the differential entropy $h(X)$ only requires the estimation of density function at the collection of points $\{X_1, \ldots, X_n\}$, it doesn’t require the estimation of density function for almost all points $x$ in $\mathbb{R}^m$, the condition $k_n \to \infty$ is not necessary to the estimation of the differential entropy. Proving this idea, Goria et al. [2005] gave the following estimator of the differential entropy and the following result:

Let $\rho^{(k)}_i$ be the Euclidean distance from $X_i$ to its $k$-th nearest point $X_j$, $j \neq i$. Given a positive integer number $k$, the differential entropy $h(X)$ is estimated by

$$
\hat{h}_k(X^n_1) = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{k-1}{(n-1) c_m [\rho^{(k)}_i]^m} = \log c_m + \log(n-1) - \psi(k) + \frac{m}{n} \sum_{i=1}^{n} \log \rho^{(k)}_i,
$$

(3.17)

where $\psi$ is the *digamma function*,

$$
\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)} = \int_0^{\infty} \left( \frac{e^{-t}}{t} - \frac{e^{-zt}}{1-e^{-t}} \right) dt, \quad \forall z \in \mathbb{C} \text{ with } \Re(z) > 0.
$$

(3.18)

**Proposition 3.5** (Kozachenko and Leonenko [1987], Goria et al. [2005]). Suppose that there exists an $\epsilon > 0$ such that

$$
\int_{\mathbb{R}^m} |\log f(x)|^{2+\epsilon} f(x) < \infty
$$

$$
\int_{\mathbb{R}^m} \int_{\mathbb{R}^m} |\log \rho(x,y)|^{2+\epsilon} f(x)f(y)dx dy < \infty.
$$

(3.19)

If density function $f$ is bounded then for any fixed positive integer $k$, the $k$-nearest neighbor estimator $\hat{h}_k(X^n_1)$ is a weak consistency for $h(X)$.

The proof of Proposition 3.5 is presented clearly in the original papers by introducing many useful techniques to study the $k$-nearest neighbor estimation. However, as Póczos and Schneider [2011] warned in the first page of their paper, there was an apparent error in the proof when the author tried to apply the reverse Fatou’s lemma. Finding out some conditions of random vector $X$ to confirm the condition for applying the reverse
Fatou's lemma is not difficult, but remediing these error for the general case is really very difficult.

**Estimation of the information dependency:** When the differential entropy of $X$ exists, the information dependency of $X$ is given by formula:

$$H(X) = \sum_{j=1}^{m} h(X_j) - h(X).$$

Thus, the estimation of the information dependency could be obtained by the estimations $\hat{h}_k(X)$ and $\hat{h}_k(X_j), j = 1, \ldots, m$ which are gotten from the $k$-nearest neighbor entropy estimation. However, for any fixed $k$, the distance from the fixed point to the $k$-th nearest neighbor in the joint space will be larger than the distances to the $k$-th nearest neighbor in the marginal spaces. Since the bias in formula (3.17) depends on these distances, the biases in $\hat{h}_k(X_1), \ldots, \hat{h}_k(X_m)$, and in $\hat{h}_k(X)$ would not cancel (see in [Kraskov et al., 2004]).

To avoid this, Kraskov et al. [2004] proposed different ways to estimate the information dependency based on the $k$-nearest neighbor method as follows: Kraskov et. al. considered the maximum distance which replaced for the Euclidean distance

$$\eta(x, y) = \max_{j=1,\ldots,m} |x_j - y_j|,$$

where $x = (x_1, \ldots, x_m), y = (y_1, \ldots, y_m) \in \mathbb{R}^m$, to replace the hypersphere $S_k(r)$ by the hypercubic with center $x$ and radius $r$. Given a fixed integer number $k$, let $\eta_{i}^{(k)}/2$ be the distance from $X_i$ to its $k$-th nearest neighbor, and let $\eta_{i,j}^{(k)}/2$ be the maximum distances between the same points projected into the $j$-th coordinate, $j = 1, \ldots, m$,

$$\frac{1}{2}\eta_{i,j}^{(k)} = \max\{|X_{i,j} - X_{u,j}| : X_u \text{ be the } u\text{-th nearest neighbor of } X_i, \ u = 1, \ldots, k\}.$$ 

Noting that the volume of hypercubic with radius one equals one, then the joint differential entropy is estimated by:

$$h(X) \approx \log(n - 1) - \psi(k) + \frac{m}{n} \sum_{i=1}^{n} \log \eta_{i}^{(k)}.$$ 

In order to estimate the marginal differential entropies $h(X_j)$, Kraskov et. al. proposed two different estimators. Denote $X_{i,j}$ be the $j$-th element of the vector $X_i$. In the first estimation, let $n_{i,j}$ be the number of samples $X_{u,j}$ of which distance from $X_{i,j}$ is strictly less than $\eta_{i}^{(k)}/2$ for all $j = 1, \ldots, m$. Then, a sequence $\{k, n_{i,1}, \ldots, n_{i,m}\}$ is a characterization of the information of mutual independent of $X$ on the hypercubic center.
3.1. ESTIMATIONS

\( X_i \) with radius \( \eta_i^{(k)} \). Thus, the marginal differential entropies are estimated by

\[
h(X_j) \approx \log(n - 1) - \frac{1}{n} \sum_{i=1}^{n} \psi(n_{i,j} + 1) + \frac{1}{n} \sum_{i=1}^{n} \log \eta_i^{(k)}.
\]

And the first estimator of the information dependency is given by

\[
\widehat{I}^{(1)}_{kNN}(X^n_1) = (m - 1) \log(n - 1) + \psi(k) - \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} \psi(n_{i,j} + 1). \tag{3.20}
\]

In the second estimation, they replace \( n_{i,j} \) by the number of samples \( X_u \) with \( |X_{u,j} - X_{i,j}| \leq \eta_i^{(k)}/2 \). In this case, the sequence \( \{k, n_{i,1}, \ldots, n_{i,m}\} \) is a characterization of the information of mutual independent of \( X \) on the other rectangle with center \( X_i \) and the length of the edge on the \( j \)-th coordinate \( \eta_i^{(k)} \). The authors show that the marginal differential entropies should be estimated as follows:

\[
h(X_j) \approx \log(n - 1) - \frac{1}{k} \frac{1}{n} \sum_{i=1}^{n} \psi(n_{i,j}) + \frac{1}{n} \sum_{i=1}^{n} \log \eta_i^{(k)},
\]

and the second estimator is given by

\[
\widehat{I}^{(2)}_{kNN}(X^n_i) = (m - 1) \log(n - 1) + \psi(k) - \frac{m - 1}{k} - \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} \psi(n_{i,j}). \tag{3.21}
\]

Note that, the original formulae proposed by Kraskov et. al., are different from the formulae (3.20) and (3.21) in replacing the \( \log(n - 1) \) by \( \psi(n) \). This difference comes from the different estimation of \( \log f^n_i(X_i) \) in (3.16). It is not important because when \( n \) is large enough \( (n \geq 1000) \), \( \log(n - 1) \approx \psi(n) \). The unbiasedness and consistency of Kraskov’s \( k \)-NN information dependency estimation are still an unsolved problem. This is our favourite works and we will study it in the near future.

Examples of Kraskov’s \( k \)-nearest neighbor estimations for 64 samples in \( \mathbb{R}^2 \) are given in figure 3.2. In this example, we choose \( k = 2 \).

Estimation of the information dependency based on estimating the \( \alpha \) divergence: Let \( X, Y \) be two \( m \)-dimensional random vectors corresponding with joint density functions \( f, g \). We denote \( \text{supp}(f) \) and \( \text{supp}(g) \) are the sets where the density functions \( f \) and \( g \) are strictly positive respectively. Given \( \alpha \in \mathbb{R} \setminus \{0, 1\} \), the \( \alpha \)-divergence of \( X \) and \( Y \) is defined as follows: [Póczos and Schneider, 2011]

\[
\mathcal{D}_\alpha(f||g) = \int_{\mathcal{M}} \left( \frac{g(x)}{f(x)} \right)^{1-\alpha} f(x) dx,
\]

(3.22)
3. INFORMATION DEPENDENCY ESTIMATION

Figure 3.2: Illustration of two different Kraskov’s \( k \)-nearest neighbor information dependency estimation. First Estimation estimates the information dependency based on the hypercubes. Second Estimation estimates the information dependency based on the hyper-rectangles.

Figure 3.3: Illustration of the models to estimate the information dependency based on the \( \alpha \)-divergence. The blue star-points are given 64-samples, the red diamond-points and square-points are independent samples generated by this sample.
where $M = \text{supp}(f)$. It is clear that $\alpha$-divergence is a special case of $\varphi$-divergence introduced in section 2.1.

Let $X_1^{n_1}$ be the sample of $X$ and $Y_1^{n_2}$ be the sample of $Y$. Given an integer number $k \geq 1$, let $\rho_i^{(k)}$ denote the Euclidean distance of $X_i$ and its $k$-th nearest neighbor in the sample $X_1^{n_1}$, and $\nu_i^{(k)}$ denote the Euclidean distance of $X_i$ and its $k$-th nearest neighbor in the sample $Y_1^{n_2}$. The $\alpha$-divergence $\mathcal{D}_\alpha(f\|g)$ is estimated by:

$$
\hat{\mathcal{D}}_\alpha(X_1^{n_1}\|Y_1^{n_2}) = \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)} \frac{1}{n_1} \sum_{i=1}^{n_1} \left( \frac{(n_1-1)[\rho_i^{(k)}]^\alpha}{n_2[\nu_i^{(k)}]^\alpha} \right)^{1-\alpha}.
$$

Let $\gamma = 1 - \alpha$ and a function

$$
T(x, f, \delta, \omega) = \sum_{j=0}^{k-1} (j!)^{-\omega} \Gamma(\gamma + j\omega) \left( \frac{f(x) + \delta}{f(x) - \delta} \right)^{j\omega} [f(x) - \delta]^{-\gamma} [(1-\delta)\omega]^{-\gamma - j\omega},
$$

Póczos and Schneider [2011] gave a series of results about the $L_2$-consistency of the $\alpha$-divergence estimator. For example

**Proposition 3.6.** Assume that $k \geq 2$, $0 < \gamma < \frac{k-1}{2}$, and

i. $f$ is bounded away from 0 and uniformly Lebesgue approximable,

ii. $\exists \delta_0 > 0$ such that $\forall \delta \in (0, \delta_0), \int_M T(x, f, \delta, \frac{1}{2}) f(x) dx < \infty$,

iii. $\int_M \rho(x, y) \gamma f(y) dy < \infty$ for almost all $x \in M$,

iv. $\int_M \int_M \rho(x, y) \gamma f(x) f(y) dxdy < \infty$, and

v. $g$ is bounded on $M$.

Then $\hat{\mathcal{D}}_\alpha(X_1^{n_1}\|Y_1^{n_2})$ is $L_2$-consistent, i.e.,

$$
\lim_{n_1, n_2 \to \infty} \mathbb{E} \left[ \hat{\mathcal{D}}_\alpha(X_1^{n_1}\|Y_1^{n_2}) - \mathcal{D}_\alpha(f\|g) \right]^2 = 0.
$$

**Proposition 3.7.** Assume that $k \geq 2$, $-\frac{k}{2} < \gamma < 0$, and

i. $g$ is bounded away from 0 and uniformly Lebesgue approximable,

ii. $\exists \delta_0 > 0$ such that $\forall \delta \in (0, \delta_0), \int_M T(x, g, \delta, \frac{1}{2}) f(x) dx < \infty$,

iii. $\int_M \rho(x, y) \gamma g(y) dy < \infty$ for almost all $x \in M$,

iv. $\int_M \int_M \rho(x, y) \gamma g(x) f(y) dxdy < \infty$, and

v. $f$ is bounded on $M$ and $\text{supp}(f) \subset \text{supp}(g)$. 

69
3. INFORMATION DEPENDENCY ESTIMATION

Then \( \hat{D}_\alpha(X_1^{n_1}||Y_1^{n_2}) \) is \( L_2 \)-consistent, i.e.,

\[
\lim_{n_1, n_2 \to \infty} \mathbb{E} \left[ \hat{D}_\alpha(X_1^{n_1}||Y_1^{n_2}) - D_\alpha(f||g) \right]^2 = 0.
\]  
(3.25)

The Rényi \( \alpha \)-divergence is defined by [Rényi, 1960]

\[
R_\alpha(f||g) = \frac{1}{\alpha - 1} \log \int_{M_1 \cap M_2} f^\alpha(x) g^{1-\alpha}(x) dx = \frac{1}{\alpha - 1} \log D_\alpha(f||g),
\]  
(3.26)

where \( M_1 = \text{supp}(f) \), \( M_2 = \text{supp}(g) \). When \( \alpha \to 1 \), the Rényi \( \alpha \)-divergence tends to Kullback-Leibler divergence,

\[
\lim_{\alpha \to 1} R_\alpha(f||g) = D(f||g).
\]  
(3.27)

Since the information dependency of \( X \) is the Kullback-Leibler divergence of the joint density function with respect to product marginal density function of \( X \), the information dependency \( J(X) \) can be estimated via the estimation of the \( \alpha \)-divergence as follows: Let \( X \) be a random vector with a joint density function \( f \) and a product-marginal density function \( g \). Given \( 2n \) samples \( X_1^{2n} = \{X_1, \ldots, X_{2n}\} \) from \( X \), we take the first \( n \) samples \( X_1^n \) to make the samples of the joint density function \( f \). From the last \( n \) samples \( X_2^n \), we build another samples \( Y_1^n \) by randomly permuting the order position \( (n, n + 1, \ldots, 2n) \) of samples independently for each coordinate. The new samples \( Y_1^n \) consider as the sample of the product-marginal density function \( g \). Choosing \( \alpha \) is near to one, and estimate \( \hat{D}_\alpha(X_1^n||Y_1^n) \) by the formula (3.23). Finally, the information dependency \( J(X) \) is estimated based on the formulae (3.25), (3.26).

Although Proposition 3.6 and Proposition 3.7 can confirm the \( L_2 \)-consistency of the \( k \)-nearest neighbor estimators for the \( \alpha \)-divergence, as well as a weakly consistent for Rényi \( \alpha \)-divergence, it can not infer the same result for the information dependency estimator. It is because these propositions do not conclude for the case \( \alpha = 1 \). Again, the consistency of this \( k \)-nearest neighbor estimator for the information dependency is still an unsolved problem. An example of the estimating information dependency based on the estimating \( \alpha \)-divergence is given in figure 3.3.

3.1.3 Gaussian kernel estimations

Let the standard multivariate Gaussian density function

\[
K_m(x) = (2\pi)^{-\frac{m}{2}} e^{-\frac{1}{2}x^T x}, \quad \forall x \in \mathbb{R}^m,
\]  
(3.28)

where \( x^T \) denotes the transpose of vector \( x \). Given a sequence of \( n \) samples \( X_1, \ldots, X_n \) of the random vector \( X \), the Gaussian kernel estimation of the joint density function \( f \)
of $X$ with the window width $h$, ($h \in \mathbb{R}$, $h > 0$) is defined by

$$
\hat{f}_n(x) = \frac{1}{nh^m} \sum_{i=1}^{n} K_m\left(\frac{x - X_i}{h}\right).
$$

(3.29)

The marginal density functions $f_j$ of $X_j$, $j = 1, \ldots, m$ are estimated by

$$
\hat{f}_{n,j}(x) = \frac{1}{nh_j} \sum_{i=1}^{n} K_1\left(\frac{x - X_{i,j}}{h_j}\right) = \frac{1}{\sqrt{2\pi nh_j}} \sum_{i=1}^{n} e^{-\frac{1}{2 nh_J^2} (x - X_{i,j})^2}.
$$

(3.30)

Since the density estimators $\hat{f}_n(x), \hat{f}_{n,1}(x), \ldots, \hat{f}_{n,m}(x)$ have the continuous derivatives of all orders, the Gaussian kernel estimation works well in the case the joint and marginal density functions of $X$ are continuous and have at least continuous derivatives of second order. In this case, Silverman [1986] showed that the bias and the variance of the density estimator are approximated by

$$
bias \hat{f}_n(x) \approx \frac{1}{2} h^2 \alpha_m \nabla^2 f(x) \quad \text{and} \quad Var \hat{f}_n(x) \approx \frac{1}{nh^m} \beta_m f(x),
$$

where $\nabla^2 f(x)$ denotes the second derivative of function $f$, and

$$
\alpha_m = \int_{\mathbb{R}^m} t^2 K_m(t) dt, \quad \beta_m = \int_{\mathbb{R}^m} K_m(t)^2 dt,
$$

and the bias and the variance of the marginal density estimators are approximated by,

$$
bias \hat{f}_{n,j}(x) \approx \frac{1}{2} h_j^2 \alpha_1 f''(x) \quad \text{and} \quad Var \hat{f}_{n,j}(x) \approx \frac{1}{nh_j} \beta_1 f_j(x),
$$

where

$$
\alpha_1 = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} t^2 e^{-\frac{1}{2} t^2} dt, \quad \beta_1 = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\frac{1}{2} t^2} dt.
$$

To evaluate the quality of the estimator $\hat{g}_n$ of the function $g$, we often use the mean integrated square error (abbreviated MISE) which is defined by

$$
MISE(\hat{g}_n) = \mathbb{E} \int \left[\hat{g}_n(x) - g(x)\right]^2 dx.
$$

The means integrated square error of $\hat{f}_n$ and $\hat{f}_{n,j}$ are approximated by their bias and variance as follows

$$
MISE(\hat{f}_n) \approx \frac{1}{4} h^4 \alpha_m \int_{\mathbb{R}^m} \left[\nabla^2 f(x)\right]^2 dx + \frac{1}{nh^m} \beta_m
$$

$$
MISE(\hat{f}_{n,j}) \approx \frac{1}{4} h_j^4 \alpha_1 \int_{\mathbb{R}} \left[f''(x)\right]^2 dx + \frac{1}{nh} \beta_1.
$$
3. INFORMATION DEPENDENCY ESTIMATION

Figure 3.4: Illustration of the Gaussian-Kernel density estimation.
3.1. ESTIMATIONS

The window width $h_m$ and $h_1$ are chosen such that their approximation of mean integrated square errors are minimum. Let $\sigma_j$ be the standard deviation of samples $\{X_{i,j}\}_{i=1}^n$, and $\sigma = 1/m \sum_{j=1}^m \sigma_j$, Silverman [1986] proved that $h = \sigma \left( \frac{4}{m+2} n^{-\frac{1}{4}} \right)$ and $h_j = 1.06 \sigma_j n^{-\frac{1}{4}}$ make the $\text{MISE}(\hat{f}_n)$ and the $\text{MISE}(\hat{f}_{n,j})$ are nearly minimized. We will use this formula of window width to construct the Gaussian kernel density estimators.

An example of Gaussian kernel estimation for joint density and marginal density function of 64 samples on the $\mathbb{R}^2$ is given in figure 3.4.

The joint differential entropy and marginal differential entropies are estimated by the Gaussian kernel estimation method as follows

$$\hat{h}(X^n) = -\frac{1}{n} \sum_{i=1}^n \log \hat{f}_n(X_i)$$

$$= \log n + m \log h + \frac{m}{2} \log 2\pi - \frac{1}{n} \sum_{i=1}^n \log \left[ \sum_{u=1}^n e^{-\frac{1}{2}h(X_u-X_i)^T(X_u-X_i)} \right]$$

(3.31)

$$\hat{h}_j(X^n) = -\frac{1}{n} \sum_{i=1}^n \log \hat{f}_{n,j}(X_{i,j})$$

$$= \log n + \log h_j + \frac{1}{2} \log 2\pi - \frac{1}{n} \sum_{i=1}^n \log \left[ \sum_{u=1}^n e^{-\frac{1}{2}h_j(X_{u,j}-X_{i,j})^2} \right].$$

And the estimation of the information dependency is given by

$$\hat{\mathcal{I}}_k(X^n) = \sum_{j=1}^m \hat{h}_j(X^n) - \hat{h}(X^n)$$

$$= (m-1) \log n + \sum_{j=1}^m \log h_j - m \log h$$

$$- \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^m \log \left[ \sum_{u=1}^n e^{-\frac{1}{2}h_j(X_{u,j}-X_{i,j})^2} \right] - \log \left[ \sum_{u=1}^n e^{-\frac{1}{2}h(X_u-X_i)^T(X_u-X_i)} \right] \right).$$

(3.32)

The difficult problem in Gaussian kernel estimation is the complexity of computation. For example, to estimate the joint differential entropy or marginal differential entropies, we need $O(n^2)$ kernel evaluations which can make the computation of the estimation very slow for the large values of $n$. Fortunately, Wand and Jones [1995] (see in Appendix D.1, p. 182) presented one way of dramatically increasing the computational speed of these estimation by using the fast Fourier transform method. Recently, Botev et al. [2010] proposed an adaptive kernel density estimator based on smoothing properties of linear diffusion processes. Their results in some simulation examples are very well-compared with the traditional kernel density estimators. In the next subsection of some numerical
experiments for the information dependency estimation, we use the Potev’s kernel density estimation to estimate the information dependency.

The unbiased and consistency of the Gaussian kernel estimations for the density function, the differential entropy and the information dependency are unsolved problems.

3.2 Numerical Experiments

3.2.1 Bivariate Experiments

We consider seven examples which we can compute or approximate the theoretical information dependency. Five examples are introduced in the section 2.4: the linear convolution of discrete distribution and Gaussian with \( r = 0.1 \), the linear convolution of diagonal uniform distribution and Gaussian with \( r = 0.2 \), the linear convolution of diagonal uniform distribution and Gaussian with \( r = 0.5 \), the linear convolution of circular uniform distribution and Gaussian with \( r = 0.2 \), and the linear convolution of circular uniform distribution and Gaussian with \( r = 0.5 \). The approximation of the information dependency of these bivariate are given in table 2.3, table 2.5 and table 2.6. Two other examples are the bivariate Gamma-Exponential distribution which are presented in [Darbellay and Vajda, 2000]. The bivariate random vector \( X = (X,Y) \) is said a *gamma-exponential distribution* vector if it has the joint density function:

\[
f(x,y) = \begin{cases} \frac{\theta_1^{\theta_2} \theta_2}{\Gamma(\theta_2)} x^{\theta_2-1} e^{-\theta_1 x} & x,y > 0 \\ 0 & \text{otherwise} \end{cases}
\]

with the parameters \( \theta_1, \theta_2, \theta_3 > 0. \) \( (3.33) \)

The random vector \( X \) has the following marginal density functions:

\[
f_1(x) = \begin{cases} \frac{\theta_1^{\theta_2} x^{\theta_2-1} e^{-\theta_1 x}}{\Gamma(\theta_2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}
\]

: Gamma distribution with parameters \( (\theta_2, \theta_1) \),

\[
f_2(y) = \begin{cases} \frac{\theta_1^{\theta_2} \theta_2}{(\theta_1 + \theta_2 y)^{\theta_2}} y > 0 \\ 0 & \text{otherwise} \end{cases}
\]

\( y > 0 \)

\approx \begin{cases} \lambda e^{-\lambda y} & y > 0 \\ 0 & \text{otherwise} \end{cases}

with \( \lambda = \frac{\theta_3 \theta_2}{\theta_1} \),

: Exponential distribution with parameter \( \lambda \). \( (3.34) \)

Darbellay and Vajda [2000] showed that the joint differential entropy of \( X \) was

\[
h(X) = 1 + \theta_2 - \theta_2 \psi(\theta_2) + \log \Gamma(\theta_2) - \log \theta_3,
\]

\( (3.35) \)
and the information dependency of $X$ was

$$I(X) = \psi(\theta_2) - \log \theta_2 + \frac{1}{\theta_2},$$  \hspace{1cm} (3.36)

which depended only on the parameter $\theta_2$. We present two simulations of these bivariate corresponding with triple parameters $(\theta_1, \theta_2, \theta_3) = (1, 0.5, 1)$ and $(\theta_1, \theta_2, \theta_3) = (1, 1, 1)$. The visual samples of two bivariate are given in figure 3.10.

For each bivariate example, we construct 500 independent trials to study the estimating information dependency. With each trial, we construct a sample of the bivariate with 30000 sample points and compute 14 different estimating information dependency values of the 14 sub-samples which their number of sample points are corresponding 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 15000, 20000, 25000 and 30000. We use nine information dependency estimators to measure the information dependency of these samples. The information dependency estimators are shortly denoted and explained as follows: Let $n$ denote the number of points in each sample.

1. **Bin-structure**: Bin-structure partition rule with parameter $r_n = \lfloor \sqrt{n/k_n} \rfloor$, where $k_n = \frac{1}{5} n^{0.65}$.

2. **Gessaman**: Gessaman’s partition rule with parameter $k_n = \lfloor \frac{1}{5} n^{0.65} \rfloor$.

3. **Tree-structure**: Tree-quantization partition rule with parameter $k_n = \lfloor \frac{1}{5} n^{0.65} \rfloor$.

4. **Gaussian-Kernel**: Using the Adaptive Kernel density estimator algorithm proposed by [Botev et al., 2010].

5. **Kraskov 1-NN**: Kraskov’s one-nearest neighbor estimator given in (3.20).

6. **Kraskov 10-NN**: Kraskov’s ten-nearest neighbor estimator given in (3.20).

7. **Kraskov log-NN**: Kraskov’s $k_n$-nearest neighbor estimator given in (3.20), where $k_n = \lfloor 2 \log n \rfloor$.

8. **Divergence 10-NN**: The ten-nearest neighbor estimator based on the estimation of $\alpha$-divergence.

9. **Darbellay-Vajda**: Using the original Darbellay-Vajda algorithm proposed by Darbellay and Vajda [1999].

The means and the standard deviations of these 14 fluctuated values estimating by above nine estimators are simulated in each figure (i.e. the first example is simulated in figure 3.5, the seventh example is simulated in figure 3.12), and their numerical values are given in each table (i.e. values of the first example are given in table 3.1, values of the seventh example are given in table 3.7).
3. INFORMATION DEPENDENCY ESTIMATION

Figure 3.5: *Mean and standard deviation* of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Discrete distribution and Gaussian with parameter $r = 0.1$.

<table>
<thead>
<tr>
<th>Est. $I(X,Y)$</th>
<th>$n = 10 000$</th>
<th>$n = 30 000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean$(\hat{I})$</td>
<td>std$(\hat{I})$</td>
</tr>
<tr>
<td>Bin-structure</td>
<td>0.5017</td>
<td>0.0058</td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.3498</td>
<td>0.0306</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.1886</td>
<td>0.0055</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>0.5207</td>
<td>0.0060</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>0.4993</td>
<td>0.0161</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.5001</td>
<td>0.0071</td>
</tr>
<tr>
<td>Kr. (2 log n)-NN</td>
<td>0.5001</td>
<td>0.0065</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.4015</td>
<td>0.0128</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.4784</td>
<td>0.0068</td>
</tr>
</tbody>
</table>

Table 3.1: *Means, standard deviations and 90% confidence intervals* for figure 3.5 corresponding with the sample sizes 10,000 and 30,000.
3.2. NUMERICAL EXPERIMENTS

Figure 3.6: Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Diagonal Uniform distribution and Gaussian with parameter $r = 0.2$.

$$\hat{I}(X,Y) = \frac{1}{n} \sum_{i=1}^{n} \hat{d}_i$$

<table>
<thead>
<tr>
<th>Method</th>
<th>$n = 10,000$</th>
<th>$n = 30,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin-structure</td>
<td>0.2569</td>
<td>0.2630</td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.2656</td>
<td>0.2635</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.2269</td>
<td>0.2622</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>0.3111</td>
<td>0.3068</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>0.2956</td>
<td>0.2951</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.2998</td>
<td>0.2968</td>
</tr>
<tr>
<td>Kr. $(2 \log n)$-NN</td>
<td>0.3025</td>
<td>0.2985</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.2950</td>
<td>0.2944</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.2972</td>
<td>0.2956</td>
</tr>
</tbody>
</table>

Table 3.2: Means, standard deviations and 90% confidence intervals for figure 3.6 corresponding with the sample sizes 10,000 and 30,000.
3. INFORMATION DEPENDENCY ESTIMATION

Figure 3.7: Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Diagonal Uniform distribution and Gaussian with parameter $r = 0.5$.

<table>
<thead>
<tr>
<th>Est. $I(X,Y)$</th>
<th>$n = 10,000$</th>
<th>$n = 30,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean($\hat{I}$)</td>
<td>std($\hat{I}$)</td>
</tr>
<tr>
<td>Bin-structure</td>
<td>0.0191</td>
<td>0.0019</td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.0321</td>
<td>0.0021</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.0196</td>
<td>0.0019</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>0.0303</td>
<td>0.0023</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>0.0158</td>
<td>0.0151</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.0161</td>
<td>0.0044</td>
</tr>
<tr>
<td>Kr. $(2 \log n)$-NN</td>
<td>0.0158</td>
<td>0.0034</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.0153</td>
<td>0.0062</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.0164</td>
<td>0.0026</td>
</tr>
</tbody>
</table>

Table 3.3: Means, standard deviations and 90% confidence intervals for figure 3.7 corresponding with the sample sizes 10,000 and 30,000.
3.2. NUMERICAL EXPERIMENTS

Figure 3.8: Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Circular Uniform distribution and Gaussian with parameter $r = 0.2$.

<table>
<thead>
<tr>
<th>Est. $I(X,Y)$</th>
<th>$n = 10,000$</th>
<th>$n = 30,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.2910$</td>
<td>mean($\hat{I}$)</td>
<td>std($\hat{I}$)</td>
</tr>
<tr>
<td>Bin-structure</td>
<td>0.2558</td>
<td>0.0056</td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.2602</td>
<td>0.0054</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.2211</td>
<td>0.0042</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>0.3053</td>
<td>0.0061</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>0.2914</td>
<td>0.0161</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.2974</td>
<td>0.0073</td>
</tr>
<tr>
<td>Kr. (2 log $n$)-NN</td>
<td>0.3006</td>
<td>0.0069</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.2937</td>
<td>0.0103</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.2925</td>
<td>0.0068</td>
</tr>
</tbody>
</table>

Table 3.4: Means, standard deviations and 90% confidence intervals for figure 3.8 corresponding with the sample sizes 10,000 and 30,000.
3. INFORMATION DEPENDENCY ESTIMATION

![Graph showing estimation of information dependency](image)

Figure 3.9: Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the bivariate Linear convolution of Circular Uniform distribution and Gaussian with parameter $r = 0.5$.

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>$n = 10,000$</th>
<th>$n = 30,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean($\hat{I}$)</td>
<td>std($\hat{I}$)</td>
</tr>
<tr>
<td>Bin-structure</td>
<td>0.0230</td>
<td>0.0020</td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.0369</td>
<td>0.0023</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.0245</td>
<td>0.0020</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>0.0335</td>
<td>0.0024</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>0.0225</td>
<td>0.0157</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.0233</td>
<td>0.0045</td>
</tr>
<tr>
<td>Kr. (2 log $n$)-NN</td>
<td>0.0237</td>
<td>0.0038</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.0233</td>
<td>0.0062</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.0218</td>
<td>0.0027</td>
</tr>
</tbody>
</table>

Table 3.5: Means, standard deviations and 90% confidence intervals for figure 3.9 corresponding with the sample sizes 10,000 and 30,000.
3.2. NUMERICAL EXPERIMENTS

Bivariate Gamma–Exponential samples with parameters \((1, 0.5, 1)\)

Bivariate Gamma–Exponential samples with parameters \((1, 1, 1)\)

Figure 3.10: Two examples of the Gamma-Exponential distribution corresponding with parameters \((1, 0.5, 1)\) and \((1, 1, 1)\), respectively. The small diagrams are the zoom in of the big diagrams at the original coordinate \((0, 0)\).
3. INFORMATION DEPENDENCY ESTIMATION

Figure 3.11: Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the Gamma-Exponential distribution with parameter (1, 0.5, 1).

<table>
<thead>
<tr>
<th></th>
<th>$n = 10,000$</th>
<th>$n = 30,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean($\hat{I}$)</td>
<td>std($\hat{I}$)</td>
</tr>
<tr>
<td>Bin-structure</td>
<td>8.10$^{-5}$</td>
<td>6.10$^{-5}$</td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.6375</td>
<td>0.0087</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.5514</td>
<td>0.0072</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>-3.591</td>
<td>0.0453</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>0.5594</td>
<td>0.0159</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.4659</td>
<td>0.0074</td>
</tr>
<tr>
<td>Kr. (2 log n)-NN</td>
<td>0.4406</td>
<td>0.0070</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.5288</td>
<td>0.0140</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.7258</td>
<td>0.0099</td>
</tr>
</tbody>
</table>

Table 3.6: Means, standard deviations and 90% confidence intervals for figure 3.11 corresponding with the sample sizes 10,000 and 30,000.
3.2. NUMERICAL EXPERIMENTS

Figure 3.12: Mean and standard deviation of 500 independent estimations of the information dependency. Estimations are done for the Gamma-Exponential distribution with parameter (1, 1, 1).

<table>
<thead>
<tr>
<th></th>
<th>$n = 10,000$</th>
<th></th>
<th>$n = 30,000$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean($\hat{I}$)</td>
<td>std($\hat{I}$)</td>
<td>90% int.</td>
<td>mean($\hat{I}$)</td>
</tr>
<tr>
<td>Bin-structure</td>
<td>$2.10^{-4}$</td>
<td>$2.10^{-4}$</td>
<td>$0.0000 - 0.0006$</td>
<td>$8.10^{-5}$</td>
</tr>
<tr>
<td>Gessaman</td>
<td>$0.3393$</td>
<td>$0.0077$</td>
<td>$0.3260 - 0.3518$</td>
<td>$0.3425$</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>$0.3106$</td>
<td>$0.0064$</td>
<td>$0.2996 - 0.3205$</td>
<td>$0.3313$</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>$-2.365$</td>
<td>$0.4972$</td>
<td>$-3.208 - -1.612$</td>
<td>$-2.795$</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>$0.3795$</td>
<td>$0.0163$</td>
<td>$0.3506 - 0.4076$</td>
<td>$0.3918$</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>$0.3365$</td>
<td>$0.0076$</td>
<td>$0.3232 - 0.3492$</td>
<td>$0.3583$</td>
</tr>
<tr>
<td>Kr. (2 log $n$)-NN</td>
<td>$0.3226$</td>
<td>$0.0070$</td>
<td>$0.3110 - 0.3341$</td>
<td>$0.3451$</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>$0.3676$</td>
<td>$0.0116$</td>
<td>$0.3489 - 0.3809$</td>
<td>$0.3832$</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>$0.4195$</td>
<td>$0.0086$</td>
<td>$0.4052 - 0.4329$</td>
<td>$0.4226$</td>
</tr>
</tbody>
</table>

Table 3.7: Means, standard deviations and 90% confidence intervals for figure 3.12 corresponding with the sample sizes 10,000 and 30,000.
3. INFORMATION DEPENDENCY ESTIMATION

From the numerical results of these examples, it seems that the Darbellay-Vajda estimator, the Kraskov 1-NN estimator and the Divergence 10-NN estimator are really the best estimators in the view of the “consistent” criterion, i.e. the mean of these estimators are nearest to the exact or estimating theoretical information dependency values. In the view of the “best” criterion, i.e. the standard deviation of these estimators are smallest, among the good-consistent estimators, the Darbellay-Vajda estimator, the Kraskov 10-NN and the Kraskov $(2 \log n)$-NN are really the best estimators, where $n$ is the number of sample points in the sample. The Darbellay-Vajda estimator converges to the theoretical values very fast, in the meaning that when the number of sample points is larger than 3000, the estimations are very near to the theoretical values. Other data-dependent partition estimators Bin-structure, Gessaman, Tree-structure often work well with the nearly-independent bivariate, but they are really not consistent estimators when the samples are dependent. The examples bivariate Gamma-Exponential distribution exhibit the disadvantage of the methods which estimate the information dependency based on estimating the density functions or the differential entropies such as the $k$-nearest neighbor methods and the Gaussian-kernel method. When the distribution has a high density in a small area but scatters in a very wide area, the $k$-nearest neighbor methods and the Gaussian-kernel methods work not well. In this cases, the Darbellay-Vajda method works very well.

3.2.2 Multivariate Experiments

A series of multivariate continuous random vectors introduced in [Darbellay and Vajda, 2000] will be presented carefully in this paragraph. Concretely, we consider three examples, the Gaussian distribution, the logistic distribution, and the exponential distribution.

Given a vector $\mu = (\mu_1, \ldots, \mu_m) \in \mathbb{R}^m$ and a positive-definite matrix $\Sigma = (\sigma_{i,j}) \in \mathbb{R}^{m \times m}$, the random vector $X$ is said an $m$-dimensional Gaussian random vector associated with the mean vector $\mu$ and the covariance matrix $\Sigma$ if it has the following joint density function: $\forall x = (x_1, \ldots, x_m) \in \mathbb{R}^m$, 
\[
 f(x) = (2\pi)^{-\frac{m}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}, \tag{3.37}
\]
where $|\Sigma|$ is the determinant of $\Sigma$. $X$ has $m$ marginal density functions: $\forall x_j \in \mathbb{R}$, $j = 1, \ldots, m$,
\[
 f_j(x_j) = \frac{1}{\sqrt{2\pi \sigma_{jj}}} e^{-\frac{x_j^2}{2\sigma_{jj}}} . \tag{3.38}
\]
The joint differential and marginal differential entropies of $X$ are given [Cover and Thomas,
3.2. NUMERICAL EXPERIMENTS

\[ h(\mathbf{X}) = \frac{1}{2} \log [(2\pi e)^{m/2} |\Sigma|], \quad h(X_j) = \frac{1}{2} \log (2\pi e \sigma_{jj}^2), \quad j = 1, \ldots, m, \]

and the information dependency of \( X \) is given by

\[ J(\mathbf{X}) = \frac{1}{2} \left( \sum_{j=1}^{m} \log \sigma_{jj}^2 - \log |\Sigma| \right). \tag{3.39} \]

Given parameters \( \mu_1, \ldots, \mu_m > 0 \), and \( \sigma_1, \ldots, \sigma_m > 0 \), the random vector \( \mathbf{X} \) is said an \textit{\( m \)-dimensional logistic random vector} if it has the following joint density function:

\[ \forall \mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m, \]

\[ f(\mathbf{x}) = \prod_{j=1}^{m} \frac{j}{\sigma_j} e^{-\frac{x_j - \mu_j}{\sigma_j}} \left( 1 + \sum_{i=1}^{m} e^{-\frac{x_i - \mu_i}{\sigma_i}} \right)^{-1} \left( \sum_{i=1}^{m} e^{-\frac{x_i - \mu_i}{\sigma_i}} \right)^{m+1}. \tag{3.40} \]

\( \mathbf{X} \) has \( m \) marginal density functions: \( \forall x_j \in \mathbb{R}, j = 1, \ldots, m, \)

\[ f_j(x_j) = \frac{1}{\sigma_j} e^{-\frac{x_j - \mu_j}{\sigma_j}} \left( 1 + e^{-\frac{x_j - \mu_j}{\sigma_j}} \right)^{-2}. \tag{3.41} \]

The joint differential and marginal differential entropies of \( \mathbf{X} \) are given (see in \cite{Darbellay and Vajda, 2000})

\[ h(\mathbf{X}) = -\sum_{j=1}^{m} \log \frac{j}{\sigma_j} + (m + 1) \sum_{j=1}^{m} \frac{1}{j}, \quad h(X_j) = \log \sigma_j + 2, \quad j = 1, \ldots, m. \]

The information dependency of \( \mathbf{X} \) is given

\[ J(\mathbf{X}) = \sum_{j=1}^{m} \log j + 2m - (m + 1) \sum_{j=1}^{m} \frac{1}{j}. \tag{3.42} \]

Given parameters \( \lambda_1, \ldots, \lambda_m > 0 \), the random vector \( \mathbf{X} \) is said an \textit{\( m \)-dimensional exponential distribution} if it has a joint density function: \( \forall \mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m, \)

\[ f(\mathbf{x}) = \begin{cases} \prod_{j=1}^{m} j \lambda_j e^{\lambda_j x_j} \left( \sum_{i=1}^{m} e^{\lambda_i x_i} - m + 1 \right)^{-(m+1)} & x_1, \ldots, x_m > 0, \\ 0 & \text{otherwise}. \end{cases} \tag{3.43} \]

\( \mathbf{X} \) has \( m \) marginal density functions: \( \forall x_j \in \mathbb{R}, j = 1, \ldots, m, \)

\[ f_j(x_j) = \begin{cases} \lambda_j e^{-\lambda_j x_j} & x_j > 0 \\ 0 & \text{otherwise}. \end{cases} \tag{3.44} \]
3. INFORMATION DEPENDENCY ESTIMATION

Figure 3.13: Mean and standard deviation of 100 independent estimations of the information dependency. Estimations are done for the Seven-variate Gaussian distribution.

<table>
<thead>
<tr>
<th></th>
<th>( n = 10\ 000 )</th>
<th>( n = 50\ 000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{I}(X,Y) )</td>
<td>mean(( \hat{I} )) std(( \hat{I} )) 90% int.</td>
<td>mean(( \hat{I} )) std(( \hat{I} )) 90% int.</td>
</tr>
<tr>
<td>Bin-structure</td>
<td>0.1281   0.0056 0.1203 - 0.1354   0.1232   0.0030 0.1192 - 0.1266</td>
<td></td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.1775   0.0056 0.1705 - 0.1839   0.1350   0.0023 0.1320 - 0.1377</td>
<td></td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.0667   0.0038 0.0628 - 0.0714   0.0655   0.0016 0.0632 - 0.0673</td>
<td></td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>0.2730   0.0199 0.2493 - 0.2972   0.2766   0.0079 0.2674 - 0.2871</td>
<td></td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.2578   0.0082 0.2476 - 0.2685   0.2700   0.0042 0.2643 - 0.2749</td>
<td></td>
</tr>
<tr>
<td>Kr. ( (2 \log n))-NN</td>
<td>0.2513   0.0078 0.2409 - 0.2616   0.2661   0.0041 0.2610 - 0.2709</td>
<td></td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.2459   0.0113 0.2318 - 0.2601   0.2523   0.0052 0.2452 - 0.2579</td>
<td></td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.2759   0.0153 0.2548 - 0.2962   0.3410   0.0085 0.3289 - 0.3514</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.8: Means, standard deviations and 90\% confidence intervals for figure 3.13 corresponding with the sample sizes 10,000 and 50,000.
3.2. NUMERICAL EXPERIMENTS

Figure 3.14: Mean and standard deviation of 100 independent estimations of the information dependency. Estimations are done for the Four-variate Exponential distribution with parameter (2, 2, 2, 2).

<table>
<thead>
<tr>
<th>Estimating Info. Dep.</th>
<th>Exact InD</th>
<th>Kraskov 1-NN</th>
<th>Kraskov 10-NN</th>
<th>Kraskov log-NN</th>
<th>Divergence 10-NN</th>
<th>Darbellay-Vajda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin-structure</td>
<td>0.0335</td>
<td>0.0022</td>
<td>0.0075</td>
<td>0.0123</td>
<td>0.0102</td>
<td>0.7604</td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.0932</td>
<td>0.0845</td>
<td>0.1021</td>
<td>0.1071</td>
<td>0.1021</td>
<td>0.7563</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.1046</td>
<td>0.1021</td>
<td>0.1071</td>
<td>0.1071</td>
<td>0.1071</td>
<td>0.7526</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>0.7604</td>
<td>0.7467</td>
<td>0.7794</td>
<td>0.7614</td>
<td>0.7546</td>
<td>0.7406</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.7563</td>
<td>0.7476</td>
<td>0.7653</td>
<td>0.7613</td>
<td>0.7571</td>
<td>0.7575</td>
</tr>
<tr>
<td>Kr. (2 log n)-NN</td>
<td>0.7526</td>
<td>0.7447</td>
<td>0.7609</td>
<td>0.7605</td>
<td>0.7565</td>
<td>0.7446</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.7575</td>
<td>0.7386</td>
<td>0.7778</td>
<td>0.7612</td>
<td>0.7515</td>
<td>0.7577</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.7406</td>
<td>0.7319</td>
<td>0.7494</td>
<td>0.7596</td>
<td>0.7549</td>
<td>0.7549</td>
</tr>
</tbody>
</table>

Table 3.9: Means, standard deviations and 90% confidence intervals for figure 3.14 corresponding with the sample sizes 10,000 and 50,000.
3. INFORMATION DEPENDENCY ESTIMATION

Figure 3.15: Mean and standard deviation of 100 independent estimations of the information dependency. Estimations are done for the Five-variate Logistics distribution with mean parameter (1, 1, 2, 2, 4) and variance parameter (2, 2, 6, 6, 9).

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>$n = 10,000$</th>
<th>$n = 50,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean(\hat{I})</td>
<td>std(\hat{I})</td>
</tr>
<tr>
<td>Bin-structure</td>
<td>0.3098</td>
<td>0.0790</td>
</tr>
<tr>
<td>Gessaman</td>
<td>0.4312</td>
<td>0.0104</td>
</tr>
<tr>
<td>Tree-structure</td>
<td>0.3527</td>
<td>0.0063</td>
</tr>
<tr>
<td>Kr. 1-NN</td>
<td>1.0628</td>
<td>0.0250</td>
</tr>
<tr>
<td>Kr. 10-NN</td>
<td>0.9630</td>
<td>0.0207</td>
</tr>
<tr>
<td>Kr. (2log n)-NN</td>
<td>0.9174</td>
<td>0.0202</td>
</tr>
<tr>
<td>Di. 10-NN</td>
<td>0.8951</td>
<td>0.0213</td>
</tr>
<tr>
<td>Darbellay-Vajda</td>
<td>0.9338</td>
<td>0.0186</td>
</tr>
</tbody>
</table>

Table 3.10: Means, standard deviations and 90% confidence intervals for figure 3.15 corresponding with the sample sizes 10,000 and 50,000.
The joint differential and marginal differential entropies of $X$ are given [Darbellay and Vajda, 2000]

$$h(X) = -\sum_{j=1}^{m} \log(j\lambda_j) + (m+1) \sum_{j=1}^{m} \frac{1}{j} - m, \quad h(X_j) = 1 - \log \lambda_j, \quad j = 1, \ldots, m.$$ 

The information dependency of $X$ is given

$$J(X) = \sum_{j=1}^{m} \log j + 2m - (m+1) \sum_{j=1}^{m} \frac{1}{j}. \quad (3.45)$$

Since the Gaussian-kernel density estimation for the multivariate case ($m \geq 4$) requires a lot of computations and often gives inaccurate results, we don’t consider this method to study the numerical results in multivariate cases. We use eight following estimators to estimate the theoretical information dependency for each samples: Bin-structure, Gessaman, Tree-structure, Kraskov 1-NN, Kraskov 10-NN, Kraskov 2 log $n$-NN, Divergence 10-NN, and Darbellay-Vajda as using in the bivariate case.

For each multivariate example, we construct 100 independent trials to study the estimating information dependency. Each trial, we construct a sample with 50000 sample points and compute 9 different estimating information dependency values of the 9 sub-samples which their number of sample points are corresponding 2000, 4000, 6000, 8000, 10000, 20000, 30000, 40000 and 50000. The means and the standard deviations of these 9 fluctuated values measuring by above 8 estimators are simulated in each figure and each table, respectively.

The numerical results of three examples show us that the class of $k$-nearest neighbor methods work better than the class of data-dependent partition methods. In the view of the “consistent” criterion, the Kraskov 1-NN is the best estimator. In the view of the “best” criterion, among the good-consistent estimators, the Kraskov 10-NN and the Kraskov (2 log $n$)-NN are the best. The $k$-nearest neighbor estimators converge to the theoretical information dependency values very fast. With the number of samples is around 4000, their estimators are very near to the theoretical values. The Darbellay-Vajda estimator works really well when the number of samples is larger ($n \geq 50000$). Other data-dependent partition estimators gives the bad estimations.

3.3 Strong consistency of Data dependent partitions

In this section, we would like to study the conditions of the data-dependent rectangle partition rules to make their estimators are strongly consistent. Let $\{\pi_n\}_{n \in \mathbb{N}}$ be a data-dependent rectangle partition rule, their information dependency estimator is given by:
3. INFORMATION DEPENDENCY ESTIMATION

For any samples $X^n_1$,

$$\hat{I}(X^n_1) = \sum_{A \in \pi_n(X^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)},$$  

(3.46)

where $\mu_n$, $\nu_n$ are the empirical joint distribution and the empirical product-marginal distribution of $X^n_1$. The rectangle partition rule $\{\pi_n\}_{n \in \mathbb{N}}$ is a strong consistency for the information dependency $I(X)$ if

$$\sum_{A \in \pi_n(X^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \rightarrow \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)} \quad \mathbb{P}\text{-a.s.},$$  

(3.47)

where $\mathcal{P}(m)$ denotes the collection of all finite partitions of $\Omega^m$, and $\mu$, $\nu$ denote the joint distribution and the product-marginal distribution of the random vector $X$.

It is clear that (3.47) occurs if

$$\sum_{A \in \pi_n(X^n_1)} \mu(A) \log \frac{\mu(A)}{\nu(A)} \rightarrow \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)} \quad \mathbb{P}\text{-a.s.},$$  

(3.48)

and

$$\sum_{A \in \pi_n(X^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \rightarrow \sum_{A \in \pi_n(X^n_1)} \mu(A) \log \frac{\mu(A)}{\nu(A)} \quad \mathbb{P}\text{-a.s.}$$  

(3.49)

We would like to study the conditions of the rectangle partition rule $\{\pi_n\}_{n \in \mathbb{N}}$ such that (3.48) occurs in the first subsection of this section, and the conditions to confirm (3.49) occurs in the second subsection. Combining the results are given in the above subsections, the final subsection will present the results to construct a class of rectangle partition rules which their information dependency estimators are strongly consistent. Also in this subsection, we explain how the Gessaman’s partition rule (Proposition 3.1), the Tree-quantization partition rule (Proposition 3.2) and the Darbellay-Vajda partition rule (Proposition 3.3) belong this class.

3.3.1 Simultaneous approximation

For each sample $X^n_1$, we denote $\mathcal{F}_{X^n_1}$ be a $\sigma$-field generated by $\pi_n(X^n_1)$. The Corollary 2.1 says that given the sequence of samples $X^n_1$, if the sequence $\{\mathcal{F}_{X^n_1}\}_{n=1}^{\infty}$ is an asymptotically generating sequence of sub-$\sigma$-fields of $\mathcal{B}^m$, then

$$\sum_{A \in \pi_n(X^n_1)} \mu(A) \log \frac{\mu(A)}{\nu(A)} \rightarrow \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)}.$$  

90
3.3. STRONG CONSISTENCY OF DATA DEPENDENT PARTITIONS

Thus, if the rectangle partition rule \( \{\pi_n\}_{n \in \mathbb{N}} \) satisfies that the sequence \( \{F_{X_n^1}\}_{n=1}^{\infty} \) is an asymptotically generating sequence of sub-\( \sigma \)-fields of \( \mathcal{B}^m \) for almost surely sample \( X_1^\infty \), then (3.48) occurs. Recently, [Liese et al., 2006] studied very carefully the conditions of the rectangle partition rule \( \{\pi_n\}_{n \in \mathbb{N}} \) such that it made the (3.48) occurring. They studied this occurring not only for the information dependency but also for other \( \varphi \)-divergencies. Precisely, let \( \Psi \) be the class of all convex functions: \( \varphi : [0, +\infty) \to \mathbb{R} \) standardized by the condition \( \varphi(1) = 0 \) and strictly convex at 1 in the sense that there is no \( \epsilon > 0 \) such that \( \varphi \) is linear on the whole interval \((1-\epsilon, 1+\epsilon)\), they defined the \textit{asymptotically sufficient partition} that the sequence of partitions \( \{\pi_n(X^n_1)\}_{n=1}^{\infty} \) is said \((\mu, \nu)\)-asymptotically sufficient if

\[
\sum_{A \in \pi_n(X^n_1)} \nu(A) \varphi \left( \frac{\mu(A)}{\nu(A)} \right) \to \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \nu(A) \varphi \left( \frac{\mu(A)}{\nu(A)} \right)
\]

for all \( \varphi \in \Psi \), and \( F_{X^n_1} \subseteq \mathcal{B} \). The main result of this paper is stated as follows:

**Proposition 3.8** ([Liese et al., 2006]). The sequence \( \{\pi_n(X^n_1)\}_{n=1}^{\infty} \) is \((\mu, \nu)\)-asymptotically sufficient if there exists a directed sequence of measurable partitions of \( \Omega^m \), \( \{A_n\}_{n \in \mathbb{N}} \) such that the \( \sigma \)-field \( \sigma(A_1 \cup A_2 \cup \ldots) \) is sufficient for \((\mu, \nu)\) and each \( A_n \) is \((\mu + \nu)\)-sufficiently refined by a sequence \( \pi_n(X^n_1) \).

Where a sequence \( \{A_n\} \) is said \textit{directed} if for all \( n \in \mathbb{N} : \sigma(A_n) \subset \sigma(A_{n+1}) \); \( \sigma \)-field \( \mathcal{A} \) is \textit{sufficient} for \((\mu, \nu)\) if

\[
\sup_{\pi \in \mathcal{P}_\mathcal{A}} \sum_{A \in \pi} \nu(A) \varphi \left( \frac{\mu(A)}{\nu(A)} \right) = \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \nu(A) \varphi \left( \frac{\mu(A)}{\nu(A)} \right)
\]

for all \( \varphi \in \Psi \), where \( \mathcal{P}_\mathcal{A} \) is collection of all finite and \( \mathcal{A} \)-measurable partitions of \( \Omega^m \); and partition \( A_n \) is \((\mu + \nu)\)-sufficiently refined by a sequence \( \pi_n(X^n_1) \) if \((\mu + \nu)(\pi_n(X^n_1) \ast A_n) \cap B) = 0 \; \forall B \in \mathcal{B} \), where \( A \ast B = \bigcup_{C \in A \ast B} C \) with \( A \ast B = A \circ B = A \cap \overline{B} - A \) and \( A \circ B = \{A \cap B : A \in A, B \in B\} \). The Proposition 3.8 is a beautiful result, however the last two conditions are difficult to check in practice.

It is not difficult to see that if the random vector \( X \) is not bounded, the bin-structure partition rule will generate an asymptotically generating sequence of sub-\( \sigma \)-fields for almost surely samples \( X_1^\infty \) of \( X \). The Gessaman’s partition rule, the Tree-quantization partition rule and the Darbellay-Vajda partition rule do not have this property. This is because these partition rules depend on the population of the samples, therefore some areas which are reached by samples with zero probability can not be partitioned deeply. Hence, it can not guarantee that the sequence \( \{F_{X^n_1}\}_{n=1}^{\infty} \) is an asymptotically generating sequence of sub-\( \sigma \)-fields of \( \mathcal{B}^m \).

Now, we would like to introduce a new notion, \textit{simultaneous approximation}, of which meaning is similar to the notion \textit{asymptotically generating sequence of sub-\( \sigma \)-fields}. This
3. INFORMATION DEPENDENCY ESTIMATION

notion depends on the distributions of the samples.

**Definition 3.6.** Let \( \mu, \nu \) be any probability measures on a measurable space \((\Omega, \mathcal{B})\), and let \( \{\pi_n\}_{n \in \mathbb{N}} \) be a partition rule on this space. Let \( X_1, X_2, \ldots \) be a sequence of i.i.d. random vectors with distribution \( \mathbb{P} \). The partition rule \( \{\pi_n\}_{n \in \mathbb{N}} \) is said to be simultaneously \((\mu, \nu)\)-approximating with respect to \( \mathbb{P} \) if, \( \forall \delta > 0 \) and for any finite measurable partition \( \pi = \{A_1, \ldots, A_r\} \) of \( \Omega \), there exists a sequence of finite measurable partitions

\[
\{\tilde{\pi}_n\}_{n \in \mathbb{N}} = \{A_{n,1}, \ldots, A_{n,r}\}_{n \in \mathbb{N}}, \quad \forall n \in \mathbb{N} : A_{n,u} \in \sigma(\pi_n(X_u^n)), \ u = 1, \ldots, r
\]

such that

\[
\limsup_{n \to \infty} \max_{u=1, \ldots, r} |\mu(A_u) - \mu(A_{n,u})| < \delta \quad \text{and} \quad \limsup_{n \to \infty} \max_{u=1, \ldots, r} |\nu(A_u) - \nu(A_{n,u})| < \delta,
\]

\( \mathbb{P} \)-almost surely, where \( \sigma(\pi) \) denotes the smallest sigma field containing the element of \( \pi \).

When \( \Omega = [0,1]^m \), \( \mathcal{B} \) is the collection of all Borel set on \([0,1]^m\), and \( \mu, \nu \) are the Lebesgue measure, the notion simultaneously \((\mu, \nu)\)-approximating is nearly equivalent to the notion asymptotically generating sequence of sub-\(\sigma\)-fields.

The notion simultaneously \((\mu, \nu)\)-approximating has a relationship with (3.48) through the following result:

**Proposition 3.9.** Let \( X \) be a random vector with a joint distribution \( \mu \) and a product-marginal distribution \( \nu \). Let \( \mathbb{P} \) be a distribution of \( X_1^\infty \), a sequence of samples of \( X \), and let \( \{\pi_n\}_{n \in \mathbb{N}} \) be a rectangle partition rule. If \( \mu \ll \nu \) and \( \{\pi_n\}_{n \in \mathbb{N}} \) is simultaneously \((\mu, \nu)\)-approximating with respect to \( \mathbb{P} \), then

\[
\lim_{n \to \infty} \sum_{A \in \pi_n(X_u^n)} \mu(A) \log \frac{\mu(A)}{\nu(A)} = \sup_{\pi \in \mathcal{P}(\mathbb{R}^m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)} \quad \mathbb{P}\text{-a.s.} \quad (3.50)
\]

**Proof.** For any \( \epsilon > 0 \), there exists a finite partition \( \pi_\epsilon = \{A_1, \ldots, A_r\} \) such that

\[
\sum_{A \in \pi_\epsilon} \mu(A) \log \frac{\mu(A)}{\nu(A)} > \sup_{\pi \in \mathcal{P}(\mathbb{R}^m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)} - \frac{\epsilon}{2}. \quad (3.51)
\]

Since \( x \log x \) is a continuous real function, for any distributions \( \mu_1, \mu_2 \), if \( \forall \epsilon > 0, \exists \delta_1 > 0: \max_{u=1, \ldots, r} |\mu_1(A_u) - \mu_2(A_u)| < \delta_1 \), then \( \sum_{i=1}^r |\mu_1(A_u) \log \mu_1(A_u) - \mu_2(A_u) \log \mu_2(A_u)| < \epsilon \).

Moreover, when \( \mu_1 \ll \nu_1, \mu_2 \ll \nu_2 \), the quantities \( \sum_{i=1}^r \mu_i(A_u) \log \mu_i(A_u) / \nu_i(A_u) \) and \( \sum_{i=1}^r \mu_i(A_u) \log \nu_i(A_u) \) are bounded, \( i = 1, 2 \). So, \( \forall \epsilon > 0, \exists \delta_1 > 0, \delta_2 > 0 \), such that if \( \max_{u=1, \ldots, r} |\nu_1(A_u) - \nu_2(A_u)| < \delta_1 \) and \( \max_{u=1, \ldots, r} |\nu_1(A_u) - \nu_2(A_u)| < \delta_2 \), then

\[
\left| \sum_{u=1}^r \mu_1(A_u) \log \frac{\mu_1(A_u)}{\nu_1(A_u)} - \sum_{u=1}^r \mu_2(A_u) \log \frac{\mu_2(A_u)}{\nu_2(A_u)} \right| < \frac{\epsilon}{2}.
\]
3.3. STRONG CONSISTENCY OF DATA DEPENDENT PARTITIONS

Hence, as a direct consequence of the hypotheses of the proposition, for any typical sequence \(x_1, x_2, \ldots\), there exists a sequence of measurable approximations of \(\pi_r\), denoted by \(\{\pi_n^r\}_{n \in \mathbb{N}}\), with \(\pi_n^r \subset \sigma(\pi_n(x_1^n))\), such that \(\exists n_0 \in \mathbb{N}, \forall n > n_0,\)

\[
\left| \sum_{A \in \pi_n^r} \mu(A) \log \frac{\mu(A)}{\nu(A)} - \sum_{A \in \pi_n} \mu(A) \log \frac{\mu(A)}{\nu(A)} \right| < \frac{\epsilon}{2}.
\]  
(3.52)

Finally, using (3.51), (3.52) and noting that \(\pi_n(x_1^n)\) is a refinement of \(\pi_n^r\), we have, \(\forall \epsilon > 0\)

\[
\liminf_{n \to \infty} \sum_{A \in \pi_n(x_1^n)} \mu(A) \log \frac{\mu(A)}{\nu(A)} > \sup_{\pi \in \mathcal{P}(m)} \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{\nu(A)} - \epsilon, \quad \mathbb{P}\text{-a.s.}
\]

Therefore, (3.50) is satisfied. \(\square\)

When \((\Omega, \mathcal{F}) \equiv (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))\), the notion simultaneously \((\mu, \nu)\)-approximating is equivalent to the notion shrinking cell condition, which is referred in [Lugosi and Nobel, 1996]. Let \(\lambda\) be the Lebesgue measure on \((\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))\), and \(\rho\) be the Euclidian norm in \(\mathbb{R}^m\). For any \(A \in \mathcal{B}(\mathbb{R}^m)\), we define its **diameter** by

\[
\|A\| = \sup_{x, y \in A} \rho(x, y).
\]

**Proposition 3.10** (Silva and Narayan [2010a]). Assuming that \(\mu, \nu \ll \lambda\), the partition rule \(\{\pi_n\}_{n \in \mathbb{N}}\) is simultaneously \((\mu, \nu)\)-approximating with respect to \(\mathbb{P}\) if it has the shrinking cell condition corresponding with \(\mu\) and \(\nu\), i.e., \(\forall \gamma > 0,\)

\[
\lim_{n \to \infty} \mu\left(\bigcup_{A \in \pi_n(x_1^n), \|A\| > \gamma} A\right) = 0 \quad \text{and} \quad \lim_{n \to \infty} \nu\left(\bigcup_{A \in \pi_n(x_1^n), \|A\| > \gamma} A\right) = 0
\]  
(3.53)

\(\mathbb{P}\)-almost surely.

**Proof.** Let us consider an arbitrary Borel measurable partition \(\pi = \{A_1, \ldots, A_r\}\). For each \(n\), let \(\{B_1^n, \ldots, B_r^n\}\) be the covering of \(\pi\) induced by \(\pi_n(x_1^n)\), i.e., \(\forall i = 1, \ldots, r,\)

\[
B_i^n = \bigcup_{A \in \pi_n(x_1^n), A \cap A_i \neq \emptyset} A,
\]

and let a new partition \(\pi^{(n)} = \{A_1^n, \ldots, A_r^n\} \subset \sigma(\pi_n(x_1^n))\) with \(A_1^n = B_1^n, A_2^n = B_2^n/B_1^n, \ldots, A_r^n = B_r^n/(\bigcup_{j=1}^{r-1} B_j^n)\). For an arbitrary \(\delta > 0\), since \(\mu\) is absolutely continuous with respect to the Lebesgue measure \(\lambda\), there exists a bounded measurable set \(B\) such that \(\mu(B) > 1 - \frac{\delta}{2}\). Let us define \(\hat{\pi} = \{\hat{A}_1, \ldots, \hat{A}_r\}\) and \(\hat{\pi}^{(n)} = \{\hat{A}_1^n, \ldots, \hat{A}_r^n\}\) with \(\hat{A}_i = B \cap A_i, \hat{A}_i^n = B \cap A_i^n, \forall i = 1, \ldots, r,\) as two partitions of \(B\). We have the inequality,
∀i = 1, . . . , r, ∀n ∈ ℤ

|µ(Ai) − µ(A_i^n)| ≤ |µ(A_i) − µ(Å_i)| + |µ(Å_i) − µ(A_i^n)| + |µ(A_i^n) − µ(A_i)| < δ + |µ(Å_i) − µ(A_i^n)|.

In addition, for any Borel measurable set A and γ > 0, let us define its γ-open covering by 

Aγ = ∪x∈A Bl(x, γ), where Bl(x, γ) is the open ball of radius γ centered at x. Let δγ(A) = Aγ/A. Since the continuity of λ under monotone set sequences [Halmos, 1950], ∀ε > 0, ∃γ > 0, such that λ(δγ(A)) < ε. This judgement is also true for the measure µ because of µ ≪ λ. Hence, ∀ε > 0 let us fix γ such that µ(δγ(Å_i) ∪ δγ(A_i^n)) < ε, ∀i = 1, . . . , r. Let us define

S_γ^n = \{x^n_i ∈ ℜ^m-n : \max_{A ∈ π_n(X^n_i)} \|A\| < γ\}.

Note that, if X^n_i ∈ S_γ^n, then Å_i ⊂ B^n_i ⊂ Å^n_i, where B^n_i = B^n_i ∩ B, ∀i = 1, . . . , r. It also infers [Å_i \ ( \bigcup_{j=1}^{l_i - 1} Å_j^n)] ⊂ Å^n_i ⊂ Å_i and [Å^n_i \ Å_i] ⊂ δγ(Å_i) & [Å^n_i \ Å_i] ⊂ δγ(Å^n_i). Then, ∀n ∈ ℤ,

|µ(Å_i) − µ(Å^n_i)| ≤ µ(Å_i Δ Å^n_i) ≤ µ[δγ(Å_i) ∪ δγ(Å^n_i)]1_{(x^n_i ∈ S_γ^n)} + µ\left(\bigcup_{A ∈ π_n(X^n_i), \|A\| > γ} A\right)1_{(x^n_i ∈ S_γ^n)} ≤ ε + µ\left(\bigcup_{A ∈ π_n(X^n_i), \|A\| > γ} A\right).

Thus, ∀ε > 0,

\lim_{n → ∞} \sup \left|\mu(A_i) − \mu(A_i^n)\right| < δ + ε, \quad \mathbb{P}\text{-a.s.}

it follows that

\lim_{n → ∞} \max_{i=1, . . . , r} \left|\mu(A_i) − \mu(A_i^n)\right| < δ, \quad \mathbb{P}\text{-a.s.}

The shrinking cell condition with distribution ν is proven similarly.

Noting that, let any F be one-to-one continuous mapping from ℜ^m to (0, 1)^m, and ě, ıt be the induced measures of µ, ν on the space (0, 1)^m, i.e., ∀A ∈ ℬ((0, 1)^m), ě(A) = µ(F^{-1}(A)), ıt(A) = ν(F^{-1}(A)), we will have ∀X^n_i

\sup_{π ∈ ℙ_{X^n_i}} \sum_{A ∈ π} μ(A) log (μ(A)/ν(A)) = \sup_{π ∈ ℙ_{(0, 1)^m}} \sum_{A ∈ π} ě(A) log (ě(A)/ıt(A)).
3.3. STRONG CONSISTENCY OF DATA DEPENDENT PARTITIONS

where \( \mathcal{P}_{\mathbb{R}^m} \) and \( \mathcal{P}_{(0,1)^m} \) denote the collections of all finite partitions on \( \mathbb{R}^m \) and \( (0,1)^m \) corresponding. Therefore, if the partition rule \( \{\pi_n\}_{n \in \mathbb{N}} \) is invariant with any one-to-one continuous mapping \( G \) from \( \mathbb{R}^m \) to \( (0,1)^m \), in the meaning that \( \forall X_t^n \),

\[
\sum_{A \in \pi_n(X_t^n)} \mu(A) \log \frac{\mu(A)}{\nu(A)} = \sum_{A \in \pi_n(G(X_t^n))} \bar{\mu}(A) \log \frac{\bar{\mu}(A)}{\bar{\nu}(A)}, \tag{3.54}
\]

we can find the conditions of \( \{\pi_n\}_{n \in \mathbb{N}} \) to confirm that (3.48) occurs by checking the shrinking cell condition of \( Y = F(X) \), a new random vector defined on the \( (0,1)^m \). For each rectangle \( A \) in \( (0,1)^m \), we know that \( \|A\| \leq \sum_{j=1}^m \|A\|_j \), where \( \|A\|_j \) denotes the Lebesgue measure of the projection of \( A \) into the \( j \)-th coordinate. Using Markov’s inequality, the shrinking cell condition with respect to the distribution \( \mu \) satisfies if

\[
\lim_{n \to \infty} E_{\mu} \left( \sum_{j=1}^m \|\pi_n(Z|Y_t^n)\|_j \right) = \lim_{n \to \infty} \sum_{A \in \pi_n(Y_t^n)} \sum_{j=1}^m \|A\|_j \mu(A) = 0, \text{ \( \mathbb{P} \)-a.s.}, \tag{3.55}
\]

where \( \pi_n(y|Y_t^n) \) is the unique cell of \( \pi_n(Y_t^n) \) containing \( y \). Based on these techniques, we have the following results:

**Lemma 3.1.** Let \( \{\pi_n\}_{n \in \mathbb{N}} \) be a Gessaman’s partition rule corresponding with a sequence of positive integer number \( \{k_n\}_{n \in \mathbb{N}} \). Let \( X \) be a random vector with the joint distribution \( \mu \) and the product-marginal distribution \( \nu \). If \( k_n/n \to 0 \) as \( n \) tends to infinity, then the Gessaman’s partition rule \( \{\pi_n\}_{n \in \mathbb{N}} \) is simultaneously \( (\mu, \nu) \)-approximating with respect to \( \mathbb{P} \), where \( \mathbb{P} \) is a distribution of \( X_t^\infty \), a sample of \( X \).

**Lemma 3.2.** Let \( \{\pi_n\}_{n \in \mathbb{N}} \) be a Tree-quantization partition rule corresponding with a sequence of positive integer number \( \{k_n\}_{n \in \mathbb{N}} \). Let \( X \) be a random vector with the joint distribution \( \mu \) and the product-marginal distribution \( \nu \). If \( k_n/n \to 0 \) as \( n \) tends to infinity, then the Tree-quantization partition rule \( \{\pi_n\}_{n \in \mathbb{N}} \) is simultaneously \( (\mu, \nu) \)-approximating with respect to \( \mathbb{P} \), where \( \mathbb{P} \) is a distribution of \( X_t^\infty \), a sample of \( X \).

The proofs of Lemma 3.1 and Lemma 3.2 can be referred in [Silva and Narayan, 2010b] and [Lugosi and Nobel, 1996].

3.3.2 Vapnik-Chervonenkis Theory and its improvement

In this subsection, we would like to introduce some results which are very useful to find out the conditions of \( \{\pi_n\}_{n \in \mathbb{N}} \) such that it has the property (3.49). These results are the Glivenko-Cantelli Theorem, the Vapnik-Chervonenkis inequality [Vapnik, 1998; Vapnik and Chervonenkis, 1971] and the Lugosi-Nobel inequality [Lugosi and Nobel, 1996].

Glivenko-Cantelli Theorem is sometimes referred to as a fundamental theorem of mathematical statistics, stating uniform almost sure convergence of the empirical dis-
3. INFORMATION DEPENDENCY ESTIMATION

tribution function to the true one. The Vapnik-Chervonenkis inequality, named *Uniform law of large number*, is a powerful generalization of the Glivenko-Cantelli Theorem. It provides upper bounds on random variables of the type

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|,$$

where $\mathcal{A}$ is a collection of measurable sets. The Lugosi-Nobel inequality is an extension of Vapnik-Chervonenkis inequality to consider the upper bounds on collection of partitions

$$\sup_{\pi \in \mathcal{P}} \sum_{A \in \pi} |\mu_n(A) - \mu(A)|,$$

where $\mathcal{P}$ is a collection of finite and measurable partitions.

Let $a = (a_1, \ldots, a_m), b = (b_1, \ldots, b_m) \in \mathbb{R}^m$, $a \leq b$ means $a_j \leq b_j, \forall j = 1, \ldots, m$.

**Theorem 3.2** (Glivenko-Cantelli Theorem). Let $X$ be a random vector with the joint distribution function $F(x) = \mathbb{P}(X \leq x)$, and let $X_1, \ldots, X_n$ be $n$ samples of $X$. Denote the standard empirical distribution function by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}.$$

Then

$$\mathbb{P}\left( \sup_{x \in \mathbb{R}^m} |F(x) - F_n(x)| > \epsilon \right) \leq 8(n + 1)^m e^{-\frac{\epsilon^2 n}{32}},$$

and, in particular, by the Borel-Cantelli lemma,

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}^m} |F(x) - F_n(x)| = 0,$$

with probability one.

**Proof.** The proof of Glivenko-Cantelli Theorem is presented carefully in [Devroye et al., 1996, p. 193]. Now we represent this proof here. Let $\mu$ be a joint distribution of $X$ and $\mu_n$ be an empirical joint distribution of $X_n$. Let $A$ be the class of sets of form $(-\infty, x) \equiv (-\infty, x_1) \times \cdots \times (-\infty, x_m)$ for $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$, then

$$\sup_{x \in \mathbb{R}^m} |F(x) - F_n(x)| = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|.$$

We prove the theorem in several steps. Let $\epsilon$ be a small positive number such that $n\epsilon^2 \geq 2$.

**Step 1. First Symmetrization by a Ghost Sample.** Define the new random vectors $X'_1, \ldots, X'_n$ such that $X_1, \ldots, X_n, X'_1, \ldots, X'_n$ are all independent and identically distributed. Denote $\mu'_n$ be the empirical distribution corresponding to the new observed
samples: $\mu'_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \in A\}}$, for all measurable set $A$. Let $A^* \in A$ be a set for which $|\mu_n(A^*) - \mu(A^*)| > \epsilon$ if such a set exists and let $A^*$ be a fixed set in $A$ otherwise. Then
\[
\mathbb{P}\left( \sup_{A \in A} |\mu_n(A) - \mu'_n(A)| > \frac{\epsilon}{2} \right) \geq \mathbb{P}\left( |\mu_n(A^*) - \mu(A^*)| > \epsilon, |\mu'_n(A^*) - \mu(A^*)| < \frac{\epsilon}{2} \right)
\]
\[
= \mathbb{E}\left[ 1_{\{|\mu_n(A^*) - \mu(A^*)| > \epsilon\}} \mathbb{P}\left( |\mu'_n(A^*) - \mu(A^*)| < \frac{\epsilon}{2} \right) \right]X_i^2
\]
The conditional probability inside is bounded by Chebyshev’s inequality, as follows
\[
\mathbb{P}\left( |\mu'_n(A^*) - \mu(A^*)| < \frac{\epsilon}{2} \right) \geq 1 - \frac{\mu(A^*)(1 - \mu(A^*))}{ne^2/4} \geq 1 - \frac{1}{n\epsilon^2}
\]
Hence, for $\epsilon > 0$ such that $n\epsilon^2 \geq 2$, we have
\[
\mathbb{P}\left( \sup_{A \in A} |\mu_n(A) - \mu'_n(A)| > \frac{\epsilon}{2} \right) \geq \frac{1}{2} \mathbb{P}( |\mu_n(A^*) - \mu(A^*)| > \epsilon)
\]
\[
\geq \frac{1}{2} \mathbb{P}\left( \sup_{A \in A} |\mu_n(A) - \mu(A)| > \epsilon \right).
\]

**Step 2. Second Symmetrization by Random Signs.** Let $\sigma_1, \ldots, \sigma_n$ be i.i.d. sign variables, independent of $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$ with $\mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = 1) = 1/2$, $i = 1, \ldots, n$. Clearly, the distribution of $\sup_{A \in A} |\sum_{i=1}^{n} [1_{\{X_i \in A\}} - 1_{\{X'_i \in A\}}]|$ is similar to the distribution of $\sup_{A \in A} |\sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}} - 1_{\{X'_i \in A\}})|$. Thus, by Step 1,
\[
\mathbb{P}\left( \sup_{A \in A} |\mu_n(A) - \mu(A)| > \epsilon \right) \leq 2\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \sum_{i=1}^{n} [1_{\{X_i \in A\}} - 1_{\{X'_i \in A\}}] > \frac{\epsilon}{2} \right)
\]
\[
\leq 2\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}} - 1_{\{X'_i \in A\}}) \right| > \frac{\epsilon}{2} \right)
\]
\[
\leq 2\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}}) \right| > \frac{\epsilon}{4} \right) + 2\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i(1_{\{X'_i \in A\}}) \right| > \frac{\epsilon}{4} \right)
\]
\[
= 4\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}}) \right| > \frac{\epsilon}{4} \right).
\]

**Step 3. Conditioning.** To bound the probability
\[
\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}}) \right| > \frac{\epsilon}{4} \right) = \mathbb{P}\left( \sup_{x \in \mathbb{R}^m} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in x\}}) \right| > \frac{\epsilon}{4} \right),
\]
we condition on $X_1, \ldots, X_n$. Fix $x_1, \ldots, x_n \in \mathbb{R}^m$, and note that as $x$ ranges over $\mathbb{R}^m$, the
number of different vectors \((1_{\{x_1 \leq x\}}), \ldots, 1_{\{x_n \leq x\}}\) is at most \((n+1)^m\). Thus, conditional on \(X_1, \ldots, X_n\), the supremum in the probability above is just a maximum taken over at most \(n+1\) random variables. Thus,

\[
\mathbb{P}\left( \sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}}) \right| > \frac{\epsilon}{4} \left| X_1^n \right| \leq (n+1)^m \sup_{A \in \mathcal{A}} \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}}) \right| > \frac{\epsilon}{4} \left| X_1^n \right).
\]

**Step 4. Hoeffding’s Inequality.** With \(x_1, \ldots, x_n\) fixed, \(\sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}})\) is the sum of \(n\) independent zero mean random variables bounded between \(-1\) and \(1\). Therefore, as a consequence of Theorem 8.1 in [Devroye et al., 1996, p. 122],

\[
\mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}}) \right| > \frac{\epsilon}{4} \left| X_1^n \right| \leq 2e^{-\frac{n \epsilon^2}{32}}.
\]

Thus,

\[
\mathbb{P}\left( \sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}}) \right| > \frac{\epsilon}{4} \left| X_1^n \right| \leq 2(n+1)^m e^{-\frac{n \epsilon^2}{32}}.
\]

Taking the expected value on both sides, we have

\[
\mathbb{P}\left( \sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1_{\{X_i \in A\}}) \right| > \frac{\epsilon}{4} \leq 2(n+1)^m e^{-\frac{n \epsilon^2}{32}}.
\]

In summary,

\[
\mathbb{P}\left( \sup_{A \in \mathcal{A}} \left| \mu_n(A) - \mu(A) \right| > \epsilon \right) \leq 8(n+1)^m e^{-\frac{n \epsilon^2}{32}}.
\]

Before presenting the Vapnik-Chervonenkis inequality, let us introduce the concept of *shatter coefficient*.

**Definition 3.7 (Vapnik and Chervonenkis [1971]).** Let \(\mathcal{A}\) be a collection of measurable sets. For \(x^n = (x_1, \ldots, x_n) \in \mathbb{R}^{m \cdot n}\), let \(S_A(x^n)\) be the number of different sets in

\[
\{ \{x_1, \ldots, x_n\} \cap A : A \in \mathcal{A} \}.
\]

The \(n\)-th shatter coefficient of \(\mathcal{A}\) is

\[
S^*_A(n) = \max_{x^n \in \mathbb{R}^{m \cdot n}} S_A(x^n).
\]

That is, the shatter coefficient is the maximal number of different subsets of \(n\) points that
can be picked out by the class of sets $A$.

The shatter coefficients measure the richness of the class $A$.

**Theorem 3.3 (Vapnik and Chervonenkis [1971]).** Let $X_1, \ldots, X_n$ be i.i.d. random vectors on the probability space $(\Omega^m, \mathcal{B}^m, \mathbb{P})$ with distribution $\mu$, i.e. $X_i \sim \mu$, $i = 1, \ldots, n$. Let $A \subset \mathcal{B}$ be a collection of measurable sets, and $\mu_n$ be an empirical distribution induced by $X^n$. For any $n \in \mathbb{N}$ and $\epsilon > 0$,

$$
\mathbb{P}\left( \sup_{A \in A} |\mu_n(A) - \mu(A)| > \epsilon \right) \leq 8S^*_A(n)e^{-\frac{n\epsilon^2}{32}}.
$$

(3.57)

**Proof.** The proof presented here is a representation of the proof in [Devroye et al., 1996, p. 197]. Assuming that $n\epsilon^2 \geq 2$, in the first two steps in proof of Theorem 3.2, we know that

$$
\mathbb{P}\left( \sup_{A \in A} |\mu_n(A) - \mu(A)| > \epsilon \right) \leq 4\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1\{X_i \in A\}) > \frac{\epsilon}{4} \right) .
$$

**STEP 3. CONDITIONING.** To bound the probability

$$
\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1\{X_i \in A\}) > \frac{\epsilon}{4} \right) ,
$$

again we condition on $X^n$. Fix $x_1, \ldots, x_n \in \mathbb{R}^m$, and observe that as $A$ ranges over $A$, the number of different vectors $(1\{x_1 \in A\}, \ldots, 1\{x_n \in A\})$ is just the number of different subsets of $\{x_1, \ldots, x_n\}$ produced by intersecting it with sets in $A$, which, by definition, cannot exceed $S^*_A(n)$. Therefore, with $X_1, \ldots, X_n$ fixed, the supremum in the above probability is a maximum of at most $S^*_A(n)$ random variables. By the union bound we get

$$
\mathbb{P}\left( \sup_{A \in A} \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1\{X_i \in A\}) > \frac{\epsilon}{4} \left| X^n \right. \right) 
\leq S^*_A(n) \sup_{A \in A} \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1\{X_i \in A\}) > \frac{\epsilon}{4} \left| X^n \right. \right) .
$$

The Step 4 in proof of Theorem 3.2 infers that

$$
\mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i(1\{X_i \in A\}) > \frac{\epsilon}{4} \left| X^n \right. \right) \leq 2e^{-\frac{n\epsilon^2}{32}} .
$$

Hence,

$$
\mathbb{P}\left( \sup_{A \in A} |\mu_n(A) - \mu(A)| > \epsilon \right) \leq 8S^*_A(n)e^{-\frac{n\epsilon^2}{32}} .
$$

$\square$
Let $\mathcal{P}$ be a collection of finite and measurable partitions of $\Omega$, the maximal cell count of $\mathcal{P}$ is defined as follows

$$\mathfrak{M}(\mathcal{P}) = \sup_{\pi \in \mathcal{P}} |\pi|,$$

where $|\pi|$ denotes the number of cells of partition $\pi$. The growth function of $\mathcal{P}$ is defined by following definition.

**Definition 3.8** (Lugosi and Nobel [1996]). For $x^n = (x_1, \ldots, x_n) \in \mathbb{R}^{m \cdot n}$, let $\Delta_{P}(x^n_\pi)$ be the number of different partitions in

$$\{ \{x_1, \ldots, x_n\} \cap \pi : \pi \in \mathcal{P}\},$$

where $B \cap \pi = \{B \cap A_1, \ldots, B \cap A_r\}$, with $\pi = \{A_1, \ldots, A_r\}$. The n-th growth function of $\mathcal{P}$ is

$$\Delta^*_{\mathcal{P}}(n) = \max_{x^n_\pi \in \mathbb{R}^{m \cdot n}} \Delta_{P}(x^n_\pi).$$

That is, the growth function is the largest number of distinct partitions of any $n$ points that can be induced by the partition in $\mathcal{P}$.

Lugosi and Nobel [1996] extended the Vapnik-Chervonenkis inequality to families of partitions by the following result:

**Theorem 3.4** (Lugosi and Nobel [1996]). Let $\mathcal{P}$ be a collection of measurable partitions of $\Omega$. For any $n \in \mathbb{N}$ and $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{\pi \in \mathcal{P}} \sum_{A \in \pi} |\mu_n(A) - \mu(A)| > \epsilon\right) \leq 8\Delta^*_{\mathcal{P}}(n)2^{m(n)}e^{-\frac{n\epsilon^2}{128\mathfrak{M}(\mathcal{P})}}.\quad (3.60)$$

**Proof.** For each partition $\pi = \{A_1, \ldots, A_r\} \in \mathcal{P}$, let $\mathcal{B}(\pi)$ be the collection of all $2^r$ sets that can be expressed as the union of cells of $\pi$. Let

$$\mathcal{B}(\mathcal{P}) = \{A \in \mathcal{B}(\pi) : \pi \in \mathcal{P}\}$$

be the collection of all such unions, as $\pi$ ranges through $\mathcal{P}$. Fix $\pi$ and define

$$\hat{A} = \bigcup_{A \in \pi, \mu_n(A) \geq \mu(A)} A.$$

Then clearly

$$\sum_{A \in \pi} |\mu_n(A) - \mu(A)| = 2(\mu_n(\hat{A}) - \mu(\hat{A})) \leq 2 \sup_{A \in \mathcal{B}(\pi)} |\mu_n(A) - \mu(A)|.$$
Consequently,

\[
\sup_{\pi \in \mathcal{P}} \sum_{A \in \pi} |\mu_n(A) - \mu(A)| \leq \sup_{\pi \in \mathcal{P}} \sup_{A \in \mathcal{B}(\pi)} |\mu_n(A) - \mu(A)|
\]

\[
= 2 \sup_{A \in \mathcal{B}(\mathcal{P})} |\mu_n(A) - \mu(A)|.
\]

Since, \( S_{\mathcal{B}(\mathcal{P})}(n) \leq 2^{\mathcal{B}(\mathcal{P})} \Delta^*_n(n) \), from Vapnik-Chervonenkis inequality, we have

\[
P\left( \sup_{\pi \in \mathcal{P}} \sum_{A \in \pi} |\mu_n(A) - \mu(A)| > \epsilon \right) \leq P\left( \sup_{A \in \mathcal{B}(\mathcal{P})} |\mu_n(A) - \mu(A)| > \frac{\epsilon}{2} \right)
\]

\[
\leq 8 \Delta^*_n(n) 2^{\mathcal{B}(\mathcal{P})} e^{-\frac{\epsilon^2}{128}}.
\]

Given a partition rule \( \{\pi_n\}_{n \in \mathbb{N}} \), a sequence of collections of partitions \( \{\mathcal{P}_{\pi_n}\}_{n \in \mathbb{N}} \) with

\[
\mathcal{P}_{\pi_n} = \left\{ \pi_n(x^n_1) : \forall x^n_1 \in \mathbb{R}^{m \cdot n} \right\} \quad \forall n \in \mathbb{N},
\]

is called an associated collection of partitions of \( \{\pi_n\}_{n \in \mathbb{N}} \). Based on the Lugosi and Nobel result (Theorem 3.4), Silva and Narayan [2010a] gave the first class of rectangle partition rules \( \{\pi_n\}_{n \in \mathbb{N}} \) which make the property (3.49) occur as follows:

**Theorem 3.5** (Silva and Narayan [2010a]). Let \( X \) be a random vector with a joint distribution \( \mu \) and a product-marginal distribution \( \nu \), and let \( X_1, X_2, \ldots \), be a sequence of samples of \( X \). Let \( \{\pi_n\}_{n \in \mathbb{N}} \) be a rectangle partition rule, and its associated collection of partitions \( \{\mathcal{P}_{\pi_n}\}_{n \in \mathbb{N}} \). If there exist a real number \( \tau \in (0, 1) \) and a sequence positive integer number \( \{k_n\}_{n \in \mathbb{N}} \) with \( (k_n) \succeq (n \frac{1+\tau}{2}) \) such that

i. \( \frac{1}{n^\tau} \mathcal{R}(\mathcal{P}_{\pi_n}) \to 0 \) and \( \frac{1}{n^\tau} \log \Delta^*_n(n) \to 0 \) as \( n \to \infty \), and

ii. \( \forall n \in \mathbb{N}, \forall x^n_1 \in \mathbb{R}^{m \cdot n}, \min_{A \in \pi_n(x^n_1)} \mu_n(A) \geq \frac{k_n}{n} \),

then (3.49) occurs, i.e.,

\[
\sum_{A \in \pi_n(x^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \to \sum_{A \in \pi_n(x^n_1)} \mu(A) \log \frac{\mu(A)}{\nu(A)} \quad \mathbb{P}\text{-a.s.}
\]

**Lemma 3.3.** Let \( \{A_n\}_{n \in \mathbb{N}} \) be a sequence of measurable sets on the probability space \((\Omega, \mathcal{B}, \mathbb{P})\). If there exists a negative number \( C_0 \) and \( \tau \in (0, 1) \) such that

\[
\limsup_{n \to \infty} \frac{1}{n^\tau} \log \mathbb{P}(A_n) \leq C_0 \quad (3.61)
\]
3. INFORMATION DEPENDENCY ESTIMATION

then \( P(\limsup_{n \to \infty} A_n) = 0 \).

**Proof.** The condition (3.61) infers that \( \forall \epsilon > 0, \exists n_0 \in \mathbb{N}, \forall n \geq n_0, 1/n^\tau \log P(A_n) \leq C_0 + \epsilon \). Since \( C_0 < 0 \), it can choose \( \epsilon \) such that \( C = C_0 + \epsilon < 0 \). And then, since \( \tau \in (0, 1) \), we have \( \forall n \geq n_0 \)

\[
\frac{1}{n^\tau} \log P(A_n) \leq C \Rightarrow P(A_n) \leq \sum_{n=n_0}^{\infty} P(A_n) \leq \sum_{n=n_0}^{\infty} e^{C n^\tau} < \infty
\]

\[
\Rightarrow \sum_{n=1}^{\infty} P(A_n) < \infty.
\]

From the Borel-Cantelli lemma [Halmos, 1950], \( P(\limsup_{n \to \infty} A_n) = 0 \). \( \square \)

**Proof of Theorem 3.5.** To make it simpler, we replace \( \pi_n(X^n_1) \) by \( \pi_n \). We consider

\[
\left| \sum_{A \in \pi_n} \left( \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} - \mu(A) \log \frac{\mu(A)}{\nu(A)} \right) \right| \leq \left| \sum_{A \in \pi_n} \left[ \mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A) \right] \right|
\]

\[
+ \left| \sum_{A \in \pi_n} \left[ \mu_n(A) \log \nu_n(A) - \mu(A) \log \nu(A) \right] \right| \equiv (3.62)
\]

The first expression in the right-hand side of (3.62) is upper bounded by

\[
\sum_{A \in \pi_n} \left| \mu_n(A) - \mu(A) \right| \log \frac{1}{\mu_n(A)} + \sum_{A \in \pi_n} \left| \log \mu_n(A) - \log \mu(A) \right| \mu(A)
\]

\[
\leq \log \frac{n}{k_n} \sup_{\pi \in \mathcal{P}_{\pi_n}} \sum_{A \in \pi} \left| \mu_n(A) - \mu(A) \right| + \sup_{A \in \pi_n} \left| \log \mu_n(A) - \log \mu(A) \right|. \quad (3.63)
\]

The Lugosi-Nobel inequality and condition (i) of the theorem show that, \( \forall \epsilon > 0, \)

\[
\frac{1}{n^\tau} \log P\left( \log \frac{n}{k_n} \sup_{\pi \in \mathcal{P}_{\pi_n}} \sum_{A \in \pi} \left| \mu_n(A) - \mu(A) \right| > \epsilon \right)
\]

\[
= \frac{1}{n^\tau} \log P\left( \sup_{\pi \in \mathcal{P}_{\pi_n}} \sum_{A \in \pi} \left| \mu_n(A) - \mu(A) \right| > \frac{\epsilon}{\log n - \log k_n} \right)
\]

\[
\leq \frac{1}{n^\tau} \left[ \log 8 + W(\mathcal{P}_{\pi_n}) + \log \Delta_{\mathcal{P}_{\pi_n}}^* (n) - \frac{n \epsilon^2}{128 (\log n - \log k_n)^2} \right]
\]

\[
\rightarrow -\infty \quad \text{as} \quad n \to \infty.
\]

Therefore, the Lemma 3.3 infers the first term in the second explanation of (3.63) tends
to zero $\mathbb{P}$-almost surely. Concerning the second term in (3.63), we see that

$$\sup_{A \in \pi_n} \left| \log \mu_n(A) - \log \mu(A) \right| \leq \max \left\{ \sup_{A \in \pi_n} \left| \frac{\mu(A)}{\mu_n(A)} - 1 \right|, \sup_{A \in \pi_n} \left| \frac{\mu_n(A)}{\mu(A)} - 1 \right| \right\}. $$

Once more, since $\mu_n(A) \geq k_n/n$, $\forall A \in \pi_n$, we have

$$\mathbb{P}\left( \sup_{A \in \pi_n} \left| \frac{\mu(A)}{\mu_n(A)} - 1 \right| > \epsilon \right) \leq \mathbb{P}\left( \sup_{A \in \pi_n} \left| \mu(A) - \mu_n(A) \right| > \frac{\epsilon k_n}{n} \right) \leq \mathbb{P}\left( \sup_{A \in \pi_n} \left| \mu(A) - \mu_n(A) \right| > \frac{\epsilon k_n}{n} \right) \leq 8\Delta^*_{\pi_n} (n) 2^{2M(p_n)} e^{-\epsilon^2 k_n^2/128}. $$

Because of $(k_n) \geq (n^{1/4})$, we also have

$$\lim_{n \to \infty} \sup_{A \in \pi_n} \left| \frac{\mu_n(A)}{\mu(A)} - 1 \right| = 0, \quad \mathbb{P}\text{-a.s.} \hspace{1cm} (3.64)$$

(3.64) concludes that $\lim_{n \to \infty} \sup_{A \in \pi_n} \frac{\mu(A)}{\mu_n(A)} = 1$, $\mathbb{P}$-almost surely, and $\lim_{n \to \infty} \sup_{A \in \pi_n} \frac{\mu_n(A)}{\mu(A)} = 1$, $\mathbb{P}$-almost surely. Then

$$\lim_{n \to \infty} \sup_{A \in \pi_n} \left| \frac{\mu_n(A)}{\mu(A)} - 1 \right| \leq \lim_{n \to \infty} \sup_{A \in \pi_n} \left| \frac{\mu(A) - \mu_n(A)}{\mu_n(A)} \right| \to 0, \quad \mathbb{P}\text{-a.s.}$$

Hence, the second term in the second explanation of (3.63) also tends to zero $\mathbb{P}$-almost surely.

The same arguments present to prove that $\sum_{A \in \pi_n} \left[ \mu(A) \log \nu(A) - \mu_n(A) \log \mu_n(A) \right]$ tends to zero because of $\nu_n(A) \geq (k_n/n)^m$, $\forall A \in \pi_n$, $\forall n \in \mathbb{N}$. For example,

$$\sum_{A \in \pi_n} \left| \mu_n(A) - \mu(A) \right| \log \frac{1}{\nu_n(A)} \leq m \log \frac{n}{k_n} \sup_{\pi_n} \sum_{A \in \pi} \left| \mu_n(A) - \mu(A) \right| \to 0, \quad \mathbb{P}\text{-a.s.,}$$

and

$$\mathbb{P}\left( \sup_{A \in \pi_n} \left| \frac{\nu(A)}{\nu_n(A)} - 1 \right| > \epsilon \right) \leq 8\Delta^*_{\pi_n} (n) 2^{2M(p_n)} e^{-\epsilon^2 k_n^2/128}. $$

Because of $(k_n) \geq (n^{1/4})$, we have

$$\lim_{n \to \infty} \sup_{A \in \pi_n} \left| \frac{\nu(A)}{\nu_n(A)} - 1 \right| = 0, \quad \mathbb{P}\text{-a.s.}$$

The formula of bound (3.62) and the convergence to zero $\mathbb{P}$-almost surely of two terms in the right hand of (3.62) give the proof of Theorem 3.5. \hfill \square
3. INFORMATION DEPENDENCY ESTIMATION

Fortunately, the condition (ii) of Theorem 3.5 is easily adaptable by the Gessaman’s partition rule and the Tree-quantization rule. Thus, the two following results will prove that there are sub-classes in the Gessaman’s partition rule and the Tree-quantization partition rule of which elements satisfy the conditions in Theorem 3.5.

Lemma 3.4. Let \( \{ \pi_n^G \}_{n \in \mathbb{N}} \) be a Gessaman’s partition rule corresponding with a sequence of positive integer number \( \{ k_n \}_{n \in \mathbb{N}} \). If there exists a real number \( \tau \in (\frac{1}{3}, 1) \) such that \( (k_n) \approx (n^{\frac{1}{2\tau}}) \), then \( \{ \pi_n^G \}_{n \in \mathbb{N}} \) satisfies all the conditions in Theorem 3.5.

Proof. The condition (ii) of Theorem 3.5 is satisfied by the construction of the Gessaman’s data-driven rule. Once more, from the construction of this partition rule, \( \forall \mathbf{x}_1^m \in \mathbb{R}^{m \times n}, |\pi_n^G(\mathbf{x}_1^m)| = r_n^m \leq \frac{n}{k_n} \Rightarrow M(\mathcal{P}_{\pi_n^G}) \leq \frac{n}{k_n} \). Hence, combining with conditions \( (k_n) \approx (n^{\frac{1}{2\tau}}) \) and \( \tau \in (\frac{1}{3}, 1) \), we have

\[
\lim_{n \to \infty} \frac{1}{n^\tau} M(\mathcal{P}_{\pi_n^G}) = 0.
\]

From the inductive method, it is not difficult to show that the number of different ways \( n_0 \) fixed points in \( \mathbb{R} \)-line can be partitioned by \( r, r \geq 1 \) intervals is

\[
\binom{n_0 + r - 2}{n_0 - 1}.
\]

When \( n_0 = 0 \), we consider the above formula equals 1. Given a sequence order of coordinates, such as \( \{ j_1, \ldots, j_m \} \), to construct the partitions of \( \mathcal{P}_{\pi_n^G} \), it is clear that at the first step, there are

\[
\binom{n + r_n - 2}{n - 1}
\]
different ways to partition \( n \) given points \( \mathbf{X}_1^n \). After the first step, assuming there is a partition which has \( r_n \) cells \( A_1, \ldots, A_{r_n} \) which \( A_i \) contains \( n_i \) samples of \( \mathbf{X}_1^n \), \( i = 1, \ldots, r_n \), \( n = \sum_{i=1}^{r_n} n_i \), then at the second step, there are

\[
\binom{n_1 + r_n - 2}{n_1 - 1} \binom{n_2 + r_n - 2}{n_2 - 1} \cdots \binom{n_{r_n} + r_n - 2}{n_{r_n} - 1}
\]
different ways to continuously partition. From the simple combinatorial inequality, \( \forall a_1, \ldots, a_k, b_1, \ldots, b_k \in \mathbb{N} \) with \( a_i \leq b_i \), \( i = 1, \ldots, k \),

\[
\prod_{i=1}^{k} \binom{a_i}{b_i} \leq \binom{a}{b}
\]

with \( a = \sum_{i=1}^{k} a_i \) and \( b = \sum_{i=1}^{k} b_i \), the different ways to partition \( A_1, \ldots, A_{r_n} \) in the
3.3. STRONG CONSISTENCY OF DATA DEPENDENT PARTITIONS

second step is upper bounded by

\[
\left( \frac{n + r_n^2 - 2r_n}{n - r_n} \right).
\]

And then, the different ways to separate \( n \) points \( X^n_i \) by the partitions in \( \mathcal{P}_{\pi^n_G} \) after the first step and the second step is

\[
\left( \frac{n + r_n - 2}{n - 1} \right) \left( \frac{n + r_n^2 - 2r_n}{n - r_n} \right).
\]

Analyzing similarly for \( k \)-th step, \( k = 3, \ldots, m \), the growth function of \( \mathcal{P}_{\pi^n_G} \), the number of different partitions of any \( n \) points partitioned by partitions in \( \mathcal{P}_{\pi^n_G} \), is upper bounded by

\[
\Delta_{\mathcal{P}_{\pi^n_G}}^*(n) \leq m! \prod_{k=1}^m \left( \frac{n + r_n^k - 2r_n^{k-1}}{n - r_n^{k-1}} \right),
\]

where \( m! \) is the number of ways to choose the sequence order of coordinates \( \{j_1, \ldots, j_m\} \). Devroye et al. [1996] showed that \( \log \binom{w}{v} \leq vh(v - u/v) \), with \( h(x) = -x \log x - (1 - x) \log(1 - x) \) for \( x \in [0, 1] \). And, since \( r_n = \lfloor (n/k_n)^{1/m} \rfloor \), \( n + r_n^k - 2r_n^{k-1} \leq 2n \) and \((r_n^k - r_n^{k-1})/(n + r_n^k - 2r_n^{k-1}) \leq 1/k_n \leq 1/2 \). So,

\[
\lim_{n \to \infty} \frac{1}{n^\tau} \log \Delta_{\mathcal{P}_{\pi^n_G}}^*(n) \leq \lim_{n \to \infty} \frac{1}{n^\tau} \sum_{k=1}^m \log \left( \frac{n + r_n^k - 2r_n^{k-1}}{n - r_n^{k-1}} \right)
\]

\[
\leq \lim_{n \to \infty} \frac{2n}{n^\tau} \sum_{k=1}^m h \left( \frac{r_n^k - r_n^{k-1}}{n + r_n^k - 2r_n^{k-1}} \right) \leq \lim_{n \to \infty} 2mn^{1-\tau} h \left( \frac{1}{k_n} \right) = 0.
\]

The proof of Lemma 3.4 is finish. \( \square \)

**Lemma 3.5.** Let \( \{\pi^n_T\}_{n \in \mathbb{N}} \) be a Tree-quantization partition rule corresponding with a sequence of positive integer number \( \{k_n\}_{n \in \mathbb{N}} \). If there exists a real number \( \tau \in (\frac{1}{3}, 1) \) such that \( (k_n) \approx (n^{\frac{1-\tau}{3}}) \), then \( \{\pi^n_T\}_{n \in \mathbb{N}} \) satisfies all the conditions in Theorem 3.5.

**Proof.** Similar to the proof of Lemma 3.4, it is not difficult to prove with the condition \( (k_n) \approx (n^{\frac{1-\tau}{3}}) \), \( \{\pi^n_T\}_{n \in \mathbb{N}} \) has the condition (ii) of Theorem 3.5 and

\[
\lim_{n \to \infty} \frac{1}{n^\tau} \mathfrak{M}(\mathcal{P}_{\pi^n_T}) = 0.
\]

Fixed a sequential order of coordinates \( \{j_1, \ldots, j_m\} \), Step 5 of the tree-quantization partition rule says that after \( k \) steps we have \( 2^k \) disjoint rectangles of which contains more than \( k_n \) sample points. Hence, any TQVS partitions are constructed by at most \( \log(n/k_n) \) steps. In the first step, we have at most \( n \) different ways to split \( n \) points by a hyper-plane orthogonal with \( j_1 \)-th coordinate. In the second step, there are at most
3. INFORMATION DEPENDENCY ESTIMATION

\[ n_1 n_2 \leq n^2, \ \forall n_1, n_2 \in \mathbb{N} \text{ with } n_1 + n_2 = n \text{ different ways to split } n_1 \text{ points in the first new rectangle and } n_2 \text{ points in another rectangle by two hyper-planes orthogonal with } j_{2}-\text{th coordinate. And so on, we have} \]

\[
\Delta^*_P \pi_T(n) \leq m! \prod_{k=1}^{[\log(n/k_n)]+1} n^k \leq m! n^{(\log n/k_n + 1)^2}
\]

\[
\Rightarrow \lim_{n \to \infty} \frac{1}{n^\tau} \log \Delta^*_P \pi_T(n) \leq \lim_{n \to \infty} (\log n - \log k_n + 1) \frac{2 \log n}{n^\tau} = 0.
\]

The Lemma 3.5 is proven.

### 3.3.3 Strongly consistent partition rules

Let \( X \) be a random vector on the probability space \( (\Omega^m, \mathcal{B}^m, \mathbb{P}) \). Assuming that its joint distribution \( \mu \) is absolutely continuous with respect to its product-marginal distribution \( \nu \).

**Strongly consistent Gessaman’s partition rule:**

Repeat of Proposition 3.1. Let \( \{\pi^G_n\}_{n \in \mathbb{N}} \) be a Gessaman’s partition rule associated with a sequence of positive integer number \( \{k_n\}_{n \in \mathbb{N}} \). If there exists a real number \( \tau \in (\frac{1}{3}, 1) \) such that \( (k_n) \approx (n^{\frac{1+\tau}{2}}) \), then \( \{\pi^G_n\}_{n \in \mathbb{N}} \) is a strongly consistent estimation for the information dependency of \( X \).

**Proof.** Proposition 3.1 is a consequence of Lemma 3.1 and Lemma 3.4.

**Strongly consistent Tree-quantization partition rule:**

Repeat of Proposition 3.2. Let \( \{\pi^T_n\}_{n \in \mathbb{N}} \) be a Tree-structure partition rule associated with a sequence of positive integer number \( \{k_n\}_{n \in \mathbb{N}} \). If there exists a real number \( \tau \in (\frac{1}{3}, 1) \) such that \( (k_n) \approx (n^{\frac{1+\tau}{2}}) \), then \( \{\pi^T_n\}_{n \in \mathbb{N}} \) is a strongly consistent estimation for the information dependency of \( X \).

**Proof.** Proposition 3.2 is a consequence of Lemma 3.2 and Lemma 3.5.

**Strongly consistent Darbella-Vajda partition rule:** Since the empirical joint distribution of the cells of the partitions generated by the Darbella-Vajda partition rules can be received zero or converge to zero very fast when the number of samples tends to infinity, the Darbella-Vajda partition rules are difficult to have the property “making (3.49) occurs”. The similar comment is also true for the simultaneous approximation property, (3.48). Thus, we would like to study the conditions of the Darbella-Vajda
3.3. STRONG CONSISTENCY OF DATA DEPENDENT PARTITIONS

Partition rules such that they are a strongly consistent estimators in another direction. This direction is based on the following interesting result.

Theorem 3.6. Assuming that there is a strongly consistent partition rule \( \{\bar{\pi}_n\}_{n \in \mathbb{N}} \) for the estimation of the information dependency of \( X \). Let \( \{\pi_n\}_{n \in \mathbb{N}} \) be another partition rule with associated collections of partitions \( \{\mathcal{P}_{\pi_n}\}_{n \in \mathbb{N}} \). If

i. there exists a real number \( \tau \in (0, 1) \) such that

\[
\lim_{n \to \infty} \frac{1}{n^\tau} \mathcal{M}(\mathcal{P}_{\pi_n}) = 0, \quad \lim_{n \to \infty} \frac{n^{1-\tau}}{\mathcal{M}^2(\mathcal{P}_{\pi_n})} > 0 \quad \text{and} \quad \lim_{n \to \infty} \frac{1}{n^\tau} \log \Delta^s_{\mathcal{P}_{\pi_n}}(n) = 0,
\]

ii. and \( \forall x^n_1 \in \mathbb{R}^{m \cdot n} \),

\[
\sum_{A \in \pi_n(x^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \geq \sum_{A \in \mathcal{P}_{\pi_n}(x^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)},
\]

where \( \mu_n, \nu_n \) are the empirical joint and product-marginal distributions induced by \( x^n_1 \), then \( \{\pi_n\}_{n \in \mathbb{N}} \) is a strongly consistent partition rule for the information dependency of \( X \), i.e.

\[
\lim_{n \to \infty} \sum_{A \in \pi_n(x^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} = \mathcal{I}(X) \quad \mathbb{P}\text{-a.s.,} \quad (3.65)
\]

where \( X^n_1 \) is \( n \) samples of \( X \).

Proof of Theorem 3.6. In this proof, when we write the sum ‘\( \sum \)’ which takes the sum of all cells \( A \) of partition \( \pi_n(x^n_1) \), if the sum has a lot of conditions, we often do not write the condition \( \{A \in \pi_n(x^n_1)\} \) below the symbol \( \sum \). We also often replace the symbol \( \pi_n(x^n_1) \) by the simpler \( \pi_n \). Since \( \{\bar{\pi}_n\}_{n \in \mathbb{N}} \) is a strongly consistent for the information dependency, i.e., \( \forall \epsilon > 0 \),

\[
\mathcal{I}(X) - \epsilon < \lim_{n \to \infty} \sum_{A \in \bar{\pi}_n} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} < \mathcal{I}(X) + \epsilon, \quad \mathbb{P}\text{-a.s.,}
\]

and the condition (ii) of the theorem, the proof of Theorem 3.6 is finish if we can prove that \( \forall \epsilon > 0 \),

\[
\lim_{n \to \infty} \sum_{A \in \pi_n} \left( \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} - \mu(A) \log \frac{\mu(A)}{\nu(A)} \right) < \epsilon, \quad \mathbb{P}\text{-a.s.} \quad (3.66)
\]

We separated the quantity in (3.66) into six terms given in (3.68), (3.70), (3.72), (3.76),

107
3. INFORMATION DEPENDENCY ESTIMATION

(3.77) and (3.79)

\[
\sum_{A \in \pi_n} \left( \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} - \mu(A) \log \frac{\mu(A)}{\nu(A)} \right) = \xi_1 + \xi_2 + \xi_3 + \xi_4 + \xi_5 + \xi_6 \tag{3.67}
\]

and prove that these terms are non-positive or almost surely converge to zero. Firstly we consider

\[
\xi_1 = \sum_{\mu_n(A) = 0} \left[ \mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A) \right] = \sum_{\mu_n(A) = 0} |\mu(A) \log \mu(A)|. \tag{3.68}
\]

Since \( x \log x \geq -\sqrt{x}, \ \forall x \in [0, 1] \) (let us consider \( 0 \log 0 = 0 \)),

\[
\xi_1 \leq \sum_{\mu_n(A) = 0} \left[ (\mu(A))^{\frac{1}{2}} \right] \leq |\pi_n|^{\frac{1}{2}} \left( \sum_{\mu_n(A) = 0} \mu(A) \right)^{\frac{1}{2}} \leq \left[ \mathfrak{M}(\mathcal{P}_{\pi_n}) \right]^{\frac{1}{2}} \left( \sum_{\mu_n(A) = 0} \mu(A) \right)^{\frac{1}{2}}.
\]

The above second inequality is a consequence of Jensen’s inequality. From the Lugosi-Nobel inequality, \( \forall \epsilon > 0 \ \exists n_0 \in \mathbb{N}, \) such that \( \forall n \geq n_0, \)

\[
\mathbb{P} \left( \sum_{\mu_n(A) = 0} \mu(A) > \epsilon \right) \leq \mathbb{P} \left( \sum_{A \in \pi_n} |\mu_n(A) - \mu(A)| > \epsilon \right) \leq 8 \Delta^*_n(n)2^{\mathfrak{M}(\mathcal{P}_{\pi_n})}e^{-\frac{\epsilon^2}{128}}
\]

\[
\Rightarrow \limsup_{n \to \infty} \frac{1}{n^r} \log \mathbb{P}(\xi_1 > \epsilon) \leq \limsup_{n \to \infty} \frac{1}{n^r} \log \mathbb{P} \left( \sum_{\mu_n(A) = 0} \mu(A) > \frac{\epsilon}{\mathfrak{M}(\mathcal{P}_{\pi_n})} \right) \\
\leq \lim_{n \to \infty} \left( \frac{\log \Delta^*_n(n)}{n^r} + \frac{\mathfrak{M}(\mathcal{P}_{\pi_n})}{n^r} - \frac{\epsilon^4n^{1-r}}{1289^2(\mathfrak{M}(\mathcal{P}_{\pi_n}))} \right) \\
< 0 \quad \text{(from the condition (i) of the theorem).}
\]

Applying the Lemma 3.3, we have \( \forall \epsilon > 0, \)

\[
\mathbb{P} \left( \lim_{n \to \infty} \xi_1 > \epsilon \right) = 0 \quad \mathbb{P}\text{-a.s.} \tag{3.69}
\]

Secondly, the term \( \xi_2 \) is given by

\[
\xi_2 = \sum_{\mu(A) \geq \mu_n(A) > 0} \left[ \mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A) \right] \tag{3.70}
\]

\[
\leq \sum_{\mu(A) \geq \mu_n(A) > \frac{1}{n}} \log \mu_n(A) [\mu_n(A) - \mu(A)] \leq \log n \left[ \sum_{A \in \pi_n} |\mu_n(A) - \mu(A)| \right].
\]

We have the condition in the second inequality because \( \mu_n(A) > 0 \) infers that \( A \) has at
3.3. STRONG CONSISTENCY OF DATA DEPENDENT PARTITIONS

At least one sample point, so \( \mu_n(A) \geq 1/n \). Thus, for all \( \epsilon > 0 \), we have

\[
P(\xi_2 > \epsilon) \leq P\left( \sum_{A \in \pi_n} |\mu_n(A) - \mu(A)| > \frac{\epsilon}{\log n} \right) \leq 8\Delta^*_{\pi_n}(n)^2 2^{\log(\pi_n)} e^{-\frac{\epsilon^2}{128(\log n)^2}}
\]

Hence, \( \exists C < 0 \) such that \( \lim_{n \to \infty} \frac{1}{n} \log P(\xi_2 > \epsilon) < C \). Again, from Lemma 3.3,

\[
P\left( \lim_{n \to \infty} \xi_2 > \epsilon \right) = 0 \quad \text{P-a.s.} \tag{3.71}
\]

Thirdly, the term \( \xi_3 \) is given and separated as follows

\[
\xi_3 = \sum_{\mu_n(A) > \mu(A) > 0} [\mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A)] = \left( \sum_{\frac{1}{2e} > \mu_n(A) > \mu(A)} + \sum_{\frac{1}{e} > \mu_n(A) > \frac{1}{2e} > \mu(A)} \right) [\mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A)]
\]

(3.72)

\[
+ \left( \sum_{\mu(A) > \frac{1}{e}} + \sum_{\mu_n(A) > \mu(A) > \frac{1}{2e}} \right) [\mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A)].
\]

Since the function \( x \log x \) decreases in the interval \([0, e^{-1}]\), the first and the second terms in the right hand side of (3.72) are non-positive. Consider the third term in (3.72), note that \( |x \log x - y \log y| \leq e^{-1} \forall x, y \in [0, 1] \), we have

\[
\sum_{\mu_n(A) > \frac{1}{2e} > \mu(A)} [\mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A)]
\]

\[
\leq \frac{1}{e} \# \left\{ A \in \pi_n : |\mu_n(A) - \mu(A)| > \frac{1}{2e} \right\}
\]

\[
\leq \frac{1}{e} 2^{\log(\pi_n)} P \left( \sum_{A \in \pi_n} |\mu_n(A) - \mu(A)| > \frac{1}{2e} \right) \to 0
\]

as \( n \) tends to infinity, where \( \# \{B\} \) is the number of elements in the set \( B \), and the last convergence is from the Lugosi-Nobol inequality. Thus, for all \( \epsilon > 0 \),

\[
P\left( \lim_{n \to \infty} \sum_{\mu_n(A) > \frac{1}{2e} > \mu(A)} [\mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A)] > \epsilon \right) = 0 \quad \text{P-a.s.} \tag{3.73}
\]

The last term in (3.72) is bounded by

\[
\sum_{\mu_n(A) > \mu(A) > \frac{1}{2e}} [\mu_n(A) \log \mu_n(A) - \mu(A) \log \mu(A)]
\]

109
3. INFORMATION DEPENDENCY ESTIMATION

\[ \leq \sum_{\mu_n(A) > \mu(A) > \frac{1}{\epsilon}} \mu(A) \log \frac{\mu_n(A)}{\mu(A)} \leq \sup_{\mu_n(A) > \mu(A) > \frac{1}{\epsilon}} \left| \log \frac{\mu_n(A)}{\mu(A)} \right| \]

\[ \leq \max \left\{ \sup_{\mu_n(A) > \frac{1}{\epsilon}} \left| \frac{\mu_n(A)}{\mu(A)} - 1 \right|, \sup_{\mu_n(A) > \frac{1}{\epsilon}} \left| \frac{\mu(A)}{\mu_n(A)} - 1 \right| \right\}. \]

Now we prove that for any \( \epsilon > 0 \)

\[ P \left( \lim_{n \to \infty} \sup_{\mu(A) > \frac{1}{\epsilon}} \left| \frac{\mu_n(A)}{\mu(A)} - 1 \right| > \epsilon \right) = P \left( \lim_{n \to \infty} \sup_{\mu_n(A) > \frac{1}{\epsilon}} \left| \frac{\mu(A)}{\mu_n(A)} - 1 \right| > \epsilon \right) = 0 \quad P \text{-a.s.} \quad (3.74) \]

Thus,

\[ P \left( \sup_{\mu(A) > \frac{1}{\epsilon}} \left| \frac{\mu_n(A)}{\mu(A)} - 1 \right| > \epsilon \right) = P \left( \sup_{\mu(A) > \frac{1}{\epsilon}} \left| \mu_n(A) - \mu(A) \right| > \epsilon \mu(A) \right) \]

\[ \leq P \left( \sup_{A \in \pi_n} \left| \mu_n(A) - \mu(A) \right| > \frac{\epsilon}{2\epsilon} \right) \leq 8 \Delta^*_{\pi_n} (n) 2^{\Theta(P_{\pi_n})} e^{-\frac{\epsilon^2 n}{512e^2}} \]

\[ \Rightarrow \frac{1}{n^\tau} \log P \left( \lim_{n \to \infty} \sup_{\mu(A) > \frac{1}{\epsilon}} \left| \frac{\mu_n(A)}{\mu(A)} - 1 \right| > \epsilon \right) \leq - \lim_{n \to \infty} \frac{\epsilon^2 n^{1-\tau}}{512e^2} < C, \]

where \( C \) is a negative number. Then, we have the first statement of (3.74). The second statement of (3.74) is proven similar with replacing \( \mu(A) > \frac{1}{\epsilon} \) by \( \mu_n(A) > \frac{1}{\epsilon} \). The results (3.73), (3.74) and the result that the first and the second terms in the separation of \( \xi_3 \) are non-positive infer that for all \( \epsilon > 0 \),

\[ P \left( \lim_{n \to \infty} \xi_3 > \epsilon \right) = 0 \quad P \text{-a.s.} \quad (3.75) \]

Fourthly, it is clear that \( \nu_n(A) = 0 \Rightarrow \mu_n(A) = 0 \), and then \( \mu_n(A) \log \nu_n(A) = 0 \). Thus, the term \( \xi_4 \) is given by

\[ \xi_4 = \sum_{\nu_n(A)=0} [\mu(A) \log \mu(A) - \mu_n(A) \log \nu_n(A)] = \sum_{\nu_n(A)=0} \mu(A) \log \nu(A) \leq 0. \quad (3.76) \]

Fifthly, if \( \nu_n(A) > 0 \) then \( \nu_n(A) \geq 1/n^m \). Hence, the term \( \xi_5 \) is given by

\[ \xi_5 = \sum_{\nu_n(A) > \mu(A) > 0} [\mu(A) \log \nu(A) - \mu_n(A) \log \nu_n(A)] \]

\[ \leq \sum_{\nu_n(A) > \frac{1}{n^m}} \log \nu_n(A) [\mu(A) - \mu_n(A)] \leq m \log n \left[ \sum_{A \in \pi_n} [\mu(A) - \mu_n(A)] \right]. \quad (3.77) \]
3.3. STRONG CONSISTENCY OF DATA DEPENDENT PARTITIONS

As explaining in the lines between equations (3.70) and (3.71), \( \forall \epsilon > 0 \)
\[
\mathbb{P}\left( \lim_{n \to \infty} m \log n \left[ \sum_{A \in \pi_n} |\mu(A) - \mu_n(A)| \right] > \epsilon \right) = 0 \Rightarrow \mathbb{P}\left( \lim_{n \to \infty} \xi_6 > \epsilon \right) = 0 \quad \text{P-a.s. (3.78)}
\]

Sixthly, the last term \( \xi_6 \) is given by
\[
\xi_6 = \sum_{\nu(A) \geq \nu_n(A) > 0} \frac{\log \nu(A) [\mu(A) - \mu_n(A)]}{\log \nu_n(A)} \leq m \log n \sum_{A \in \pi_n} |\mu(A) - \mu_n(A)|.
\]

(3.79)

Again, \( \forall \epsilon > 0 \)
\[
\mathbb{P}\left( \lim_{n \to \infty} m \log n \left[ \sum_{A \in \pi_n} |\mu(A) - \mu_n(A)| \right] > \epsilon \right) = 0 \Rightarrow \mathbb{P}\left( \lim_{n \to \infty} \xi_6 > \epsilon \right) = 0 \quad \text{P-a.s. (3.80)}
\]

Finally, from (3.67), (3.69), (3.71), (3.75), (3.76), (3.78) and (3.80), we conclude that for every \( \epsilon > 0 \),
\[
\mathbb{P}\left[ \lim_{n \to \infty} \sum_{A \in \pi_n} \left( \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} - \mu(A) \log \frac{\mu_n(A)}{\nu(A)} \right) > \epsilon \right] = 0.
\]

The proof of Theorem 3.6 is finished. \( \square \)

**Repeat of Proposition 3.3.** Let \( \{\pi_n^{DV}\}_{n \in \mathbb{N}} \) be a Darbellay-Vajda partition rule associated with a sequence of integer numbers \( \{h_n\}_{n \in \mathbb{N}} \) and a sequence of positive real numbers \( \{\epsilon_n\}_{n \in \mathbb{N}} \). If \( \epsilon_n \approx (n^{-\frac{1}{2}}) \), and there exists a real number \( \tau \in (\frac{1}{2}, 1) \) such that \( (2^{m h_n}) \approx (n^{-\frac{1}{2}\tau}) \), then \( \{\pi_n^{DV}\}_{n \in \mathbb{N}} \) is a strongly consistent estimation for the information dependency of \( X \).

The proof of Proposition 3.3 requires the following result: Given a rectangle partition \( \pi \) and its cell \( A \) which contains at least two sample points. Given a coordinate \( i \) in the space \( \mathbb{R}^m \), we would like to construct a new rectangle partition which is finer than the partition \( \pi \) by using a hyperplane orthogonal with \( i \)-th-coordinate to split \( A \). Let \( A_1 \) and \( A_2 \) be new cells of a new partition.

**Lemma 3.6.** The information dependency estimation based on a partition of which \( \mu_{n,i}(A_1) = \mu_{n,i}(A_2) \) is not smaller than the one based on a partition of which \( \mu_n(A_1) = \mu_n(A_2) \), where \( \mu_{n,i} \) is the empirical \( i \)-th-marginal distribution, and \( \mu_n \) is the empirical joint distribution.

**Proof.** Denote \( a = |A|, \ (a \geq 2), \ b = |A_1 \times \mathbb{R}^{m-1}|, \) where \( A_1 \) is the projection of \( A \) into the \( i \)-th-coordinate. Let \( \hat{\pi} \) be a finer partition of \( \pi \) of which \( \mu_{n,i}(A_1) = \mu_{n,i}(A_2) \), and denote
3. INFORMATION DEPENDENCY ESTIMATION

\[ a_1 = |A_1|, \ a_2 = |A_2|. \] Let \( \pi \) be a finer partition of \( \pi \) of which \( \mu_n(A_1) = \mu_n(A_2) \), and denote \( b_1 = |A_{1,i} \times \mathbb{R}^{m-1}|, \ b_2 = |A_{2,i} \times \mathbb{R}^{m-1}| \), where \( A_{1,i}, A_{2,i} \) are the projections of \( A_1, A_2 \) into \( i \)th-coordinate respectively. Clearly, \( a_1 + a_2 = a, \ b_1 + b_2 = b \) and \( b_1, b_2 \geq a/2 \geq 1 \). It is not difficult to have

\[ n(\hat{\mathcal{I}}_\pi(X^n_1) - \hat{\mathcal{I}}_\pi(X^n_2)) = (a_1 \log a_1 + a_2 \log a_2 - a \log \frac{a_1}{a_2}) + a(\log b_1 + \log b_2 - \log \frac{a_1}{a_2}). \]

Since the function \( x \log x \) is the convex function, we have \( a_1 \log a_1 + a_2 \log a_2 - a \log \frac{a_1}{a_2} \geq 0, \ \forall a_1, a_2 \). Since \( b_1 + b_2 = b \) and \( b_1, b_2 \geq 1 \), we also have \( \log b_1 + \log b_2 - \log \frac{a_1}{a_2} \geq 0 \). Lemma 3.6 is proven.

\[ \square \]

**Proof of Proposition 3.3.** Let \( \{\pi^n_n\}_{n \in \mathbb{N}} \) be a Tree-structure partition rule associated with a sequence \( \{k_n\}_{n \in \mathbb{N}}, \ k_n = \lfloor n/2^{mh_n} \rfloor \approx n^{\frac{1+\tau}{4}} \). From Proposition 3.2, \( \{\pi^n_n\}_{n \in \mathbb{N}} \) is a strongly consistent estimation for the information dependency of \( X \). Let \( \{\bar{\pi}_n\}_{n \in \mathbb{N}} \) be a partition rule constructing similarly with Darbellay-Vajda partition rule, but we replace the stopping partition condition in Step 3 by the new condition: a terminal node \( t(u,v) \) is not partitioned if \( \text{depth}(t(u,v)) = h_n \). From the constructions of \( \{\pi^n_n\}_{n \in \mathbb{N}} \) and \( \{\bar{\pi}_n\}_{n \in \mathbb{N}} \), when \( n \) is large enough, without loss of generality, we can assume that

1. \( \forall A_1, A_2 \in \pi^n_n(X^n_1) : \mu_n(A_1) = \mu_n(A_2), \)
2. \( \forall B_1, B_2 \in \bar{\pi}_n(X^n_1) : \nu_n(B_1) = \nu_n(B_2), \) and
3. \( |\pi^n_n(X^n_1)| = |\bar{\pi}_n(X^n_1)| = 2^{mh_n}. \)

As a consequence of Lemma 3.6, we have \( \forall X^n_1, \)

\[ \sum_{A \in \pi^n_n(X^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} > \sum_{A \in \bar{\pi}_n(X^n_1)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)}, \]

the condition (ii) of Theorem 3.6. How the partition rule \( \{\bar{\pi}_n\}_{n \in \mathbb{N}} \) fulfills the condition (i) of Theorem 3.6 is explained as follows: Since, \( \forall X^n_1, |\pi^n_n(X^n_1)| = |\bar{\pi}_n(X^n_1)| = \frac{n}{k_n}, \ \{\bar{\pi}_n\}_{n \in \mathbb{N}} \) satisfies the first and second sub-conditions of condition (i),

\[ \lim_{n \to \infty} \frac{1}{n^\tau} \mathfrak{M}(\mathcal{P}_{\bar{\pi}_n}) = 0 \quad \text{and} \quad \lim_{n \to \infty} \frac{n^{1-\tau}}{\mathfrak{M}^2(\mathcal{P}_{\bar{\pi}_n})} = \lim_{n \to \infty} \frac{k_n^2}{n^{1+\tau}} > 0. \]

Let us denote \( \mathcal{P}_{\bar{\pi}_n} = \{\bar{\pi}_n(X^n_1) : X^n_1 \subset \mathbb{R}^m\}. \) From the construction of \( \bar{\pi}_n \), we can give the bounded of \( \Delta_{\bar{\pi}_n}(n) \) similar to the proof of Lemma 3.5. Thus, in the first step of the construction of \( \bar{\pi}_n(X^n_1) \), we have at most \( n^m \) different ways to split \( n \) sample points. In the second step, there are at most \( (n^2)^m \) different ways, and so on, in the \( k \)th step, there are at most \( (n^{2^k-1})^m \) different ways. Since the construction of \( \bar{\pi}_n(X^n_1) \) will stop at the
3.3. STRONG CONSISTENCY OF DATA DEPENDENT PARTITIONS

In the $h_n$th step, we have the bound

$$\Delta^*_\pi(n) \leq \prod_{k=1}^{h_n} \left( n^{2^{k-1}} \right)^m < n^{m2^{h_n}}.$$  

Because $2^{mh_n} \approx n^{1-\frac{\tau}{2}} \Rightarrow 2^{h_n} \approx n^{\frac{1-\frac{\tau}{2}}{2m}}$. Since $\tau \geq \frac{1}{3}$, $m \geq 2$, we have

$$\lim_{n \to \infty} \frac{1}{n} \log \Delta^*_\pi(n) \leq \lim_{n \to \infty} m2^{h_n} \frac{\log n}{n^{\frac{1}{2}}} = \lim_{n \to \infty} n^{\frac{1-(2m+1)\tau}{2m}} \log n = 0.$$  

Hence, $\hat{\pi}_n$ satisfies the third sub-condition of (i) of Theorem 3.6, and then it is a strongly consistent estimation for the information dependency of $X$, i.e.

$$\lim_{n \to \infty} \sum_{A \in \hat{\pi}_n(X^n)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} = I(X), \quad \mathbb{P}\text{-a.s.} \quad (3.81)$$

On the other hand, based on the definition of local independence of cells, the difference between the estimation from $\pi_{DV}^n$ and the estimation from $\hat{\pi}_n$ is bounded by

$$\left| \sum_{A \in \pi_{DV}^n(X^n)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} - \sum_{A \in \hat{\pi}_n(X^n)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} \right| \leq 2^{mh_n} \epsilon_n. \quad (3.82)$$

Since $(2^{mh_n}) \approx (n^{\frac{1-\tau}{2m}})$, $(\epsilon_n) \approx (n^{-\frac{1}{2}})$, we have $2^{mh_n} \epsilon_n \approx n^{-\frac{1}{2}} \to 0$ as $n$ tends to infinity. Combining with (3.81) and (3.82), we have

$$\lim_{n \to \infty} \sum_{A \in \pi_{DV}^n(X^n)} \mu_n(A) \log \frac{\mu_n(A)}{\nu_n(A)} = I(X), \quad \mathbb{P}\text{-a.s.}$$

This completes the proof. \qed
3. INFORMATION DEPENDENCY ESTIMATION
Chapter 4

Independent Component Analysis

4.1 The problem

The *Independent Component Analysis* (ICA) [Comon, 1994; Comon and Jutten, 2010; Hyvärinen and Oja, 2001] is shortly introduced by the following sentence

"The independent component analysis of a random vector consists of searching for a linear transformation that minimizes the statistical dependence between its components."

[Comon, 1994]

More precisely, let $X$ be a random vector on the space $\mathbb{R}^m$, the ICA problem would like to find out the matrix $W \in \mathbb{R}^{m \times m}$ such that a new random vector

$$Y = WX$$

(4.1)

has its components $Y_1, \ldots, Y_m$ getting the smallest statistical dependency.

The ICA can be stated for the general cases $W \in \mathbb{R}^{m \times m'}$. When $m' > m$ the problem is often known as the name *under-determined mixtures*, when $m' < m$ the problem is known as the name *upper-determined mixtures*, and when $m' = m$, the problem is known as the name *determined mixtures*. In this chapter, we would like to study the *determined ICA problem*.

Around 1983, Herault and Jutten addressed the ICA problem as a general framework to solve the *Blind source separation* (BSS) problem. The motivation of the ICA, originated from the motivation of the BSS, is stated as follows: Assuming that the values of the *observed sources* $X$ are generated by a linear mixture of *blind original sources* $S = (S_1, \ldots, S_m)$. The BSS problem try to find out some hypotheses of the original sources such that it is sufficient to determine the original sources based on the values of the observed sources. The ICA problem proposes the hypothesis that the original sources
$S_1, \ldots, S_m$ are mutual independent, and try to determine the sources $Y$ in (4.1) to approximate the original sources $S$. The famous example of the ICA is the cocktail party problem, which is described as follows: In a party, there is a discussion of $m$ people. We take $m$ recorders at $m$ different positions in the room to record the discussion. After the party finish, if someone want to recover the speech of each person, he or she can use the $m$ observed signals of $m$ recorders. The problem is how they recover the speech of each person based on $m$ recording signals. If the speeches of each person are nearly independent, the signals recorded by each recorder are nearly the linear mixture of $m$ original speeches, then this is the ICA problem.

The ICA has many useful applications in various domains of science such as signal processing, statistics, and neural networks, etc. According to [Comon and Jutten, 2010, Chapter 1], in June 2009, there are around 22,000 scientific papers published on the subject of BSS or ICA. Since 1993, many international workshops and conferences are organized per year to develop the problem in ICA, for instance GRETSI since 1993 (France), NOLTA’95 (Las Vegas, USA), ISCAS (Atlanta, USA), EUSIPCO since 1996, NIPS’97 post workshop (Denver, USA), ESANN’97 (Bruges, Belgium), IWANN’99 (Alicante, Spain), MaxEnt2006 (Paris, France) (see in [Comon and Jutten, 2010]).

4.1.1 The properties of ICA

The linear transformation and the statistical dependency are two essences of the ICA problem. In this subsection, we would like to introduce some basic results about the relationship between the linear transformation and the statistical dependency.

A random variable $X$ is said to be Gaussian with mean $\mu$ and variance $\sigma^2$, if its distribution function is $\Phi(\frac{x-\mu}{\sigma})$, where $\sigma \geq 0$ and $\mu$ are constants, and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$ If a random variable $X$ has a distribution function $F(x)$, a characteristic function $\varphi(t)$ of $X$ is defined as follows

$$\varphi(t) = \mathbb{E}(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} dF(x) \quad \forall t \in \mathbb{C}. $$

If $X$ is Gaussian with mean $\mu$ and variance $\sigma^2$, $X$ has the characteristic function:

$$\varphi(t) = e^{it\mu - \frac{1}{2} \sigma^2 t^2} \quad \forall t \in \mathbb{C}. $$

An integral function $f(z)$, $z \in \mathbb{C}$ is said to be of finite order if there is a positive
4.1. THE PROBLEM

number $A$ such that

$$\lim_{|z| \to \infty} e^{-|z|^4} f(z) < \infty.$$  

The lower bound $\rho$ of number $A$ for which this is true is said the order of the function $f$.

Thus, if $f(z)$ is of order $\rho$, $\forall \epsilon > 0$

$$\lim_{|z| \to \infty} e^{-|z|^\rho + \epsilon} f(z) < \infty.$$  

It is clear that a characteristic function of a Gaussian has an order 2.

**Proposition 4.1** ([Cramér, 1962], Theorem 19). *If the sum of two independent random variables is Gaussian, then each variable is itself Gaussian.*

**Proof.** Let $X_1$ and $X_2$ be two independent variables with distribution functions $F_1$ and $F_2$, and the characteristic functions $\varphi_1$ and $\varphi_2$, respectively. Assuming that $X_1 + X_2$ has the distribution function $\Phi\left(\frac{x-\mu}{\sigma}\right)$. Since $\{X_1 \leq x, X_2 \leq y\} \subset \{X_1 + X_2 \leq x + y\}$, then for all $x$ and $y$,

$$F_1(x)F_2(y) \leq \Phi\left(\frac{x+y-\mu}{\sigma}\right). \quad (4.2)$$

We see that, for all $x < 0$

$$\frac{d}{dx}\left(\frac{1}{|x|} e^{-\frac{x^2}{2}}\right) = e^{-\frac{x^2}{2}} + \frac{1}{x^2} e^{-\frac{x^2}{2}} \geq e^{-\frac{x^2}{2}}$$

$$\Rightarrow \quad \Phi(x) \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \left(\frac{d}{dt}\left[\frac{1}{|t|} e^{-\frac{t^2}{2}}\right]\right) dt = \frac{1}{\sqrt{2\pi|x|}} e^{-\frac{x^2}{2}}.$$  

From (4.2), since $X_2$ is independent with $X_1$, we can determine $A, B \in \mathbb{R}$ independent of $x$, such that for all $x < 0$

$$F_1(x) < Ae^{-\frac{x^2}{2\sigma^2} - Bx}. \quad (4.3)$$

Similarly, we can determine $A', B' \in \mathbb{R}$ such that for all $x > 0$

$$1 - F_1(x) < A' e^{-\frac{x^2}{2\sigma^2} + B'x}. \quad (4.4)$$

The inequalities (4.3) and (4.4) confirm the integral

$$J = \int_{-\infty}^{\infty} e^{\frac{x^2}{4\sigma^2}} dF_1(x) \quad (4.5)$$

is convergent. The convergence of (4.5) also confirms that the characteristic function $\varphi_1(t)$ is absolutely and uniformly convergent in every finite domain in the complex-plane.
Thus $\varphi_1(t)$ is an integral function of the complex variable $t$. Since

$$|tx| \leq \sigma^2|t|^2 + \frac{x^2}{4\sigma^2} \quad \Rightarrow \quad |\varphi_1(t)| \leq \int_{-\infty}^{\infty} e^{\sigma^2|t|^2 + \frac{x^2}{4\sigma^2}} dF_1(x) = Je^{\sigma^2|t|^2},$$

so that the order of the integral function $\varphi_1(t)$ does not exceed 2. In the same way it is proved that $\varphi_2(t)$ is an integral function of $t$ and has an the order smaller than 2. Because $X_1 + X_2$ is a Gaussian with the mean $\mu$ and the variance $\sigma^2$,

$$\varphi_1(t)\varphi_2(t) = e^{it\mu - \frac{1}{2}\sigma^2 t^2}$$

which shows that $\varphi_1$ and $\varphi_2$ are integral functions without zeros. In the Theory of Functions, we have an interesting result

**Lemma 4.1** ([Titchmarsh, 1932], p. 250, Hadamard’s factorization theorem). If $f(z)$ is an integral function of order $\rho$, with zeros at $z_1, z_2, \ldots, \neq 0$, $f(0) \neq 0$, then

$$f(z) = e^{Q(z)}P(z),$$

where $P(z)$ is the canonical product formed with the zeros of $f(z)$, and $Q(z)$ is a polynomial of degree not greater than $\rho$.

Applying the Lemma 4.1, we have

$$\varphi_1(t) = e^{q_1(t)} \quad , \quad \varphi_2(t) = e^{q_2(t)},$$

(4.6)

where $q_1(t)$ and $q_2(t)$ are polynomials of degree not greater than 2. The convergence of (4.5) infers that the random variables $X_1$ and $X_2$ have finite means and variances denoted $\mu_1$, $\mu_2$ and $\sigma_1^2$, $\sigma_2^2$, respectively. Then, as a consequence of (4.6),

$$\varphi_1(t) = e^{it\mu_1 - \frac{1}{2}\sigma_1^2 t^2} \quad , \quad \varphi_2(t) = e^{it\mu_2 - \frac{1}{2}\sigma_2^2 t^2} \quad \forall t \in \mathbb{C}.$$ 

The random variables $X_1$ and $X_2$ are Gaussian. The proposition is proven. 

**Corollary 4.1.** Let $X_1, \ldots, X_m$ are $m$ independent random variables. If $Y = \sum_{i=1}^{m} \omega_i X_i$ is Gaussian, then all the variables $X_i$ for which $\omega_i \neq 0$ are Gaussian.

**Proof.** Assuming that $\omega_i \neq 0$. Since $X_1, \ldots, X_m$ are independent, $X_i$ is independent of any linear transformations of $\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m\}$. Thus, if for all $\omega_j = 0$ with $j \neq i$, then $Y = \omega_i X_i$. $Y$ is Gaussian so $X_i$ is also Gaussian. The proof of corollary is finish. If there exists $\omega_j \neq 0$, $j \neq i$, then $X_i$ and $Z = \sum_{j \neq i} \omega_j X_j$ are independent. Because $\omega_i X_i + Z$ is Gaussian, from the Proposition 4.1 $X_i$ and $Z$ are Gaussian.
Continuing for other $X_j$, $j \neq i$ based on the hypothesis $Z$ be Gaussian, the proof of the corollary is finish. 

**Proposition 4.2 (Skitovich-Darmois theorem).** Let $X_1, \ldots, X_m$ be non-deterministic and independent random variables, and suppose that

$$Y_1 = \sum_{i=1}^{m} a_i X_i \quad \text{and} \quad Y_2 = \sum_{i=1}^{m} b_i X_i$$

are independent. Then all the variables $X_i$ for which $a_i b_i \neq 0$ are Gaussian.

The proof of the Proposition 4.2 can be referred in [Darmois, 1953; Skitovich, 1954]. The main idea of this proof is similar to the proof of Proposition 4.1.

**Corollary 4.2.** Let $X$ be an $m$-dimensional random vector and $Y$ be a linear transformation of $X$, $Y = WX$, $W = (\omega_{i,j})_{m \times m} \in \mathbb{R}^{m \times m}$. Suppose that $X$ has its components to be independent, and $Y$ has pairwise independent components, i.e. $\forall u, v \in \{1, \ldots, m\}$, $u \neq v$ such that $Y_u$ and $Y_v$ are independent. If $W$ has two non-zero entries in the same column $i$, then $X_i$ is either Gaussian or deterministic.

**Proof.** Assuming that there exist two non-zero entries in the same column $i$. They are $\omega_{u,i}$ and $\omega_{v,i}$. The hypothesis pairwise independent components of $Y$ infers that $Y_u$ and $Y_v$ are independent. Since

$$Y_u = \sum_{j=1}^{m} \omega_{u,j} X_j \quad \text{and} \quad Y_v = \sum_{j=1}^{m} \omega_{v,j} X_j,$$

and $\omega_{u,i} \omega_{v,i} \neq 0$, the Proposition 4.2 shows that $X_i$ is either Gaussian or deterministic. 

**Corollary 4.3.** Let $X$ be a random vector with its components non-deterministic and independent. Assuming that at most one of its components is Gaussian. Let $Y$ be any linear orthogonal transformations of $X$, $Y = WX$ with $W$ be an orthogonal matrix. Then the following three properties are equivalent:

i. The components $Y_i$ of $Y$ are pairwise independent.

ii. The components $Y_i$ of $Y$ are mutually independent.

iii. The transforming matrix has the form $W = \Lambda P$, where $\Lambda$ is a diagonal and $P$ is a permutation.

**Proof.** [Comon, 1994]: Implications (iii) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (i) are obvious. We shall prove the last one, (i) $\Rightarrow$ (iii). Assume $Y$ has pairwise independent components, and suppose $W$ is not of the form $\Lambda P$. Since $W$ is orthogonal, it necessarily has two non-zero entries in
4. INDEPENDENT COMPONENT ANALYSIS

at least two different columns. The Corollary 4.2 said that \( Y \) will have two components Gaussian or deterministic, which is contrary to the hypothesis of this corollary. Therefore, \( W \) should have the from \( \Lambda P \).

Three above corollaries give us the condition of non-Gaussian of the ICA problem as follows: Assuming that the observed sources \( X \) is an independent linear transformation of original sources \( S \), i.e. \( X = AS \) with an invertible matrix \( A \), and the components of the original sources \( S \) are mutually independent. The ICA will find a new random vector \( Y \) from the observed sources \( X \) to approximate the original sources \( S \). The new sources \( Y \) is a linear transformation of \( X \), i.e. \( Y = WX \), and its components are mutually independent. From this model, we have

\[
Y = (WA)S. \tag{4.7}
\]

We wish the multiply matrix \( (WA) \) having the form \( \Lambda P \), where \( \Lambda \) is a diagonal and \( P \) is a permutation, so the sources \( Y \) is exactly the original sources \( S \) with different the scale and the permutation of components. The Corollary 4.3 says that if \( S \) has its components are non-deterministic and at most one of them is Gaussian, then \( WA \) will have the form \( \Lambda P \), where \( \Lambda \) is a diagonal and \( P \) is a permutation. Moreover, the Corollary 4.1 proposes that the observed sources \( X \) should be non-Gaussian. Thus, the ICA is more reasonable if we add the condition that the original sources \( S \) are non-deterministic and has at most one of its components is Gaussian, and then the observed sources \( X \) has also at most one of its components is Gaussian.

4.1.2 Contrast functions and Information dependency

Let \( \delta \) be a measure of dependency, and \( X \) be a given \( m \)-dimensional random vector. We define the contrast function \( C_X \) from the space of \( m \times m \) real matrices to \( \mathbb{R} \)

\[
C_X(W) = \delta(WX), \tag{4.8}
\]

the dependency of a new random vector \( WX \). The ICA is stated in a mathematical view as follows:

Finding \( W^* \in \mathbb{R}^{m \times m} \) such that \( C_X(W^*) = \min_{W \in \mathbb{R}^{m \times m}} C_X(W) \). \( \tag{4.9} \)

From the properties of the statistical dependency and the optimization, the contrast function of ICA should has some properties:

1. \( \forall W : C_X(WP) = C_X(W) \), for all permutation \( P \): invariant with any permuted transformation.
4.1. THE PROBLEM

2. $\forall W : \mathcal{C}_X(W\Lambda) = \mathcal{C}_X(W)$, for all diagonal $\Lambda$: invariant with any scale changing of components.

3. If $X$ has independent components, then $\mathcal{C}_X(WA) = \mathcal{C}_X(W)$ $\forall W$, if and only if $A$ has the form $\Lambda P$, where $\Lambda$ is a diagonal and $P$ is a permutation: discriminating property.

4. If $X$ has independent components, then $\mathcal{C}_X(A) \geq \mathcal{C}_X(I_m)$, for all invertible matrix $A$, where $I_m$ is an identity matrix.

In the non-parametric field, all most researchers use the information dependency for construct the contrast function of ICA. For instance, the Infomax ICA algorithm [Bell and Sejnowski, 1995; Lee et al., 1999] takes the information dependency between $Y$ and $X$ be the contrast function. They would like to find out the optimal solution $W^*$ by maximizing the quantity $J(Y, X)$. Since

$$J(Y, X) = h(Y) - h(Y|X),$$

in addition the differential conditional entropy $h(Y|X)$, the amount information of $Y$ which did not come from $X$, does not depend on $W$, the maximization of $J(Y, X)$ is equivalent with maximizing the differential entropy $h(Y)$. Other algorithms such as JADE [Cardoso, 1999] and Non-Parametric ICA [Boscolo et al., 2004] try to minimize the information dependency $J(Y)$ for finding $W^*$. We know that

$$J(Y) = \sum_{i=1}^{m} h(Y_i) - h(Y) = \sum_{i=1}^{m} h(Y_i) - \log|\det W| - h(X).$$

Then, the minimizing $J(Y)$ is equivalent with minimizing the sum of the marginal differential entropies $h(Y_i)$ minus log $|\det W|$. Cardoso [1999] estimate the marginal differential entropies by cumulants, and Boscolo et al. [2004] estimate these quantities by using Gaussian kernel estimation. There are some other ideas using to construct the contrast function such as minimizing non-linear cross-correlation of $Y$ [Cichocki and Unbehauen, 1994; Jutten and Herault, 1991] and maximizing non-Gaussian of components of $Y$ [Hyvarinen and Oja, 2000, 2001]. The Central Limit Theorem said that under certain conditions, the sum of independent random variables tends toward a Gaussian. Thus, a sum of $m$ independent random variables is usually closer to Gaussian than any of the $m$ original random variables. Hence, if (4.7) occurs, the possible way to find $Y$ which is near as possible as to the original sources $S$ is making the components of $Y$ are maximized non-Gaussian. Hyvarinen and Oja [2000] suggested measuring the non-Gaussian of random variable by
4. INDEPENDENT COMPONENT ANALYSIS

its kurtosis or its negentropy. The kurtosis of random variable $X$ is defined by

$$
kurt(X) = \mathbb{E}(X^4) - 3[\mathbb{E}(X^2)]^2.
$$

The kurtosis is zero when $X$ is a Gaussian, negative when $X$ is a sub-Gaussian, and positive when $X$ is a super-Gaussian. The kurtosis is more far zero, the variable $X$ is more non-Gaussian. The negentropy of the random variable $X$ is defined as follows:

$$
h_N(X) = h(\tilde{X}) - h(X),
$$

where $\tilde{X}$ is a Gaussian vector of the same covariance matrix as $X$. Since, the Gaussian has the largest differential entropy among all random vector of equal covariance matrix, the negentropy $h_N$ is always nonnegative, and larger when $X$ is more non-Gaussian.

In our works on the ICA problem, we find interested in using the minimizing information dependency $I(Y)$ to construct the contrast function. Concretely, we define the following contrast function

$$
C_X(W) = I_X(W) = I(WX) \quad \forall \ W \in \mathbb{R}^{m \times m}.
$$

(4.10)

**Proposition 4.3.** Given a random vector $X$ which at most one of its components is Gaussian, the function $I_X$ satisfies all four above proposed properties of the contrast function.

**Proof.** From the definition of the information dependency (Definition 2.12), the formula of the information dependency are independent with the scale of each random variables, and the ordering index of them. Hence, the function $I_X$ has the properties (1) and (2). The satisfaction of (4) is a consequence of Proposition 2.12.

Now, we prove $I_X$ having the property (3). Assuming that $X$ has independent components and there exists matrix $A$ such that $I_X(WA) = I_X(W) \ \forall \ W$. Since $I_X(I_mA) = I_X(I_m) = 0$, the random vector $AX$ has independent components. Therefore, the Corollary 4.3 infers that $A$ has the form $\Lambda P$, where $\Lambda$ is a diagonal and $P$ is a permutation. Vice versa, assuming that $A$ has the form $\Lambda P$, the properties (1) and (2) confirm $I_X(WA) = I_X(W) \ \forall \ W$. \hfill $\square$

In next subsection, we would like to study an interested property of the information dependency contrast function, $I_X$ for the simplest case of ICA. Assuming that the observed 2-dimensional random vector $X$ is the linear transformation of two independent random variables $X$ and $N$, where $N$ is Gaussian and $X$ is non-Gaussian. Next subsection study conditions to confirm that this ICA problem is solvable with the contrast function $I_X$. This result is presented carefully in [Kien, 2011].
4.2 The Simplest case of ICA

Let $\Omega$ be any metric space with a norm $\| \cdot \|$. Given an optimization problem defined on $\Omega$

$$\min_{w \in \Omega} f(w),$$

(4.11)

the point $w^*$ is said a local solution of (4.11) if

$$\exists \delta > 0 : \forall w \text{ such that } \|w - w^*\| < \delta, \text{ then } f(w) \geq f(w^*),$$

the point $w^*$ is said a global solution of (4.11) if

$$\forall w \in \Omega \quad f(w) \geq f(w^*).$$

The optimization (4.11) is said solvable if it has an unique local solution, and of course this solution is also a global solution. It is called a solvable optimization because in the theory of optimization, with a general condition, to find local solutions of the optimization we have some feasible methods, but to find global solutions it is really impossible.

Let $X = (X_1, X_2)$ be an 2-dimensional random vector which is an independent linear transformation of $S = (X, N)$, where $X$ is a non-deterministic and non-Gaussian with zero mean, finite variance, and continuously differentiable density function, $N$ is a standard Gaussian independent with $X$, $X = AS$,

(4.12)

where $A$ is an invertible $(2 \times 2)$-matrix. Since $X, N \neq O$ almost everywhere, where $O$ is a deterministic variables receiving zero-value, to rebuild these random variables we consider the ICA problem:

Finding $W^* \in \mathbb{R}^{2 \times 2}$ such that $I(Y^*) = \min_{W \in \mathbb{R}^{2 \times 2}} I(Y)$,

(4.13)

where $Y^* = W^* X$, $Y = W X$, and $\emptyset$ is the collection of all $(2 \times 2)$-matrices which have at least one of its rows be zeros-entries. This problem is called the simplest case of ICA. The matrix $W^*$ and the random vector $Y^*$ are said solutions of this problem.

Proposition 4.3 confirms that $I(WBX) = I(WX)$ if $B$ has the form $\Lambda P$ with $\Lambda$ be a diagonal and $P$ be a permutation. Hence, the optimization (4.13) is equivalent with a following optimization problem

Finding $t_1^*, t_2^* \in \mathbb{R}$ such that $I(Y^*) = \min_{t_1 \leq t_2 \in \mathbb{R}} I(X_1 + t_1 X_2, X_1 + t_2 X_2)$,

(4.14)

where $\mathbb{R} = \mathbb{R} \cup \{+\infty\}$, $Y^* = (X_1 + t_1^* X_2, X_1 + t_2^* X_2)$. Note that, we refer to $I(X_1 +
4. INDEPENDENT COMPONENT ANALYSIS

\( t_1X_2, X_1 + t_2X_2 \) be \( \mathcal{I}(X_1 + t_1X_2, X_2) \) when \( t_2 = \infty \). Because \( A \) is an invertible matrix, we have an one-to-one map from a set of solutions of (4.14) to a set of solutions of the following optimization problem

\[
\text{Finding } t^*_1, t^*_2 \in \mathbb{R} \text{ such that } \mathcal{I}(Y^*) = \min_{t_1 \leq t_2 \in \mathbb{R}} \mathcal{I}(X + t_1N, X + t_2N), \tag{4.15}
\]

where \( Y^* = (X + t_1^*N, X + t_2^*N) \).

**Theorem 4.1** ([Kien, 2011]). The optimization problem (4.15) is solvable with a unique solution be \((t^*_1, t^*_2) = (0, +\infty)\) or \(Y^* = (X, N)\). Hence, the simplest case of ICA is a solvable optimization problem and \(Y^* = (X, N)\) is a unique solution of its.

Note that, the word “unique” has the meaning that if \(Y^*\) is a solution of this optimization problem then all the solutions of this problem will have the form \(\Lambda P Y^*\), where \(\Lambda\) is a diagonal and \(P\) is a permutation.

Two basic notations in Information theory and Estimation theory are needed in the proof of the Theorem 4.1. They are Fisher’s information and Minimum mean square error (mmse). Let \(X\) be a random variable with continuously differentiable density \(f_X\). The Fisher’s information of \(X\) is defined \(\mathcal{I}(X) = \mathbb{E}\rho^2(X)\), where \(\rho = f'_X / f_X\) the score function for \(X\). Let two random variables \(X, Y\) with finite variances. The minimum mean square error in estimating \(X\) with condition of appearing \(Y\) is given by formula \(\text{mmse}(X|Y) = \mathbb{E}\{[X - \mathbb{E}(X|Y)]^2\}\), where the expectation is taken over the joint distribution of \(X\) and \(Y\).

Now, we present four lemma about the relationships between the first derivative of the differential entropy, the Fisher’s information and the minimum mean square error. This results will be used in the proof of Theorem 4.1. Let \(X\) be any non-Gaussian with continuously differentiable density function \(f_X\). Let \(N\) be a standard Gaussian independent with \(X\).

**Lemma 4.2.** For all \(t \in \mathbb{R}\), the convolution \(X + tN\) has an interesting property

\[
\frac{d}{dt} h(X + tN) = t\mathcal{I}(X + tN). \tag{4.16}
\]

**Proof.** Let \(f_t\) be a density of variable \(X + tN\). We have

\[
f_t(y) = \int f_X(x) \frac{1}{\sqrt{2\pi t}} e^{-\frac{||y-x||^2}{2t^2}} dx.
\]
4.2. THE SIMPLEST CASE OF ICA

\[ 
\frac{\partial}{\partial t} f_t(y) = \int f_X(x) \frac{1}{\sqrt{2\pi t^2}} \left( \frac{||y - x||^2}{t^2} - 1 \right) e^{-\frac{||y - x||^2}{2t^2}} dx
\]

1. \[ 
\frac{\partial}{\partial y} f_t(y) = \int f_X(x) \frac{1}{\sqrt{2\pi t^2}} \left( \frac{||y - x||^2}{t^2} \right) e^{-\frac{||y - x||^2}{2t^2}} dx
\]

2. \[ 
\frac{\partial^2}{\partial y^2} f_t(y) = \int f_X(x) \frac{1}{\sqrt{2\pi t^3}} \left( \frac{||y - x||^2}{t^2} - 1 \right) e^{-\frac{||y - x||^2}{2t^2}} dx.
\]

It infers \( \frac{\partial}{\partial y} f_t(y) = t \frac{\partial^2}{\partial y^2} f_X(y) \). Hence, the first derivative of the differential entropy in \( t \) is computed as follows

\[ 
\frac{d}{dt} h(X + tN) = - \int \log f_t(y) \frac{\partial}{\partial t} f_t(y) dy = - \int t \log f_t(y) \frac{\partial^2}{\partial y^2} f_t(y) dy + t \int \left( \frac{\partial}{\partial y} \log f_t(y) \right) \left( \frac{\partial}{\partial y} f_t(y) \right) dy.
\]

Since the Fisher’s information of \( X + tN \) is existed, then a value \(( \frac{\partial}{\partial y} f_t(y)) / \sqrt{f_t(y)} \) is bounded when \( ||y|| \to \infty \). By \( \sqrt{f_t(y)} \log f_t(y) \to 0 \) as \( ||y|| \to \infty \), we have

\[ 
\frac{d}{dt} h(X + tN) = t \int \left[ \frac{f_t(y)}{f_t(y)} \right]^2 f_t(y) dy = t F(X + tN).
\]

The Lemma 4.2 is proven. \( \square \)

Lemma 4.3 ([Guo et al., 2005; Rioul, 2007]). If \( X \) has a finite variance, then

1. \[ t^2 F(X + tN) + \text{mmse}(N)X + tN = 1, \quad \forall t \in \mathbb{R}. \]

2. \[ t^2 F(X + tN) \text{ decreases in } t \in (-\infty, 0], \text{ increases in } t \in [0, +\infty), \text{ and bounded by zero and one.} \]

Proof. Let any random variable \( Z \) with a continuously differentiable density \( f_Z \). Denote \( S(Z) = f_Z / f_Z \), which called score function for \( Z \), be a zero mean variable. Blachman [1965] showed us that \( S(X + tN) = E[S(tN)X + tN] \). By the law of total variance, we have

\[ 
F(X + tN) = \text{Var}[S(X + tN)] = \text{Var}(E[S(tN)X + tN])
\]

\[ = \text{Var}[S(tN)] - E(\text{Var}[S(tN)X + tN])
\]

\[ = F(tN) - \text{mmse}[S(tN)X + tN],
\]

where \( \text{Var}(\cdot) \) denotes a variance operator of random variable. \( tN \) is Gaussian variable with variance \( t^2 \), so \( S(tN) = -tN/t^2 \) and \( F(tN) = t^{-2} \). Therefore

\[ 
F(X + tN) = 1 - \frac{1}{t^2} \text{mmse}(tN)X + tN = 1 - \text{mmse}(N)X + tN.
\]
4. INDEPENDENT COMPONENT ANALYSIS

Of course, \(mmse(N|X+tN)\) increases from zero to one in \(t \in (-\infty, 0]\), and decreases from one to zero in \(t \in [0, +\infty)\). Then \(t^2\mathcal{F}(X+tN)\) decreases from one to zero in \(t \in (-\infty, 0]\), and increases from zero to one in \(t \in [0, +\infty)\). \(\square\)

**Lemma 4.4.** If \(X\) has a finite variance, then, for all \(t \in \mathbb{R}, t \neq 0\)

\[
mmse(N|X+tN) = \frac{1}{t^2} mmse(X|\frac{1}{t}X + N).
\] (4.18)

**Proof.** By the law of total variance,

\[mmse(N|X+tN) = 1 - Var[\mathbb{E}(N|X+tN)] = 1 - \int \frac{\kappa_i(y,t)}{\kappa_0(y,t)} dy,
\]

where \(\kappa_i(y,t) = \int f_X(y-xt) \frac{x_i}{\sqrt{2\pi}} e^{-\frac{|x_i|^2}{2t^2}} dx, \ i = 0, 1\). By changing variable \(z = \frac{y-x}{t}\), we have

\[
\kappa_1(y,t) = \frac{y}{t^2} \int f_X(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{|y-z|^2}{2t^2}} dz - \int f_X(z) \frac{z}{\sqrt{2\pi t^2}} e^{-\frac{|y-z|^2}{2t^2}} dz
\]

\[
= \frac{1}{t} \left( yf_t(y) - \mathbb{E}(X|X_t = y)f_t(y) \right)
\]

\[
\Rightarrow \int \frac{\kappa_1^2(y,t)}{\kappa_0(y,t)} dy = \frac{1}{t^2} \left( Var(X_t) - 2\mathbb{E}[X_t\mathbb{E}(X|X_t)] + \mathbb{E}[(\mathbb{E}(X|X_t))^2] \right)
\]

\[
= \frac{1}{t^2} \left( 1 + t^2 - 2\mathbb{E}[\mathbb{E}X_tX|X_t] + Var[\mathbb{E}(X|X_t)] \right)
\]

\[
= 1 - \frac{1}{t^2} mmse(X|X_t),
\]

where \(X_t = X + tN\). It is not difficult to see \(mmse(X|X_t) = mmse(X|\frac{1}{t}X + N)\). Thus, \(mmse(N|X+tN) = \frac{1}{t^2} mmse(X|\frac{1}{t}X + N)\). \(\square\)

The function \(mmse(N|X+tN)\) decreases from one to zero, and continuous in \(t \in [0, +\infty)\), so an equation \(mmse(N|X+tN) = 1/2\) always exists a unique solution which denoted \(t^*\). Since, \(mmse(N) = 1, \ mmse(N|X+tN) \to 0\) as \(t \to +\infty\), then \(t^* \in (0, +\infty)\). We also define a function \(g: [0, +\infty) \cup \{+\infty\} \to [0, +\infty) \cup \{+\infty\}\) satisfied \(mmse(N|X+g(t)N) = 1 - mmse(N|X+tN) \forall t \in (0, +\infty)\), and \(g(0) = +\infty, g(+\infty) = 0\). The function \(g\) is decreasing and continuous. The last result in this section will focus on a structure of \(mmse(N|X+tN)\) related to new notations \(t^*\) and \(g(t)\), which is the important idea for proving Theorem 4.1.

**Lemma 4.5.** Assume that \(X\) has a finite variance. Let \(t^*\) and \(h(t)\) defined as above. Then

1. \(mmse(N|X+tN) \leq 1 - \frac{t}{2\pi t}, \ \forall t \in [0, t^*], \) and \(mmse(N|X+tN) > 1 - \frac{t}{2\pi t}, \ \forall t \in (t^*, +\infty)\).
4.2. THE SIMPLEST CASE OF ICA

The simplest case of ICA

\[ 1 \equiv B_6 \quad B_5 \quad C_1 \]
\[ 1 \equiv B_3 \quad B_2 \quad B_1 \]
\[ 0 \equiv O_t \equiv A_1 t^* \equiv A_2 A_3 A_4 A_5 A_6 \equiv g(t) \]
\[ C_3 \quad C_4 \quad C_6 \]
\[ +\infty \quad 1 - t^* \quad 2t^* \]

\[ \text{Figure 4.1: Interpreting a geometric view of the proof of Lemma 4.5.} \]

2. An equation

\[ \text{tmmse}(N|X + tN) = g(t)\text{mmse}(N|X + h(t)N) \]

has only three solutions \( t = 0, t^*, \) and \( +\infty \) on the interval \([0, +\infty]\).

\[ \text{Proof.} \quad \text{Let } \xi(t) = \text{mmse}(N|X + tN) - 1 + t/(2t^*), \text{ then } \xi(t^*) = 0. \text{ Denote } M_\delta = \text{Var}(X|\sqrt{\delta}X + N) \text{ for all } \delta \geq 0. \text{ In} [\text{Guo et al., 2005, 2008}], \text{D.Guo, et al. showed that} \]
\[ \frac{d}{dt} \text{mmse}(X|\sqrt{\delta}X + N) = -E(M_\delta^2). \text{ Of course, } \text{mmse}(X|\sqrt{\delta}X + N) = E(M_\delta). \text{ From the Lemma 4.4, we know } \text{mmse}(N|X + tN) = \delta \text{mmse}(X|\sqrt{\delta}X + N) \text{ with } \delta = 1/t^2. \text{ So} \]
\[ \xi'(t) = \frac{d}{dt} \left[ \delta \text{mmse}(X|\sqrt{\delta}X + N) \right] + \frac{1}{2t^*} = \frac{2}{t^3} \left[ \delta E(M_\delta^2) - E(M_\delta) + \frac{t}{4\delta t^*} \right] \]
\[ = \frac{2}{t^3} \left[ \left( \sqrt{\delta}E(M_\delta) - \frac{1}{2\sqrt{\delta}} \right)^2 + \delta \left( E(M_\delta^2) - [E(M_\delta)]^2 \right) + \frac{1}{4\delta} \left( \frac{t}{t^*-1} \right) \right]. \]

Since \( E(M_\delta^2) \geq [E(M_\delta)]^2, \) and \( t/t^* > 1, \forall t > t^* \Rightarrow \xi \text{ strictly increases in } t \in (t^*, +\infty). \)

Because \( \xi(t^*) = 0, \text{ then } \text{mmse}(N|X + tN) > 1 - t/(2t^*) \text{ for all } t > t^* \). Moreover, if exists \( t_1 \in (0, t^*) \text{ such that } \xi(t_1) > 0 \Rightarrow E(M_{\delta_1}) > 1/\delta_1 - 1/(2\delta_1)\sqrt{\delta_2/\delta_1}, \text{ where} \]

127
\[ \delta_1 = t_1^{-2}, \quad \delta^* = t^*-2. \] Apply in formula of \( \xi'(t_1) \), we have

\[
\xi'(t_1) \geq \left[ \sqrt{\delta_1} E(M_{\delta_1}) - \frac{1}{2\sqrt{\delta_1}} \right]^2 + \frac{1}{4\delta_1} \left( \frac{t_1}{t^*} - 1 \right)
\]

\[
> \left[ \frac{1}{\sqrt{\delta_1}} - \frac{1}{2\sqrt{\delta_1}} \left( 1 + \sqrt{\frac{\delta^*}{\delta_1}} \right) \right]^2 + \frac{1}{4\delta_1} \left( \sqrt{\frac{\delta^*}{\delta_1}} - 1 \right)
\]

\[
= \frac{1}{4\delta_1} \left( \sqrt{\frac{\delta^*}{\delta_1}} - 1 \right) \left( \sqrt{\frac{\delta^*}{\delta_1}} - 2 \right).
\]

\( \delta_1 > 0 \) and \( \delta^*/\delta_1 < 1 \) \( \Rightarrow \) \( \xi'(t_1) > 0 \) \( \Rightarrow \exists \varepsilon > 0 : \xi(t_1 + \varepsilon) > \xi(t_1) > 0. \) It means that \( \xi(t) > 0, \forall t \in [t_1, t^*], \) and \( \xi(t^*) > 0. \) Contradict with \( \xi(t^*) = 0. \) Thus, the hypothesis \( \exists t_1 \in (0, t^*) : \xi(t_1) > 0 \) is wrong. Hence, \( \text{mmse}(N|X + tN) \leq 1 - t/(2t^*) \) for all \( 0 \leq t \leq t^*. \)

The second result in this lemma is a consequence of the first result. The explanation is given in figure 4.1. Of course \( t = 0, t^*, +\infty \) are solutions of equation \( \text{mmse}(N|X + tN) = g(t)\text{mmse}(N|X + g(t)N). \) Now we prove that if \( t \in (0, t^*), \) then \( \text{mmse}(N|X + tN) < g(t)\text{mmse}(N|X + g(t)N). \) Indeed, denoting \( |\cdot| \) be a length of line segment in geometric view as in figure 4.1. Denoting \( O \equiv (0, 0), \) \( A_1 \equiv (t, 0), B_4 \equiv (0, \text{mmse}(N|X + tN)), C_2 \equiv (t, \text{mmse}(N|X + tN)), \) and other points \( A_2, A_3, \ldots, C_6 \) as simulating in figure 4.1. Because \( \text{mmse}(N|X + tN) \leq 1 - t/(2t^*), \forall t \in (0, t^*) \Rightarrow |A_1C_2| < |A_1C_1| \Rightarrow |B_5B_6| < |B_4B_6|. \) Let's \( |OB_1| = |B_5B_6| = t/(2t^*), \)

\[
|OB_2| = |B_4B_6| = 1 - \text{mmse}(N|X + tN), \quad \text{we have } |OB_1| \leq |OB_2|.
\]

Because \( \text{mmse}(N|X + tN) > 1 - t/(2t^*), \forall t \in (t^*, +\infty) \Rightarrow |B_2C_4| < |B_2C_6|. \) Note that, since \( C_1, C_3, C_4, C_5 \) are in straight line \( B_6A_5, |B_6B_3| = |OB_1| \leq |OB_2| \leq |OB_3| = 1/2 \Rightarrow S_{OA_1C_1B_5} = S_{OA_3C_3B_5} \leq S_{OA_2C_4B_3}, \) where \( S \) denotes an area of rectangle. Therefore, \( \text{mmse}(N|X + tN) = S_{OA_1C_2B_4} \leq S_{OA_1C_1B_5} \leq S_{OA_3C_3B_5} < S_{OA_2C_4B_3} = g(t)\text{mmse}(N|X + g(t)N). \) It means \( \text{mmse}(N|X + tN) < g(t)\text{mmse}(N|X + g(t)N) \) for all \( t \in (0, t^*). \)

Since, \( g[g(t)] = t, \) then the function \( \text{mmse}(N|X + tN) = g(t)\text{mmse}(N|X + g(t)N) \) also doesn’t have a solution in \( t \in (t^*, +\infty). \) Thus, the equation \( \text{mmse}(N|X + tN) = g(t)\text{mmse}(N|X + g(t)N) \) does not have any solution excepting \( 0, t^*, +\infty. \) The proof of Lemma 4.5 is finish.

\[ \square \]

**Proof of Theorem 4.1.** We know that

\[ \mathcal{I}(X + t_1N, X + t_2N) = h(X + t_1N) + h(X + t_2N) - \log |t_2 - t_1| - h(X, N). \]

Since the differential entropy \( h(X, N) \) is independent with \( t_1, t_2, \) the local solutions of (4.15) are the local solutions of a function \( L(t_1, t_2) = h(X + t_1N) + h(X + t_2N) - \log |t_2 - t_1|, t_1, t_2 \in \mathbb{R}. \) Then, \( (t_1, t_2) \) is a local solution of the optimization problem (4.15) if and
only if it is a solution of the following equations
\[
\begin{align*}
\frac{\partial}{\partial t_1} L(t_1, t_2) &= \frac{\partial}{\partial t_2} L(t_1, t_2) = 0 \\
\left(\frac{\partial^2}{\partial t_i \partial t_j} L(t_1, t_2)\right)_{i,j \in \{1,2\}} &\text{ is positive.}
\end{align*}
\] (4.19)

From the Lemma 4.2, we compute the first derivation of \( L \) as follows
\[
\frac{\partial}{\partial t_1} L(t_1, t_2) = t_1 F(X + t_1 N) + \frac{1}{t_2 - t_1} = \frac{a(t_1)(t_2 - t_1) + t_1}{t_1(t_2 - t_1)}
\]
\[
\frac{\partial}{\partial t_2} L(t_1, t_2) = t_2 F(X + t_2 N) + \frac{1}{t_1 - t_2} = \frac{a(t_2)(t_1 - t_2) + t_2}{t_2(t_1 - t_2)}
\] (4.20)

where \( a(t) = t^2 F(X + t N) \). Equation (4.20) shows us that if \( t_1 = t_2 \), then \( \frac{\partial}{\partial t_1} L(t_1, t_2) = \frac{\partial}{\partial t_2} L(t_1, t_2) = \infty \). Indeed, we only consider \( t_1 \neq t_2 \). Consider the case \( t_1, t_2 \in (0, +\infty) \). Since \( t_1 \leq t_2, a(t_1), t_2 - t_1, t_1 > 0 \Rightarrow \frac{\partial}{\partial t_1} L(t_1, t_2) > 0 \). Similarly for the case \( t_1, t_2 \in (-\infty, 0) \). Hence, \((t_1, t_2)\) is a local solution of (4.15) only if \( t_1 t_2 \leq 0 \) and then \( t_1 \in (-\infty, 0] \), \( t_2 \in [0, +\infty) \). Next, we see that
\[
\begin{align*}
\frac{\partial}{\partial t_1} L(t_1, t_2) = 0 &\quad \Leftrightarrow \quad a(t_1) = \frac{t_1}{t_2 - t_1} \\
\frac{\partial}{\partial t_2} L(t_1, t_2) = 0 &\quad \Rightarrow \quad a(t_2) = \frac{t_2}{t_2 - t_1} \Rightarrow a(t_1) + a(t_2) = 1.
\end{align*}
\] (4.21)

Lemma 4.3 shows that \( a(t) = t^2 F(X + t N) = 1 - \text{mmse}(N|X + t N) \) is a continuous function and decreasing from one to zero in \( t \in \mathbb{R}^+ \). We define a value \( t^* \) and a function \( g : [0, +\infty] \to [0, +\infty] \) satisfy \( a(t^*) = 1/2 \) and \( \text{mmse}(N|X + t N) + g(t)\text{mmse}[N|X + g(t)N] = 1 \) as in Lemma 4.5. Because \( N \) is a standard Gaussian variable, the functions \( h(X + t N) \) and \( \text{mmse}(N|X + t N) \) will be symmetric at \( t = 0 \), i.e \( h(X + t N) = h(X - N), \text{mmse}(N|X + t N) = \text{mmse}(N|X - t N), \forall t \in \mathbb{R} \). Therefore the equation \( a(t_1) + a(t_2) = t_1 \text{mmse}(N|X + t_1 N) + t_2 \text{mmse}(N|X + t_2 N) = 1 \) with \( t_2 \geq 0 \geq t_1 \) infers \( t_1 = -g(t_2) \). So, the function \( L(t_1, t_2) \) can rewrite according parameter \( t \equiv t_2 \) as follows
\[
L(t_1, t_2) \equiv L(t) = h(t) + h[-g(t)] - \log[t + g(t)] = h(t) + h[g(t)] - \log[t + g(t)],
\]
where \( h(t) = h(X + t N) \). Note that, \( t h'(t) + g(t) h'[g(t)] = 1 \) comes from \( a(t) + a[g(t)] = 1, \)
and \( g \) is a nonnegative decrease function. Then

\[
L'(t) = 0 \iff h'(t) + h'[g(t)]g'(t) - \frac{1 + g'(t)}{t + g(t)} = 0
\]

\[
\iff \frac{1 - g(t)h'[g(t)]}{t} + h'[g(t)]g'(t) - \frac{1 + g(t)}{t + g(t)} - \frac{g'(t) - g(t)}{t + g(t)} = 0
\]

\[
\iff \left(g'(t) - \frac{g(t)}{t}\right) \left(g^2(t)f[X + g(t)N] - \frac{g(t)}{t + g(t)}\right) = 0
\]

\[
\iff \text{mmse}[N | X + g(t)N] = 1 - \frac{g(t)}{t + g(t)} = \frac{t}{t + g(t)}
\]

\[
\iff \text{mmse}[N | X + tN] = g(t)\text{mmse}[N | X + g(t)N].
\]

The second result in Lemma 4.5 states that the equation \( \text{mmse}(N | X + tN) = g(t)\text{mmse}(X | X + g(t)N) \) has only three solutions \( t = 0, t^*, +\infty \) on \([0, +\infty]\). By the continuousness property, \( L(t) \) has two local minimum points \( \{0, +\infty\} \), and one local maximum point \( \{t^*\} \). It infers the function \( L(t_1, t_2) \) has only two local minimum points \( (t_1, t_2) \equiv (-\infty, 0) \) and \( (t_1, t_2) \equiv (0, +\infty) \). In other words, \( Y^* = (X, N) \) is a unique local solution of the optimal problem (4.15). The conclusion for the simplest case of ICA is as a consequence of the conclusion of (4.15). Theorem is proven. □

**Simulation of the simplest case of ICA:** We present three examples to simulate the result in Theorem 4.1. The random variable \( X \) in the three examples have corresponding the uniform distribution on \([0, 1]\), the gamma distribution with parameters \((\theta_1, \theta_2) = (2, 1)\), and the exponential distribution with parameter \( \lambda = 1 \). In each example, we use invertible matrices \( A \) to transform the random vector \((X, N)\) to a new random vector \( Y = A(X, N)^T \), where \( N \) is Gaussian independent with \( X \). Now, we would like to simulate that the optimization problems

\[
\min_{-1 \leq \omega_1 < \omega_2 \leq 1} \mathcal{J}(WY) \quad \text{with} \quad W = \begin{pmatrix}
\sqrt{1 - \omega_1^2} & \omega_1 \\
\sqrt{1 - \omega_2^2} & \omega_2
\end{pmatrix}
\]

(4.22)

has a unique local solution. Let

\[
W^* = \begin{pmatrix}
\sqrt{1 - \omega_1^2} & \omega_1^* \\
\sqrt{1 - \omega_2^2} & \omega_2^*
\end{pmatrix},\quad -1 \leq \omega_1^* < \omega_2^* \leq 1, \quad \text{such that} \quad W^*A \quad \text{is a diagonal matrix.}
\]

It is clear that \( W^* \) is a global solution of the above optimization problems. Thus, it is sufficient if we show that in each example, \( W^* \) is the unique local solution. To do that, we construct 20 000 samples of \( Y \), and use the Darbellay-Vajda partition rule to estimation the information dependency of a bivariate \( WY \), where \( W \) is given as (4.22).
4.2. THE SIMPLEST CASE OF ICA

Figure 4.2: Three simulations the unique local solution of the Simplest case of ICA. The star in each sub-figure are the positions of the local solutions. We use the operator (4.23) to simulate the values of Darbellay-Vajda information dependency estimation.
4. INDEPENDENT COMPONENT ANALYSIS

The invertible matrices $A$ used in the three examples are

\[
A_1 = \begin{pmatrix} 1 & -1 \\ 3 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & -3 \\ 2 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 2 \\ -5 & 1 \end{pmatrix},
\]

respectively. The global solution $W^*$ of each example are corresponding

\[
W_1^* = \begin{pmatrix} 0.9487 & -0.3162 \\ 0.7071 & 0.7071 \end{pmatrix}, \quad W_2^* = \begin{pmatrix} 0.8944 & -0.4472 \\ 0.3162 & 0.9487 \end{pmatrix}, \quad W_3^* = \begin{pmatrix} 0.4472 & -0.8944 \\ 0.9806 & 0.1961 \end{pmatrix}.
\]

The Darbellay-Vajda information dependency estimations and the positions of global solutions of each example are given in figure 4.2. Since the logarithmic function is increasing, the figure 4.2 presents the quantities

\[
\log \left[ \log \left[ \log \left[ \log \left[ \log \left( I_{DV} + 1 \right) + 1 \right] + 1 \right] + 1 \right] + 1 \right] + 1 \tag{4.23}
\]

which are replaced to the quantities $I_{DV}$ to exhibit clearly the unique local solution in the domain $-1 \leq \omega_1 < \omega_2 \leq 1$. The star points are positions of two global solutions of the optimization problems in the domain $-1 \leq \omega_1, \omega_2 \leq 1$. These examples exhibit clearly the result of Theorem 4.1. Note that, in the sub-figures of the second example, there seem other local solutions in near the points $(\omega_1 = 0, \omega_2 = -1)$ and $(\omega_1 = -1, \omega_2 = 0)$; however the local solutions of this areas are $(\omega_1^*, \omega_2^*)$ and $(\omega_1^*, \omega_2^* - 1)$ respectively. A note of the third example is similar to the second example.

4.3 Blind Audio Sources Separation

In this section, we would like to introduce an application of ICA. This is the blind audio sources separation. Normally, any sounds in our life are kept as archives by using one or several recorders to record. Each recorder will receive all possible sounds in the near environment and generates a digital audio signal which is a mixture of the possible original sources of sounds. These digital audio signals are often the mixtures of desired sounds and undesired sounds, hence the separation of audio signals is needed keeping the desired sounds.

Fortunately, the physical characteristic of the mixture of digital sound-signals is nearly simple which is shortly stated as follows:

"The source mixing phenomenon results from the simultaneous propagation of sound waves through air from the sources to the recorders and the acoustic-electric conversion performed by the recorders. In normal situations, these phenomena are both linear... the contribution of each source to each recorder
is obtained by linear filtering, i.e. convolution, of the original source signal by a source-recorder filter.” [Comon and Jutten, 2010, Chapter 19, p. 780]

The physical property of the sound-mixture is understood in the mathematical view as follows. Assuming that \( X(t) \) is a group of \( m \) observed digital signals which are the mixtures of a group of \( m \) original audio signals, denoted \( S(t) \). Since an audio signal is closer a time series than a random variable, we would like to use the parameter \( t \) to remind that \( X(t) \) and \( S(t) \) are the values of these signals at the time \( t \). Then there exists a sequence of \((m \times m)\)-matrices \( H = \{H_k\}_{k \in \mathbb{Z}} \) for which \( X(t) \) is a convolution of \( S(t) \) and \( H \),

\[
X(t) = H \ast S(t) = \sum_{k=-\infty}^{\infty} H_k S(t-k) \quad \forall t, \tag{4.24}
\]

where \( \ast \) denotes the linear convolution operator.

Since the audio sound-signals have a wave form, they should be presented on the frequency domain where their characteristic frequencies are exploited. For example, in general a main frequency of a man voice is lower than a woman voice, and a main frequency of a sound from a flute is higher than a sound from a drum. The presentations of signals on the frequency domain are constructed as follows: Let \( X_i(t), S_j(t) \) be the \( i \)-th components of \( X(t) \), \( S(t) \), and \( H_k(i,j) \) be the \( (i,j) \)-th entries of the matrix \( H_k \), \( i,j = 1, \ldots, m \). Given a frequency space \( F = \{0, 1, 2, \ldots, L-1\} \subset \mathbb{N} \) with a frame size \( L \), applying the discrete Fourier transform, the transformations of the signals \( X_i(t), S_j(t) \) and a sequence of convoluted coefficients \( \{H_k(i,j)\}_{k \in \mathbb{Z}} \) in the frequency domain are given: for all \( f \in F \), for all \( t \)

\[
\hat{X}_i(f,t) = \sum_{k=0}^{L-1} e^{-i2\pi \frac{ft}{L}} w(k) X_i(t+k),
\]

\[
\hat{S}_j(f,t) = \sum_{k=0}^{L-1} e^{-i2\pi \frac{ft}{L}} w(k) S_j(t+k), \tag{4.25}
\]

\[
\hat{H}_f(i,j) = \sum_{k=0}^{L-1} e^{-i2\pi \frac{ft}{L}} w(k) H_k(i,j),
\]

where \( w(k) \) is any window function, for example the Hanning window \( w(k) = \frac{1}{2}[1 + \cos(2\pi k/L)] \).

The signals \( \hat{X} = (\hat{X}_1, \ldots, \hat{X}_m) \), \( \hat{S} = (\hat{S}_1, \ldots, \hat{S}_m) \), and the sequence of matrices \( \hat{H} = \{\hat{H}_f\}_{f \in F} \) are called the discrete Fourier transforms of \( X, S \) and \( H \), respectively. They are also reminded as the time-frequency presentations of \( X, S \) and \( H \) corresponding on the frequency space \( F \).

Considering on the time-frequency domain, the equation (4.24) is equivalent, for all
\[ f \in \mathcal{F}, \quad \tilde{X}(f, t) = \hat{H}_f \hat{S}(f, t) \quad \forall t. \quad (4.26) \]

**Dependency in Audio Signals:** Naturally, we often assume that the audio signals coming from independent audio sources should be statistically independent. Recently, Puigt et al. [2009] have confirmed this assumption by using the Kraskov's k-nearest neighbor method to measure thirty independent pairs of speech signals and thirty pairs of music signals. These audio signals are sampled at 22.05 kHz. These results show that when the signals have the durations be larger than 0.75 seconds (around \(2^{14}\) sample points), the information dependency estimations of each pair are very small (< 0.03).

We make three experiments of the human speech signals to study the dependencies between mixtures signals and original signals. In the first experiment, we used two recorders to record the voices of a man and a woman speaking in the same time from one to ten by Vietnamese and English respectively. These signals are sampled at 22.05 kHz. The second and third experiments are borrowed in [Nion et al., 2010], which are available at [http://www.telecom.tuc.gr/~nikos/BSS_Nikos.html](http://www.telecom.tuc.gr/~nikos/BSS_Nikos.html). The second experiment are the record of a man and a woman reading different English stories. The third experiment add to the second experiment another voice of a woman reading another English stories. All these signals are sampled at 16kHz. The structures of these speech signals are given in figure 4.3.

For each experiment, we compute the Darbellay-Vajda information dependency estimations of a group of original signals, a group of mixture signals, and group of original and mixture signals. We compute for 7 different lengths of samples corresponding with 7 durations 200ms, 400ms, 600ms, 800ms, 1s, 1.5s and 2s, and for each duration, we compute for 500 different samples with same length but different starting time points. The mean and the standard deviation of these values are given in the part (A) of the figure 4.4.

These results show us that in most experiments the information dependency estimations of the group of the original signals have the lowest values and the information dependency estimation of the group of the mixture signals have the highest values. The differences between two above groups are clear when the length of durations is large (length > 1.5s).

We also study the dependency of these speech signals in the frequency domain. To do that, we consider the frequency domain \( \mathcal{F} = \{0, 1, \ldots, 511\} \), use the discrete Fourier transform with the Hanning window and the overlapping parameter 0.5, 0.75 and 0.9. The part (B), (C) and (D) of the figure 4.4 present the Darbellay-Vajda information dependency estimations of each groups of signals for each first 80 frequencies. In most of the important frequencies of the human voice (4-60), the information dependency
Figure 4.3: The visual samples of original sound-signals (original) and observed sound-signals (mixture) in 5 seconds of three experiments.
Figure 4.4: The Darbellay-Vajda information dependency estimations in the time domain and the frequency domain. Red: original signals, blue: mixture signals, black: man-mixtures, green: woman-mixtures, and pink: second woman-mixtures (in the third experiment).
estimations of the group of the original signals are lower than of the other groups.

Thus, we can give an idea that any speech signals from original sources should be less dependency than any mixtures of them, especially linear mixtures, and based on this idea we propose a model to solve the blind audio sources separation as follows.

Separating Audio Signals: Assuming that the hypothesis “the audio signals from some independent audio sources should be less dependency than any mixtures of them”, the recovering original sources $S(t)$ based on the observed sources $X(t)$, named as the Blind Audio Sources Separation, is stated as:

Finding a frequency space $\mathcal{F}$ and a sequence $(m \times m)$-complex matrices $\{\hat{W}_f\}_{f \in \mathcal{F}}$ such that for each frequency $f$,

$$\hat{Y}(f, t) = \hat{W}_f \hat{X}(f, t) \quad (4.27)$$

has its components as statistically independent as possible, and

$$Y(t) : \text{an inverse discrete Fourier transform of } \hat{Y}(f, t), \quad (4.28)$$

has its components as statistically independent as possible,

where $\hat{X}(f, t)$ is a discrete Fourier transform of the observed sources $X(t)$ with respect the frequency space $\mathcal{F}$. The new sources $Y(t)$, named separated sources, use to approximate the original sources $S(t)$.

In practice, we often choose the frequency space $\mathcal{F} = \{0, 1, 2, \ldots, L - 1\}$ with $L = 256, 512$ or 1024. Note that, we always have $\hat{Y}_i(f, t) = \hat{Y}_i^H(L - f, t), \forall \ 0 \leq f \leq L/2 - 1, 1 \leq i \leq m$, where $^H$ denote the transpose and conjugate operator. Therefore, the optimization problem (4.27) is simplified by: For each frequency $f \in \{0, 1, \ldots, \frac{L}{2} - 1\}$, finding complex matrix $W_f$ such that $\hat{Y}(f, t) = \hat{W}_f \hat{X}(f, t)$ has its components as statistically independent as possible. Clearly, this is a sequence of $\frac{L}{2}$ Independent Component Analysis problems.

The optimization problem (4.27) can not independently solve the blind audio sources separation. The reason is in the characteristic of the ICA problem, that is for any sequence $\{\hat{W}_f\}_{f \in \mathcal{F}}$ is the optimal solution of (4.27), so for any sequence of diagonal matrices $\{\Lambda_f\}_{f \in \mathcal{F}}$ and any sequence of permutations $\{P_f\}_{f \in \mathcal{F}}$, the sequence $\{P_f \Lambda_f \hat{W}_f\}_{f \in \mathcal{F}}$ is also another optimal solution of (4.27). Nevertheless, a separated source $Y(t)$ generated by $\{\hat{W}_f\}_{f \in \mathcal{F}}$ is different to a separated source $Y_1(t)$ generated by $\{P_f \Lambda_f \hat{W}_f\}_{f \in \mathcal{F}}$. We do not know that $Y(t)$ or $Y_1(t)$ is an optimal solution to approximate the original source $S(t)$. Thus, combining the optimal problem (4.28) with (4.27) to solve the blind audio sources separation problem is needed.
Both problems (4.27) and (4.28) can be done if we have a good and simple tool to measure the dependency. Chapter 2 and Chapter 3 of this thesis show us that the Darbellay-Vajda partition rule estimation and the Kraskov’s $k$-nearest neighbor estimations can be used to estimate the information dependency, the good tool to measure the dependency. Then, we can use these estimations to solve the blind audio sources separation. This is our works in the future.
Chapter 5

Conclusions

In this thesis, we focus on studying the problem “how to express the dependency of a group of random variables”. This problem has two main points:

i. Studying an approximate measure to express the theoretical dependency.

ii. Based on the samples of a group of random variables, estimating the dependency-values expressed by the chosen measure.

In our work, we choose the information dependency to evaluate the dependency, and use the Darbella-Vajda partition rule method (see in [Darbellay and Vajda, 1999]) and Kraskov’s $k$-nearest neighbor method (see in [Kraskov et al., 2004]) to estimate the information dependency.

In chapter 2, we explain why the information dependency is chosen. Firstly, we agree that the best way to express the dependency of a random vector is the way to evaluate the difference between the joint distribution and the product-marginal distribution of this random vector. In information theory, the information dependency is a well-known measure to express the above difference. Secondly, among the popular measures of dependency, there are only the information dependency and the maximal correlation which satisfy the Rényi axioms (see in table 2.1). In contrast to the information dependency, the theoretical values of the maximal correlation cannot be approximated by the samples without a knowledge of the joint distribution of the random vector. Hence, we can say that the information dependency is the unique measure of dependency which satisfies all the conditions of Rényi. Finally, the information dependency can be easily applied for the multivariate cases which keeps the same meaning as applying it for the bivariate case.

In chapter 3, we prove that the Darbellay-Vajda partition estimators are strongly consistent for the information dependency. This is the main result of our work in this thesis. The result bases on some results in Vapnik-Chervonenkis theory (see in [Lugosi and Nobel, 1996; Vapnik, 1998; Vapnik and Chervonenkis, 1971]) and the data-dependent...
5. CONCLUSIONS

partition theory (see in [Breiman et al., 1984; Devroye et al., 1996]). From some simulations of the information dependency estimation, the Darbellay-Vajda partition estimator is the best method for the case of small dimension (i.e. the number of random variables is smaller than 3). In the large dimension cases (i.e. the number of random variables is larger than 4), Kraskov’s $k$-nearest neighbor estimator is the best method. However, the proof of the consistence of Kraskov’s $k$-nearest neighbor estimator for the information dependency is missing. Giving the proofs of the weak or strong consistence of this method for the information dependency is our work in the nearly future.

Finally, in chapter 4, we present the interesting application of our works in the two previous chapters. This application is named the Independent Component Analysis problem. Our result, Theorem 4.1, shows that the simplest case of Independent Component Analysis (see in [Kien, 2011]) can be solved by the information dependency. Three simulations of this result also exhibit that the simplest case of ICA can be solved not only theoretically but also in practical examples by using the Darbellay-Vajda partition rule method to estimate the information dependency. Since, the Independent Component Analysis has many useful applications in the signal processing problem, we would like to explain how the information dependency estimated by the Darbellay-Vajda partition rule works on the Blind Audio Sources Separation in the last section of this chapter. From some simple simulations with speech signals, we believe that the information dependency estimated by the Darbellay-Vajda partition rule method or Kraskov’s $k$-nearest neighbor method can solve well the Blind Speech Sources Separation. These also will be our work in the future.
Appendix

We would like to construct the theoretical estimation of the mutual information of Example 1 in section 2.4 for the cases “\(r\) is positive and small” (\(0 < r \leq 0.15\)) and “\(r\) is large” (\(r \geq 2\)). In these cases, \((X,Y)\) is a continuous random vector, then

\[
I(X,Y) = h(X) + h(Y) - h(X,Y),
\]

where \(h\) is a differential entropy operator. The density functions of \((X,Y)\) is given by

\[
f(x,y) = \frac{1}{5} \left[ \phi_r(x)\phi_r(y) + \phi_r(x-1)\phi_r(y-1) + \phi_r(x-1)\phi_r(y+1) + \phi_r(x+1)\phi_r(y-1) + \phi_r(x+1)\phi_r(y+1) \right], \quad \forall x, y \in \mathbb{R},
\]

\[
f_1(x) = \frac{1}{5} \left[ \phi_r(x) + 2\phi_r(x-1) + 2\phi_r(x+1) \right], \quad \forall x \in \mathbb{R},
\]

\[
f_2(y) = \frac{1}{5} \left[ \phi_r(y) + 2\phi_r(y-1) + 2\phi_r(y+1) \right], \quad \forall y \in \mathbb{R},
\]

Let us consider the case \(0 < r \leq 0.15\): We give the bounds of these entropies.

The upper bounds: Let

\[
g_1(x) = \begin{cases} 
\frac{2}{5}\phi_r(x+1) & x \leq -\frac{1}{2} \\
\frac{4}{5}\phi_r(x) & -\frac{1}{2} < x \leq \frac{1}{2} \\
\frac{2}{5}\phi_r(x-1) & \frac{1}{2} < x 
\end{cases},
\]

\[
g(x,y) = \begin{cases} 
\frac{1}{5}\phi_r(x+1)\phi_r(y+1) & |x|, |y| > \frac{1}{2}, x, y \leq 0 \\
\frac{1}{5}\phi_r(x+1)\phi_r(y-1) & |x|, |y| > \frac{1}{2}, x \leq 0, y > 0 \\
\frac{1}{5}\phi_r(x-1)\phi_r(y-1) & |x|, |y| > \frac{1}{2}, x > 0, y \leq 0 \\
\frac{1}{5}\phi_r(x-1)\phi_r(y) & |x|, |y| > \frac{1}{2}, x, y > 0 \\
\frac{1}{5}\phi_r(x)\phi_r(y) & |x|, |y| \leq \frac{1}{2}.
\end{cases}
\]
APPENDIX

It is clear \( g_1(x) \leq f_1(x) \) and \( g(x, y) \leq f(x, y) \), then

\[
h(X) \leq -\int f(x) \log g_1(x) \quad \text{and} \quad h(X, Y) \leq -\int f(x, y) \log g(x, y) dx dy.
\]

Using the formulae

\[
\int x \phi_r(x) dx = -r^2 \phi_r(x) + C \quad \text{and} \quad \int x^2 \phi_r(x) dx = r^2 \Phi_r(x) - r^2 x \phi_r(x) + C,
\]

we compute the series integrals

\[
A_1 = -\int_{-\infty}^{1/2} \phi_r(x - 1) \log \phi_r(x + 1) dx
\]
\[
= \log(\sqrt{2\pi r}) \Phi_r(-\frac{3}{2}) + \frac{1}{2r^2} \left[ (r^2 + 4) \Phi_r(-\frac{3}{2}) - \frac{5}{2} r^2 \phi_r(\frac{3}{2}) \right]
\]

\[
A_2 = -\int_{-\infty}^{1/2} \phi_r(x) \log \phi_r(x + 1) dx
\]
\[
= \log(\sqrt{2\pi r}) \Phi_r(-\frac{1}{2}) + \frac{1}{2r^2} \left[ (r^2 + 1) \Phi_r(-\frac{1}{2}) - \frac{3}{2} r^2 \phi_r(\frac{1}{2}) \right]
\]

\[
A_3 = -\int_{-\infty}^{1/2} \phi_r(x + 1) \log \phi_r(x + 1) dx
\]
\[
= \log(\sqrt{2\pi r}) \Phi_r(\frac{1}{2}) + \frac{1}{2r^2} \left[ r^2 \Phi_r(\frac{1}{2}) - \frac{1}{2} r^2 \phi_r(\frac{1}{2}) \right]
\]

\[
A_4 = -\int_{-\frac{1}{2}}^{1} \phi_r(x - 1) \log \phi_r(x) dx = \log(\sqrt{2\pi r}) \left[ \Phi_r(-\frac{1}{2}) - \Phi_r(-\frac{3}{2}) \right]
\]
\[
+ \frac{1}{2r^2} \left[ (r^2 + 1) \left( \Phi_r(-\frac{1}{2}) - \Phi_r(-\frac{3}{2}) \right) - \frac{3}{2} r^2 \phi_r(\frac{1}{2}) + \frac{1}{2} r^2 \phi_r(\frac{3}{2}) \right]
\]

\[
A_5 = -\int_{-\frac{1}{2}}^{1/2} \phi_r(x) \log \phi_r(x) dx
\]
\[
= \log(\sqrt{2\pi r}) \left[ \Phi_r(\frac{1}{2}) - \Phi_r(-\frac{1}{2}) \right] + \frac{1}{2r^2} \left[ r^2 \Phi_r(\frac{1}{2}) - \Phi_r(-\frac{1}{2}) \right] - \frac{1}{2} r^2 \phi_r(\frac{1}{2})
\]

\[
A_6 = -\int_{-\frac{1}{2}}^{1/2} \phi_r(x + 1) \log \phi_r(x) dx = \log(\sqrt{2\pi r}) \left[ \Phi_r(\frac{3}{2}) - \Phi_r(\frac{1}{2}) \right]
\]
\[
+ \frac{1}{2r^2} \left[ (r^2 + 1) \left( \Phi_r(\frac{3}{2}) - \Phi_r(\frac{1}{2}) \right) - \frac{3}{2} r^2 \phi_r(\frac{1}{2}) + \frac{1}{2} r^2 \phi_r(\frac{3}{2}) \right]
\]

\[
A_7 = -\int_{1/2}^{\infty} \phi_r(x - 1) \log \phi_r(x - 1) dx
\]
\[
= \log(\sqrt{2\pi r}) \left[ 1 - \Phi_r(-\frac{1}{2}) \right] + \frac{1}{2r^2} \left[ r^2 \left( 1 - \Phi_r(-\frac{1}{2}) \right) - \frac{1}{2} r^2 \phi_r(\frac{1}{2}) \right]
\]

\[
A_8 = -\int_{1/2}^{\infty} \phi_r(x) \log \phi_r(x - 1) dx
\]
\[
= \log(\sqrt{2\pi r}) \left[ 1 - \Phi_r(\frac{1}{2}) \right] + \frac{1}{2r^2} \left[ (r^2 + 1) \left( 1 - \Phi_r(\frac{1}{2}) \right) - \frac{3}{2} r^2 \phi_r(\frac{1}{2}) \right]
\]
\[ A_9 = - \int_{-\infty}^{\infty} \phi_r(x + 1) \log \phi_r(x - 1) \, dx \]
\[ = \log(\sqrt{2\pi}r)[1 - \Phi_r\left(\frac{3}{2}\right)] + \frac{1}{2r^2}\left[(r^2 + 4)\left(1 - \Phi_r\left(\frac{3}{2}\right)\right) - \frac{5}{2}r^2\phi_r\left(\frac{3}{2}\right)\right]. \]

Then

\[- \int_{-\infty}^{\infty} f_1(x) \log g_1(x) \, dx = \log 5 - \int_{-\infty}^{-\frac{1}{2}} f_1(x) \log 2\phi_r(x + 1) \, dx - \int_{\frac{1}{2}}^{\frac{1}{2}} f_1(x) \log 2\phi_r(x - 1) \, dx \]
\[= \log 5 - \log 2\left[1 - \int_{-\frac{1}{2}}^{\frac{1}{2}} f_1(x) \, dx\right] + \frac{2}{5}(A_1 + A_3 + A_4 + A_6 + A_7 + A_9) + \frac{1}{5}(A_2 + A_5 + A_8) \]
\[= \log 5 - \frac{2}{5}\log 2\left[1 + 2\Phi_r\left(-\frac{3}{2}\right) + \Phi_r\left(\frac{1}{2}\right)\right] + \log(\sqrt{2\pi}r) \]
\[+ \frac{1}{10r^2}\left[5r^2 + 12\Phi_r\left(-\frac{3}{2}\right) + 6\Phi_r\left(-\frac{1}{2}\right) - 8r^2\phi_r\left(\frac{3}{2}\right) - 12r^2\phi_r\left(\frac{1}{2}\right)\right]. \]

\[- \int_{\mathbb{R}^2} f(x, y) \log g(x, y) \, dx \, dy = \log 5 - \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x, y) \log [\phi_r(x)\phi_r(y)] \, dx \, dy \]
\[= \log 5 - 4\int_{-\infty}^{0} \int_{-\infty}^{0} f(x, y) \log [\phi_r(x + 1)\phi_r(y + 1)] \, dx \, dy + 4\int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x, y) \log [\phi_r(x + 1)\phi_r(y + 1)] \, dx \, dy \]
\[= \log 5 + 2\log(\sqrt{2\pi}r) \]
\[+ \frac{1}{5r^2}\left(4(r^2 + 4)\Phi_r\left(-\frac{3}{2}\right) + 5r^2\Phi_r\left(\frac{1}{2}\right) + 2\Phi_r\left(-\frac{1}{2}\right)\Phi_r\left(\frac{1}{2}\right)(r^2 + 2) + r^2\Phi_r^2\left(-\frac{1}{2}\right) \right) \]
\[= 8\Phi_r\left(-\frac{1}{2}\right)r^2\phi_r(0) - \left[\frac{9}{2} - 5\Phi_r\left(-\frac{1}{2}\right)\right]r^2\phi_r\left(\frac{1}{2}\right) - 10r^2\phi_r\left(\frac{3}{2}\right). \]

Thus the different entropy \(h(X), h(Y)\) and \(h(X,Y)\) have following upper bounds:

\[h(X), h(Y) \leq C_1 + C_2,\]

where
\[C_1 = \log 5 - \frac{2}{5}\log 2\left[1 + 2\Phi_r\left(-\frac{3}{2}\right) + \Phi_r\left(\frac{1}{2}\right)\right] + \log(\sqrt{2\pi}r) \]
\[C_2 = \frac{1}{10r^2}\left[5r^2 + 12\Phi_r\left(-\frac{3}{2}\right) + 6\Phi_r\left(-\frac{1}{2}\right) - 8r^2\phi_r\left(\frac{3}{2}\right) - 12r^2\phi_r\left(\frac{1}{2}\right)\right], \]
Figure 1: Differences between the density functions $f_1$, $f$ and their lower and upper bounds. The case $0 < r \leq 0.15$.

and

$$h(X, Y) \leq D_1 + D_2,$$

where

$$D_1 = \log 5 + 2 \log(\sqrt{2\pi r})$$

$$D_2 = \frac{1}{5r^2} \left(4r^2 + 4\Phi_r\left(-\frac{3}{2}\right) + 5r^2\Phi_r\left(\frac{1}{2}\right) + 2\Phi_r\left(-\frac{1}{2}\right)\Phi_r\left(\frac{1}{2}\right)(r^2 + 2) + r^2\Phi_r^2\left(-\frac{1}{2}\right)
- 8\Phi_r\left(-\frac{1}{2}\right)r^2\Phi_r(0) - \left[\frac{9}{2} - 5\Phi_r\left(-\frac{1}{2}\right)\right]r^2\Phi_r\left(\frac{1}{2}\right) - 10r^2\Phi_r\left(\frac{3}{2}\right)\right).$$

The lower bounds: Since $r$ is small, we can find a small positive number $\epsilon$ such that with
\[ r_1 = (1 + \epsilon) r, \] the functions
\[
k_1(x) = \begin{cases} 
\frac{2}{5 \sqrt{2\pi}} e^{-\frac{1}{2}r(x+1)^2} & x \leq -\frac{1}{2} \\
\frac{1}{5 \sqrt{2\pi}} e^{-\frac{1}{2}r(x-1)^2} & -\frac{1}{2} < x \leq \frac{1}{2} \\
\frac{2}{5 \sqrt{2\pi}} e^{-\frac{1}{2}r(x-1)^2} & \frac{1}{2} < x 
\end{cases}
\]

\[
k(x, y) = \begin{cases} 
\frac{1}{5 \sqrt{2\pi}e} e^{-\frac{1}{2}r[(x+1)^2+(y+1)^2]} & |x|, |y| > \frac{1}{2}, x, y \leq 0 \\
\frac{1}{5 \sqrt{2\pi}e} e^{-\frac{1}{2}r[(x+1)^2+(y-1)^2]} & |x|, |y| > \frac{1}{2}, 0, x > 0, y \leq 0 \\
\frac{1}{5 \sqrt{2\pi}e} e^{-\frac{1}{2}r[(x-1)^2+(y+1)^2]} & |x|, |y| > \frac{1}{2}, x, y > 0 \\
\frac{1}{5 \sqrt{2\pi}e} e^{-\frac{1}{2}r[(x-1)^2+(y-1)^2]} & |x|, |y| > \frac{1}{2}, x, y \leq 0 \\
\frac{1}{5 \sqrt{2\pi}e} e^{-\frac{1}{2}r|x^2+y^2|} & |x|, |y| \leq \frac{1}{2}, 
\end{cases}
\]

makes
\[
h(X) \geq -\int_\mathbb{R} f_1(x) \log k_1(x) dx \quad \text{and} \quad h(X, Y) \geq -\int_{\mathbb{R}^2} f(x, y) \log k(x, y) dxdy. \quad (1)
\]

We do not give out an exact formula for \( \epsilon \), however finding \( \epsilon \) which satisfies two above inequalities is not difficult. For example, let us consider \( r = 0.1 \) and choose \( \epsilon = 5 \times 10^{-3} \).

The differences between \( k_1(x) \) and \( f_1(x) \), and \( k(x, y) \) and \( f(x, y) \) are given in (C) and (D) of figure 1. Since \( f_1(x), f(x, y) \) are continuous, the figure 1 can confirm that (1) occurs.

The integrals \( \int_\mathbb{R} f_1(x) \log k_1(x) dx \) and \( \int_{\mathbb{R}^2} f(x, y) \log k(x, y) dxdy \) are given by

\[
\int_\mathbb{R} f_1(x) \log k_1(x) dx = \log 5 - \frac{2}{5} \log 2 \left[ 1 + 2\Phi_r \left( -\frac{3}{2} \right) + \Phi_r \left( \frac{1}{2} \right) \right] + \log(\sqrt{2\pi}r)
\]
\[
= C_1 + \frac{1}{(1+\epsilon)^2} C_2,
\]

\[
\int_{\mathbb{R}^2} f(x, y) \log k(x, y) dxdy = \log 5 + 2 \log(\sqrt{2\pi}r)
\]
\[
+ \frac{1}{5r^2} \left( 4(r^2 + 4)\Phi_r \left( -\frac{3}{2} \right) + 5r^2\Phi_r \left( \frac{1}{2} \right) + 2\Phi_r \left( -\frac{1}{2} \right) \Phi_r \left( \frac{1}{2} \right) (r^2 + 2) + r^2\Phi_r^2 \left( -\frac{1}{2} \right) \right.
\]
\[
- 8\Phi_r \left( -\frac{1}{2} \right) r^2\phi_r(0) - \left[ \frac{9}{2} \Phi_r \left( -\frac{1}{2} \right) r^2\phi_r \left( \frac{1}{2} \right) - 10r^2\phi_r \left( \frac{3}{2} \right) \right]
\]
\[
= D_1 + \frac{1}{(1+\epsilon)^2} D_2,
\]

Then
\[
h(X), h(Y) \geq C_1 + \frac{1}{(1+\epsilon)^2} C_2 \quad \text{and} \quad h(X, Y) \geq D_1 + \frac{1}{(1+\epsilon)^2} D_2.
\]
Therefore

\[(2C_1 - D_1) + \left(\frac{2}{(1 + \epsilon)^2}C_2 - D_2\right) \leq f(X, Y) \leq (2C_1 - D_1) + \left(\frac{2C_2}{(1 + \epsilon)^2}D_2\right)\].

Let us consider the large case \( r \geq 2 \):

\textit{The bounds for univariate entropies:} Finding small real numbers \( 0 < \alpha_1 < \alpha_2 \) such that with \( r_1 = r + \alpha_1, \; r_2 = r + \alpha_2 \), the functions

\[ g_1(x) = \frac{\phi_r(0)}{f_1(0)} \frac{1}{r \sqrt{2\pi}} e^{-\frac{1}{2r^2} x^2} \quad \text{and} \quad k_1(x) = \frac{\phi_r(0)}{f_1(0)} \frac{1}{r \sqrt{2\pi}} e^{-\frac{1}{2r^2} x^2}, \; \forall x \in \mathbb{R}. \]

satisfy \( g_1(x) - f_1(x) < 10^{-6} \) and \( f_1(x) - k_1(x) < 10^{-6} \) for all \( x \in \mathbb{R} \). For example, with \( r = 2 \), we choose \( \alpha_1 = 0.16, \; \alpha_2 = 0.23 \) and have

\[ \min_{-20 \leq x \leq 20} \{ f_1(x) - g_1(x) \} = -2.27 \times 10^{-6}, \quad \min_{-20 \leq x \leq 20} \{ k_1(x) - f_1(x) \} = 0. \]

The difference between \( g_1, k_1 \) and \( f_1 \) for the case \( r = 2 \) are given in the figure 2. With the conditions \( g_1(x) - f_1(x) < 10^{-6} \) and \( f_1(x) - k_1(x) < 10^{-6} \) for all \( x \in \mathbb{R} \), we can conclude

\[-\int_{\mathbb{R}} f_1(x) \log k_1(x) dx \leq h(X) = h(Y) \leq -\int_{\mathbb{R}} f_1(x) \log g_1(x) dx\]
Figure 3: Differences between the joint density \( f \) and their bounds. The cases \( r \geq 2 \).

\[
\Rightarrow \quad \log f_1(0) + \frac{5r^2 + 2}{10(r + \alpha_2)^2} \leq h(X) = h(Y) \leq \log f_1(0) + \frac{5r^2 + 2}{10(r + \alpha_1)^2}.
\]

The bounds for bivariate entropy: Finding small real numbers \( 0 < \beta_1 < \beta_2 \) such that with \( R_1 = r + \beta_1, \ R_2 = r + \beta_2 \), the functions

\[
g(x,y) = \frac{\phi_r^2(0)}{f(0,0)} \frac{1}{2\pi r^2} e^{-\frac{1}{2}R^2(x^2+y^2)} \quad \text{and} \quad k(x,y) = \frac{\phi_r^2(0)}{f(0,0)} \frac{1}{2\pi r^2} e^{-\frac{1}{2}R^2(x^2+y^2)}, \quad \forall x,y \in \mathbb{R}
\]

satisfy \( g(x,y) - f(x,y) < 10^{-6} \) and \( f(x,y) - k(x,y) < 10^{-6} \) for all \( x,y \in \mathbb{R} \). For example, with \( r = 2 \), we also choose \( \beta_1 = 0.16, \ \beta_2 = 0.23 \) and have

\[
\min_{-20 \leq x,y \leq 20} \{ f(x,y) - g(x,y) \} = -6.73 \times 10^{-7}, \quad \min_{-20 \leq x,y \leq 20} \{ k(x,y) - f(x,y) \} = 0.
\]

The difference between \( g, k \) and \( f \) for the case \( r = 2 \) are given in the figure 3. The conditions \( g(x,y) - f(x,y) < 10^{-6} \) and \( f(x,y) - k(x,y) < 10^{-6} \) for all \( x,y \in \mathbb{R} \) conclude that

\[
\Rightarrow \quad \log f(0,0) + \frac{5r^2 + 2}{5(r + \beta_2)^2} \leq h(X,Y) \leq \log f(0,0) + \frac{5r^2 + 2}{5(r + \beta_1)^2}.
\]
Finally, the bounds of the information-dependency are given as follows:

\[
\frac{5r^2 + 2}{5} \left( \frac{1}{(r + \alpha_2)^2} - \frac{1}{(r + \beta_1)^2} \right) \leq I(X, Y) \leq \frac{5r^2 + 2}{5} \left( \frac{1}{(r + \alpha_1)^2} - \frac{1}{(r + \beta_2)^2} \right).
\]
Bibliography


BIBLIOGRAPHY


B. Schweizer and A. Sklar. Operations on distributions not derivable from operations on random variables. *Studia Mathematica*, 52:43–52, 1974. 9, 10


Erklärungen

Hiermit erkläre ich, dass diese Arbeit bisher von mir weder an der Mathematisch-Naturwissenschaftlichen Fakultät der Ernst-Moritz-Arndt-Universität Greifswald noch einer anderen wissenschaftlichen Einrichtung zum Zwecke der Promotion eingereicht wurde.

Ferner erkläre ich, daß ich diese Arbeit selbständig verfasst und keine anderen als die darin angegebenen Hilfsmittel und Hilfen benutzt und keine Textabschnitte eines Dritten ohne Kennzeichnung übernommen habe.

Unterschrift des Promovenden

Le Trung Kien
Curriculum Vitae

Name: Le Trung Kien
Date of birth: 12.07.1982
Place of birth: Thua Thien Hue, Vietnam
Nationality: Vietnamese

Education and work experience:
1988 - 2000: Primary school, secondary school and high school in Hue, Vietnam
2000 - 2003: B.S. in Mathematics, Hue University, Hue, Vietnam
2003 - 2004: Student of Osaka University Short-term Student Exchange Programme
Osaka University, Osaka, Japan
2004 - 2005: Continue B.S. in Mathematics, Hue University, Hue, Vietnam
2005 - 2007: Researcher and Teaching Assistance - Faculty of Mathematics,
Hue College of Science, Hue University, Vietnam
2007 - 2008: Diploma of Equivalent in Mathematics - Joint Educational Training
Center Hanoi - Greifswald program
2008 - date: Ph.D. in Institut für Mathematik und Informatik,
Ernst-Moritz-Arndt Universität, Greifswald, Germany.

Publications:
1. Le Trung Kien, Tran Loc Hung and Le Anh Vu. Applying probabilistic model for
ranking Webs in multi-context. *VNU Journal of Science, Mathematics - Physics*,

2. Le Trung Kien, Le Trung Hieu, Tran Loc Hung and Nguyen Duy Tien. Deeper

3. Le Trung Kien, Le Trung Hieu, Tran Loc Hung and Le Anh Vu. MP ageRank: The

4. Le Trung Hieu, Le Anh Vu and Le Trung Kien. An Unsupervised Learning and


Unterschrift des Promovenden

Le Trung Kien
Acknowledgements

I owe my deepest debt of gratitude to my supervisor, Prof. Dr. Christoph Bandt. Not only has he instructed me the knowledge used herein, but he has also taught me the way to think and study Mathematics.

I would like to thank all my colleagues: Dr. Petra Gummelt, Dr. Nguyen Van Son, Dr. Mai The Duy, Katharina Wittfeld, Nguyen Hoang Huy, Stefan Frenzel, Marcus Vollmer, Rüdiger Zeller, Helena Pena and Hoang Thi Thanh Giang for the time to make discussions in seminars and share joking and drinking Vietnamese coffee together.

I appreciate all my Vietnamese friends for sharing their thoughts and having fun with me in Greifswald. It is always in my mind about the time we play football, drink Vietnamese tea, and enjoy the weekend parties together. Helping me relax a lot after hard-working weeks. My warm thanks are specially sent to Nguyen Hoai Bac for taking care of me during my very first time of staying in Greifswald.

Last but not least, I wish to express my deepest gratitude to my family, especially my mother and my wife, who always inspire and support me a lot. My family’s love does encourage me to overcome all obstacles for all of my endeavors on my way to PhD target. Skype software has been interestingly useful in helping me close up my wife during the time of studying PhD.

Thanks for the silence, the ancientness and the beauty of Greifswald!