2011-01-21

A Correlation-Based Method to Detect Weak Dependence

Yabing Luo
Brigham Young University - Provo

A CORRELATION-BASED METHOD TO DETECT WEAK DEPENDENCE

Yabing Luo

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Wynn C. Stirling, Chair
Randal W. Beard
Brian D. Jeffs
Brian A. Mazzeo
Michael D. Rice

Department of Electrical & Computer Engineering

Brigham Young University

April 2011

ABSTRACT

A CORRELATION-BASED METHOD TO DETECT WEAK DEPENDENCE

Yabing Luo

Department of Electrical & Computer Engineering

Doctor of Philosophy

The focus of this thesis is an investigation of ways to detect weak dependence between two random variables X and Y. Our approach is to design tests for correlation rather than testing for dependence directly, since X and Y are not independent if they are not uncorrelated. We examined the magnified Pearson correlation after the Box-Cox transformation to determine whether X and Y are dependent. The results indicated that our approach not only has the potential to detect and evaluate the weak dependence cases that have previously been intractable, but also is conceptually simple and easy to implement.

Keywords: Yabing Luo, Weak dependence, Box-Cox transformation

ACKNOWLEDGMENTS

I am heartily thankful to my supervisor, Wynn C. Stirling, who has supported me throughout my dissertation with his patience and knowledge. I attribute the level of my Ph.D degree to his encouragement and effort.

I would also like to thank Dr. James K. Archibald and my graduate committee members, Dr. Randal W. Beard, Dr. Brian D. Jeffs, Dr. Brian A. Mazzeo, and Dr. Michael D. Rice for their insights into my research progress.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

In probability theory, conditional probability $P(A|B)$ is introduced to capture the partial information that event B provides about event A. If knowing B occurs makes it neither more probable nor less probable that A occurs and $P(A|B)$ gives no information that could affect the probability that A has occurred, then $P(A|B) = P(A)$. We say A is *independent* of B, or A and B are *independent*, when this equality holds. Interesting and important cases of independence arise everywhere in the real world. One typical way of analyzing an application system model is to assume that the probabilistic behaviors of its components are independent, which usually simplifies the calculations and the analysis. For example, two samples collected from a test population at two different times are assumed to have no effect on each other so that the experiment concerning each sample can be counted as an independent trial. If the noise is independent from the signal when it passes through a communication channel, the ordinary approaches such as linear filtering and nonlinear median filtering can be applied to gain accurate insight into the underlying behavior of a communication system before the approximate decoupling of signal and noise is considered. The independence assumption underlying some stochastic processes (Bernoulli, Poisson) has important implications such as a memorylessness property: whatever has happened in past trials provides no information on the outcomes of future trials, and thus allows much simpler solutions of many problems that would otherwise be difficult if independence were not assumed.

The independence between two random variables X and Y has been well investigated in the literature [1]. However, very few studies have addressed a special case known as *weak dependence*, where the random variables have such a low degree of dependence that it is hard to distinguish that case from genuine independence in finite sample applications. To illustrate, consider the following scenario: Let X be time I arrive at home after work, and let Y be the number of customers who shop in a mall. It seems that these two variables are independent because we intuitively think that

knowing the time I open the door of my home does not yield any information about how many people are shopping in the mall. But the truth is that these two variables are both affected by weather. If it is sunny, I may go home earlier and more people may go shopping. If it is raining, I may walk a longer time and arrive at home late while more people may choose not to go outside and thus fewer people are in the mall. The influence of weather is so small that it is possible that no relevance can be found using traditional methods by checking the records of X and Y. In other words, the dependence between X and Y is very weak and cannot be easily detected.

1.1 Motivation

The focus of this study is to investigate ways to detect weak dependence, more specifically, to separate weak dependence from independence. Our work is motivated by the following issues. The first objective is to rule out ambiguity or inconsistency and verify the validity of the independence assumption. The independence principle is the basis of nearly all modern applications, thus its credibility is fundamental when modeling practical phenomena with statistical methods. As a statistically strong and often physically plausible assumption, independence is easy to grasp intuitively. If the occurrence of two events is controlled by distinct and noninteracting processes, such events are then considered as independent. However, independence is not easily illustrated in terms of probabilistic models. For example, to interpret the relationship between two random variables based on n observations, the sample correlation, Kullback-Leibler distance, or other statistics are often computed. The determination of the dependence between random variables requires a considerable amount of data and an appropriate analysis approach. Therefore, the approaches used to rule out ambiguity or inconsistency and to validate the credibility of the independence assumption become necessary and important.

The second motivation is the limitation faced by most of the current independence tests in detecting weak dependence. In spite of the vast literature of unconditional independence research, there still exist some open questions regarding testing independence or dependence relationship among random variables. One of such issues is weak dependence. When finite data samples are collected from weakly dependent random variables, many independence tests fail to separate this extremely low dependence from absolute independence in real applications due to the limitation of sample size and estimation errors. In other words, the results of many traditional approaches

tend to show no difference between the behaviors of weak dependence and genuine independence. Therefore, alternative methods should be investigated.

The third motivation of our study is the direct impact of weak dependence on regression analysis, an essential tool in data analysis in statistics. When two variables are weakly dependent, they are usually considered as independent using the standard regression analysis methods such as the stepwise method. Hence, one variable cannot be viewed as the explanatory entry of another variable and no regressing relationship is built between them. In fact, the explanatory variable, which should have entered into the regressing procedure, becomes a part of regression noise. The bias caused by the omission from a regression of some variables that affect the response variable may lead to a violation of fundamental assumptions required for the unbiasness, consistency, and efficiency of a regression estimator. If weak dependence between random variables can be found, the variable omission and the resulting bias will be avoided, and a more accurate regression model will result.

The fourth motivation is the serious challenge faced by current methods when attempting to represent dependence relationships statistically. To evaluate the distributions or densities of random variables, most classic estimation approaches, such as the nonparametric Nadaraya-Watson kernel density estimator or the parametric polynomial regression estimator, require either the availability of a large number of data samples or the precise estimation of crucial parameters, or both. However, in many actual conditions, a large amount of data may not be available. Moreover, the coarse estimation of parameters could introduce some additional and unnecessary uncertainties to the model which is already burdened with statistical uncertainty. As a consequence, the reliability of an inference coming from such approaches is questionable. It is imperative to find an approach to address such difficulties and to ensure the feasibility and effectiveness of distribution estimators.

Last but not least of the motivations are the practical applications in detecting and differentiating weak dependence. In many applications, the identification of weak dependence is usually the first and necessary step leading to the recognition and confirmation of moderate and even strong dependence. This is extremely helpful when there is no clue or preliminary knowledge of the variables. Through the discovery of weak dependence, investigators will adjust and reassign the input parameters, or reset the experimental conditions, and then reevaluate the new results. The relationship among the variables will be reconsidered through these steps, and the moderate and strong

dependence can be established. As in the case of diagnosis of Parkinson disease, most current techniques rely on the strong signals from the late stage of the disease or anatomization might be required to confirm the occurence of disease. Current direction of the research in the field is to establish the link of some protein level in blood as a sign of the disease at early stage. Considering the large number of candidate proteins in blood, the first-hand experimental data might provide no information and weak links might be dominated by the high noise level caused by other unrelated proteins. Techniques to identify and amplify the weak links then become necessary and extremely useful. As in this case and in many other scientific investigations, the identification of weak dependence is the first step that leads to the important conclusions.

1.2 Related Work

The independence assumption helps simplify problems, facilitate calculations, and find a feasible way to solve complicated problems. More importantly, it is only a simple and convenient condition to obtain a powerful result in many practical applications or theories. For these reasons, independence has drawn great attention in the literature. The following is a brief review of the approaches that are commonly used to measure dependence and to construct a corresponding independence test.

A primary work pertaining to measuring dependence and testing of independence is centered around the Cramer-Von Mises distance. For random variables X and Y, let the joint distribution of X,Y be F_{XY} and the product of marginal distribution of X,Y be $F_X F_Y$. The Cramer-Von Mises type distance is defined as

$$d_c(F_{XY}, F_X F_Y) = \int (F_{XY} - F_X F_Y)^2 \, dF_{XY}.$$

This distance is a measurement of the closeness of F_{XY} and $F_X F_Y$ in the L^2 norm. Hoeffding [2] first investigates such closeness with finite sample distributions in some special cases. Blum, Kiefer and Rosenblatt [3] provide the asymptotic theory about the Cramer-Von Mises distance in the case of having observations (X_i, Y_i) where each pair of $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ are independent and identically-distributed (iid). Their work is also extended to the time-series case where the serial independence is studied. Skaug and Tjøstheim [4] discuss the Cramer-Von Mises distance with

the distributions estimated empirically, while Rosenblatt [5] estimates distributions with a smoothing kernel. The results show that the latter test behaves better in small sample situations with increasing complexity. In addition, An and Cheng [6] use another similar L^1 norm measurement of Kolmogorov-Smirnov distance, $d_k(F_{XY}, F_X F_Y) = \sup |F_{XY} - F_X F_Y|$, to test the independence. All these tests use quadratic metrics and are generally applied to continuous distributions. Skaug and Tjøstheim [7] prove that, in some cases, this kind of test is less powerful than the tests based on the Kullback-Leibler information and Hellinger distance.

If the joint density f_{XY} and the product of marginal densities $f_X f_Y$ exist, an independence test in terms of generalized entropy measure can be constructed. Robinson [8] proposes an independence test measuring the discrepancy between f_{XY} and $f_X f_Y$ with the Kullback-Leibler distance

$$p_K = \int \log \left[\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} \right] f_{XY}(x,y)\, dxdy.$$

It is quite appealing to have an information-theoretic interpretation, especially for communication applications. Skaug and Tjøstheim [7, 9] make an advancement to Robinson's work and adopt the Hellinger distance difference

$$p_H = \int [\sqrt{f_{XY}(x,y)} - \sqrt{f_X(x) f_Y(y)}]^2\, dxdy$$

as the test statistic. It is known that the Hellinger distance is more robust with respect to contaminated data since it is less sensitive to the outliers [10]. Motivated by the fact that both the Kullback-Leibler information p_K and the Hellinger distance p_H are special cases of the generalized Tsallis entropy measure

$$p_{q(f,g)} = \frac{1}{1-q} \left\{ 1 - \int [g(u)/f(u]^{1-q} f(u)\, du \right\},$$

when $q \to 1$ and $q = 1/2$ respectively, Fernandes [11] develops a family of independence tests directly based on this entropy measure. Asymptotic normality and local power are derived using the functional delta method, but they perform poorly in finite samples. Moreover, there is no clear way to select the entropy index to maximize the power of the test. Zhang and Taniguchi [12]

employ the independence tests using another entropy related scale of Renyi entropy measure:

$$R_\lambda = \frac{2}{\lambda(\lambda+1)} \log \int \left[\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right]^{1+\lambda} - (1+\lambda)\left[\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} - 1\right] f_X(x)f_Y(y)\, dxdy, \lambda \neq 0, -1.$$

Simulation results indicate that their tests outperform the ordinary Kullback-Leibler independence test in examining goodness of fit between the densities.

There are other types of independence tests. The BDS test named after its originators: Brock, Dechert, Scheinkman and LeBaron [13], examines serial independence using the correlation integral of chaos theory. It is effective against a wide range of econometric applications including autoregressive conditional heteroscedasticity (ARCH) and generalized ARCH (GARCH) models. Rank tests serve as another valuable alternative in some situations. Hallin, Jurečková, Pciek and Zahaf [14] apply a rank independence test for autoregression rank scores. Pinkse [15] suggests a consistent nonparametric test for serial independence based on the characteristic functions. It performs better than the correlation dimension test of Brock [16].

All of the works cited above have the common focus of finding an accurate and efficient way to determine the independence between random variables. However, the emphasis in this dissertation is to differentiate weak dependence from the genuine independence, and therefore an alternative method is needed. In 2008, Su and White [17] provide a test which aims at conditional independence verification regarding the weak dependence case. With a modified Hellinger metric statistic, this test is superior to other conditional independence tests in that it overcomes the difficulty of distinguishing the difference between weak dependence and independence. However, the test concerns conditional independence and is rather complicated.

1.3 Technical Approach

To investigate the weak dependence relationship between two random variables X and Y, our first step is to transform X and Y to form new random variables U and V that are more amenable to analysis. Let F_X and F_Y denote the cumulative distribution functions of X and Y, respectively, and define U and V as

$$U = F_X(X)$$

$$V = F_Y(Y).$$

Since X and Y are independent if and only if U and V are independent, it suffices to study the dependency of U and V. The motivation for this transformation is that it converts the original X and Y to nonnegative variables and, as a result, the Box-Cox transformation, which is based on nonnegative variables, can be applied.

Rather than testing for dependence directly, the approach of this paper is to design tests for correlation. As is well known, if X and Y are independent, then they are uncorrelated. Although the converse is not true (except in the Gaussian case), the contrapositive can be a useful tool. That is, if X and Y are not uncorrelated, then they are not independent.

We construct a significance test to establish the hypothesis that X and Y are correlated. Our approach is to determine whether $U = F_X(X)$ and $V = F_Y(Y)$ are dependent by computing the Pearson correlation between U and V. The main emphasis of our study is to discriminate weak dependence from genuine independence through a hypothesis test. Weak dependence represents extremely low dependence between random variables in both linear relations (measured by Pearson correlation) and other type of relations (measured by Spearman correlation, Kendall correlation, etc). We choose to adopt the Pearson correlation instead of other correlation measures as the measurement of dependence and the hypothesis test statistic because the Pearson correlation can be amplified using the Box-Cox transformation of the nonnegative functions $F_X(\cdot)$ and $F_Y(\cdot)$, and hence can be more easily detected.

Our approach is to apply the Box-Cox transformation to enhance the linear relationship between U and V, perform a test to determine the significance of the Pearson correlation coefficient, and then make a decision as to whether U and V are correlated. If U and V are tested as correlated, it is certain that X and Y are correlated and, hence, dependent. However, even if the test result that U and V are uncorrelated may give a clue that X and Y are uncorrelated, it certainly does not establish the independence between X and Y unless they are jointly normally distributed. Strictly speaking, we are proposing a dependence test instead of an independence test. The essence of our approach is to detect the correlation and determine weak dependence. As a supplement, we also introduce a way to construct an independence test based on U and V. Distinct types of test statistics and distribution estimation methods are applied to demonstrate the performance of independence tests.

1.4 Contribution

In this study, we propose a correlation test based on an after-transform Pearson correlation coefficient. Our method has three notable features. First, it is designed to address the inability of current independence tests to examine weak dependence, due to the difficulty in distinguishing the difference between weak dependence(*almost* no dependence among events) and independence(*absolutely* no dependence among events). The traditional independence tests behave rather satisfactorily in most cases, but they are not designed to address the special case of weak dependence. Su and White investigate the weak conditional dependence problem in detail. Their test performs well, but at the price of complexity. Our method is built on a basic rule: if X and Y are independent, then the arbitrary functions of X and Y, $g(X)$ and $h(Y)$, are also independent; if $g(X)$ and $h(Y)$ are not independent, then X and Y are not independent. Our approach is to define transformations $g(X)$ and $h(Y)$ whose correlation relationship is enlarged, and then detect this magnified correlation. We use the Box-Cox function as our desired transformation. The Box-Cox transformation can improve the linear fit between two random variables and therefore enhance the correlation test based on an after-transform Pearson correlation [18,19]. Our method not only has the potential to detect and evaluate the weak dependence cases that have previously been intractable, but also is conceptually simple and easy to implement.

Through out this study we apply current statistical techniques. We ensure the validity and accuracy of distribution estimation by using both parametric and nonparametric estimation methods. The nonparametric Nadaraya-Watson kernel density estimator is applied to obtain the marginal distributions and the performance properties of this one-dimensional estimator have been clearly illustrated in many areas. The combination of parametric copula estimation and nonparametric Nadaraya-Watson kernel density estimation is used to obtain the joint distributions for distinct test statistics such as Cramer-Von Mises distance. This semi-parametric method not only avoids the curse of dimensionality of fully nonparametric estimation but also requires less dependency on the underlying distribution assumption which constitutes the foundation of parametric estimation and sometimes is the major reason for an unsatisfactory estimation result.

Third, rather than restricting the whole study in a pure numerical analysis environment, we provide some real applications to demonstrate the benefit of weak dependence detection as accomplished by our after-transform correlation test.

1.5 Organization

The dissertation is organized as follows.

- Chapter 2 provides a definition of weak dependence for stochastic variables in terms of the covariance.
- Chapter 3 introduces different approaches to estimate distribution and density functions. The Nadaraya-Watson(NW) kernel density estimation is presented to compute marginal distributions, the double kernel local linear estimation method is used to estimate the conditional densities, and the copula theory is introduced to make probabilistic inferences of joint distributions.
- Chapter 4 introduces the bootstrap method and provides the way to compute the cutoff value of our hypothesis tests regarding weak dependence.
- Chapter 5 illustrates the test statistics of our dependence tests. The Pearson correlation coefficient of the Box-Cox transformed random variables is tested using the bootstrap approach. We pay particular attention to weak dependence cases, since many previous independence tests cannot successfully separate the weak dependence from the genuine independence.
- Chapter 6 shows Monte Carlo simulation results of our hypothesis tests based on the Pearson correlation. First, correlation tests before and after the Box-Cox transformation are implemented. Then, independence tests using different test statistics and different distribution estimation methods are conducted and their performance in detecting weak dependence is displayed. Finally, we compare and summarize the simulation results of each test graphically.
- Chapter 7 provides two practical applications, regression analysis and weak signal identification, where weak dependence detection is helpful.
- Chapter 8 concludes our work with a brief summary remark.

CHAPTER 2. WEAK DEPENDENCE

Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. Various measures of dependence have been introduced, such as the Hellinger distance p_H, the Kullback-Leibler distance p_K, or the Cramer-Von Mises distance d_c mentioned in Chapter 1, to attempt to explain any of a broad class of statistical relationships between random variables X and Y. On the one hand, conducting statistical tests based on dependence measures answers one essential question, that is, whether or not X and Y are independent, by telling us if these measures are significantly different from zero. On the other hand, another important question arises: is there a way to describe the strength of the dependence relationship between X and Y according to the dependence measures? In other words, which cases should be considered weakly, moderately or strongly dependent, respectively, by summarizing the values of dependence measures? More precisely, we are interested in finding the mathematical framework of weak dependence because it will be shown later that weak dependence can easily get confused with genuine independence in many practical applications and therefore is of critical importance.

In the literature, the notion of weak dependence is introduced to describe explicitly the asymptotic independence between "past" and "future" of a time series. According to Doukhan and Louhichi, weak dependence represents a concept of fading memory [20]. That is,

The "past" is progressively forgotten. In terms of the initial time series, "past" and "future" are elementary events given through finite dimensional marginals. Roughly speaking, for convenient functions f and g, we shall assume that

$$cov(f(\text{“}past\text{”}), g(\text{“}future\text{”}))$$

is small when the distance between the "past" and the "future" is sufficiently large. Such inequalities are significant only if the distance between indices of the initial time series in the "past" and "future" terms grows to infinity.

Weak dependence conditions of a time series arise in a variety of economic and financial applications and such deep econometric motivations finally lead to a plurality of powerful theorems and results in reality [20–22]. This concept of weak dependence emphasizes the diminishing dependence between the "past" and the "future" when this time gap becomes larger and larger. However, instead of focusing on the extent that the past influences the future in a time series, we are more interested in exploiting the strength of statistical dependence between two random variables X and Y. Conceptually, the focus of both the classical weak dependence and the weak dependence we care about is to investigate the imperceptible or subtle relationship between some specific variables. Mathematically, the classical weak dependence represents the dependence relationship between the "past" and the "future" of a time series through a function of covariance. Assume $\{X_t\}_{t\in\mathbb{Z}}$ is a time series. Let $\{X_i, i \in [i_1, i_u]\}$ be a sequence of random variables which denotes the the "past", and let $\{X_j, j \in [j_1, j_v]\}$ be a sequence of random variables which denotes the the "future", where $i_1 < \cdots < i_u \leq i_u + k \leq j_1 < \cdots < j_v$. If $|\mathrm{cov}(f(X_{i_1}, \cdots, X_{i_u})), g(X_{j_1}, \cdots, X_{j_v}))|$, where f and g are some functions, decreases to zero when k goes to infinity, then $\{X_t\}_{t\in\mathbb{Z}}$ is said to be weakly dependent. We borrow this useful mathematical tool, the covariance, to interpret the new concept of weak dependence between two stochastic variables X and Y. What we are doing, in some sense, is to simplify the definition of classical weak dependence in order to fit our needs. Using a single random variable X and Y to replace, respectively, the random sequence X_i and X_j, which denote the the "past" and the "future", respectively, we define *weak independence* in terms of the covariance as follows.

Definition 2.1 Let X and Y be two random variables which take values $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, where $\mathscr{X}, \mathscr{Y} \subseteq \mathbb{R}$. Let f and g be any arbitrary real-valued functions of X and Y respectively, that is, $f : \mathscr{X} \to \mathbb{R}$ and $g : \mathscr{Y} \to \mathbb{R}$. Define

$$c = \sup_{f,g} |\mathrm{cov}(f(X), g(Y))|.$$

If $c = 0$, then, we say, X and Y are weakly independent. □

Obviously, if we can find some functions which satisfy $c > 0$, then X and Y are dependent. Furthermore, can we determine that X and Y are weakly dependent based on the value of c? To answer this question, we apply the normalized covariance: the Pearson correlation coefficient as the measurement and therefore we can use the Cohen scale as the reference.

The Pearson product-moment correlation coefficient is measured on a standard scale – it can only range between -1.0 and +1.0. As such, we can interpret the correlation coefficient as a representing tool which tells us the strength of the linear relationship between the two variables. Cohen [23] suggests a convention to interpret this relationship. A correlation coefficient less than 0.10 in magnitude is trivial and considered as a very weak association; a correlation coefficient of 0.10 to 0.29 in magnitude is thought to represent a weak or small association; a correlation coefficient of 0.30 to 0.49 in magnitude is considered as a moderate or medium correlation; and a correlation coefficient of 0.50 or larger in magnitude is thought to represent a strong or large correlation. In brief, Cohen scale suggests:

- $|\rho| < 0.30$ small/weak correlation;
- $0.30 \leq |\rho| < 0.50$ medium/moderate correlation;
- $|\rho| \geq 0.50$ large/strong correlation,

where ρ represents the Pearson correlation between two stochastic random variables. In practice, ρ is usually replaced by the sample Pearson correlation between the samples of two random variables.

The Cohen scale really depends on the context. A correlation of 0.5 might be regarded as strong in social science situations, but it may also be considered as weak in some physical science situations where instrumentation can be extremely precise. Therefore, the cutoff criteria of the Cohen scale are somewhat arbitrary and may not be applied too strictly in a real scenario. However, in our opinion, it is sufficient to view the Cohen scale as a convenient criterion to separate a special weak dependence case from the genuine independence.

Let

$$\rho_{sup} = \frac{c}{\sigma(f(X))\sigma(g(Y))},$$

where $f(X), g(Y)$ and c are defined in 2.1, and $\sigma(\cdot)$ is the standard deviation. Derived from the Cohen's scale, we claim:

- $\rho_{sup} < 0.30$ small/weak dependence;
- $0.30 \leq \rho_{sup} < 0.50$ medium/moderate dependence;
- $\rho_{sup} \geq 0.50$ large/strong dependence.

ρ_{sup}, also known as Renyi correlation, is studied by Sethuraman [24] and Kumar [25]. Sethuraman provides the asymptotic distribution of Renyi correlation, while Kumar suggests an upper bound of Renyi correlation. However, the calculation of ρ_{sup} is still practically unattractive on the basis of these results. It is difficult to obtain ρ_{sup} without specific information, such as the closed form probability distributions of both X and Y. In the simulations of Chapter 6, we use the maximal sample correlation r_{sup} rather than ρ_{sup} to roughly, however, robustly, divide the level of dependence between two random variables.

Example 1: assume $Y = X^2$, where X has a density function that is symmetric about 0 and the third moment of X exists. We know that $\text{cov}(X,Y) = 0$ and the corresponding correlation coefficient $\rho = 0$. But are X and Y weakly dependent? It is apparent that the correlation coefficient $\rho' = \text{cov}(X,g(Y))/[\sigma(X)\sigma(g(Y))] = \text{cov}(X,X)/\sigma^2(X) = 1$ if $g(Y) = \sqrt{Y}$. Therefore, even if X does not linearly relate to Y due to $\rho = 0$, $\rho_{sup} = \rho' = 1$ represents that X and Y are strongly dependent, nonlinearly.

Example 2: suppose $Y = X + Z/k$, where X,Z are i.i.d standard normally distributed and k is a nonzero integer. By simple calculations, we know that the correlation coefficient $\rho = c/[\sigma(X)\sigma(Y)] = 1/\sqrt{1+1/k^2}$. When $k \to 0$, $\rho \to 0$. Let $g_1(Y) = Y^2$ and $g_2(Y) = Y^3$, then

$$\rho_1 = \frac{\text{cov}(X,g_1(Y))}{\sigma(X)\sigma(g_1(Y))} = \frac{E(X^3) - E(X)E(X^2)}{\sigma(X)\sigma(Y^2)} = 0,$$

and

$$\rho_2 = \frac{\text{cov}(X,g_2(Y))}{\sigma(X)\sigma(g_2(Y))} = \frac{E(X^4)}{E(X^6) + 15/k^2 E(X^4Z^2) + 15/k^4 E(X^2Z^4) + 1/k^6 E(Z^6)}.$$

ρ_2 also goes to zero when k is extremely small. In fact, we can come to a conclusion intuitively: the correlation between the newly generated random variables $f(X)$ and $g(Y)$ can be enlarged, that is, the absolute value of ρ is great than zero, when f and g are randomly picked. But the maximum magnitude of ρ, or ρ_{sup}, as $k \to 0$, should always be small no matter what forms f and g take. When k is small, only a tiny piece of knowledge about X is contained in Y comparing to the

amount of knowledge of Z contained in Y. Therefore, by making modifications to X and Y with some functions $f(X)$ and $g(Y)$, it is rather difficult to grasp and summarize the characteristic of the little amount of information of X in Y and increase the linear relationship between $f(X)$ and $g(Y)$. In other words, no matter what functions $f(X)$ and $g(Y)$ are applied, it is not likely that the dependence relationship of Y is greatly improved in the direction of X instead of in the direction of Z. Therefore, in this case, we say that X and Y are weakly dependent. Theoretically, the value of ρ_{sup} should be confirmed to classify the level of dependence. However, practically, the value of the sample maximal correlation ρ_{sup} is used to find the weak dependence situations.

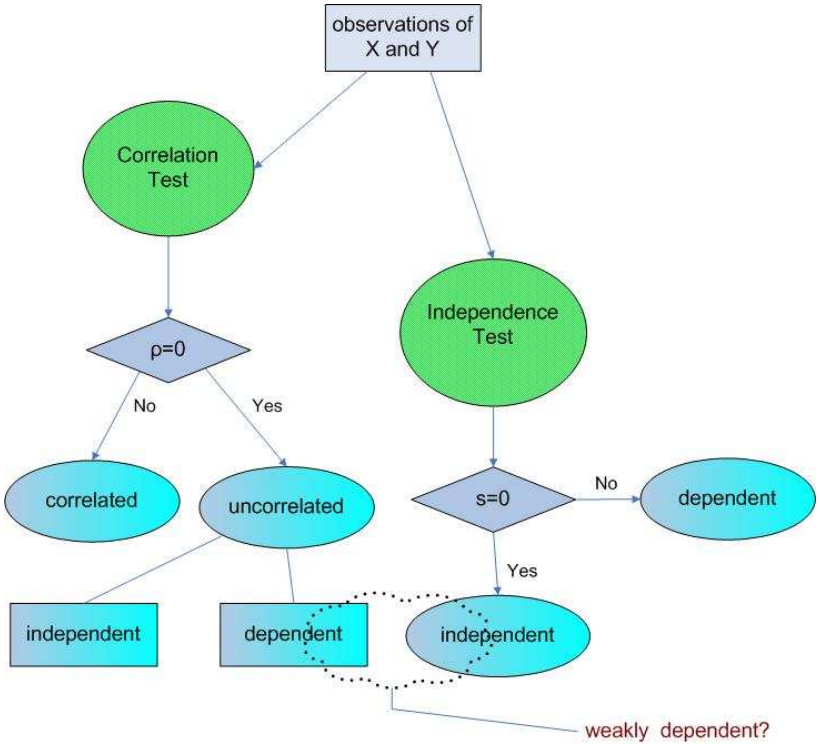

Figure 2.1: Statistical test illustration diagram

In real scenarios, the general statistical tests tend to improperly determine that X and Y in the cases like Example 2 are independent by examining the statistics which measure the dependence relationship between the observations of X and Y. Figure 2.1 provides the illustration of two kind of general statistical tests: the correlation test and the independence test. Both tests are performed to decide if random variables X and Y are independent. As shown in the left side of

Figure 2.1, the correlation test determines whether the population correlation ρ statistically equals zero and yields the result that X and Y are either uncorrelated or correlated. Correlated X and Y certainly means dependent X and Y, but uncorrelated X and Y does not necessarily imply independent X and Y. We wish to identify the cases where X and Y are weakly dependent if X and Y are uncorrelated. The independence test is demonstrated in the right side of Figure 2.1. By determining if the test statistic s, which might be a Kullback-Leibler distance or Cramer-Von Mises distance, is significantly different from zero, it seems that the independence test directly provides the conclusion whether X and Y are independent or not and thus no further test is needed. This statement is theoretically true, but biased test results are sometimes present in practice. As we will see in Chapter 6, due to the restrictions of the sample size, test algorithms, and the computation accuracy, the general independence tests sometimes confuse the genuine independence and weak dependence cases. In Figure 2.1, the independence result might falsely include weak dependence cases. Therefore, a further examination to separate these two cases is indispensable. Our study is to find an efficient way to solve this problem, that is, differentiating the weak dependence from genuine independence in real applications.

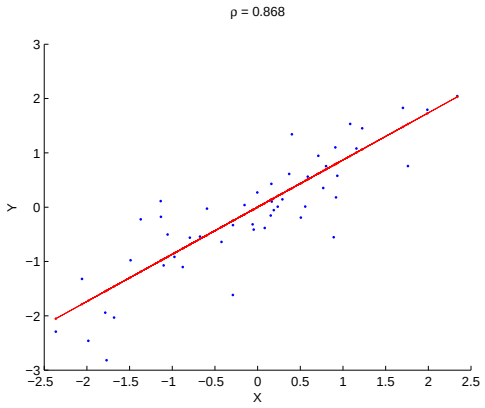

Figure 2.2: Linear correlation $\rho = 0.868$

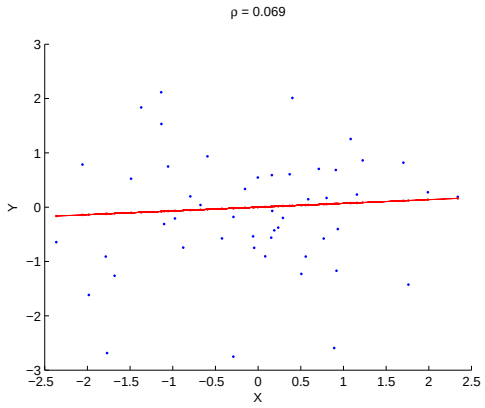

Figure 2.3: Linear correlation $\rho = 0.069$

As is well known, the Pearson correlation ρ only represents the degree of linearity between random variables X and Y. $|\rho(X,Y)| = 1$ if and only if X and Y are related by an affine transformation, i.e., $Y = aX + b$, where a and b are arbitrary constants. Figures 2.2 and 2.3 are two scatter plots of variables X and Y. Notice in Figure 2.2, as X increases, Y also tends to increase. If this

were a perfect positive correlation, all of the points would fall on the solid straight line which is the best linear fit based on some regression criterion. The closer the data points to the linear fit line, the more linear relationship between X and Y and the larger the Pearson correlation coefficient. Actually, $\rho = 0.868$ indicates a strong linear relationship between X and Y and such a relationship is easily discovered in practical applications. However, in Figure 2.3, there seems to be no relationship between X and Y. The data points spread around the solid linear fit line, and X does not linearly interact with Y. The corresponding Pearson correlation coefficient is $\rho = 0.069$, and this weak relationship might be difficult to detect in practical applications. Suppose there were to exist a nonlinear transformation which would transform the random variables X and Y in Figure 2.3 to the random variables in Figure 2.2. As a result, the closer linear relationship in Figure 2.2 would be easier to observe. We utilize this idea to implement the weak dependence detection in the following study.

Figure 2.4 indicates our method to detect the weak dependence based on the Pearson correlation coefficient. With the Box-Cox transformation, the original random variables X and Y are transformed into two new random variables $g(U)$ and $h(V)$, where $g(\cdot)$ and $h(\cdot)$ are chosen to obtain the maximal linear relationship in terms of the Pearson correlation coefficient. Performing the correlation test over the newly generated random variables $g(U)$ and $h(V)$, we come to a conclusion whether $g(U)$ and $h(V)$ are correlated, as depicted in Figure 2.4. The dependence relationship between $g(U)$ and $h(V)$, rather than that between X and Y, is tested because the linear dependence relationship is increased and therefore more detectable. In this way, we can find the small dependence relationship between random variables instead of incorrectly considering then as being independent.

How do the test results of $g(U)$ and $h(V)$ shown in Figure 2.4 relate to the correlation relationship between the original random variables X and Y? Figure 2.5 reveals the dependence relationship between variables with and without the Box-Cox transformation. If $g(U)$ and $h(V)$ are correlated, it is certain that $g(U)$ and $h(V)$ are dependent. Dependent $g(U)$ and $h(V)$ indicate that X and Y are dependent, because both the Box-Cox transformation and the cumulative distribution transformation are one-to-one functions. If $g(U)$ and $h(V)$ are uncorrelated, then U and V are uncorrelated, since $|\text{corr}(g(U),h(V))| \geq |\text{corr}(U,V)|$, where $g(\cdot)$ and $h(\cdot)$ are the Box-Cox transformations. Uncorrelated U and V implies independent X and Y when U and V are jointly

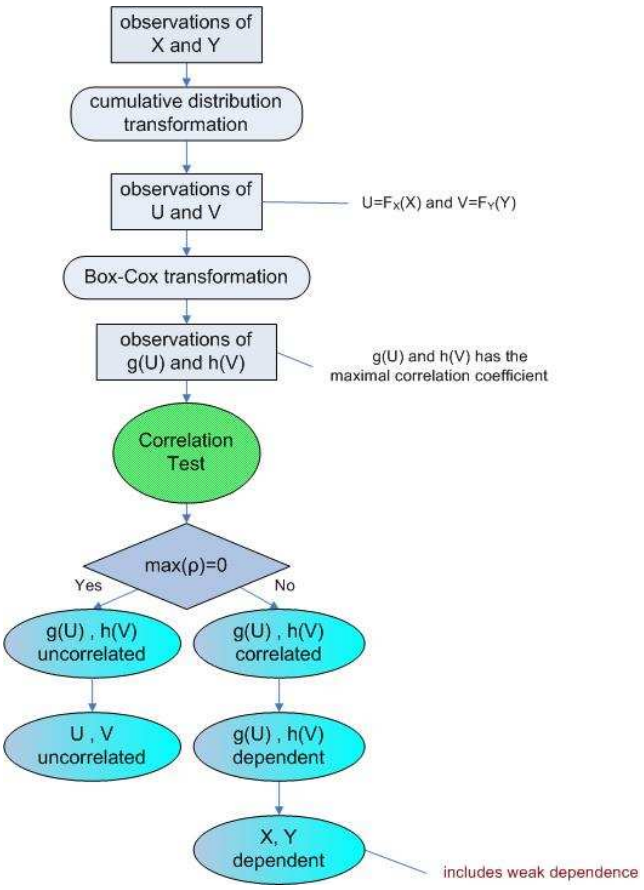

Figure 2.4: Correlation test after the Box-Cox transformation

normal. Although, in general, no direct conclusion whether X and Y are uncorrelated is made, it does not really affect the validity of our test, since our focus is to determine whether X and Y are dependent, especially weakly dependent.

In the following chapters, we will mainly study ways to detect weak dependence by implementing the simulations based on the working principle shown in Figure 2.4. In addition, from Figure 2.4 and 2.5, the test results of $g(U)$ and $h(V)$ are correlated or uncorrelated can only suggest that X and Y are dependent. There is no definite answer to the question whether X and Y are independent. Therefore, as a supplement and comparison, we will also introduce the independence tests performed on the Box-Cox transformed random variables in Chapter 6.

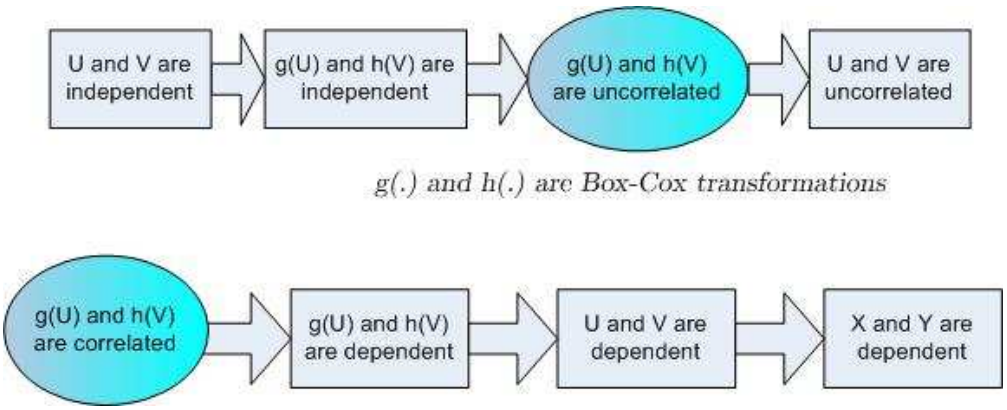

Figure 2.5: Correlation equivalence with/without Box-Cox transformation

CHAPTER 3. DISTRIBUTION ESTIMATION

The first step in building a probabilistic model and making a statistical inference is to obtain sufficient knowledge of the distributions of the component random variables. Generally, such information becomes visible through analysis and summarization of data samples collected from real scenarios, i.e., *distribution estimation*. Estimation of a distribution can be univariate, multivariate, or both. For example, estimation of a conditional distribution might include specifying both multivariate joint distributions and univariate marginal distributions.

Distribution estimation can be divided into two types: parametric estimation and nonparametric estimation. Conventional parametric distribution estimation methods, from polynomial regression, AR model in engineering to ARCH, GARCH models [26] in economics, are established on a common fundamental assumption that the underlying distribution structure is known or given. However, such a rigid restriction is usually inappropriate in practical applications. For instance, a transmitter sends a message through a wireless channel which is noisy because of electrical noise, atmospheric disturbances and other distortions. In many cases, the white noise assumption is possible and reasonable, while the message itself usually could not be expressed by a known distribution such as the normal distribution, since ignorance of the skewness, kurtosis or other features between the hypothetical distribution and the true distribution may bias the results.

Without a preliminary distribution hypothesis, nonparametric distribution estimation methods, such as the nearest neighbor method, the orthogonal series estimation, and the kernel density estimation, evaluate the distribution entirely from the data. To ensure the validity and accuracy of the inference made on the basis of observations, a large amount of data is required. In general, the nonparametric distribution estimation method is superior for the low dimension cases where the number of random variables is $n = 1$ or 2. For higher dimension ($n > 2$), the nonparametric

distribution estimation method is not a good choice because both its computation complexity and performance reduction grow rapidly [27, 28].

We adopt the commonly used nonparametric Nadaraya-Watson (NW) kernel density estimator to obtain univariate marginal distributions. An overview of the NW kernel density estimation is provided in this chapter. We start with an illustration of the basic idea and then follow with a discussion of two important issues: the bandwidth selection and the boundary effect.

With regard to the multivariate distributions, we introduce two kinds of estimation methods: the double kernel local linear method and the copula method. The double kernel local linear method is a nonparametric way to estimate the conditional density function, while the copula is a parametric statistical tool used for estimating both joint and conditional distributions. The former requires no extra information beyond the data, while the latter picks the functional form as a specified parameter. In our study, the statistics discussed in Chapter 6, such as the Cramer-Von Mises distance or Kolmogorov-Smirnov distance, require the calculation of multivariate distributions. The essential ideas of two methods are described in Section 3.2 and 3.3.

Finally, the asymptotic behavior of different distribution estimation methods are briefly discussed in Section 3.4.

3.1 Nadaraya-Watson Kernel Density Estimation

When constructing the estimator of an unknown distribution function F from the observations, nonparametric estimation is motivated to satisfy the need to allow the data to influence the estimate more than would be the case if F were constrained to a given parametric family. The estimation could be used as a stand-alone system or an elemental component of a statistical inference procedure.

To reduce the computation burden and the loss of accuracy due to the dimensional increase, a nonparametric estimation method is always used in low dimensions. We apply the well-studied nonparametric technique called the Nadaraya-Watson kernel density estimator, to determine the univariate distributions of X and Y.

3.1.1 Overview

The univariate Nadaraya-Watson (NW) kernel density estimator is defined as [27–29]:

$$\hat{f}(x;h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - x_{s_i}}{h}\right), \tag{3.1}$$

where the *kernel* K is a function satisfying $\int_{-\infty}^{\infty} K(x)\, dx = 1$, the *bandwidth* h is a positive number, and $x_{s_1}, \cdots, x_{s_n}$ are observed samples taken from a continuous univariate density f.

K is usually chosen to be a unimodal probability density function that is symmetric about zero. This ensures that $\hat{f}(x;h)$ itself is also a density. Some common kernels are: (1) the normal kernel $K(x) = 1/\sqrt{2\pi}e^{-x^2/2}$; (2) the triangular kernel $K(x) = 1 - |x|,\ |x| < 1$; and (3) the beta kernel $K(x) = [2^{2p+1}B(p+1, p+1)]^{-1}(1-x^2)^p,\ |x| < 1$, where B is the beta function.

Figure 3.1 shows the kernel density estimation $\hat{f}(x;h)$ constructed by nine observations with the normal kernel. The working principle is explicit: $\hat{f}(x;h)$ at a given point x is the average of weighted kernels centering at n observation points $x_{s_1}, \cdots, x_{s_n}$, where the kernel at x_{s_i} takes a relatively large value if x_{s_i} is near x and it takes a relatively small value if x_{s_i} is far from x. In other words, a larger estimate $\hat{f}(x;h)$ is achieved if more observation points in the neighbor of x are counted and thus contribute to the average, and vice versa. For example, the highest peak value of f at $x \approx 1.5$ mainly comes from the five observation points between $[1, 2]$, while the small value of $f \approx 0.05$ at $x = -2$ results from the fact that there is only one observation point $x \approx -1$ in the neighbor of $x = -2$.

3.1.2 Important Issues of Nadaraya-Watson Estimation

Two issues can not be ignored in estimating a density function with the Nadaraya-Watson estimator. One is to obtain an appropriate bandwidth parameter h, and the other is to remove the boundary effect. These two factors determine the performance of NW estimation.

Bandwidth Selection

For the Nadaraya-Watson kernel density estimator, the choice of the shape of the kernel function K is not particularly important, but the choice of value for the bandwidth h is crucial [27]. That is, the normal kernel and the beta kernel with same h may not make notable differences in the

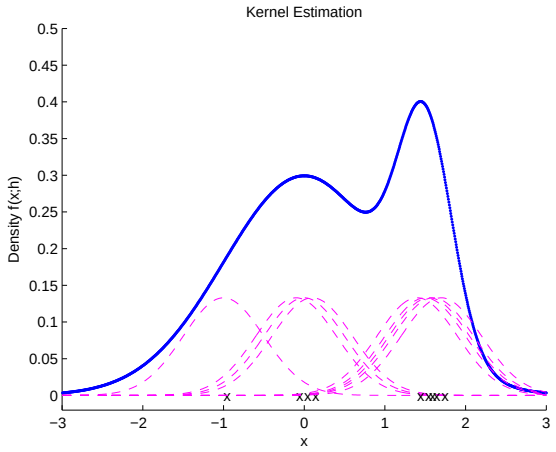

Figure 3.1: Nadaraya-Watson kernel density estimate based on nine observations

estimation of an unknown density f, while a kernel function with distinct h values will definitely generate different estimates of f.

The selection of the bandwidth h is based on an appropriate error criterion which represents the performance of the NW kernel density estimator. A common error criterion is the measurement of the closeness between the estimated density $\hat{f}(x;h)$ and the target density $f(x)$ denoted as the asymptotic mean integrated squared error(AMISE). According to [27], $E\hat{f}(X;h) = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{X-x_{s_i}}{h})\right] = \int K(z)f(x-hz)\,dz$. Then, expand $f(x-hz)$ in a Taylor series about x as

$$f(x-hz) = f(x) - hzf'(x) + \frac{1}{2}h^2z^2f''(x) + o(h^2).$$

Since $\int K(z)\,dz = 1$ and $\int zK(z)\,dz = 0$, we obtain:

$$E\hat{f}(X;h) = f(x) + \frac{1}{2}h^2f''(x)\int z^2K(z)\,dz + o(h^2).$$

Denote $\int z^2K(z)\,dz = \mu_2(K)$ and $\int K(z)^2\,dz = R(K)$. The resulting bias and variance are

$$\text{bias } = E\hat{f}(X;h) - f(x) = \frac{1}{2}h^2\mu_2(K)f''(x) + o(h^2),$$

$$\text{variance } = \text{Var}\{\hat{f}(X;h)\} = (nh)^{-1}R(K)f(x) + o\{(nh)^{-1}\}.$$

and hence

$$\begin{aligned}\mathrm{MISE}\{\hat{f}(x;h)\} &= \int \mathrm{MSE}\{\hat{f}(x;h)\}\,dx = \int E\,[\hat{f}(X;h)-f(x)]^2\,dx \\ &= \int [\text{variance} + (\text{bias})^2]\,dx = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') + o\{(nh)^{-1}+h^4\},\end{aligned}$$

or

$$\mathrm{AMISE}\{\hat{f}(\cdot;h)\} = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') \ \text{ as } n\to\infty. \tag{3.2}$$

The AMISE is the large sample approximation of MISE. When h is small, the first term (variance term) of (3.2) increases and the second term (bias term) decreases. When h is large, the variance term decreases and the bias term increases. Considering the inconsistency between variance and bias, a trade-off is needed. The optimal bandwidth h is the compromise value that minimizes the AMISE. There are various bandwidth selection rules derived from the AMISE criterion, such as normal scale (NS), direct plug-in (DPI) or solve-the-equation (STE) introduced in Appendix A.

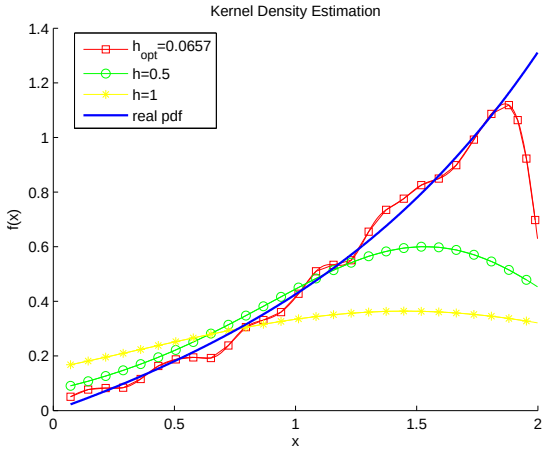

Figure 3.2: Nadaraya-Watson kernel density estimate with different bandwidth

We provide an intuitive interpretation of the influence of distinct h values to the NW estimate $\hat{f}(x;h)$ with the following example. Suppose the target density function is

$$f(x) = \frac{5}{16}x + \frac{3}{32}x^2 + \frac{5}{256}x^4,\ x\in[0,2]. \tag{3.3}$$

We first generate data samples $x_{s_1}, \cdots, x_{s_n}$ from the density function $f(x)$ in (3.3). We then estimate $\hat{f}(x;h)$ entirely from $x_{s_1}, \cdots, x_{s_n}$ as if the real density $f(x)$ is unknown. Figure 3.2 clearly shows the difference of three normal-kernel estimation curves which represent three cases that h takes the value of the calculated optimal DPI bandwidth $h_{opt} = 0.0657$ and two other arbitrary numbers 0.5 and 1. The solid line represents the real density $f(x)$.

Boundary Correction

The boundary effect, or boundary bias, which represents the severe performance reduction in density tails, results from the inherent structure limitation of the kernel density estimator, and therefore is inevitable.

One cause of the boundary effect is that the kernel density estimator lacks the knowledge of the boundary region and consequently is incapable of assigning the probability to the end points. Suppose f has a finite support $x \in [a,b]$. Then the knowledge of the boundary region near a or b is restricted. Actually, one side is known, since only samples $x_s \geq a$ or $x_s \leq b$ are observable. As a result, the estimation is more biased near the endpoints than in the interior.

Another cause of the boundary effect is that the kernel K itself has a finite support. Even if some kernels, such as the normal kernel, are defined on the entire real line, they are always truncated onto a finite interval for the convenience of implementation. Without loss of generality, assume that the support of $f(x)$ is [0,1] and the support of K is [-1,1]. From (3.1), the support of $\hat{f}(x;h)$ is $[\min(x_{si}) - h, \max(x_{si}) + h]$, since $-1 \leq \frac{x - x_{s_i}}{h} \leq 1$. According to Marron [30], this support of $\hat{f}(x;h)$ is typically much larger than [0,1] and the expected value of $\hat{f}(0)$ is approximately $\frac{1}{2} f(0)$. A similar bias exists at the right-hand boundary $x = 1$ as well.

Furthermore, the boundary effect is a serious problem, since the boundary region usually occupies $20\% - 50\%$ of the entire support of unknown density function in real scenarios. Therefore, an efficient boundary correction approach is imperative. A variety of boundary correction approaches exist, from simple 'reflection' at the boundary or the generation of pseudodata beyond the extremes of the density support, to more adaptive and complicated methods of boundary (edge) kernel or transformation kernel [31–33]. Some representative methods are introduced in Appendix A. In this study, the reflection method and the generalized jack-knifing method are applied.

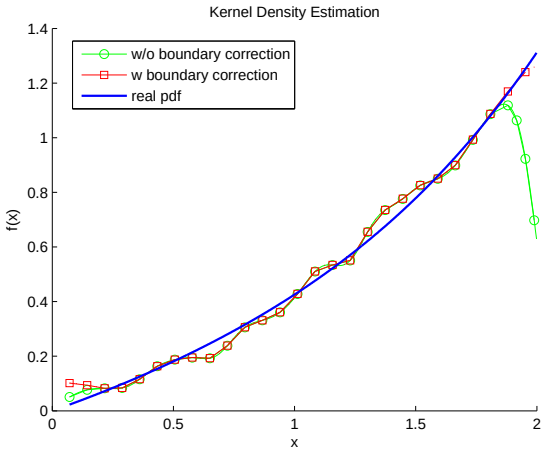

Figure 3.3: Nadaraya-Watson kernel density estimate with boundary correction

We also use the same example (3.3) to show the boundary effect visually. Figure 3.3 compares the estimation with/without boundary correction to the real density curve. Obviously, the boundary correction greatly improves the tail behaviors.

3.2 Double Kernel Local Linear Method

A conditional density provides the most direct or intuitive dependence relationship between random variables. It characterizes the probabilistic aspect of a time series and thus can determine the likelihood function of statistical prediction in some financial models [34], or, it is simply explored to construct specific statistics such that the dependent or independent random variables can be distinguished.

One nonparametric way to estimate the conditional density function of Y given X is the double-kernel local linear method (LLM), in which the estimation of the conditional density is considered as a nonparametric regression problem [35, 36].

Let $f_{Y|X}(y|x)$ denote the conditional density of Y given $X = x$, evaluated at $Y = y$. Fan makes the following observation of $f_{Y|X}(y|x)$ [35],

$$E\{K_{h_2}(Y-y)|X=x\} \to f_{Y|X}(y|x) \text{ as } h_2 \to 0, \tag{3.4}$$

where the kernel K is a nonnegative density function and $K_h(y) = \frac{1}{h}K(\frac{y}{h})$, h is the bandwidth. Take y an arbitrary value y_0 and let $y' = y - y_0$, then

$$E\{K_{h_2}(Y-y_0)|X=x\} = \int K_{h_2}(y-y_0)f_{Y|X}(y|x)dy = \int K_{h_2}(y')f_{Y|X}(y'+y_0|x)dy'$$
$$= \int \delta(y')f_{Y|X}(y'+y_0|x)dy = f_{Y|X}(y_0|x) \text{ as } h_2 \to 0,$$

where $\delta(\cdot)$ denotes the Dirac delta function and $K_h(y) \to \delta(y)$ as $h \to 0$. Using the normal kernel K as an example, $K_h(y) = e^{\frac{-y^2}{2h^2}}/(\sqrt{2\pi}h)$. When $h \to 0$ and $y \neq 0$, according to the ĹHôpital's rule, $K_y(y) = 0$; When $h \to 0$ and $y = 0$, $K_y(y) = 1/(\sqrt{2\pi}h) \to \infty$. In other words, the Dirac delta function δ is the limit of the normal sequence K_h when the bandwidth h goes to zero. Therefore, $E\{K_{h_2}(Y-y)|X=x\}$ legitimately represents the conditional density $f_{Y|X}(y|x)$ as $h_2 \to 0$.

Regression analysis addresses the problem of predicting a response variable Y given the value x of an explanatory variable X. As is well known, the best prediction one could possibly hope for, in the sense of minimizing the mean square error (MMSE), would be the function : $m(x) = E[Y|X=x]$. We say, this expression $m(x)$ defines the regression function of Y on X. Therefore, the left side of (3.4) is the regression function of the random variable $K_{h_2}(Y-y)$ given $X=x$. The estimation of the conditional density $f_{Y|X}(y|x)$ becomes the problem to find the regression function $m(x)$.

Sometimes the form of the MMSE estimator or the regression function $m(x)$ is based on knowledge about the relationship between Y and X that does not rely on the data. However, in many real cases where no such knowledge is available, it is not possible to determine a closed form of $m(x)$. As a result, a flexible or convenient form is usually chosen to compute $m(x)$. A few various methods, such as kernel, spline, and orthogonal series, have been proposed for estimating $m(x)$ [37, 38]. We discuss a design-adaptive regression estimator suggested by Fan [36], which derives $m(x)$ with a principle that overcomes the disadvantage of kernel methods and adapts to more widely extended density cases .

Suppose that the second derivative of $m(x) = E[Y|X=x]$ exists. By Tayler expansion, $m(y) \approx m(x) + m'(x)(y-x) \equiv \alpha + \beta(y-x)$ if y is in a small neighborhood of the point x. When $y = x$, $m(y) = m(x) = \alpha$. Hence, estimating the regression function $m(x)$ is equivalent to estimating

the intercept α. In this way, the problem of estimating $m(x)$ is turned into a local linear regression problem to find the regression coefficient α, i.e.,

$$\text{Find } \min_{\alpha,\beta} e(\alpha,\beta), \text{ with } e(\alpha,\beta) = \sum_{t=1}^{T} \hat{\varepsilon}_t{}^2 = \sum_{t=1}^{T} (y_t - \alpha - \beta(x_t - x))^2 W_{h_1}(x_t - x),$$

where the weight kernel W is a nonnegative density function, $\{x_t, y_t\}_{t=1}^{T}$ are observations, and $\hat{m}(x) = \hat{\alpha}$.

As to the conditional density function $E\{K_{h_2}(Y-y)|X=x\}$ when $h_2 \to 0$, the corresponding $\hat{\alpha}$ is estimated by minimizing

$$\sum_{t=1}^{T} \hat{\varepsilon}_t{}^2 = \sum_{t=1}^{T} [K_{h_2}(y_t - y) - \alpha - \beta(x_t - x)]^2 W_{h_1}(x_t - x). \tag{3.5}$$

To represent the minimization problem of (3.5) in a more convenient matrix form, let

$$\mathbf{y} = \begin{bmatrix} K_{h_2}(y_1 - y)|X = x \\ K_{h_2}(y_2 - y)|X = x \\ \vdots \\ K_{h_2}(y_T - y)|X = x \end{bmatrix}, \ \mathbf{A} = [\mathbf{p_1}\ \mathbf{p_2}] = \begin{bmatrix} 1 & x_1 - x \\ 1 & x_2 - x \\ \vdots & \vdots \\ 1 & x_T - x \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} W_{h_1}(x_1 - x) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & W_{h_1}(x_T - x) \end{bmatrix},$$

and $\mathbf{c} = [\alpha\ \beta]^H$, where $(.^H)$ denotes the transpose. We wish to determine $\mathbf{c}$ which minimizes the L^2 norm $||\varepsilon||_2^2 = ||\mathbf{y} - \mathbf{Ac}||_W^2$, where the weighted inner product is defined as $\langle \mathbf{x}, \mathbf{y} \rangle_W = \mathbf{y^H W x}$. According to [39], the minimum $||\varepsilon||_2^2$ occurs when ε is orthogonal to each vector, i.e.,

$$\langle \mathbf{y} - \mathbf{Ac}, \mathbf{p_i} \rangle_W, \ i = 1, 2,$$

or

$$\begin{bmatrix} \mathbf{p}_1^H \\ \mathbf{p}_2^H \end{bmatrix} \mathbf{W}(\mathbf{Ac}-\mathbf{y}) = 0 \Rightarrow \mathbf{A}^H\mathbf{W}(\mathbf{Ac}-\mathbf{y}) = 0.$$

Therefore,

$$\mathbf{c} = (\mathbf{A}^H\mathbf{WA})^{-1}\mathbf{A}^H\mathbf{Wy},$$

or

$$\hat{f}_{Y|X}(y|x) = \hat{\alpha} = \mathbf{e}(\mathbf{A}^H\mathbf{WA})^{-1}\mathbf{A}^H\mathbf{Wy}, \tag{3.6}$$

where $\mathbf{e} = [1\ 0]^H$.

Fan and Yao [34] also give the equivalent expression in the form of the kernel,

$$\hat{f}_{Y|X}(y|x) = \frac{1}{nh_1h_2}\sum_{t=1}^{T} W_T\left(\frac{x_t - x}{h_1}; x\right) K\left(\frac{y_t - y}{h_2}\right), \tag{3.7}$$

where

$$W_T(z;x) = W(z)\frac{s_{T,2}(x) - zh_1 s_{T,1}(x)}{s_{T,0}(x)s_{T,2}(x) - s_{T,1}(x)^2} \text{ and } s_{T,j}(x) = \frac{1}{T}\sum_{t=1}^{T}(x_t - x)^j W_{h_1}(x_t - x), \text{ for } j = 0,1,2.$$

Equations (3.6) and (3.7) of LLM provide us the way to compute the conditional density function given a pair of observations $\{x_t, y_t\}_{t=1}^{T}$. However, similarly to the discussion of the NW kernel density estimator, an important problem has not yet been solved: the selection of the bandwidth $h = (h_1, h_2)$. The bandwidth h controls the degree of smoothing applied to the density estimate and it contains the smooth parameters in two directions, both x and y direction. There are several ways to obtain the approximate bandwidth h [40–43]. Two representative data-driven bandwidth selection methods which are used in this study, the crossvalidation rule and the penalized average rule, are listed in Appendix B.

3.3 Copula Estimation

3.3.1 Copula Basics

A copula is a function that couples univariate marginal distributions with dependence relationships among variables to construct a joint distribution. It was first introduced by Sklar in 1959

and became well-known in late 1980s. It has been widely used in finance and econometrics. The related research include portfolio value-at risk studies [44], aggregate loss models [45], insurance loss and expense applications [46], etc. When the univariate marginal distributions are computed through the Nadaraya-Watson kernel density estimation, we can use the copula to assess the joint distribution.

Definition 3.1 A *copula C* is a joint cumulative distribution function of standard uniform random variables. Let $U_i, i = 1, \cdots, n$ be a set of uniform random variables, i.e., $U_i \sim \mathscr{U}(0,1)$. The copula is defined as

$$C(u_1, u_2, \cdots, u_n) = Pr\{U_1 \leq u_1, U_2 \leq u_2, \cdots, U_n \leq u_n\}.$$

□

By this definition, C links univariate marginal distribution $F_i(x_i)$ to their full multivariate distribution. Suppose X_i are random variables with arbitrary marginal distributions $F_i(x_i), i = 1, \cdots, n$ and define $U_i = F_i(X_i)$. Define a generalized inverse $F_i^{-1}(u) = \min\{x : F_i(x) \geq u,\ 0 < u < 1\}$[1]. Since $U_i = F_i(X_i) \sim \mathscr{U}(0,1)$, then

$$C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n)) = F(x_1, x_2, \cdots, x_n).$$

To see, we observe that

$$\begin{aligned} C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n)) &= Pr(U_1 \leq F_1(x_1), U_2 \leq F_2(x_2), \cdots, U_n \leq F_n(x_n)) \\ &= Pr(F_1^{-1}(U_1) \leq x_1, F_2^{-1}(U_2) \leq x_2, \cdots, F_n^{-1}(U_n) \leq x_n) \\ &= Pr(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n) = F(x_1, x_2, \cdots, x_n). \end{aligned}$$

Theorem 3.3.1. *For random variables $X_1, \cdots, X_n$, let the joint distribution function be $F(x_1, x_2, \cdots, x_n)$ and the corresponding univariate marginal distribution function be $F_i(x_i)$, there exists a copula*

[1] If F_i is continuous, $F_i^{-1}(u)$ reduces to the usual inverse function. Suppose F_i is discontinuous at $x = x_0$. For any $\varepsilon > 0$, let $u_1 = \lim_{\varepsilon \to 0} F_i(x_0 - \varepsilon)$ and $u_2 = \lim_{\varepsilon \to 0} F_i(x_0 + \varepsilon) = u_1 + P(x = x_0)$. Then the inverse $F_i^{-1}(u)$ with $u \in [u_1, u_2]$ maps to a single point $x = x_0$.

function C such that

$$F(x_1, x_2, \cdots, x_n) = C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n)).$$

Theorem 3.3.1 is known as the *Sklar theorem* [47]. It establishes that any joint distribution F can be written in a copula form. The copula provides a general way to formulate arbitrary probability distributions characterizing multivariate dependent data. A trivial example is that the copula of independent random variables takes the form of product copula

$$C(F_1(x_1), \cdots, F_n(x_n)) = F_1(x_1) \cdots F_n(x_n).$$

Since C is unique only if each $F_i(x_i)$ is continuous [48, 49], we assume each $F_i(x_i)$ is continuous and differentiable in this study. Another important concept is the *copula density*.

Definition 3.2 Given that $F_i(x_i)$ and C are differentiable, its *copula density* c is

$$c(u_1, u_2, \cdots, u_n) = \frac{\partial C^n(u_1, u_2, \cdots, u_n)}{\partial u_1 \cdots \partial u_n}.$$

The joint density function $f(x_1, x_2, \cdots, x_n)$ can be written as

$$f(x_1, x_2, \cdots, x_n) = c[F_1(x_1), F_2(x_2), \cdots, F_n(x_n)] \prod_{i=1}^{n} f_i(x_i), \tag{3.8}$$

where $f_i(x_i)$ is the density corresponding to $F_i(x_i)$. □

It is clear that only c accounts for the dependence relationship among X_i. Therefore, according to the Sklar theorem, a joint distribution may be constructed by specifying the univariate marginals and copula separately.

The specification of a copula involves two elements: the functional forms and the parameters. On the one hand, the functional forms are some known and fixed distributions such as the normal, t-student or Archimedean distributions, which construct the copula families complying with the essential properties of copula concept [48, 49]. The copula is said to be parametric in this sense. On the other hand, the parameters, such as the Pearson correlation, Spearman's correlation or Kendall's correlation, contain dependence information among random variables. In next sec-

tion, we will introduce two kinds of well-known copulas: the normal copula and the Archimedean copula.

3.3.2 Copula Functional Forms

The normal copula is the most widely used and convenient copula in many studies, including our study. Let $X_i, i = 1, \cdots, n$ be continuous random variables with marginal distributions $F_i(x_i)$. Denote $u_i = F_i(x_i)$ and $\mathbf{u} = (u_1, \cdots, u_n)^H$, the symbol $(.^H)$ represents the transpose of a vector or a matrix. Let $\Phi_{\mathbf{R}}$ and $\phi_{\mathbf{R}}$ denote the standard multivariate normal distribution function and density function with correlation matrix $\mathbf{R}$, and let Φ and ϕ denote the standard univariate normal distribution function and density function. The normal copula is defined as:

$$C^N(\mathbf{u}) = \Phi_{\mathbf{R}}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_n)), \tag{3.9}$$

and thus the normal copula density is

$$\begin{aligned} c^N(\mathbf{u}) &= \frac{\phi_{\mathbf{R}}^{(n)}\{\Phi^{-1}[F_1(x_1)], \Phi^{-1}[F_2(x_2)], \cdots, \Phi^{-1}[F_n(x_n)]\}}{\phi\{\Phi^{-1}[F_1(x_1)]\} \times \phi\{\Phi^{-1}[F_2(x_2)]\} \cdots \times \phi\{\Phi^{-1}[F_n(x_n)]\}} \\ &= \exp\{-\xi^H(\mathbf{R}^{-1} - \mathbf{I})\xi/2\}/|\mathbf{R}|^{1/2}, \end{aligned} \tag{3.10}$$

where $\xi = (\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_n))^H$, $\mathbf{I}$ is the $n \times n$ identity matrix and $\phi_{\mathbf{R}}^{(n)}$ is n-th order derivative of $\phi_{\mathbf{R}}$. The joint density function of $\mathbf{X} = \{X_1, \cdots, X_n\}$ is

$$\begin{aligned} f(x_1, \cdots, x_n) &= f_1(x_1) \cdots f_n(x_n) \exp\{-\xi^H(\mathbf{R}^{-1} - \mathbf{I})\xi/2\}/|\mathbf{R}|^{1/2} \\ &= f_1(x_1) \cdots f_n(x_n) \exp\{-(\Phi^{-1}[F_1(x_1)], \cdots, \Phi^{-1}[F_n(x_n)]) \\ &\quad (\mathbf{R}^{-1} - \mathbf{I})(\Phi^{-1}[F_1(x_1)], \cdots, \Phi^{-1}[F_n(x_n)])^H/2\}/|\mathbf{R}|^{1/2}. \end{aligned} \tag{3.11}$$

Let $\mathbf{R}$ and $\mathbf{u}$ be partitioned as follows:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{n-1} & \mathbf{r} \\ \mathbf{r}^H & 1 \end{pmatrix} \text{ and } \mathbf{u} = (\mathbf{u}_{n-1}, u_n),$$

where $\mathbf{u}_{n-1} = (u_1, \cdots, u_{n-1})^H$, $\mathbf{R}_{n-1}$ is the $(n-1)\times(n-1)$ correlation matrix and $\mathbf{r} = (r_{1n}, \cdots, r_{(n-1)n})^H$. The conditional density is:

$$
\begin{aligned}
& f(x_n|x_1, \cdots, x_{n-1}) \\
&= \frac{\phi^{(n)}(\Phi^{-1}[F_1(x_1)], \Phi^{-1}[F_2(x_2)], \cdots, \Phi^{-1}[F_n(x_n)]|\mathbf{R})}{\phi^{(n-1)}(\Phi^{-1}[F_1(x_1)], \cdots, \Phi^{-1}[F_{n-1}(x_{n-1})]|\mathbf{R}_{n-1})} \cdot \frac{f_n(x_n)}{\phi(\Phi^{-1}([F_n(x_n)])} \\
&= f_n(x_n)(1-\mathbf{r}^H\mathbf{R}_{n-1}^{-1}\mathbf{r})^{-\frac{1}{2}} \exp\left\{-0.5\left[\frac{(\Phi^{-1}[F_n(x_n)] - \mathbf{r}^H\mathbf{R}_{n-1}^{-1}\mathbf{u}_{n-1})^2}{(1-\mathbf{r}^H\mathbf{R}_{n-1}^{-1}\mathbf{r})} - (\Phi^{-1}[F_n(x_n)])^2\right]\right\}.
\end{aligned}
\tag{3.12}
$$

In general, the correlation matrix $\mathbf{R}$ could be evaluated by the maximum likelihood (ML) method. Regarding the normal distribution, the ML correlation matrix is equal to the sample matrix[2]. That is, given T observations,

$$\hat{\mathbf{R}}_{ML} = \frac{1}{T}\sum_{t=1}^{T} \xi_t \xi_t^H,$$

where $\xi_t = (\Phi^{-1}(u_{1t}), \cdots, \Phi^{-1}(u_{nt}))^H$ of the observed data samples $\{x_{1t}, \cdots, x_{nt}\}_{t=1}^T$.

Another copula applied in this study is the Archimedean copula. Generally, the Archimedean copula is viewed as a bivariate copula and can be constructed as follows [50]:

$$C_\varphi(u,v) = \varphi^{-1}[\varphi(u) + \varphi(v)],$$

for all marginal distributions $u = F_{X_1}(x_1)$ and $v = F_{X_2}(x_2)$. φ is a *generator* which satisfies the following three conditions:

- $\varphi(1) = 0$;
- for all $t \in (0,1)$, $\varphi'(t) < 0$. i.e., φ is decreasing;
- for all $t \in (0,1)$, $\varphi''(t) \geq 0$. i.e., φ is convex;

where $\varphi'(\cdot)$ and $\varphi''(\cdot)$ represent the first and the second order derivative of $\varphi(\cdot)$.

[2] A brief proof is presented in Appendix B.

The conditional density is computed as [46]:

$$f_{X_2|X_1}(x_2|x_1) = \varphi'(v)v' \frac{\varphi^{-1}[\varphi(u)+\varphi(v)]}{\varphi^{-1}[\varphi(u)]}.$$

The choice of the generator determines the type of Archimedean copulas. Several important families of Archimedean copulas which are yielded by distinct generators are listed in Table 3.1 and 3.2[3] .

Table 3.1: Archimedean copulas and generators

Family	Generator $\varphi(t)$	Copula parameter α	Bivariate copula $C_\varphi(u,v)$
Independence	$-\ln t$	Not applicable	uv
Clayton,Cook-Johnson	$t^{-\alpha}-1$	$\alpha > 1$	$(u^{-\alpha}+v^{-\alpha}-1)^{-1/\alpha}$
Gumbel,Hougaard	$(-\ln t)^{\alpha}$	$\alpha \geq 1$	$\exp\{-[(-\ln u)^{\alpha}+(-\ln v)^{\alpha}]^{1/\alpha}\}$
Frank	$\ln \frac{e^{\alpha t}-1}{e^{\alpha}-1}$	$-\infty < \alpha < \infty$	$\frac{1}{\alpha}\ln(1+\frac{(e^{\alpha u}-1)(e^{\alpha v}-1)}{e^{\alpha}-1})$

Table 3.2: Archimedean copulas and dependence measures

Family	Bivariate Copula $C_\varphi(u,v)$	Kendall's τ	Spearman's ρ_S
Independence	uv	0	0
Clayton,Cook-Johnson	$(u^{-\alpha}+v^{-\alpha}-1)^{-1/\alpha}$	$\frac{\alpha}{\alpha+2}$	Complicated form
Gumbel, Hougaard	$e^{\{-[(-\ln u)^{\alpha}+(-\ln v)^{\alpha}]^{1/\alpha}\}}$	$1-\alpha^{-1}$	No close form
Frank	$\frac{1}{\alpha}\ln(1+\frac{(e^{\alpha u}-1)(e^{\alpha v}-1)}{e^{\alpha}-1})$	$1-\frac{4}{\alpha}(D_1(-\alpha)-1)$	$1-\frac{12}{\alpha}(D_2(-\alpha)-D_1(-\alpha))$

[3] In the Frank copula,"Debye" function is:

$$D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t-1}dt$$

for $k = 1,2$. To evaluate negative arguments of the Debye function D_k, use $D_k(-x) = D_k(x) + \frac{kx}{k+1}$.

One question arises naturally: which type of copula better fits the empirical data? There exist criteria for selecting the best copula [46, 51, 52]. However, we choose the Frank copula without exploring such criteria because only the Frank copula, as shown in Table 3.1, has no limitation to the range of the copula parameter α. According to Table 3.2, arbitrary α is equivalent to saying that Kendall's τ can take any values in [-1,1]. Therefore, the Frank copula is feasible in all situations, while the Gumbel and the Clayton copulas are only available in some specific cases among all possible situations.

3.4 Asymptotic Property

In this section, we discuss the asymptotic behavior of the mentioned distribution estimation methods. As is well known, the estimation performance is always restricted by the finite sample size. For example, if the information of whole population were known, it is possible that the weak dependence between variables can be observed using traditional statistical approaches and thus no advanced study is needed. Since there is no way to have the knowledge of population, the asymptotic analysis helps us to obtain a reliable understanding of estimation performance in large samples.

NW Kernel Density Estimation

Assume $\int z^2 K(z)\, dz = \mu_2(K)$ and $\int K(z)^2\, dz = R(K)$, where K is the kernel. According to [27], the smallest possible asymptotic mean integrated squared error (AMISE) for the estimation of f using the NW estimator is

$$\inf_{h>0} \text{AMISE}\{\hat{f}(\cdot;h)\} = \frac{5}{4}\{\mu_2(K)^2 R(K)^4 R(f'')\}^{1/5} n^{-4/5} \quad \text{as } n \to \infty.$$

Suppose b is the binwidth of the histogram $\hat{f}_H(\cdot : b)$, the smallest possible AMISE for estimation of f using the histogram (empirical method) is

$$\inf_{b>0} \text{AMISE}\{\hat{f}_H(\cdot;b)\} = \frac{1}{4}\{36R(f')\}^{1/3} n^{-2/3} \quad \text{as } n \to \infty.$$

Both $\hat{f}(\cdot;h)$ and $\hat{f}_H(\cdot;b)$ converge to the real pdf f as the sample size n goes to infinity. However, the MISE of the histogram is asymptotically inferior to the NW kernel density estimator since its convergence rate is $O(n^{-2/3})$ compared to the kernel estimator's $O(n^{-4/5})$ rate, i.e., the histogram is mathematically inefficient.

Double Kernel Local Linear Method (LLM)

LLM is to estimate the conditional density $f_{Y|X}(y|x)$ by finding the regression function $m(x) = E\{K_{h_2}(Y-y)|X=x\}$. Assume $\int z^2K(z)\,dz = \mu_2(K)$ and $\int K(z)^2\,dz = R(K)$, where K is the kernel. Also assume the following four conditions,

1. The regression function $m(x)$ has a bounded and continuous second derivative.
2. The conditional variance $\sigma^2(x) = \text{var}(Y|X=x)$ is bounded and continuous.
3. The marginal density f_X is bounded and continuous away from zero in an interval (a_0, b_0).
4. The kernel K is a bounded density function with $\int_{-\infty}^{\infty} xK(x)\,dx = 0$ and $\int_{-\infty}^{\infty} x^4K(x)\,dx < \infty$.

The bandwidth h_1 is a non-random sequence of positive numbers with respect to the sample size n. If $h_1 \to 0$ and $nh_1 \to \infty$, that is, h_1 approaches zero, but at a rate slower than n^{-1}. We have AMISE of LLM as follows [36]:

$$\inf_{b>0} \text{AMISE}\{\hat{m}(x)\} = \frac{\mu_2(K)^2}{4}\int_{-\infty}^{\infty}[m''(x)]^2w(x)\,dxh_1^4 + \frac{R(K)}{nh_1}\int_{-\infty}^{\infty}\frac{\sigma^2(x)}{f_X(x)}w(x)\,dx + O(h_1^4 + (nh_1)^{-1}) \quad \text{as } n \to \infty,$$

where $w(x)$ is a weight function which has a support containing in (a_0, b_0). The AMISE shows that the regression function $\hat{m}(x)$ converges to the real conditional density $f_{Y|X}(y|x)$ at a rate $O(n^{-1})$.

Copula Estimation

The asymptotic properties of copulas are different with the above two density function estimation methods. Recall the joint density function $f(x_1, x_2, \cdots, x_n)$ defined in (3.8),

$$f(x_1, x_2, \cdots, x_n) = c[F_1(x_1), F_2(x_2), \cdots, F_n(x_n)]\prod_{i=1}^{n} f_i(x_i),$$

where $f_i(x_i)$ is the univariate density corresponding to the univariate distribution $F_i(x_i)$, and c is the copula density. Therefore, the asymptotic performance of copulas include the performance of two parts: the copula density and the univariate densities. We have seen that the estimation of the univariate densities using the NW estimator converges to the real density when the sample size n goes to infinity.

The asymptotic performance of the copula density, more specifically, the copula parameters, such as the correlation matrix $\mathbf{R}$ in the normal copula and the Kendall's τ in the Frank copula, is related to the ways to obtain these parameters [49].

Let $\{x_{1t}, \cdots, x_{nt}\}_{t=1}^{T}$ be observations of n variables. From (3.8), the log-likelihood function is expressed as:

$$l(\theta) = \sum_{t=1}^{T} \ln c(F_1(x_{1t}), \cdots, F_n(x_{nt})) + \sum_{t=1}^{T} \sum_{i=1}^{n} \ln f_i(x_{it}), \tag{3.13}$$

where θ is the set of all parameters of both the marginals and the copula. In our study, θ only represents copula parameters since the bandwidth of the NW kernel density estimation is known.

Letting $\partial l / \partial \theta = 0$, we have the maximum likelihood estimator $\hat{\theta}_{ML} = \text{argmax } l(\theta)$. The asymptotic normality of $\hat{\theta}_{ML}$ has been verified, that is,

$$\sqrt{T}(\hat{\theta}_{ML} - \theta_0) \rightarrow \mathcal{N}(\mathbf{0}, \mathscr{J}^{-1}(\theta_0)),$$

where $\mathscr{J}(\theta_0)$ is the usual Fisher information matrix and θ_0 is the true value.

The IFM estimator is obtained by using the method of inference functions for the margins (IFM) [53]. It also has the property of asymptotic normality, that is,

$$\sqrt{T}(\hat{\theta}_{IFM} - \theta_0) \rightarrow \mathcal{N}(\mathbf{0}, \mathscr{V}^{-1}(\theta_0)),$$

where $\mathscr{V}(\theta_0)$ is the information matrix of Godambe [54].

The asymptotic normality shows that the estimation of the copula parameters is a consistent estimator whose distribution around the true parameter θ_0 approaches a normal distribution with standard deviation shrinking in proportion to $1/\sqrt{T}$ as the sample size T grows. Therefore, the estimation of copula reaches the true value in the form of (3.8).

Through the asymptotic analysis, we may predict the estimation performance in a large sample environment. But our simulation can only result from the finite samples. We use an example to illustrate the performance of copula estimation methods, double kernel local linear method, and univariate NW density estimator with 100 samples.

Suppose the distributions associating with two random variables X_1, X_2 are known as:

$$f_{X_1}(x_1) = \frac{\pi}{4}\sin(\frac{\pi}{2}x_1),\ 4 \le x_1 \le 6,$$

$$f_{X_2|X_1}(x_2|x_1) = x_1\frac{(x_2-a)^{x_1-1}}{(b-a)^{x_1}},\ x_1 > 0,\ 0 \le a \le x_2 \le b < \infty.$$

Assume $a = 1, b = 5$. Generate n sample pairs $(x_{11}, x_{21}), \cdots, (x_{1n}, x_{2n})$ from the above given distributions. Compute the estimated distributions with the double kernel local linear method (LLM), the copula, and the NW density estimator, based on n samples. Define the mean integrated square error as MISE $= \int(\hat{f}(x) - f(x))^2\ dx$, where $f(x)$ denotes the true, analytical distribution and $\hat{f}(x)$ denotes the estimation. We explore the estimation performance in terms of the MISE.

Figure 3.4 represents the univariate NW kernel density estimations of $f(x_1)$ and $f(x_2)$ with the normal kernel, which are the square solid lines denoted by 'kernel density estimation'. The dashed lines represent the histograms of the sample data. Comparing the square solid lines and the solid lines of real pdf, we derive the MISE of $f(x_1), f(x_2)$ as 6.0567e-005 and 9.8268e-005, respectively.

The conditional density function $f(x_2|x_1)$ is estimated by both the LLM, the normal copula and the Frank copula, which correspond to the star solid line, square solid line and circle solid line, respectively, in Figure 3.5. The solid line represents the real density. The MISE of $f(x_2|x_1)$ using the LLM, the normal copula and the Frank copula are calculated as 3.0571e-004, 5.3776e-004 and 5.4018e-004, respectively.

Interpreting with respect to the value of MISE, we verify a well-known result that the NW density estimators is an appropriate way in estimating univariate density. Furthermore, good fits of both the LLM and the copulas are also observed in Figure 3.5. Therefore, we conclude that the combinations of the LLM, the copulas and the NW density estimator perform well as long as the

corresponding parameters are precisely picked according to some specific criteria. As a result, we adopt these methods to compute the required statistics discussed in Chapter 6.

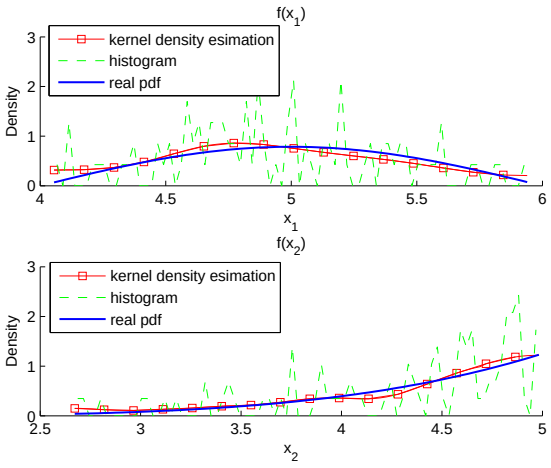

Figure 3.4: Univariate density estimation of $f(x_1), f(x_2)$

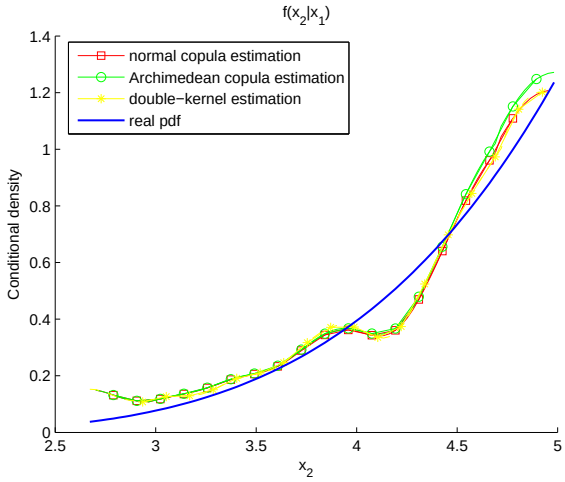

Figure 3.5: Conditional density estimation of $f(x_2|x_1)$

CHAPTER 4. THE BOOTSTRAP METHOD

Suppose $F_X(x)$ and $F_Y(y)$ have been estimated using the NW kernel density estimation based on the observations of X and Y. A hypothesis test can then be conducted to support or reject the claim that $U = F_X(X)$ is dependent of $V = F_Y(Y)$. We adopt the Pearson correlation coefficient as the test statistic. The traditional formula-based hypothesis tests, such as the t-test or Fisher's z test, may not be valid for examining the Pearson correlation coefficient when the sampling distribution is unknown. Therefore, the adaptable alternatives, the bootstrap and permutation methods, are introduced when traditional theory fails. Because the bootstrap and permutation methods take advantage of the resampling idea and collect the information about the sampling distribution of a statistic entirely from one sample, they allow us to perform the hypothesis test based on the Pearson correlation coefficient.

First, we introduce the basics of statistical hypothesis testing which lays the foundation for statistical inference. Then, we present the bootstrap method. Finally, the permutation test is introduced. It has same methodology as the bootstrap method but resamples in a different way.

4.1 Terminology of Statistical Hypothesis Test

A *statistical hypothesis test*, or *significance test*, is a way to make statistical decisions using observations. It is a procedure to analyze the evidence in favor of or against some claims about a population parameter and to make an inference to support or reject the claims based on the sample data drawn from this population.

Every statistical hypothesis test starts with the formulation of the claims, which requires a *null hypothesis* H_0 and an *alternative hypothesis* H_a. Both claims are stated in terms of a population parameter such as the mean. The null hypothesis H_0 represents the basis of the argument and is

established for the purpose of testing, while the alternative hypothesis H_a is the complement of the null hypothesis H_0. If H_0 is rejected, H_a should be accepted.

Because the hypothesis test is based on incomplete information, i.e., a sample rather than the entire population, there always exists the possibility that the wrong decision is made. A Type I error occurs if the null hypothesis is rejected when it is true, while a Type II error occurs if the null hypothesis is not rejected when it is not true. The maximum allowable probability of making a Type I error is called *size*, or *the level of significance*, α. In practice, α represents the risk we are willing to take by controlling the probability of a Type I error.

Once the claims and the level of significance have been set up, the next step is to compute the *test statistic*, a quantity coming from the sample information which represents the population parameter of interest. The test statistic is used to make decision of rejecting or accepting H_0.

We apply two decision rules for deciding to reject H_0 or fail to reject H_0. One is based on the *p-value* and the other is based on the rejection region or *critical value*. The p-value is the probability of obtaining a value at least as extreme as the test statistic would be if observed under the null hypothesis, while the critical value is the value that a test statistic must exceed in order for the the null hypothesis to be rejected. Both the p-value and the critical value are determined according to the level of significance and the sampling distribution of the test statistic.

The most common hypothesis testing method is to find a test statistic whose distribution under the null hypothesis is a well-known result, such as a normal distribution or a t-student distribution, and make an inference upon it. Since no data are exactly normal, the t-procedures based on t-student distribution are more useful in real applications due to the property of insensitivity to data deviations from normality. Nevertheless, t-procedures may lead to bias in the cases involving strongly skewed data, small amounts of data, etc. A new method, the bootstrap method, has been introduced to meet the needs that simple traditional methods do not satisfy [55, 56].

4.2 The Bootstrap Idea

Statistical inference using a hypothesis test is based on the sampling distribution of a statistic. In practice, the explicit sampling distributions are sometimes hard to obtain and therefore the traditional inference methods cannot apply in such settings. The bootstrap was introduced by Efron in 1979 as an alternative to provide answers when traditional methods are known to fail [57].

The bootstrap method relaxes strict conditions such as distribution restrictions or sample size, and sets us free from the need for normal data or large samples. Consequently, it is particularly useful in the cases where the central limit theorems are inapplicable or difficult to employ. Moreover, with sufficient computing power, the bootstrap method could give results that are more accurate than those from traditional methods.

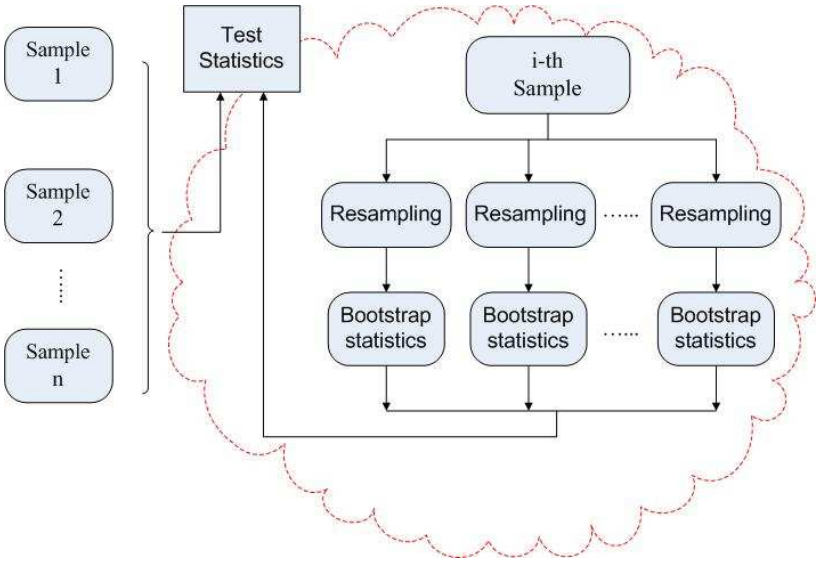

Figure 4.1: Bootstrap idea illustration diagram

When you 'pull up by your own bootstraps', you succeed by relying on your own limited resources. The bootstrap works by sampling within a sample. The essential idea of the bootstrap is shown in the cloud of Figure 4.1. The outside of the cloud is the traditional way to get the test statistic based on n collected data samples, while the inside of the cloud applies the bootstrap process to obtain a test statistic using only one sample. The bootstrap process basically includes doing resampling of a certain sample, computing the statistics from each resample, and eventually evaluating the statistic of the original population associating with the bootstrap distribution based on these resamples.

The bootstrap resampling is to sample with replacement from one sample.

Instead of getting many samples from the population, bootstrap resampling creates many resamples by repeatedly sampling with replacement from one random sample. That is, a

resample contains multiple duplicates of some observations and no other observations at all. Each resample has the same size as the original sample and is independent and identically distributed (iid). For example, the top box in Figure 4.2 is an arbitrary sample of size $n = 8$. The four lower boxes are four resamples from this original sample. Some values from the original are repeated in the resamples because each resample is obtained by sampling with replacement.

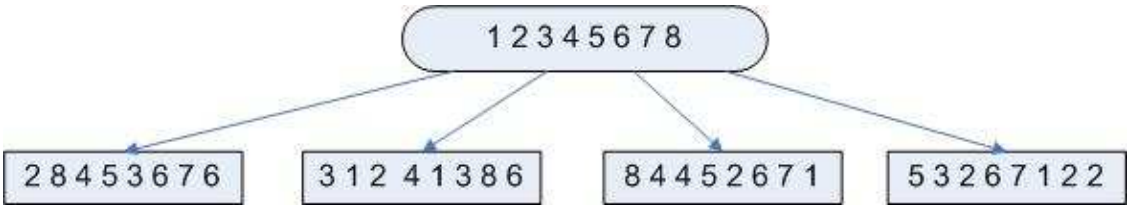

Figure 4.2: Resampling illustration

The bootstrap distribution represents the sampling distribution.

Moore [58] defined the key idea of bootstrap as: '*bootstrap is first of all a way of finding the sampling distribution, at least approximately, from just one sample*.'

We assume the only available sample is a valid representative of the population. If we do resampling many times to this sample, these resamples should represent the characteristics we could get when we draw many samples from the population. As a result, the bootstrap distribution of a statistic coming from many resamples represents the sampling distribution of the statistic coming from many samples.

It is proved that the bootstrap distribution is consistent with the sampling distribution in shape and spread [58], but does not have the same center. The bootstrap distribution is centered at the computed statistic for the original sample, plus any bias. On the other hand, the sampling distribution is centered at the actual value of the parameter in the population, plus any bias. The important thing is that the two biases are similar. A small bias of the bootstrap distribution means that it is centered at the statistic of the original sample, and suggests that the sampling distribution of the statistic is centered at the population parameter.

A hypothesis test using both traditional inference method and bootstrap idea is illustrated in the following example.

Example 1: A bus company claims that the passenger's mean waiting time is less than 10 minutes. An investigation of the waiting times of 35 randomly chosen passengers displayed in Table 4.1 shows that the waiting time has a sample mean of 9.4457 minutes and a standard deviation of 1.8705 minutes. Is there enough evidence to support the claim at $\alpha = 0.01$?

Table 4.1: Sample statistics for bus waiting time

Bus waiting time of 35 passengers(in minutes)											
10.4	8.5	12.9	13.12	6.9	7.8	6.34	8.2	8.1	13.3	8.32	7.1
9.02	9.5	8.98	10.9	11.81	8.5	9.1	8.65	9.2	11.5	9.4	14.2
11.4	8.2	8.4	9.22	9.24	9.3	9.2	9.1	8.7	8.2	7.9	

The claim is "the mean waiting time is less than 10 minutes", and therefore the null and alternative hypotheses are

$$H_0 : \mu \geq 10 \text{ minutes},$$

$$H_a : \mu < 10 \text{ minutes } (claim).$$

1. Traditional z-test

   According to the central limit theorem, the z-score $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, where μ is the mean and σ is the standard deviation, is approximately standard normal distributed if n is large enough. The test statistic, or the z-score, of this example is $z = \frac{9.4457-10}{1.8705/\sqrt{35}} \approx -1.7531$.

   The two decision rules based on the p-value and the rejection region are indicated in Figure 4.3. Since the alternative hypothesis H_a contains the less-than inequality symbol($<$), the hypothesis test is a left-tailed test. Consequently, the p-value p is equal to the area to the left of $z = -1.7531$ as shown in the right side graph. According to the standard normal distribution table, $p = 0.04$. The fact that p-value is greater than $\alpha = 0.01$ leads us not to reject the null hypothesis because small p-values are evidence against H_0. In the left side graph, the critical value z_α, which defines the rejection region where the null hypothesis is

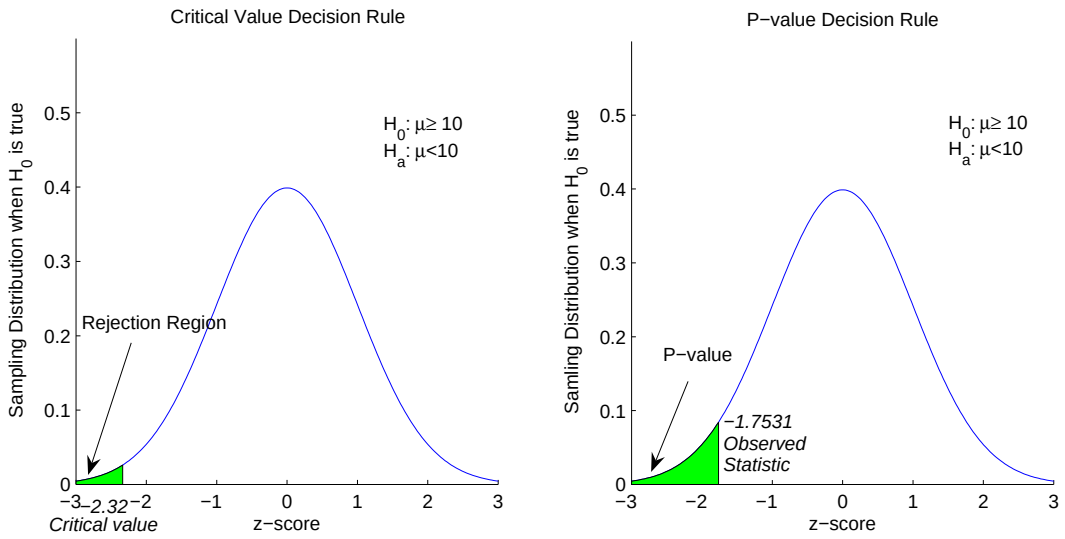

Figure 4.3: Two decision rules for example 1

not probable, is computed. If the test statistic falls in the rejection region, H_0 is rejected. For a left-tailed test, the region left of the critical value z_α is the rejection region. The inverse of the standard normal cumulative distribution function at $\alpha = 0.01$ yields the critical value $z_\alpha = -2.32$. Since $z = -1.7531 > z_\alpha$, the test statistic does not fall in the rejection region, and thus H_0 is not rejected. Both decision rules turn out the same conclusion, that is, at the 1% level of significance, we do not have sufficient evidence to say that the mean waiting time is less than 10 minutes.

2. Bootstrap method

The above traditional z-test is derived from the assumption that the distribution of the sample mean can be approximated as a normal distribution according to the central limit theorem. This approximation is improved as the sample size increases. The left side graph of Figure 4.4 is the actual histogram of the 35 samples in Table 4.1, and it is not surprising that there is obvious deviation from normality due to the relatively small sample size. The hypothesis test conclusion based on a biased distribution assumption tends to be unreliable.

The right side graph of Figure 4.4 shows the histogram of 1000 resample means derived from bootstrap resampling. Define the mean and the standard deviation of B resamples $x_{r1}, \cdots, x_{rB}$

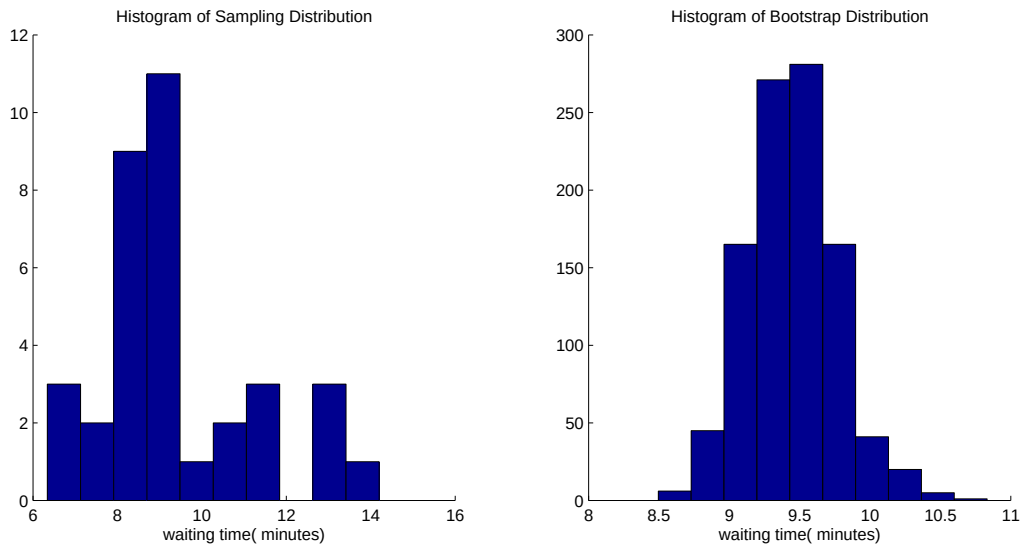

Figure 4.4: Comparison between traditional method and bootstrap method for example 1

as

$$\bar{x}_{\text{boot}} = \frac{1}{B}\sum_{i=1}^{B} x_{ri}$$

$$\sigma_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(x_{ri} - \bar{x}_{\text{boot}})^2}.$$

The bootstrap distribution is centered at $\bar{x}_{\text{boot}} = 9.4341$ and has a standard deviation $\sigma_{\text{boot}} = 0.3053$. This bootstrap distribution is roughly normal. It suggests that the sampling distribution of mean waiting time is also roughly normal. The estimate bias is $9.4341 - 9.4457 = -0.0116$. The small bias of -0.0116 of the bootstrap statistic suggests a small bias in the estimate of population mean.

Since the bootstrap distribution is approximately normal and has a small bias, it is now safe to apply the z-test. The corresponding z-score is:

$$z = \frac{\bar{x}_{\text{boot}} - \mu}{\sigma_{\text{boot}}} = \frac{9.4341 - 10}{0.3053} \approx -1.8532.$$

$z = -1.8532$ is not located in the rejection region of Figure 4.3 and therefore we arrive at the same conclusion: there is not sufficient evidence to say that the mean waiting time is less than 10 minutes.

In Example 1, both the traditional inference method based on theory and the bootstrap method are viable, since the sampling distribution of the statistic is known. Although they both end up with the formula-based z-test, the result of the bootstrap method is considered more accurate than that of traditional inference method since the bootstrap process greatly reduces the distribution deviation from normality.

The bootstrap method is useful in settings where the sampling distribution of the statistic is not known. The bootstrap method does not rely on central limit theory or other theory to give information about the sampling distribution of a statistic, and therefore allows us to learn about the unknown sampling distribution in many settings where traditional theory fails. Furthermore, it has been established that the bootstrap method is more adaptable and precise as long as the distribution is computed with sufficiently many resamples [55, 56].

4.3 The Permutation Test

A permutation test is a statistical significance test in which a sample distribution is obtained by calculating all possible values of the test statistic under rearrangements of the observations within a sample. It shares the same methodology as the bootstrap method to improve the finite sample approximation, but resamples in a different way.

A permutation test involves two distinct procedures of getting *permutation samples* and constructing a *permutation distribution*. The key is to choose permutation samples from the data *without replacement in a way that is consistent with the null hypothesis*, since the p-value is calculated as if the null hypothesis were true. We outline the permutation process using the following example.

Example 2: The student union wants to know whether the student's feedback (0-100) rating of its performance has increased after a series of activities were held. Table 4.2 shows the feedback ratings of 15 randomly chosen students before these activities and 12 randomly chosen students

after these activities. At $\alpha = 0.05$, is there enough evidence to conclude that the student's feedback rating has changed?

Table 4.2: Feedback statistics for the performance of student union

Feedback for student union's performance														
Before activities								After activities						
42	52	58	32	50	25	40	53	57	48	70	61	84	43	
68	61	75	46	62	79	54	70	55	80	75	78	88		

Let the difference between "before activities" ratings and "after activities" ratings be μ_d, the null and alternative hypotheses are claimed as

$$H_0 : \mu_d = 0,$$
$$H_a : \mu_d > 0.$$

According to the hypotheses, the test statistic is $\bar{x}_{\text{after}} - \bar{x}_{\text{before}}$. As shown in Figure 4.5, the feedback rating distributions are far from a normal distribution, so that we choose to apply a permutation test rather than a traditional t-test.

To resample in a manner that agrees with the null hypothesis, permutation resampling, which scrambles the assignment of feedback ratings to 'after activities' group and 'before activities' group, is used. That is, randomly choose 15 of 27 feedback ratings as 'before activities' ratings and the others as 'after activities' ratings. Such permutation resamples are chosen without replacement from the original sample. Calculate $\bar{x}_{\text{after}}$ and $\bar{x}_{\text{before}}$, the difference between them is the statistic.

Repeat the permutation resampling 1000 times. These 1000 calculated statistics form a permutation distribution. The actually observed statistic from Table 4.2 is: $\bar{x}_{\text{after}} - \bar{x}_{\text{before}} = 67.4167 - 53.1333 = 14.2833$. We can get the p-value by locating the observed statistic in the permutation distribution as shown in Figure 4.6.

Viewing the permutation distribution as the sampling distribution, the p-value is the probability that the statistic takes a value at least as large as the observed statistic 14.2833 under the hy-

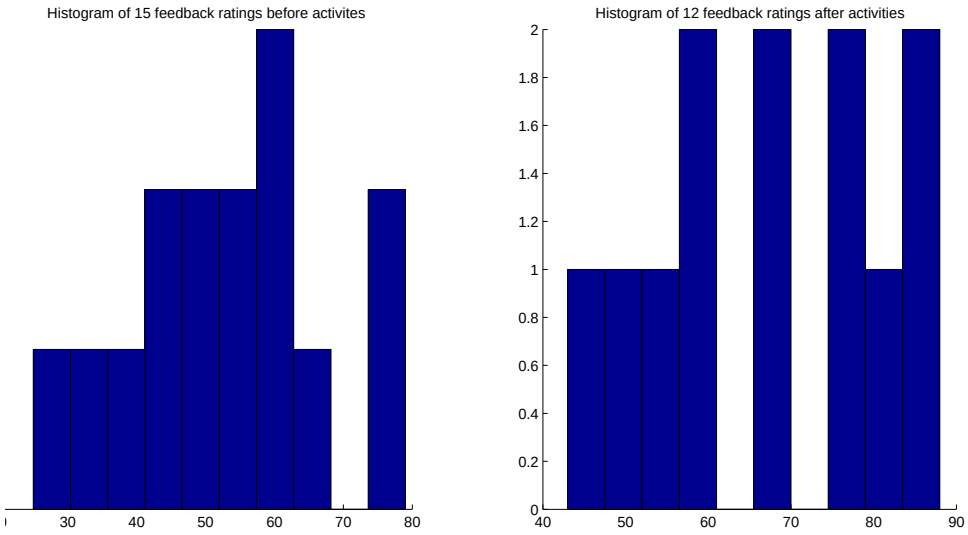

Figure 4.5: Histogram of feedback ratings for example 2

pothesis of $\mu_d > 0$. The resampling result counts 13 points which are larger than 14.2833 and thus the p-value is $p = \frac{13}{1000} = 0.013 < \alpha = 0.05$, which means that at the significance level $\alpha = 0.05$, we have strong evidence to say that the feedback rating to student union's performance has increased after a series of activities.

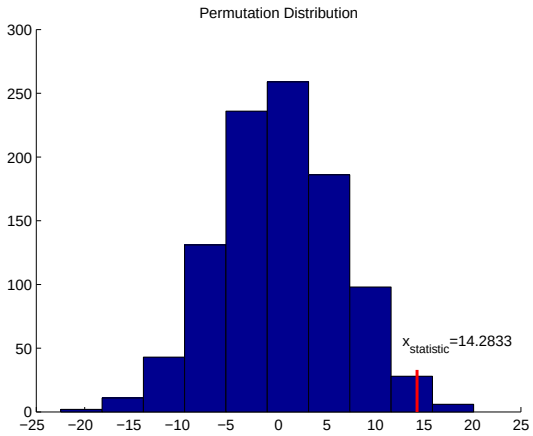

Figure 4.6: Permutation distribution of 1000 resamples for example 2

To find the p-value of a permutation test, the only data sample must be resampled in a way consistent with the null hypothesis. In practice, the permutation test could be applied in cases such as matched-pair problem, two-sample comparison, and the relationships between two random variables. In this study, we use the bootstrap idea and the permutation test to investigate the dependence relationship between two distribution functions U and V.

CHAPTER 5. TEST STATISTICS

We determine whether $U = F_X(X)$ and $V = F_Y(Y)$ are dependent by measuring the correlation between them. Several typical dependence measurements, such as the Spearman correlation and the Kendall correlation, are introduced in Section 5.1. It is well known that the Pearson correlation coefficient is an appropriate and strong dependence measurement only in Gaussian (normal) models and to some degree in general linear models [59]. An alternative dependence measurement, such as the Kendall correlation, might be needed to satisfy the requirement of non-Gaussian and nonlinear models. However, in this study, we still consider the traditional Pearson correlation function as a reasonable dependence measurement choice. The reason is as follows.

Our main emphasis is to discriminate weak dependence from genuine independence through a significance test. The weak dependence represents extremely low dependence between random variables in both linear relation (measured by the Pearson correlation) and other types of relation (measured by the Spearman correlation, the Kendall correlation, etc). Our approach is to apply the Box-Cox transformation to enhance the linear relationship between U and V, perform a test to determine the significance of the Pearson correlation coefficient, and then make a decision as to whether U and V are correlated. If U and V are tested as correlated, it is certain that X and Y are correlated and dependent. However, if U and V are tested as uncorrelated, it only implies that X and Y are uncorrelated. The independence between X and Y holds only if they are jointly normally distributed. In this sense, we test dependence rather than independence.

In Section 5.2, we discuss the hypothesis tests based on the Pearson correlation coefficient. It is then applied in practical simulations to show us the performance in distinguishing weak dependence from independence.

Another key point is how to find the weak dependence between U and V. Our solution is to apply the Box-Cox transformation to enhance the linear relationship between U and V. Basically,

the extremely low dependence will be magnified by this transformation so that the independence test can better detect it. In Section 5.3, we will introduce the Box-Cox transformation and its application in our study. Finally, we summarize the test statistics in Section 5.4.

5.1 Dependence Measurements

The most common measures of dependence are based on the classification of pairs of observations as concordant or discordant, i.e., *concordance measurements*. The Kendall's correlation τ, Spearman's correlation ρ_S and Pearson correlation ρ are several concordance measures [48,49].

A pair of observations is concordant if the observation with the larger value of X has also the larger value for Y. The pair is discordant if the observation with the larger value of X has the smaller value of Y. If (X_1,Y_1) and (X_2,Y_2) denote distinct copies of (X,Y) then the (X_i,Y_i)'s are said to be concordant if $(X_1-Y_1)(X_2-Y_2)>0$ holds true whereas they are said to be discordant when the reverse inequality is valid. Hence, Pr(concordance) $=\Pr[(X_1-X_2)(Y_1-Y_2)>0]$ and Pr(discordance) $=\Pr[(X_1-X_2)(Y_1-Y_2)<0]$.

The Kendall's correlation is defined as:

$$\tau = \Pr[(X_1-X_2)(Y_1-Y_2)>0]-\Pr[(X_1-X_2)(Y_1-Y_2)<0],$$

and the Spearman's correlation is defined as:

$$\rho_S = 3\{\Pr[(X_1-X_2)(Y_1-Y_3)>0]-\Pr[(X_1-X_2)(Y_1-Y_3)<0]\},$$

where $(X_1,Y_1),(X_2,Y_2)$ and (X_3,Y_3) are three distinct copies of (X,Y). The Kendall's correlation and the Spearman's correlation share the properties of concordance measurements: they range from -1 to 1 and are equal to zero when X and Y are independent. Concordance measurements are intended to measure the strength of dependence between X and Y. But different measurements measure the strength of dependence in different ways. It is not easy to assign an operational interpretation to Spearman's ρ_S. Kendall's τ, on the other hand, has a simple interpretation. It is the difference between the probability of concordance and discordance. $\tau=1$ means that X and Y are comonotonic while $\tau=-1$ means that X and Y are countermonotonic. Moreover, the

estimation of both ρ_S and τ using practical observations are not difficult. For example, using T observations $\{x_t, y_t\}$, the sample Kendall's τ can be computed by

$$\hat{\tau} = \frac{2}{T(T-1)} \sum_{i=1}^{T} \sum_{j>i} \text{sgn}(x_i - x_j)(y_i - y_j).$$

On the other hand, the Pearson correlation is defined as:

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}.$$

$|\rho| = 1$ if and only if X and Y are related by an affine transformation [60], i.e., $|\rho| = 1 \Leftrightarrow Y = bX + a$ for some constants a and b. Thus, an interpretation of the Pearson correlation is the degree of linearity between X and Y. The closer $|\rho|$ is to unity, the closer X and Y are to being related by an affine transformation. Consider the bivariate normal distribution,

$$f_{XY}(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det \mathbf{A}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^H \mathbf{A}^{-1}(\mathbf{x}-\mathbf{m})},$$

where $\mathbf{x} = [x, y]^H, \mathbf{m} = [E(X), E(Y)]^H$, and

$$\mathbf{A} = E\left[\begin{bmatrix} X - E(X) \\ Y - E(Y) \end{bmatrix} \quad [X - E(X), Y - E(Y)] \right] = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{\sigma_Y^2}{\sigma_X^2\sigma_Y^2(1-\rho^2)} & \frac{-\rho\sigma_X\sigma_Y}{\sigma_X^2\sigma_Y^2(1-\rho^2)} \\ \frac{-\rho\sigma_X\sigma_Y}{\sigma_X^2\sigma_Y^2(1-\rho^2)} & \frac{\sigma_X^2}{\sigma_X^2\sigma_Y^2(1-\rho^2)} \end{bmatrix}.$$

When $|\rho| = 1$, the matrix $\mathbf{A}$ becomes singular. The singularity of $\mathbf{A}$ means that linear dependency exists between the columns of $\mathbf{A}$, or X and Y are linearly related.

In Chapter 6, we give a comparison of the sample dependence measurements ρ_S, τ and ρ based on the random variables generated from the data generating processes of our simulation. The results intuitively indicate that weak dependence represents not only a low linear relationship but also some low relationships of other types. In other words, ρ_S and τ do not provide any extra dependence information than ρ does. Therefore, it is sufficient to use the Pearson correlation ρ as a test statistic to distinguish weak dependence from independence.

5.2 Hypothesis Tests based on the Pearson Correlation

To judge if $U = F_X(X)$ and $V = F_Y(Y)$ are dependent through a statistical hypothesis test of the Pearson correlation coefficient, the corresponding null and alternative hypotheses are claimed as:

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0,$$

where ρ is the population Pearson correlation coefficient between U and V.

According to the discussion in Chapter 4, we calculate r, the sample correlation coefficient, to determine whether there is enough evidence to decide whether the population correlation coefficient ρ is significant at a specified level of significance level α. If $|r|$ is greater than the critical value or in the rejection region, it is said that the null hypothesis is not probable, or there is enough evidence to decide that the correlation is significant, and thus we could conclude that U and V are not independent. If $|r|$ is less than the critical value or not in the rejection region, we fail to reject the null hypothesis, or there is not enough evidence to conclude that the correlation is significant, or U and V are uncorrelated. Therefore, X and Y are uncorrelated and independent under the assumption of joint normality. The specification of the critical value or rejection region depends on the distinct setup of various tests. The feasible tests with respect to the Pearson correlation include: the standard t-test, Fisher's z transformation test and the bootstrap test.

1. t-test for Pearson correlation coefficient

If the population distribution is normal or nearly normal, a t-test can be used to test whether the Pearson correlation between two random variables is significant. The standardized test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}},$$

where r is the sample correlation coefficient and n is the sample size. The sampling distribution of the test statistic t is a t-distribution with $n-2$ degrees of freedom. Given a level of significance α (0.05 or 0.1), the critical value r_0 is obtained by solving $\Pr(r > r_0) = \frac{\alpha}{2}$ due to the symmetry of the t-distribution. When $|r| > |r_0|$, we reject the null hypothesis and say U and V are dependent.

When $|r| < |r_0|$, we accept the null hypothesis.

2. Fisher's z transformation test

Fisher's z transformation converts the sample Pearson correlation coefficient r to a normally distributed variable z,

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

where ln is the natural logarithm. This robust transformation was suggested by Ronald Fisher in 1915 and proved to have an approximately normal distribution when the samples come from a bivariate normal distribution with sample sizes of 10 or more [61]. The mean and the standard deviation of z are

$$\mu = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) + \frac{\rho}{2(n-1)} \text{ and } \sigma = \frac{1}{\sqrt{n-3}},$$

respectively, where n is the sample size and ρ is the population correlation coefficient. Under the null hypothesis $H_0 : \rho = 0$, $z \sim \mathcal{N}(0, \frac{1}{n-3})$, and the corresponding critical value z_0 is obtained by solving $\Pr(z > z_0) = \frac{\alpha}{2}$, where α is the level of significance. $|z| > |z_0|$ means that U and V are dependent.

In some cases, the Fisher's z transformation test is believed to be more reliable than the t-test in that it improves the normality substantially, especially for small sample sizes and extreme sample correlations [62]. That is, the Fisher's z transformation test works even if the t-test might be invalid because the sampling distribution of the Pearson correlation coefficient r fails to meet the normality assumption. When the absolute value of the correlation in the population is low (say $|\rho|$ is less than about 0.4), then the sampling distribution of r is approximately normal. However, with high values of correlation, the distribution has an apparent negative skew. In such case, the t-test for the Pearson correlation coefficient may lead to a biased result. The t-test of the correlation gives reliable results only if the sampling distribution of the Pearson correlation coefficient is at least roughly normal.

3. The bootstrap test for the Pearson correlation

Under asymptotic theories, the statistic of the above two tests is either t-distributed or normally distributed. However, such explicit asymptotic distributions are hard to achieve under realistic conditions, or they are inferred with certain rigid population distribution assumptions. In many cases where the central limit theorems are impracticable or the specified distribution assumptions are not met, the bootstrap test exhibits its strong capability and applicability. In fact, since the bootstrap test performs better than the traditional t-tests and z-tests which require strict distribution restrictions, it has become an efficient alternative in current research [58] and also an important test approach in our study.

As discussed in Chapter 4, the essential principle of any hypothesis test is that the sampling distribution of the test statistic is estimated under the assumption that the null hypothesis is true. Therefore, the resampling of the bootstrap should comply with the same rule, that is, resampling the data in a manner that is consistent with the null hypothesis. We claim that $H_0 : \rho = 0$, i.e., U and V are uncorrelated. Let $\{u_t, v_t\}_{t=1}^{T}$ be observations. Since resampling u_t and v_t separately will destroy the relation between U and V, according to [58], we do resampling by randomly permutating u_t among v_t. The procedure is a permutation test which resamples the data observation without replacement according to the null hypothesis and is shown as follows:

- Calculate the sample correlation r_0 between original u_t and v_t;
- Resample the data. Let u_t^*, v_t^* denote the resample observations, then $v_t^* = v_t$ in their original order and $u_t^* = u_t$ in the reshuffled order;
- Repeat step 2 B times, where B is a large number. We obtain B resamples and each resample contains observations $\{u_t^*, v_t^*\}_{t=1}^{T}$;
- For every resample, calculate the sample correlation r^*;
- Calculate the p-value, which is the proportion of the resamples with sample correlation r^* larger than r_0, i.e.,

$$\hat{p} = \Pr(r^* > r_0).$$

Since the p-value is the observed probability of a Type I error and the level of significance α is the maximum allowable probability of making a Type I error, the null hypothesis is rejected if $\hat{p} < \alpha$ and the null hypothesis is accepted if $\hat{p} > \alpha$. The Type I error occurs if the null hypothesis is rejected when it is actually true.

To make the test results based on the correlation coefficient more reliable, we introduce a simple correction formula to compensate for the estimation bias caused by the small sample size [63, 64]. The sample Pearson correlation coefficient, r, is a biased estimation of the population correlation coefficient, ρ. The bias could reach $0.03 - 0.04$ under some real scenarios and it decreases as the sample size increases. This small discrepancy could be critical when we are concerned about the accuracy of a non-zero correlation coefficient estimation which discriminates an independence relationship from a weak dependence relationship and therefore an estimation correction is necessary. Two approximately unbiased estimators of the population correlation are provided: $\hat{\rho} = r(1 + (1 - r^2)/2n))$ (Fisher, 1915) and $\tilde{\rho} = r[1 + (1 - r^2)/2(n - 3)]$ (Olkin and Pratt, 1958). Since the latter one is a more nearly-unbiased estimator of population correlation, we accept it as an estimation correction in our study.

5.3 Box-Cox Transformation

We apply the Pearson correlation coefficient as the test statistic to examine the dependence relationship between $U = F_X(X)$ and $V = F_Y(Y)$. It is not surprising that the test performance reduces rapidly in weakly dependent cases. Due to limited sample size and measurement accuracy, the value of the sample correlation coefficient measurement r representing the weak dependence can easily be small enough to be considered as zero statistically. However, zero correlation under the normality assumption theoretically means absolutely no dependence between two random variables. In other words, such simple correlation tests cannot differentiate between weak dependence and independence.

One of the essential principles of probability and statistical theory is that the Pearson correlation between any two random variables or their transformations is zero if they are independent. Let the symbol '$\perp\!\!\!\perp$' denote stochastic independence. $X \perp\!\!\!\perp Y \Rightarrow \rho(X,Y) = \rho(g(X),h(Y)) = 0$, where $g(\cdot), h(\cdot)$ are some transformation functions and ρ is the population correlation coefficient. The contrapositive is that, if the Pearson correlation between any two random variables

or their transformations is nonzero, then the two random variables are not independent. i.e., $\rho(X,Y) \neq 0$ or $\rho(g(X),h(Y)) \neq 0 \Rightarrow X \not\perp\!\!\!\perp Y$. This principle constructs the foundation of our test mechanism. If U and V are independent, $\rho(g(U),h(V))$ must be zero regardless of the formation of transformations $g(\cdot)$ and $h(\cdot)$. We attempt to find some particular transformations $g(\cdot)$ and $h(\cdot)$ such that $|\rho(g(U),h(V))|$ is large enough to allow the corresponding sample correlation $r(g(U),h(V))$ be considered as nonzero statistically. Once $r(g(U),h(V))$ is tested as nonzero, we conclude that X and Y are not independent and hence the weak dependence is separated from independence.

The Box-Cox transformation is particularly useful in improving the linearity between two random variables [18, 19]. We investigate the hypothesis that if U and V are correlated by testing the hypothesis that if the maximal Pearson correlation coefficient between the Box-Cox transformations of these two random variables is significant. Basically, we magnify the measured correlation through the Box-Cox transformation such that it can be better detected and thus make a distinct conclusion between weak dependence and independence.

5.3.1 Overview

The Box-Cox transformation addresses two issues: normality and linearity. It was first suggested and originated from the idea to improve distribution normality [18]. As we know, the statistical tests based on the assumption of normality are usually more simple, powerful and mathematically tractable than tests without the normality assumption. Therefore, a transformation, such as the Box-Cox transformation, which could produce an approximately normally distributed data set, is necessary and useful.

Suppose Y is a nonnegative random variable, the Box-Cox transformation is defined as:

$$T_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0. \end{cases} \tag{5.1}$$

Note that $\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log(y)$, y is an observation of Y.

With traditional approaches, the transformation parameter λ is chosen such that $T_\lambda(Y) \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2)$, where $\mathbf{X}$ is a known matrix. In linear regression analysis, the column vector of $\mathbf{X}$ rep-

resents the individual input variable. For example, $T_\lambda(Y) \sim \mathcal{N}([X_1, X_2, \cdots, X_n][\beta_1, \beta_2, \cdots, \beta_n]^T, \sigma^2)$. Usually, the maximum likelihood (ML) is applied to get the optimal λ.

Box and Cox assume that $T_\lambda(Y)$ is normally distributed without explanation in [18]. Many current studies using the Box-Cox transformation rely on this normality [65, 66]. For example, Freeman and Modarres assume two bivariate normally distributed random variables Y_1, Y_2 can be obtained by making a Box-Cox transformation to two arbitrary nonnegative random variables X_1, X_2 and then show that the independence test based on Y_1, Y_2 is more efficient than the independence test of X_1, X_2 [65]. However, it is clear that not all data can be power transformed to exactly normal and linear with the known matrix $\mathbf{X}$. Draper and Cox [67] investigate this problem and conclude that the distribution of $T_\lambda(Y)$ is usually symmetric even if no estimation of λ can lead to an exactly normal distribution. One example in [67] is that $T_\lambda(Y)$ has a Weibull distribution by transforming the raw data with exponential distribution, and a Weilbull distribution looks very like a normal distribution.

More importantly, the Box-Cox transformation improves the linearity. The construction of the Box-Cox transformation itself is also a linear fitting process between two random variables. It is this linearity concept that is used in our study to increase the correlation between U and the transformation of V. As a result, the enlarged correlation could be probed more easily and the weak dependence could be distinguished from the independence [68, 69].

5.3.2 The Box-Cox Linearity

When performing a linear fit of X against Y, an appropriate transformation of Y can often significantly improve the fit. The Box-Cox transformation is based on the model $T_\lambda(Y) = X\beta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We wish to achieve the best regressing relationship of $T_\lambda(Y)$ on X by choosing the transformation parameter λ. As we have seen in Figures 2.2 and 2.3 , the closer the data points to the linear fit line, the more linear relationship between variables and the larger the Pearson correlation coefficient. When λ is chosen to maximize the correlation, the distance between the linear fit $T_\lambda(Y) = X\beta + \varepsilon$ and the data points is the smallest in the sense of minimizing the mean square error. Therefore, the Box-Cox transformation is a useful transformation to improve the linearity between $T_\lambda(Y)$ and X [68, 69].

Since there is no closed form solution for the value of λ that maximizes the correlation coefficient, the Box-Cox linearity plot, which plots the correlation between X and the transformed Y for given values of λ, provides a convenient way to find the appropriate transformation parameter λ without engaging in trial and error fitting. The optimal choice for λ is the value of λ leading to the maximal magnitude of the correlation on the plot.

Figure 5.1 in [69] illustrates the linearity of the Box-Cox transformation. The upper-left scatter plot of the original data shows some deviation from a linear fit and indicates that a quadratic fit might be preferable. The upper-right plot is the Box-Cox linearity plot, where λ is the coordinate for the x-axis and the value of the correlation between X and the transformed Y is the coordinate for the y-axis. This Box-Cox linearity plot shows the optimal $\lambda = 2$ and the corresponding Pearson correlation coefficient $\rho = -1$. Therefore, in the third plot which shows the spread of the transformed data, we see a better linear fit. We observe not only a perfect negative linear relationship, but also a significant reduction in the residual standard deviation.

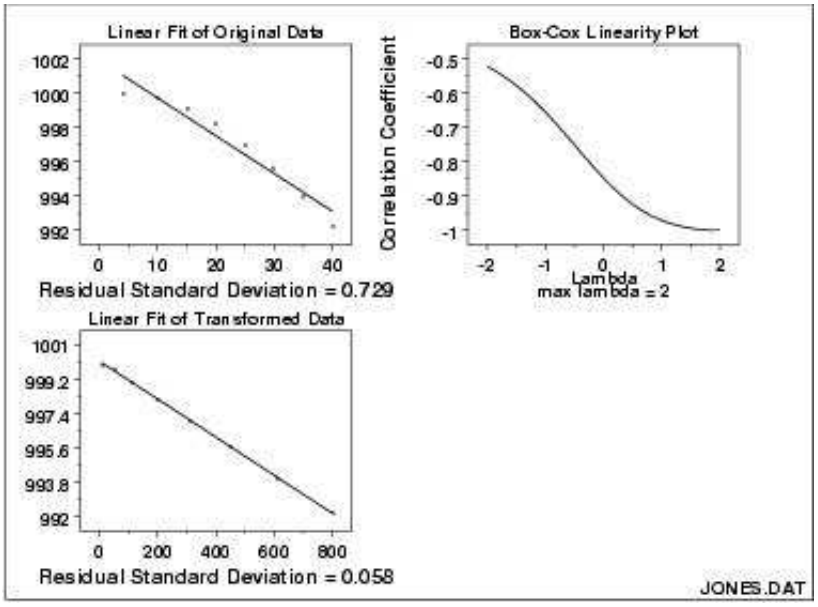

Figure 5.1: Linearity of the Box-Cox transformation

Numerous papers have been published to attest to the effectiveness of the Box-Cox transformation in achieving the normal scenario and allowing thereby application of linearity ever since

the Box-Cox transformation was introduced in 1964. However, there is currently no theory to explain why a power transformation is so effective in achieving the normal and linear situations, and indeed the original derivation of the power transformation was based solely one empirical evidence [18, 70]. In spite of the lack of the theoretical explanation for the normality and linearity of the Box-Cox transformation, this transformation has been widely used in engineering and econometric fields. As to the Box-Cox linearity, we see applications in Alaska pipeline problem of [69], the linear mixed model of [71], etc.

In this study, we are concerned about how to increase the linear relation or magnify the Pearson correlation between $U = F_X(X)$ and $V = F_Y(Y)$ after transformation. Based on the model $T_\lambda(U) = V\beta + \varepsilon$, we wish to improve the linear fit between $T_\lambda(U)$ and V by choosing λ to maximize the correlation $\rho(T_\lambda(U), V)$.

The Box-Cox linearity plot is generated as follows: Choose different values $\lambda_i = \lambda_0 + \frac{i}{m}c$, where c is a positive constant and m is a very large number to ensure the small interval between λ_i and λ_{i+1}. We have $-c = \lambda_0 < \lambda_1 < \cdots < \lambda_{2m} = c$. Let $c = 4$ to save the computation time. Then draw a Box-Cox linearity plot with $\lambda_0, \cdots, \lambda_{2m}$ and find the λ corresponding to the maximal correlation as the optimal choice.

In the Box-Cox linearity plots of Figures 5.2 and 5.3, λ is the coordinate for the abscissa and the value of the correlation between V and the transformed $T_\lambda(U)$ is the coordinate for the ordinate. The value of λ corresponding to the maximum correlation (or minimum for negative correlation) on the plot is then the optimal choice for λ.

Figures 5.2 and 5.3 are produced with 100 observation pairs $\{x_t, y_{t-1}\}$ collected from the data generating processes 2i and 4d, which are introduced in Chapter 6 as follows:

$$2\text{i} : x_t = 0.5x_{t-1} + \varepsilon_{1,t},\ y_t = 0.5y_{t-1} + \varepsilon_{2,t}$$

$$4\text{d} : x_t = 0.5x_{t-1} + 0.5y_{t-1}\varepsilon_{1,t},\ y_t = 0.5y_{t-1} + \varepsilon_{2,t},$$

where both $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ and the initial values $\{x_0, y_0\}$ are i.i.d $\mathcal{N}(0, I_2)$. As discussed in Appendix D, data generated from the processes 2i and 4d imply that X and Y are independent and dependent, respectively. The optimal values of λ are around -0.2 and -1.6 respectively in Figure 5.2 and 5.3. Also, note that the original correlation between U and V could be obtained when $\lambda = 1$,

since the linear transformation $T_1(U) = U - 1$ does not affect the Pearson correlation. In Figure 5.2, the measured sample Pearson correlation $\hat{r}$ between V and the transformed $T_\lambda(U)$ increases from original 0.09184 ($\lambda = 1$) to the maximal 0.121 ($\lambda = -0.2$). The small gain of 0.0292 in this case does not provide a strong evidence to conclude that the sample correlation is statistically significant, or U and V are dependent. In Figure 5.3, the correlation gain achieved by the Box-Cox transformation is 0.1531. The corresponding maximal correlation 0.2212 ($\lambda = -1.6$) is sufficient to make the judgement that U and V are weakly dependent through a Pearson correlation test. Such result is consistent with the known fact that 4d generates data which possess the dependent relationship between X and Y.

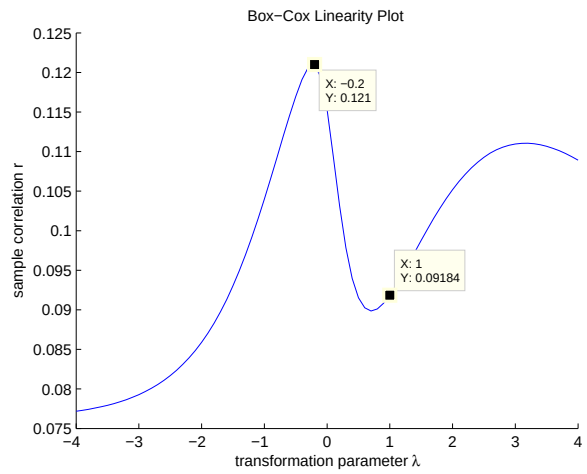

Figure 5.2: Box-Cox linearity plot of 2i, $\lambda_{opt} = -0.2$

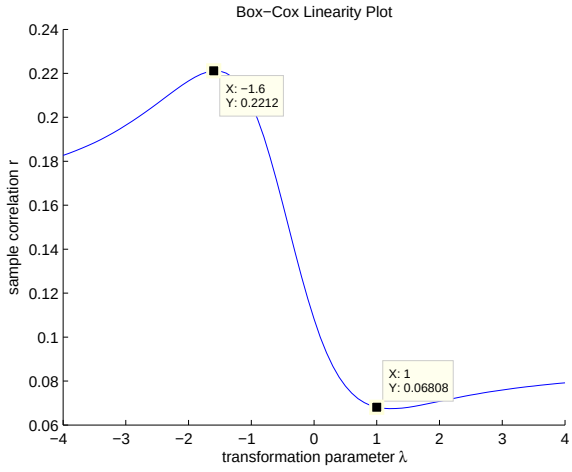

Figure 5.3: Box-Cox linearity plot of 4d, $\lambda_{opt} = -1.6$

5.4 Test Statistics

To examine the independence relationship among random variables X and Y when they are weakly dependent or independent, the underlying test statistics of U and V could be summarized:

1. Perform a Box-Cox transformation $T_\lambda(\cdot)$ to one of the two random variables, the absolute value of newly obtained sample correlation $\hat{r}(U, T_\lambda(V))$ and $\hat{r}(T_\lambda(U), V)$ are increased because of the linearity of Box-Cox transformation, that is, $|\hat{r}| - |r| \geq 0$. If the resulting $|\hat{r}|$

is still close to zero, we conclude that U and V are uncorrelated. If the resulting $|\hat{r}|$, which represents the maximal correlation after the Box-Cox transformation, is greatly enlarged, we conclude that U and V are dependent because any nonzero population correlation coefficient between two random variables or their transformations means they are dependent.

2. In addition to the elementary assumption that samples should be chosen randomly, theoretically, the t-test and the Fisher's z transformation test also require the premise of bivariate normal distribution of population. However, this assumption is not strictly checked in reality and a lot of illustrations of real applications do not pay much attention on it. For example, no check of normality is done in the dice problem ([72, p. 478]), the Boats and Manatees problem ([73, p. 506]), and the college grades and income problem ([74, p. 315]), etc. The more reliable Fisher's z transformation test is deployed in our study only as the reference and validation of the bootstrap test, where the latter test is what we are really interested in since it does not require any rigid distribution restriction and consequently the calculation of the corresponding critical value or reject region does not rely on any formula. The numerical results are presented in Chapter 6.

CHAPTER 6. NUMERICAL RESULTS

In this chapter, we present Monte Carlo simulation results for the hypothesis tests based on the test statistics of the Pearson correlation coefficient. The finite-sample performance of dependence tests such as the Fisher's z transformation test and the bootstrap test are illustrated through Monte Carlo experiments. We create a nonlinear model which involves two random variables X and Y as inputs which meet specified requirements, and then evaluate this model iteratively to investigate if independence can be determined.

Our emphasis is to discriminate weak dependence from genuine independence. When X and Y are weakly dependent, neither the conventional independent tests nor the general correlation test works consistently. Therefore, a new approach is necessary. Our approach is to apply hypothesis tests for the Pearson correlation coefficient over the Box-Cox transformed variables of $U = F_X(X)$ and $V = F_Y(Y)$. We introduce the correlation test before and after the Box-Cox transformation. It turns out that the dependence, probably weak dependence, between X and Y can be found by testing the significance of the Pearson correlation after the Box-Cox transformation. However, the correlation test cannot make the inference as to whether X and Y are independent. As a comparison and complement, we then introduce a variety of independence tests. They are conducted over the variables before and after the Box-Cox transformation with distinct distribution estimation methods and distinct test statistics. It is observed that the performance gets better when the Box-Cox transformation is applied.

In past literature, Su and White [17] propose a conditional independence test regarding weak dependence in time series. They suggest several data generating processes to construct Monte Carlo models. These data generating processes represent various typical instances of linear and nonlinear stochastic processes in time series analysis such as (G)ARCH models, and thus could be widely used in practical applications. More importantly, these processes provide descriptions of

weak dependent variables, which are of great interest to our study. Therefore, we use these data generating processes in our Monte Carlo simulations.

Suppose there are T identically distributed observations $w_t = \{x_t, y_t\}$, $t = t_0, \cdots, t_0 + T - 1$, where w_t denotes the realization of the random variables X and Y. Note that the observation pairs w_t of the data generating processes are collected at different time points, i.e., their ordering is through time. w_t can be considered as the realizations of two random variables X and Y because the processes x_t and y_t reach an equilibrium after a sufficiently long time, independent of the initial conditions. Put another way, the processes are asymptotically strictly stationary processes. The stationarity of the processes is demonstrated in Appendix E.

The following data generating processes, $*$i, $*$d, $*$h and $*$z (non time-series) are applied to test the independence and weak dependence(see [17]):

1. Independent case $*$i

   1i : $w_t = \{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, $t = 0, 1, \cdots$, are i.i.d. $\mathcal{N}(0, I_2)$.

   Let $w_t = \{x_t, y_{t-1}\}$ in the processes of 2i, 3i and 4i, where $y_t = 0.5y_{t-1} + \varepsilon_{2,t}$, $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, $t = 0, 1, \cdots$, are i.i.d $\mathcal{N}(0, I_2)$, and the initial values $\{x_0, y_0\}$ are also i.i.d $\mathcal{N}(0, I_2)$.

   2i : $x_t = 0.5x_{t-1} + \varepsilon_{1,t}$;

   3i : $x_t = \sqrt{h_t}\varepsilon_{1,t}$, $h_t = 0.01 + 0.5x_{t-1}^2$;

   4i : $x_t = \sqrt{h_{1,t}}\varepsilon_{1,t}$, $y_t = \sqrt{h_{2,t}}\varepsilon_{2,t}$, $h_{1,t} = 0.01 + 0.9h_{1,t-1} + 0.05x_{t-1}^2$, $h_{2,t} = 0.01 + 0.9h_{2,t-1} + 0.05y_{t-1}^2$;

   Since data that are generated according to the processes $*$i, where $* \in \{1, \cdots, 4\}$, possess the independent relationship between X and Y, the covariance or correlation between x_t and y_{t-1} is zero.

2. Dependent case $*$d

   With the following processes, $w_t = \{x_t, y_{t-1}\}$ and $y_t = 0.5y_{t-1} + \varepsilon_{2,t}$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, $t = 0, 1, \cdots$, are i.i.d $\mathcal{N}(0, I_2)$ and the initial values $\{x_0, y_0\}$ are also i.i.d $\mathcal{N}(0, I_2)$.

   1d : $x_t = 0.5x_{t-1} + 0.5y_{t-1} + \varepsilon_{1,t}$;

   2d : $x_t = 0.5x_{t-1} + 0.5y_{t-1}^2 + \varepsilon_{1,t}$;

3d : $x_t = 0.5x_{t-1}y_{t-1} + \varepsilon_{1,t}$;

4d : $x_t = 0.5x_{t-1} + 0.5y_{t-1}\varepsilon_{1,t}$;

5d : $x_t = \sqrt{h_t}\varepsilon_{1,t}$; $h_t = 0.01 + 0.5x_{t-1}^2 + 0.25y_{t-1}^2$.

6d : $x_t = \sqrt{h_{1,t}}\varepsilon_{1,t}$, $y_t = \sqrt{h_{2,t}}\varepsilon_{2,t}$, $h_{1,t} = 0.01 + 0.1h_{1,t-1} + 0.4x_{t-1}^2 + 0.5y_{t-1}^2$, $h_{2,t} = 0.01 + 0.9h_{2,t-1} + 0.05y_{t-1}^2$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d $\mathcal{N}(0, I_2)$.

Data that are generated according to the processes $*$d, where $* \in \{1, \cdots, 6\}$, relate with each other, and therefore X and Y are dependent.

As shown in Appendix D, x_t and y_{t-1} are dependent and uncorrelated for the cases 2d-6d, while 1d generates data that are correlated.

1d: $\text{cov}(x_t, y_{t-1}) = \sum_{i=1}^{t} 0.5^{2i-1}\text{var}(y_{t-i}) \neq 0$, where $\text{var}(y_t)$ is a nonzero constant because the process y_t can be considered as an asymptotically strictly stationary process.

2d: $\text{cov}(x_t, y_{t-1}) = \sum_{i=1}^{t} 0.5^{2i-1}E(y_{t-i}^3) = 0$.
Since $y_t = 0.5y_{t-1} + \varepsilon_{2,t}$, $y_0 \sim \mathcal{N}(0,1)$ means that $E(y_t) = 0$ and y_t is normal for all t. Therefore, $E(y_{t-i}^3) = 0$ because of the symmetric property of the normal distribution.

3d: $\text{cov}(x_t, y_{t-1}) = a_t E(x_0 y_0^{t+1})$, where $a_t = a_{t-1}0.5^{t+1}$ and $a_1 = 0.5$. Therefore, $\text{cov}(x_t, y_{t-1}) = 0$ when $\{x_0, y_0\}$ are i.i.d $\mathcal{N}(0, I_2)$.

4d, 5d, 6d: $\text{cov}(x_t, y_{t-1}) = 0$.

3. High frequency case $*$h: $y_t = 0.5z_t + 4\alpha\varphi(x_t/\alpha) + 0.5\varepsilon_t$,

where $\{x_t, z_t, \varepsilon_t\}$, $t = 0, 1, \cdots$, are i.i.d. $\mathcal{N}(0, \mathbf{I}_3)$, α is a constant, and φ is the standard normal density function. The term "high frequency" in econometrics usually means that observations are taken at fine time intervals [75, 76]. For example, $*$h can represent some transaction-by-transaction or trade-by-trade data in financial markets.

Let the processes 1h-4h correspond to α taking values 0, 0.5, 1 and 2, respectively. That is,

1h : $x_t \sim \mathcal{N}(0,1)$, $y_t = 0.5z_t + 0.5\varepsilon_t$,

2h-4h : $y_t = 0.5z_t + 4\alpha\varphi(x_t/\alpha) + 0.5\varepsilon_t$, when $\alpha \in \{0.5, 1, 2\}$.

For the $*$h processes, $w_t = \{x_t, y_t\}$. The process 1h indicates the independence relationship between X and Y, and therefore the covariance or correlation between x_t and y_t is

zero. Furthermore, data generated from the processes 2h-4h are dependent. As discussed in Appendix D, the covariance or correlation between x_t and y_t is also zero. For 2h-4h, $\text{cov}(x_t, y_t) = \text{E}(4\alpha x_t \varphi(x_t/\alpha)) = 0$ since $x_t \varphi(x_t/\alpha)$ is odd.

4. Zero correlation non time series case $*$z

   In this case, X and Y are two simple random variables which are not formed from a time series. In addition, the dependent X and Y are related nonlinearly rather than linearly, that is, the Pearson correlation coefficient equals zero.

   $1\text{z} : Y = X^2,$

   where X has a density function that is symmetric about 0 and the third moment of X exists. $\text{cov}(X,Y) = E(X^3) - E(X)E(X^2) = 0.$

   $2\text{z} : Y = ZX,$

   where $X \sim \mathcal{N}(0,1)$, Z and X are independent with $\text{P}(Z=1) = \text{P}(Z=-1) = \frac{1}{2}$. Therefore, $\text{cov}(X,Y) = E(ZX^2) - E(Z)E(X)^2 = E(Z)E(X^2) - E(Z)E(X)^2 = E(Z)\text{var}(X) = 0.$

Based on data generated from the processes $*$i, $*$d, $*$h and $*$z, we investigate the dependence relationship between X and Y by performing hypothesis tests over the observations of X and Y or $U = F_X(X)$ and $V = F_Y(Y)$. The Monte Carlo simulations of the statistical tests and their performances are organized as follows. First, we introduce the correlation test in Section 6.1. In Section 6.1.1, we examine how the dependence test of the Pearson correlation coefficient between X and Y behaves without any modification to the test statistic. It is not surprising that the test results show deficiency in weak dependence cases where the sample correlations are too small to be considered as nonzero statistically. In Section 6.1.2, we illustrate the performance of the same bootstrap correlation test based on the newly generated random variables $g(U)$ and $h(V)$ which are the Box-Cox transformations of U and V. The test results indicate that the power of the test is greatly improved.

As is well known, the correlation test developed in Section 6.1 can only determine the correlation relationship, and not the independence relationship between two random variables. Therefore, one further test is necessary to examine whether two random variables are independent. As a supplement and comparison, we provide the simulation results of general independence tests

in Section 6.2. The tests are performed on the original X and Y and on the after Box-Cox transformations of X and Y. To show the validity and efficiency, four distinct test statistics and four different density or distribution estimation approaches are applied.

6.1 Correlation Test

The hypothesis of the correlation test shown in Figure 2.1 is:

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0, \end{aligned} \tag{6.1}$$

where H_0 is the null hypothesis and H_a is the alternative hypothesis. ρ is the population Pearson correlation coefficient between two random variables. In simulations, a sample correlation coefficient r is calculated to determine which claim H_0 or H_a should hold.

A Type I error occurs if the null hypothesis is rejected when it is true, while a Type II error occurs if the null hypothesis is not rejected when it is not true. The level of significance, α, which is defined as the maximum allowable probability of making a Type I error, represents the sensitivity of the test and is generally adopted as the cutoff value to decide to reject or accept H_0. A test statistic, for example, the sample correlation coefficient r, is computed from the sample data. The p-value is the probability of obtaining a value at least as extreme as this test statistic would be if observed under the null hypothesis. If the p-value is less than or equal to α, we reject H_0. Otherwise, we do not reject H_0. A small p-value, that is, a value less than or equal to α, means that it is highly unlikely that the observed test statistic would occur under the null hypothesis. We would thus conclude that the null hypothesis is incorrect, that is, we reject H_0.

As shown in the examples of Chapter 4, a single hypothesis test is conducted when only a single sample is available. However, under the conditions where there are two samples available, two tests of the same hypothesis will be conducted. It is possible that we will have two different test results. A natural question then arises: which significance test is more accurate? In practice, the problem of a possibly spurious single significance test is of great concern to researchers. Therefore, in many cases of statistical analysis, it is common and sometimes required to conduct multiple tests on the basis of multiple samples and interpret the results from multiple tests. In our study,

multiple samples can be produced from the data generating processes, and thus multiple tests can be considered. Based on the multiple tests, we define the rejection rate as in [17] and use it to present our simulation results. The rejection rate of test γ is the proportion of the null hypothesis H_0 being rejected in N repetitive simulations, i.e.,

$$\gamma = \frac{\sum_{i=1}^{N} \Upsilon_i}{N}, \tag{6.2}$$

where $\Upsilon_i = 1$ if H_0 is rejected, while $\Upsilon_i = 0$ if H_0 is not rejected. The empirical probability γ represents the frequency that the null hypothesis H_0 is rejected in multiple tests. For every single hypothesis test simulation run, a p-value less than or equal to α means $\Upsilon_i = 1$, while a p-value greater than α is equivalent to saying that $\Upsilon_i = 0$.

We now perform the correlation tests based on data samples generated from the processes $*$i, $*$d, $*$h and $*$z. The simulation procedure is as follows:

1. Obtain observation pairs $w_t = \{x_t, y_t\}$, $t = 1, \cdots, T$, from the data generating processes $*$i, $*$d, $*$h and $*$z for $T = 100$.

2. Compute the marginal density functions $\hat{f}_X(x_t)$ and $\hat{f}_Y(y_t)$ with a Nadaraya-Watson estimator and integrate $\hat{f}_X(x_t)$ and $\hat{f}_Y(y_t)$ to compute the cumulative distributions $\hat{F}_X$ and $\hat{F}_Y$. Finally, generate $u_t = \hat{F}_X(x_t)$ and $v_t = \hat{F}_Y(y_t)$.

3. Compute the test statistics. For the correlation test performed before the Box-Cox transformation defined in Section 6.1.1, we measure the Pearson correlation coefficient between x_t and y_t to check if X and Y are dependent. For the correlation test performed after the Box-Cox transformation defined in Section 6.1.2, we use the Pearson correlation coefficient between the Box-Cox transformations of u_t and v_t as the test statistic to decide if U and V are dependent, or equivalently, X and Y are dependent. The results in Section 6.1.1 and 6.1.2 are compared to see how the test performances in weak dependence cases.

4. Conduct the correlation tests. Given the level of significance $\alpha = 0.05$ and 0.1, we choose the running repetition times $N = 100$. For the bootstrap test, we set the bootstrap resample times $B = 200$ so that there are sufficient data samples to ensure the validity of the sampling distribution.

5. Interpret the test results. In Section 6.1.1 and 6.1.2, the rejection rate γ is resulted from the simulations. The larger γ is, the more likely X and Y are dependent. The smaller γ is, the more likely X and Y are independent. We check the consistency between the simulation results and the dependence relationship of the original data generating processes to determine if the test mechanism works correctly and efficiently.

As shown in the correlation computation of Appendix D, except 1d, the statistical correlation coefficients between x_t and y_{t-1} ($*$i, $*$d) or x_t and y_t ($*$h) are all equal to zero as long as the initial values x_0 and y_0 obey i.i.d standard normal distributions. Furthermore, data generated from non time series $*$z are also dependent and uncorrelated. Since the processes $*$i, $*$d and $*$h are asymptotically strictly stationary processes, we collect 100 samples between $t = 8901$ and $t = 9000$ to eliminate the influence of the initial values in simulations. For non time series $*$z, 100 samples are arbitrarily chosen.

6.1.1 Correlation Test Results before the Box-Cox Transformation

In this simulation, ρ in (6.1) denotes the population Pearson correlation coefficient between X and Y. Correspondingly, the sample correlation coefficient r between x_t and y_t, $t = 1, \cdots, T$, is computed as a test statistic to decide to reject H_0 or H_a, or equivalently, to say whether X and Y are correlated.

Table 6.1 presents the simulation results of the correlation tests on the observations of x_t and y_t, which are labeled as 'before Box-Cox transformation' as a contrast to the test in Section 6.1.2. The columns in Table 6.1 indicate the rejection rate of the Fisher's z transformation test and the bootstrap test for $\alpha = 0.05$ and $\alpha = 0.1$. As mentioned in Chapter 5, the bootstrap test requires no population distribution assumption and the Fisher's z test does not require strict normality of the population distribution, and therefore the simulation results in Table 6.1 are reliable.

As is well known, the two random variables generated from $*$z are uncorrelated. Moreover, given that the initial values $\{x_0, y_0\}$ are i.i.d $\mathcal{N}(0, I_2)$, the variables X and Y generated from $*$i, $*$h and 2d-6d are also uncorrelated. Therefore, the simple correlation test based on x_t and y_t, except for 1d, should support the claim $H_0 : \rho = 0$. The simulation results of rejection rate γ are shown in Table 6.1.

The rejection rate γ in (6.2) represents the frequency of H_0 being rejected in N iterative simulations. A large γ means that $H_0 : \rho = 0$ is not accepted in most simulation runs, while a small γ means that $H_0 : \rho = 0$ is accepted in most simulation runs. Consider $\alpha = 0.05$. In Table 6.1, the rejection rate γ is between 0.05 and 0.09 for $*$i, γ is between 0.02 and 0.08 for $*$h, and γ is between 0.03 and 0.05 for $*$z. These values, which are close to zero, indicate that the null hypothesis $H_0 : \rho = 0$ is accepted in more than 90% simulation runs. We therefore conclude that X and Y are uncorrelated with strong confidence. These results are consistent with the facts: either X and Y are independent ($*$i and 1h) or X and Y are nonlinearly related (2h-4h,$*$z), that is, $\rho = 0$. On the other hand, except for 1d, the test results show relatively low rejection rates, 0.12-0.30, for $*$d where X and Y are uncorrelated as well. For 2d-6d, the null hypothesis $H_0 : \rho = 0$ is rejected 12 to 30 times among 100 experiments. Although the evidence is not as strong as above in $*$i,$*$h, and $*$z cases, $\gamma \in [0.12, 0.30]$ will not lead us to conclude that X and Y are correlated. As for 1d, where X and Y are linearly related and a nonzero correlation exists. Since a large sample Pearson correlation coefficient r can be detected, the null hypothesis $H_0 : \rho = 0$ should be rejected in most simulation runs. Actually, $\gamma = 0.96$ or 1 means that, in more than 96% simulation runs, the correlation test supports the alternative hypothesis $H_a : \rho \neq 0$. Therefore, we say that X and Y are correlated with strong confidence.

Table 6.1 shows that the correlation test performs well in determining whether X and Y are correlated. However, our emphasis is to know whether X and Y are dependent. If X and Y are dependent and correlated, such as 1d, this simple correlation test works. When the claim $H_a : \rho \neq 0$ is tested to hold, we can safely say that X and Y are dependent. If X and Y are dependent but uncorrelated, such as in 2d-6d and 2h-4h, the above correlation test does no help to identify the dependence among variables. The test results turn out to determine that $\rho = 0$ in both cases: X and Y are independent or X and Y are dependent but uncorrelated. Therefore, the traditional correlation test is not adequate for discovering such dependence relationships.

It is natural for us to look for alternative ways to determine the dependence relationship between X and Y [4, 8, 11]. The conventional independence tests work efficiently in detecting the usual dependency among variables as mentioned in Chapter 1. However, as discussed in Section 6.2, many typical independence tests lack the ability to identify an extremely low dependence relationship, i.e., weak dependence or 'almost independence,' which appears more like 'absolute

independence.' In addition, we wish to know whether distinct dependence measurements, such as the Spearman correlation ρ_S, can make notable difference in distinguishing weak dependence and independence.

Dependence refers to any relationship between two measured random variables. The Pearson correlation coefficient particularly estimates the linear relationship. As shown in Section 5.1, alternative measurements, such as Spearman correlation and Kendall's tau, evaluate different types of associations. Table 6.2 presents the measurements of Pearson correlation ρ, Spearman's ρ_S, and Kendall's τ between x_t and y_t when 100 data samples are observed. Small values of $\hat{\rho}$, $\hat{\rho}_S$, and $\hat{\tau}$ indicate that there is no apparent linear or other types of relationships existing in the data generating processes 2d-6d,2h-4h and 1z-2z, that is, none of these coefficients is sufficient to measure any kind of dependence relationships that might exist in these processes. Compared to ρ, ρ_S and τ do not add extra information of dependence. Therefore, the selection of either one of ρ, ρ_S or τ as test statistic will not strongly influence a dependence test. Our improved hypothesis test shown in the next section will then be performed on the basis of the Pearson correlation coefficient ρ.

Furthermore, no significant difference is observed in the comparison of dependence measurements in $*$i, $*$h, $*$z and 2d-6d. This again substantiates the simulation results in Table 6.1. No matter what kind of dependence measurement is applied, the dependence test based on examining such value will be deficient in identifying weak dependence. A new test mechanism is imperative to distinguish the weak dependence and independence.

6.1.2 Correlation Test Results after the Box-Cox Transformation

As shown in Figure 2.4, we provide a correlation test based on the Box-Cox transformation of U and V to improve the test performance in the weak dependence cases, that is, ρ in (6.1) denotes the population Pearson correlation coefficient between the new random variables yielded from the Box-Cox transformation. Our test statistics are built on a basic fact that $f(U)$ and $g(V)$ are independent if U and V are independent, where $f(\cdot)$, $g(\cdot)$ are arbitrary functions of U and V, respectively. The contrapositive is certainly true. Therefore, if the test shows that $f(U)$ and $g(V)$ are correlated, then we conclude that the original U and V are correlated. In our test, $f(\cdot), g(\cdot)$ are specified as the Box-Cox transformation. We check the amplified after-transformation correlation

between the transformed samples u_t and v_t to decide if the random variables X and Y are correlated, and hence, are dependent.

Our test statistic is to measure the sample correlation coefficient after the Box-Cox transformation. We wish to enlarge the correlation difference between weak dependence and independence, therefore U and V are Box-Cox transformed under the assumption of improving the linearity. That is, the transformation parameter λ is chosen to maximize the linear relation between V and transformed U, or U and transformed V. The Box-Cox transformation is deployed in the following two different ways:

- Only U or V is transformed. In simulations, V is transformed and $\hat{r}(u_t, T_\lambda(v_t))$ is used as the test statistic. By choosing the optimal λ from the linearity plot, the Box-Cox transformed variable $T_\lambda(V)$, in the form of (5.1), fits best to the linear model $T_\lambda(V) = U\beta + \varepsilon$, where β is a constant and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

- Both U and V are transformed, but not simultaneously. U is Box-Cox transformed by letting $T_\lambda(U) = V\beta + \varepsilon$, while V is Box-Cox transformed by letting $T_\lambda(V) = U\beta + \varepsilon$. $\hat{r}_1(u_t, T_\lambda(v_t))$ and $\hat{r}_2(T_\lambda(u_t), v_t)$ are computed according to the corresponding linearity plots. We then choose $\hat{r} = \max(\hat{r}_1, \hat{r}_2)$ as the test statistic.

The Nadaraya-Watson marginal distribution estimator with a normal kernel is applied to obtain $u_t = \hat{F}_X(x_t)$ and $v_t = \hat{F}_X(v_t)$. The normal kernel is used because it is convenient and the shape of the kernel function does not strongly influence the estimation result. Let the simulation repetitive time be $N = 100$ and the sample size be $T = 100$. Consider the situation that V is Box-Cox transformed and U is invariant. We calculate $\hat{r}(u_t, T_\lambda(v_t))$, where $T_\lambda(v_t)$ is the Box-Cox transformation of V which best fits U linearly. By choosing an appropriate parameter λ, the Box-Cox transformation allows that the magnitude of the after-transformation sample correlation $\hat{r}$ between u_t and $T_\lambda(v_t)$ is larger than or equal to the magnitude of the original r between u_t and v_t in both dependence cases ($*$d, $*$z and 2h-4h) and independence cases ($*$i and 1h). Let Δr_{dep} and Δr_{indep} denote the positive sample correlation difference $|\hat{r}| - |r|$ in dependence cases and independence cases, respectively. If Δr_{dep} represents a considerable increase, then the corresponding $\hat{r}(u_t, T_\lambda(v_t))$ provides evidence that $\rho \neq 0$, or that U and V are dependent. On the other hand, a small increase in Δr_{indep} may allow us to view $\hat{r}(u_t, T_\lambda(v_t))$ as statistically zero. In this case, we can make the

inference that X and Y are uncorrelated. Such a small Δr_{indep} occurs because the transformed $T_\lambda(V)$ and U is theoretically independent no matter what λ is chosen, as long as the underlying X and Y are independent. As a result, dependence, especially weak dependence, can be differentiated from independence, while independence will not be mistakenly tested as dependence. In one simulation, for 1i and 4d, the sample correlation increases are recorded as $\Delta r_{1i} = 0.0683$ and $\Delta r_{4d} = 0.1653$. Obviously, $\hat{r}(u_t, T_\lambda(v_t))$ of 4d with a 0.1653 increase is more likely to be the evidence to support the claim $H_a : \rho \neq 0$, and hence to say X and Y are dependent.

Since the bootstrap method does not require any specific assumptions, it is the only test deployed in the simulations. Consider the case that V is Box-Cox transformed and U is invariant. First, find the optimal λ^* using the linearity plot of the Box-Cox transformation. Then, get the new data set $T^*_\lambda(v_t)$. Finally, based on the observation pairs $\{u_t, T^*_\lambda(v_t)\}_{t=1}^T$, compute the bootstrap p-value in the way shown in Section 5.2.

The rejection rates for $\alpha = 0.05$ and $\alpha = 0.1$, are listed in Table 6.3. The first column labeled as "1-var Box-Cox" shows the rejection rate γ when V is Box-Cox transformed and U is invariant. $\hat{r}(u_t, T_\lambda(v_t))$ is calculated for $N = 100$ times to obtain the empirical frequency γ. The second column labeled as "2-var Box-Cox" represents the maximal rejection rate when both U and V are Box-Cox transformed, i.e., the test statistic is $\hat{r} = \max(\hat{r}_1(u_t, T_\lambda(v_t)), \hat{r}_2(T_\lambda(u_t), v_t))$.

The rejection rate γ in (6.2) represents the frequency of H_0 being rejected in N iterative simulations. A large γ means that $H_0 : \rho = 0$ is not accepted in most simulation runs, thus $H_a : \rho \neq 0$ is accepted in most simulation runs, which tends to say that X and Y are dependent. If the multiple tests show that H_0 is rejected 100 times among 100 experiments, or $\gamma = 1$, then we can make a definite statement that $\rho \neq 0$, or equivalently X and Y are dependent. On the other hand, a small γ means that $H_0 : \rho = 0$ is accepted in most simulation runs, in which case we conclude that X and Y are uncorrelated with strong confidence. If γ is computed as 0, then we are definitely sure that $\rho = 0$, or equivalently X and Y are uncorrelated.

We notice that the performance of our test decreases in $*$i. Consider $\alpha = 0.05$. Before the transformation, the rejection rate γ is between 0.05 to 0.09 as shown in Table 6.2. But Table 6.3 shows that the reject rate after the Box-Cox transformation ranges from 0.12 to 0.17. This means that 12 to 17 out of 100 simulation runs support that $H_0 : \rho = 0$ is not true when it is actually true. It is natural because the possible maximal sample correlation between two finite data samples by

doing Box-cox transformation are detected, and therefore more simulations justify a statistically nonzero correlation. However, in these cases, the resulting values $\gamma \in [0.12, 0.17]$ will definitely not lead us to an opposite conclusion that U and V are dependent and hence X and Y are dependent.

On the other hand, the rejection rate of $*$d provides strong evidence that X and Y are dependent. In Table 6.3, the rejection rate γ of $*$d after the Box-Cox transformation ranges from 0.55 to 1.00. In particular, when both U and V are transformed (the second column), γ is between 0.72 and 1. Therefore, we conclude that $H_a : \rho \neq 0$ holds, i.e., X and Y are correlated, and hence dependent, with strong confidence.

Consider the processes 1z and 2z. The rejection rate γ before the Box-Cox transformation shown in Table 6.2 is between 0.03 and 0.05 and indicates that X and Y are uncorrelated. The rejection rate γ after the Box-Cox transformation shown in Table 6.3 is between 0.94 and 1.00 and indicates that in more than 94% simulation runs, the transformations of X and Y are correlated. Consequently, we say X and Y are dependent.

We now look at the simulation result of processes $*$h in Table 6.3. Consider $\alpha = 0.05$. For 1h, γ is 0.18 or 0.22. Although the performance is weaker than $\gamma = 0.02$ in Table 6.2, it definitely will not lead us to conclude that X and Y are dependent, similarly as in $*$i. However, for 2h-4h, γ in Table 6.3 only ranges from 0.30 to 0.46. Clearly, this test result is not sufficient to derive a conclusive inference as to whether X and Y are dependent. In 2h-4h, y_t and z_t are linearly related, and y_t is related to x_t with the standard normal density function φ. Because $\varphi(x_t/\alpha) \to 0$ when $|x_t/\alpha| > 3$, the variation of y_t is mainly caused by z_t instead of x_t. Therefore, it is not likely that the dependence relationship of y_t is greatly improved in the direction of x_t instead of in the direction of z_t. The significance level α represents the tolerance level of committing a Type I error. The smaller the α, the more significant the result is said to be. In other words, the data must diverge more from the null hypothesis to fall in the rejection region to be significant. In this sense, the 0.05 level is always considered to be more conservative than the 0.10 level. Therefore, in Table 6.3, we should see a better result when $\alpha = 0.10$: the rejection rate γ is between 0.51 to 0.67. These values allow us to conclude with more confidence that X and Y are dependent.

In Section 6.1.1 and 6.1.2, the correlation tests before and after the Box-Cox transformation are implemented. The correlation test over the original variables X and Y in Section 6.1.1 shows low rejection rates in $*$i,$*$h and $*$z, relatively low rejection rates in 2d-6d, and high rejection rate

in 1d. This leads to a conclusion that all the variables obtained from the data generating processes, except 1d, are uncorrelated. However, the result of $\rho = 0$ gives us no clue whether the variables are dependent, i.e., all the dependent cases, except 1d, cannot be successfully identified. In the correlation test after the Box-Cox transformation of Section 6.1.2, the rejection rates γ in the dependence cases, $*$d, 2h-4h and $*$z, are greatly improved. The larger γ, the more likely $\rho \neq 0$ between the transformed variables of X and Y. As a result, $\gamma \to 1$ constructs a strong evidence for us to determine that X and Y are dependent.

6.1.3 Power Analysis

Our study at this point focuses on is how to identify the weak dependent variables. Comparing with the independence cases $*$i and 1h, we are more concerned about the dependence cases $*$d, 2h-4h and $*$z. Considering the following general hypothetical claims of independence test, i.e.,

$$\begin{aligned} H_0' &: X \perp\!\!\!\perp Y \\ H_a' &: X \not\perp\!\!\!\perp Y, \end{aligned} \tag{6.3}$$

we say H_a' holds for the dependence cases $*$d, 2h-4h and $*$z. Power analysis is then carried out for these cases.

The power of a statistical test is the probability that the test will reject a false null hypothesis or the probability not to make a Type II error, that is, [77]

$$\text{Power} = \Pr(\text{ Reject } H_0 | H_a \text{ is true}) = 1 - \Pr(\text{Make Type II error}).$$

As power increases, the chance of a Type II error decreases, and vice versa. The analytical approach to calculate power requires the knowledge of the underlying distributions of the observations, which is usually intractable in practice. Another way is to get a simulated power via Monte Carlo simulation [78]. Given that H_a is true, the simulated power of the test is the frequency that H_0 is rejected in multiple experiments, that is, the proportion of bootstrapped p-value which are less than or equal to the significance level α [79]. For the dependent processes $*$d, 2h-4h and $*$z, let us consider the simulations of correlation test after the Box-Cox transformation of Section 6.1.2.

The rejection rate γ represents the frequency of $H_0 : \rho = 0$ being rejected in N iterative simulations, i.e. only in the remaining $1-\gamma$ times, the claim $H_0 : \rho = 0$ is accepted. However, there is no way to know how many times among these $1-\gamma$ experiments the claim $H_0' : X \perp\!\!\!\perp Y$ is accepted. As a result, the simulated power of the test shown in (6.3) is at least equal to the rejection rate γ obtained in Section 6.1.2.

Figure 6.1 shows the simulated power of hypothesis test in (6.3) by using the after Box-Cox transformation correlation coefficient as the test statistic in Section 6.1.2. Although there are no formal standards for power, statistical studies aim for powers of 0.80 or 0.90 and typically assess 0.80 as a standard for adequacy [80]. When two variables are both Box-Cox transformed, the power of the test are basically above 0.80 except in 2h-4h cases. Our test does not behave perfectly, but it does show efficiency in identifying the weak dependence. More details, such as the comparison between the power of the tests before and after the Box-Cox transformation, will be discussed in next section.

6.2 Independence Test

From Figure 2.1-2.5, the correlation test can only tell us if X and Y are correlated, and hence dependent. There is no definite answer to the question whether X and Y are independent. As shown in Section 6.1.1, if X and Y are dependent but uncorrelated, such as 2d-6d, 2h-4h and $*$z, the correlation test on the original variable X and Y can only lead to the conclusion that X and Y are uncorrelated and no inference whether X and Y are dependent can be made. As a result, a modified correlation test based on the transformed variables of X and Y is implemented in Section 6.1.2 such that the dependence relationship in the variables generated from the processes 2d-6d, 2h-4h and $*$z can be found. However, one may say that the correlation test before the Box-Cox transformation in Section 6.1.1 does not work because it is not an appropriate hypothesis test choice for detecting the dependent but uncorrelated relationship in the first place, and other tests, for example, the one measures the closeness between the joint density $f_{XY}(x,y)$ and the product of marginals $f_X(x)f_Y(y)$, could work. The general independence tests can efficiently discover whether the variables are independent as mentioned in Chapter 1. But we wish to know if these tests also show good performance in the processes 2d-6d, 2h-4h and $*$z, some of which might be weak

dependence cases. In this section, as a supplement and comparison, we will show performances of some conventional independence tests via Monta-Carlo simulations.

There exists a variety of independence tests in the literature whose performances vary. We investigate the difference among the tests by the following ways:

- The measures of dependence, or the test statistic, are often considered as an intermediate step to implement the test. In practice, the test statistic is carefully chosen to be adaptable and effective in a specific real scenario. For example, the statistics of entropy type are more preferred in communication applications [8, 12, 81]. The correlation integral is a common test statistic in a wide range of econometrical applications including ARCH, GARCH models [26,82,83]. Independence between stochastic variables is defined in terms of distribution functions, or, if they exist, in terms of density functions. According to this classification, there are two kinds of dependence measures: the measures based on the distribution functions and the measures based on the density functions. Since the density functions of the data generating processes $*$i,$*$d,$*$h and $*$z exist, in our study, four test statistics are chosen:
  - Cramer-Von Mises distance(CM) and Kolmogorov-Smirnov distance(KS) based on the distribution functions,
  - Chan and Tran's distance functional(CT) and Rosebblatt, Rosenblatt and Wahlen's distance functional(RW) based on the density functions.
- There are two kinds of distribution estimation methods: parametric and nonparametric. As discussed in Chapter 3, parametric estimation requires a known assumption of underlying distribution structure, while nonparametric estimation requires a large amount of data. The former requirement cannot always be satisfied in real applications, while the latter asks for complicated computation and therefore has the problem of 'dimensionality curse.' In our study, four distribution methods are used:
  - nonparametric methods: the empirical distribution estimation and the double kernel local linear method,
  - semi-parametric methods: the normal copula and the Archimedean copula.

The hypotheses of a general independence test are claimed as:

$$\begin{aligned} H_0 &: s = 0 \\ H_a &: s \neq 0, \end{aligned} \tag{6.4}$$

where s is the chosen test statistic which possesses the property that $s = 0$ if and only if X and Y are independent. Four types of test statistics s: CM, KS, CT and RW will be introduced.

Conventional distance measures between two distribution functions F_X and F_Y are the Cramer-Von Mises distance [2–5, 84]

$$\text{CM} = \int (F_{XY} - F_X F_Y)^2 \, dF_{XY}, \tag{6.5}$$

and the Kolmogorov-Smirnov distance [6, 85]

$$\text{KS} = \sup |F_{XY} - F_X F_Y|. \tag{6.6}$$

Assuming that the density functions of all concerned variables exist,, the dependence measures between two density functions f_X and f_Y can be considered. A well known distance measure is of entropy type. For example, both the Kullback-Leibler distance and the Hellinger distance are special cases of generalized Tsallis entropy measure. The entropy measure is a good candidate test statistic in many independence tests, but it is not appropriate in our study due to the data processing inequality [86]. This inequality asserts that if the random variables $X \to Y \to Z$ form a Markov chain in this order, that is, X and Z are conditionally independent given Y, then the mutual information between them satisfy $I(X;Z) \leq I(X;Y)$. In other words, no processing of Y, deterministic or random, can increase the information that Y contains about X. In particular, if Z is given by a deterministic function g of Y, we have $I(X;g(Y)) \leq I(X;Y)$, i.e., the functions of Y can not increase the information about X. Note that the mutual information discussed here represents both the Shannon entropy and the Tsallis entropy [86, 87]. The Box-Cox transformation, as a result, can not increase the dependence information in the sense of the Shannon entropy and the Tsallis entropy. For this reason, neither the Kullback-Leibler distance nor the Hellinger distance is applied in our simulation where the purpose of executing the Box-Cox transformation is to increase

a certain type of dependence to some extent. Instead, we use the alternative distance functionals CT and RW, which are also based on the density functions. Chan and Tran [88] propose a distance measure

$$\mathrm{CT} = \int |f_{XY}(x,y) - f_X(x) f_Y(y)| \, dxdy, \tag{6.7}$$

and Rosenblatt and Wahlen [5, 84] provide a squared distance functional

$$\mathrm{RW} = \int (f_{XY}(x,y) - f_X(x) f_Y(y))^2 \, dxdy. \tag{6.8}$$

We now turn our attention to the estimation of distribution or density functions. Four nonparametric or semi-parametric methods are used to present the possible performance difference caused by different distribution estimation methods.

1. Empirical distribution estimation

   Recall that the k-dimensional measure on $\mathbf{R}^k$ is defined by

   $$m_k(A) = \int_A 1 \, d\mathbf{z} \text{ for } A \subseteq \mathbf{R}^k.$$

   In particular, m_1 is the length measure on $\mathbf{R}$, m_2 is the area measure on $\mathbf{R}^2$, and m_3 is the volume measure on $\mathbf{R}^3$.

   Suppose $\mathbf{Z}$ is a random vector with a continuous distribution on a subset S of $\mathbf{R}^k$ and $\{\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n\}$ denotes n random samples.

   Let $\{A_j : j \in J\}$ be a partition of S which divides S into a countable number of subsets. Then, the empirical probability of A_j based on n samples is

   $$P_n(A_j) = \#(\mathbf{Z}_i \in A_j, i \in \{1, 2, ..., n\})/n, \tag{6.9}$$

   where '$\#(a)$' denotes the number of samples which satisfy the condition a. The corresponding empirical density function is defined as follows:

   $$f_n(\mathbf{z}) = P_n(A_j)/m_k(A_j) \text{ for } \mathbf{z} \in A_j. \tag{6.10}$$

Replacing $\mathbf{Z}$ by the random variables X, Y, and $\{X, Y\}$, we can obtain marginal and joint empirical distribution and density functions. Let $w_i = \{x_i, y_i\}_{i=1}^T$ be the observations of $W = \{X, Y\}$. Adapted from Linton and Gozalo [85], the functional A_T is defined as:

$$A_T(w_i) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(w_t \le w_i) - \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(x_t \le x_i)\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(y_t \le y_i),$$

where $\mathbf{1}(a)$ is the indicator function, that is, $\mathbf{1}(a) = 1$ if the condition a is satisfied and $\mathbf{1}(a) = 0$ otherwise. The test statistics CM and KS based on the distribution functions are:

$$\text{CM} = \frac{1}{T}\sum_{t=1}^{T} A_T(w_i)^2, \ \text{KS} = \max_{1\le i\le T} |A_T(w_i)|.$$

From (6.9) and (6.10), the test statistics CT and RW based on the density functions are:

$$\text{CT} = \sum_{i=1}^{J} |f_T(x_i, y_i) - f_T(x_i) f_T(y_i)| m_1(A_{j_1}) m_1(A_{j_2}),$$

and

$$\text{RW} = \sum_{i=1}^{J} (f_T(x_i, y_i) - f_T(x_i) f_T(y_i))^2 m_1(A_{j_1}) m_1(A_{j_2}),$$

where A_{j_1} and A_{j_2} are partitions of the domain of X and Y, respectively. $\{x_i, y_i\}_{i=1}^J$ are the grid points generated from the partitions.

2. Double kernel local linear method

   As discussed in Section 3.2, this method is used to estimate the conditional density by solving a nonparametric regression problem with appropriate bandwidth selection. We first estimate the conditional density $f_{Y|X}(y|x)$ by this method and estimate the marginal density $f_X(x)$ by the NW density estimation. We then compute the joint density and distribution functions by $f_{XY}(x, y) = f_{Y|X}(y|x) f_X(x)$ and the integrals. Based on the estimations, the test statistics CM, KS, CT and RW are calculated from (6.5) to (6.8).

   In simulations, the normal kernel is chosen in both double kernel local linear method and the NW density estimation to get the conditional density and the marginal density. However, the usage of the kernel does not influence the fact that both the double kernel local linear

method and the NW density estimation method are nonparametric, since no assumption of underlying densities is made.

3. Normal copula estimation

   The normal copula is applied to obtain the joint density $f_{XY}(x,y)$. The corresponding copula parameter, which measures the dependence between X and Y, is the Pearson correlation coefficient. In the mean time, the marginal densities $f_X(x)$ and $f_Y(y)$ are estimated by the NW density estimation.

4. Archimedean copula estimation

   The Archimedean copula, more specifically, the Frank copula, is applied to obtain the joint density $f_{XY}(x,y)$. The Kendall's τ is estimated to compute the copula parameter α shown in Table 3.2. As before, the marginal densities $f_X(x)$ and $f_Y(y)$ are estimated by the NW density estimation.

   The semi-parametric estimation combined by parametric copula and nonparametric NW density estimation avoids the 'dimensionality curse', and it also requires less assumption of underlying distributions.

In addition, the bootstrap method is applied in this test to compute the p-value due to the infeasibility of the central limit theorem. To merge the null hypothesis in the resampling scheme and draw the resamples, according to [89], the bootstrap procedure is:

1. Calculate the test statistic s_T for the original observations $\{x_t, y_t\}, t = 1, \cdots, T$.

2. Since the null hypothesis is that X and Y are independent, the resamples are generated by applying a random shuffle to the data sample $\{x_t, y_t\}$, that is, $\{x_t^*, y_t^*\}$ is the randomly reordering of observations. It is apparent that the marginal distributions of x_t^* and y_t^* are identical to that of the original data x_t and y_t respectively, while the generated joint distribution will not be affected if X and Y are actually independent.

3. Resample $\{x_t, y_t\}$ in the same way as shown in step 2.

Choose the resample times $B = 200$ and calculate s_T^* for resamples $\{x_t^*, y_t^*\}$, $t = 1, \cdots, T$, then the corresponding bootstrap p-value is computed by the empirical distribution of this resam-

pled statistic, i.e., $\hat{p}_T = \Pr(s_T^* > s_T)$. If $\hat{p}_T < \alpha$, where α is the significance level, we reject H_0. Otherwise, we do not reject H_0.

In the following sections, based on the data generated from $*$i, $*$d, $*$h and $*$z, we will carry out the hypothesis test of (6.4) multiple times to get the rejection rate defined in (6.2) as a test result. Our goal is to investigate whether the general independence tests using four types of test statistics and four types of distribution estimation methods are efficient in detecting dependence, in particular, the weak dependence situations.

6.2.1 Independence Test Result before the Box-Cox Transformation

This section presents the general independence test results before the Box-Cox transformation, that is, s in (6.4) measures the dependence relationship between the original random variables X and Y. The rejection rate γ shown in Tables 6.4-6.7 are obtained by using four distribution estimation methods: empirical distribution estimation, double kernel local linear estimation, the normal copula, and the Archimedean copula, respectively.

For the process $*$i and 1h, all four distribution estimation methods show good performances. Considering $\alpha = 0.05$, the rejection rate γ is always below 0.10. The null hypothesis H_0 implying that X and Y are independent is strongly considered to be true. This result accurately reflects the underlying independence relationship between X and Y generated from $*$i and 1h. As for the other high-frequency cases 2h, 3h and 4h, Tables 6.4-6.7 also show the comparable small γ values as that in $*$i. Therefore, we conclude with strong confidence that X and Y are independent in the cases 2h-4h. However, this conclusion does not correctly reflect the actual dependence relationship of 2h-4h. The independence test, which is designed to include as much as possible diversity by using four types of statistics and four types of distribution estimation methods, is invalid and inefficient in identifying the independent relationship existing in the cases 2h-4h.

Observe the process $*$z. For 1z and 2z, the empirical distribution method and double kernel local linear estimation show that the rejection rate γ is close to 1, while both the normal copula and the Archimedean copula method show relatively small rejection rates. Actually, a smaller rejection rate is expected if the sample size is larger. Thus, the independence test results diverge in this case.

Consider the process $*$d. All four distribution methods show the rejection rate $\gamma \approx 1$ in the process 1d. It is obviously true since there exists a linear relationship between x_t and y_{t-1}.

Furthermore, all four methods also show the relatively small rejection rate, which is below 0.31, in 3d, 4d, 5d and 6d. For the case 2d, both the normal copula and the Archimedean copula tests yield relatively low rejection rates, that is, under 0.30. However, both the empirical distribution method and double kernel local linear estimation show large rejection rate. Once again, there exists some disagreement among the simulation results in this particular case of 2d.

We may summarize and explain the test results of Tables 6.4-6.7 as follows:

- If the random variables X and Y are independent ($*$i and 1h), all the independence tests conducted in our simulation perform well.

- The inconsistency of results occurs when X and Y are dependent but with zero correlation, that is, 2d-6d,2h-4h and 1z-2z.

  – A copula measures the dependence between random variables and brings this dependence into the construction of the joint distribution by:

$$f(x_1, x_2, \cdots x_n) = c[F_1(x_1), F_2(x_2), \cdots F_n(x_n)] \prod_{i=1}^{n} f_i(x_i),$$

  where $f_i(x_i)$ and $F_i(x_i)$ are the density and distribution functions of a random variable X_i. The copula density c links univariate marginals to their full multivariate distribution and only c encodes information about the dependence among X_i. Therefore, it is not surprising that the corresponding independence tests using copulas, whether the normal copula or the Archimedean copula, show low rejection rates, i.e., tend to say X and Y are independent in the processes 2d-6d, 2h-4h and $*$z. In these situations, low dependence, which are represented in terms of the Pearson correlation $\hat{\rho}$ or the Kendall's $\hat{\tau}$ as shown in Table 6.2, are measured. The form of the copula also explains why the test result in 1d clearly shows that X and Y are dependent. In this case, x_t and y_{t-1} are linearly related and therefore the corresponding c is far from unity.

  – The empirical distribution estimation test and the double kernel local linear estimation test perform differently. Their simulation results agree with the results of copula tests in 3d-6d and 2h-4h. This consistency is understandable since these situations are classified as 'weak dependence' or 'almost independence' as discussed in Section 6.3. That

is, X and Y are very weakly related and therefore the general independence tests fails in detecting this weak dependence.

Both the empirical distribution estimation test and the double kernel local linear estimation test show $\gamma \approx 1$ in 2d. Consider the 2d process, where $x_t = 0.5x_{t-1} + 0.5y_{t-1}^2 + \varepsilon_{1,t}$. Although the correlation between x_t and y_{t-1} is 0, there exists an apparent dependence relationship between x_t and y_{t-1} and this dependence is directly embodied in the estimation of $F_{XY}(x,y)$ and $f_{Y|X}(y|x)$.

6.2.2 Independence Test Result after the Box-Cox Transformation

In this section, the Box-Cox transformation is applied to increase the correlation between X and Y so that the subtle dependence relationship can be detected. The initial condition is invariant, that is, both x_0 and y_0 are i.i.d. normal distributed with zero mean and unit variance. The samples are collected between $t = 8901$ and $t = 9000$.

Tables 6.8-6.11 show the rejection rate γ of independence test after the Box-Cox transformation, where distinct test statistics and distinct distribution methods are used, and the test results are briefly displayed as follows.

$*$**i** The rejection rate is under 0.40, no matter which test statistic or distribution method is applied. This result is consistent with the fact that X and Y are independent in this case.

$*$**d** The Box-Cox transformation allows us to obtain an enlarged Pearson correlation coefficient $\hat{r}_{max}$, which is larger than the original correlation coefficient $\hat{r}$ between X and Y. The enhanced linear relationship implied by $\hat{r}_{max} > \hat{r}$ is validated through the larger values of rejection rate for the four distribution methods in Tables 6.8-6.11. However, the improvement of $\hat{r}_{max}$ might not be large enough to say that there is a strong dependence between the after-transformation variables. Therefore, no solid evidence is provided in Table 6.8,6.9 and 6.11 to show that the dependency exists in 3d-6d. Only the methods of the normal copula, which use the Pearson correlation coefficient as the measure of dependence in calculating the joint distribution, achieve high rejection rates, which clearly reveal the fact that data generated from $*$d are dependent.

∗**h** Each method performs well in 1h. We also see similar improvement of rejection rates in 2h-4h. By the normal copula, we may say that variables generated from 2h-4h are dependent, but the evidence is not as strong as in ∗d.

∗**z** As for the first three distribution estimation methods, the rejection rate is close to 1. The nonlinear dependence between X and Y is transformed, to a large extent, to linear dependence through the Box-Cox transformation and thus is easily detected. However, the strong linear relationship does not necessarily imply the strong nonlinear correlation such as Kendall's τ, and therefore no big increase of rejection rates are shown in Table 6.11 of the Archimedean copula.

6.2.3 Power Analysis

Regardless of distinct distribution estimation methods and distinct test statistics, the values of the rejection rate after the Box-Cox transformation shown in Tables 6.8-6.11 become larger. As for dependence cases ∗d, 2h-4h and ∗z, these results provide evidence that X and Y are dependent. However, only γ obtained from the normal copula method by directly using the Pearson correlation coefficient to compute the joint distribution provides conclusive evidence that X and Y are dependent.

As is well known, power $=$ Pr(Reject $H_0|H_a$ is true) $= 1 -$ Pr(Make Type II error). It is one of the most important criteria to evaluate the efficiency of a statistical test. For the hypothesis test of (6.4), its power is the probability to conclude that $H_0 : s = 0$ when $H_a : s \neq 0$ is actually true. To put it another way, the power of the test is the probability that X and Y are tested as dependent when they are actually dependent. In the simulations of dependence cases ∗d, 2h-4h and ∗z, the frequency of rejecting H_0, i.e., the power, is represented by the rejection rate γ.

Figures 6.2-6.5 show the power comparison of the independence tests before and after the Box-Cox transformation based on the test results using the normal copula estimation method in Section 6.2.1 and 6.2.2. Four types of test statistics: CM, KS, CT, and RW are shown in Figures 6.2-6.5, respectively. Except for 1d, the power of the conventional independence tests (solid lines) is low, under 0.37. Although there are no formal standards for power, we expect it to be greater than 0.80 in practice. Power less than 0.37 is far from 0.80, thus the tests without the Box-Cox

transformation behave poorly. Consider the power of the independence tests after the Box-Cox transformation (dotted lines). When $\alpha = 0.10$, the values of the power using CM, KS, CT and RW, all exceed the aimed 0.80 in the cases of $*$d and $*$z. However, the performance in 2h-4h is weaker. The power is above 0.52 but below 0.80. In the sense of the power, our independence test using the Box-Cox transformation shows great improvement compared to the traditional independence tests, although it shows some deficiency in the cases of 2h-4h. The power of a hypothesis test is its ability to distinguish between the false null hypothesis and the true alternative. Generally, the value of the power relates to the sample size and the "true difference" between H_0 and H_a. The larger the sample size, the larger the power. The larger the difference between H_0 and H_a, the larger the power. As to our independence test after the Box-Cox transformation, an increase of the power is expected as the sample size increases. However, the difference between H_0 and H_a is the problem we wish to solve. Weak dependence means there is a small difference between H_0 and H_a. In such a case, a high power is more difficult to achieve, and therefore a more intricate test method should be designed.

In order to show the test result disagreements of Section 6.2.1, the power of the test using the empirical distribution method before the Box-Cox transformation is illustrated in Figures 6.2-6.5 via the star points. The higher values of power observed in 2d and $*$z show certain efficiency of the empirical method, but the independence test after the Box-Cox transformation (dotted lines) obviously performs better in these cases. Note that the independence test (using the normal copula) and the correlation test after the Box-Cox transformation work in the same way, that is, the difference between H_0 and H_a is enhanced by the enlarged linear correlation through the Box-Cox transformation. Therefore, both tests can be considered as the correlation-based test. This explains the resemblance between the powers in Figures 6.2-6.5 and in Figure 6.1. Furthermore, as a whole, we observe higher power in the former plots because the real power should at least equal to that shown in Figure 6.1.

6.2.4 Independence Test using Different Box-Cox Transformation

Up to this point, the Box-Cox transformation we have discussed takes the following form:

$$T(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0, \end{cases} \tag{6.11}$$

where y can only take nonnegative values. This is the original Box-Cox transformation [18]. To satisfy the nonnegative condition, we compute the cumulative distribution function $F_Y(y)$ of y and use $F_Y(y)$ as the input of the Box-Cox transformation. However, since $F_Y(y)$ is between 0 and 1, the correlation enlargement might be compromised by the limitation of the range of $F_Y(y)$. Therefore, in the simulation, we also use other two modified Box-Cox transformations to increase the linear correlation. These transformations takes negative y values, thus no cumulative distribution transformation is necessary and the range problem is avoided.

The first modification is the exponential transformation proposed by Manly [90]:

$$T(\lambda) = \begin{cases} \frac{e^{\lambda y} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ y & \text{if } \lambda = 0. \end{cases} \tag{6.12}$$

The second modified Box-Cox transformation is suggested by Yeo and Johnson [91],

$$T(\lambda) = \begin{cases} \frac{(y+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ \frac{(1-y)^{2-\lambda} - 1}{\lambda - 2} & \text{if } \lambda \neq 2, y < 0 \\ -\log(1-y) & \text{if } \lambda = 2, y < 0. \end{cases} \tag{6.13}$$

In Figure 6.6-6.8, the test results of rejection rate γ using the original Box-Cox transformation (BC), the first (BC1) and the second modifications (BC2) of Box-Cox transformation in (6.11)-(6.13) are shown. The symbols of plus, circle, star and square represent the four types of test statistics, i.e., CM, KS, CT and RW, respectively. Furthermore, to simplify the simulation, we only apply the normal distribution estimation method, i.e., the joint distribution $f_{XY}(x,y)$ is computed by the normal copula, and specify the significance level $\alpha = 0.10$. In $*$z, γ is always 1 after the Box-Cox transformations, and therefore no plot is necessary. As shown in the $*$i,$*$d and $*$h

cases of Figures 6.6-6.8, small changes of rejection rates in terms of the different Box-Cox transformations are observed. For example, in the process $*$d of Figure 6.7, γ is between 0.65 and 1. No matter which Box-Cox transformation is used, high γ ($\gamma \geq 0.8$) occurs in 1d, 2d and 4d, while the relatively lower γ, i.e., γ occasionally drops to below 0.8, occurs in 3d, 5d and 6d. Moreover, the independence tests using the test statistic CM, which is labeled by '+', show generally weaker results in $*$d regardless of the type of Box-Cox transformation. However, we also notice that there exist slight improvements of rejection rate in the second and third columns labeled by "BC1" and "BC2", where the modified Box-cox transformations are applied. It substantiates the idea that the removal of the range limitation of variables may increase the Pearson correlation. In short, the independence tests using distinct Box-Cox transformations yield consistent results, i.e., $*$i and 1h are independent, and $*$d, $*$z and 2h-4h are dependent.

6.3 Simulation Summary

Both the correlation test and the independence test are performed in Section 6.1 and 6.2 to determine the dependence relationship between variables X and Y generated from the data generating processes $*$i, $*$d, $*$h and $*$z. Now we summarize the simulation results from last two sections to discuss the pros and cons of the correlation test and the independence test, and to tell why the Box-Cox transformation is useful in detecting weak dependence and which cases can be classified as weak dependence.

As discussed in Chapter 2, let $c = \sup_{f,g} |\text{cov}(f(X), g(Y))|$, where f and g are arbitrary real-valued functions of X and Y respectively. Define

$$r_{sup} = \frac{c}{\sigma(f(X))\sigma(g(Y))},$$

where $\sigma(\cdot)$ denotes the standard deviation. Then, we say:

- $r_{sup} < 0.30$ small/weak dependence;
- $0.30 \leq r_{sup} < 0.50$ medium/moderate dependence;
- $r_{sup} \geq 0.50$ large/strong dependence.

Since the closed-form mathematical expression of r_{sup} for the random variables generated from the data generating processes $*$i, $*$d, $*$h and $*$z is unknown, we observe $\hat{r}_{max}$ after the Box-Cox transformation and use it as the criterion for classifying dependence levels. Table 6.12 shows the sample correlation coefficient $\hat{r}$ before and after the Box-Cox transformation. Note that $\hat{r}$ after the Box-Cox transformation represents the possible maximal correlation coefficient, which is labeled as $\hat{r}_{max}$.

For the case 1d, X and Y are linearly related and therefore r should not be greatly increased after the Box-Cox transformation. In Table 6.12, both $\hat{r}_{org}$ and $\hat{r}_{max}$ are large, above 0.5, and there exists only a slight difference between them. As for 1z and 2z, there is an obvious dependence relationship between X and Y in spite of the zero linear correlation. Therefore, we can anticipate a significant increase in r after the Box-Cox transformation. Table 6.12 shows, in 1z and 2z, that $\hat{r}_{org}$ is close to zero, and $\hat{r}_{max}$ is much larger (above 0.5) so that it can not be statistically considered as zero. Another case which produces a relatively large value of $\hat{r}_{max}$, close to 0.4, is 2d. Based on the criterion of $\hat{r}_{max}$, we say that 1d and $*$z are strong dependence cases, while 2d is a moderate dependence case.

Consider an observation of the processes of 3d-6d and 2h-4h. Similarly to 2d and $*$z, these processes are dependent but uncorrelated. Consequently, the $\hat{r}_{org}$ values are also small. The results shown in Table 6.12 indicate a certain level of increase for $\hat{r}_{max}$, but all the magnitudes of $\hat{r}_{max}$ are less than 0.3. According to the classification in terms of $\hat{r}_{max}$, we say that 3d-6d and 2h-4h are weak dependence cases.

In addition, we know that r_{sup} should be zero for the independence cases of $*$i and 1h. The $\hat{r}_{max}$ in Table 6.12 are not zero only because the computation accuracy is limited by the sample size and estimation methods.

By acknowledging that 1d and $*$z are strong dependence cases, 2d is a moderate dependence case, and 3d-6d and 2h-4h are weak dependence cases, we now make a brief summary and explanation of the simulation results in Section 6.1 and 6.2. To demonstrate the problem graphically and intuitively, we replot Figure 2.1 and Figure 2.4. Figure 6.9 presents the simulation results before the Box-Cox transformation. Both the correlation test and the independence test are deployed on two variables X and Y. The correlation test reveals that only 1d is tested as correlated and the rest of the data generating processes are tested as uncorrelated. This outcome agrees

with our expectation, since all the cases except 1d possess zero correlation coefficients no matter whether the corresponding random variables are independent or dependent. The conventional independence tests using the test statistics CM, KS, CT and RW in Section 6.2.1, however, tell us different stories based on different estimation methods. Tables 6.4-6.7 show subtle changes of the rejection rate with respect to different test statistics. The rejection rates are comparable among various test statistics when the distribution estimation method is fixed. The factor that really matters is the estimate method of the distribution or density functions. The independence tests using four types of distribution estimation methods perform well in both independence cases $*$i and 1h, and the large dependence case 1d. However, in other large dependence cases $*$z and the moderate dependence case 2d, the test results diverge. Both the empirical distribution method and the double kernel local linear estimation method directly detect the dependence relationship between the random variables, no matter whether the dependence is linear or nonlinear, and compute the distribution functions. Therefore, the large or moderate dependence in $*$z and 2d can be identified. On the other hand, the copulas take account of the dependence relationship by the dependence measures of Pearson correlation ρ or Kendall's τ. $\hat{\rho}$ and $\hat{\tau}$ are measured small in Table 6.3, and hence are not easily detectable. As a result, $*$z and 2d are falsely determined as independence.

Now consider the cases where our interest lies: the weak dependent cases 3d-6d and 2h-4h. All the simulation results before the Box-Cox transformation show a common property in this case, that is, no weak dependence is successfully identified. All the tests before the Box-Cox transformation, the correlation test and different independence tests, lead to an unacceptable erroneous result, that is, they are short of validity and efficiency in identifying the dependence cases, more precisely, the weak dependence cases. Therefore, a correlation test based on the Box-Cox transformed variables $g(U)$ and $h(V)$ is performed and its simulation result is graphically shown in Figure 6.10. In addition to 1d, other processes 2d-6d, 2h-4h and 1z-2z are also identified as correlated, i.e, dependent. This result is consistent with the actual data generating processes. It means that our correlation test is efficient and valid in detecting dependent relationship, especially weak dependence. Also, we need to notice the correlation test is not always powerful as in 2h-4h cases where the rejection rates are between 0.51 and 0.67 as shown in Table 6.3. Further improvement or other methods to precisely identify the weak dependence in such high frequency situations will be needed.

Table 6.1: Rejection rate γ before Box-Cox transformation

	$T = 100, \alpha = 0.05$		$T = 100, \alpha = 0.1$	
	Fisher's z	bootstrap	Fisher's z	bootstrap
1i	0.06	0.05	0.10	0.11
2i	0.05	0.05	0.10	0.11
3i	0.06	0.06	0.08	0.09
4i	0.08	0.09	0.11	0.11
1d	1.00	0.96	1.00	1.00
2d	0.30	0.30	0.40	0.40
3d	0.19	0.18	0.27	0.25
4d	0.28	0.28	0.37	0.36
5d	0.13	0.12	0.21	0.21
6d	0.16	0.16	0.23	0.22
1h	0.02	0.02	0.06	0.08
2h	0.08	0.07	0.11	0.11
3h	0.05	0.05	0.11	0.11
4h	0.05	0.05	0.10	0.09
1z	0.04	0.05	0.17	0.16
2z	0.03	0.04	0.18	0.16

Table 6.2: Sample dependence measures when $T = 100$

	Pearson $\hat{\rho}$	Spearman $\hat{\rho}_S$	Kendall $\hat{\tau}$		Pearson $\hat{\rho}$	Spearman $\hat{\rho}_S$	Kendall $\hat{\tau}$
1i	0.0706	0.0445	0.0182	1d	0.5496	0.5234	0.3741
2i	-0.0856	-0.1073	-0.0721	2d	0.1368	0.1214	0.0909
3i	0.0666	0.0306	0.0186	3d	-0.0624	-0.1294	-0.0974
4i	0.0723	0.0622	0.0347	4d	-0.0698	-0.1561	-0.1099
				5d	0.0945	0.0728	0.0585
				6d	0.0939	0.0719	0.0566
1h	-0.0094	-0.0118	-0.0079	1z	-0.0126	-0.0128	-0.0121
2h	-0.0079	-0.0068	-0.0040	2z	0.0034	0.0021	0.0011
3h	-0.0176	-0.0208	-0.0138				
4h	-0.0012	-0.0074	-0.0056				

Table 6.3: Rejection rate γ comparison after Box-Cox transformation

	$T = 100, \alpha = 0.05$		$T = 100, \alpha = 0.1$	
	1-var Box-Cox	2-var Box-Cox	1-var Box-Cox	2-var Box-Cox
1i	0.12	0.16	0.27	0.31
2i	0.16	0.21	0.35	0.39
3i	0.13	0.15	0.24	0.36
4i	0.15	0.17	0.25	0.34
1d	0.96	1.00	1.00	1.00
2d	0.97	1.00	1.00	1.00
3d	0.55	0.69	0.67	0.85
4d	0.62	0.80	0.81	0.95
5d	0.58	0.72	0.75	0.91
6d	0.55	0.78	0.69	0.90
1h	0.18	0.22	0.34	0.39
2h	0.40	0.46	0.58	0.67
3h	0.34	0.41	0.56	0.63
4h	0.30	0.40	0.51	0.58
1z	0.97	1.00	1.00	1.00
2z	0.94	0.99	1.00	1.00

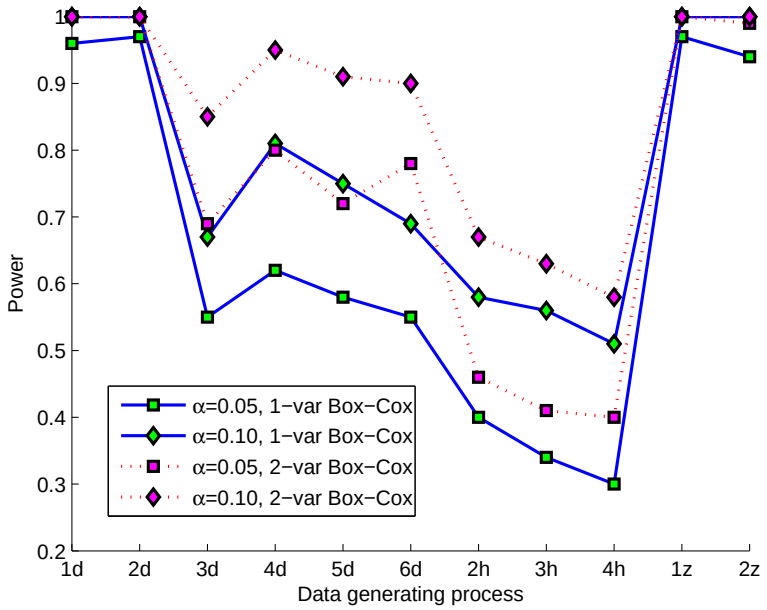

Figure 6.1: Power comparison of correlation test after Box-Cox transformation

Table 6.4: Rejection rate γ (before Box-Cox transformation, empirical distribution estimation)

	$T = 100, \alpha = 0.05$				$T = 100, \alpha = 0.1$			
	CM	KS	CT	RW	CM	KS	CT	RW
1i	0.04	0.05	0.03	0.05	0.10	0.09	0.08	0.09
2i	0.05	0.04	0.03	0.04	0.11	0.11	0.08	0.09
3i	0.04	0.06	0.05	0.05	0.08	0.09	0.08	0.08
4i	0.06	0.06	0.06	0.04	0.10	0.08	0.09	0.09
1d	0.99	0.98	0.40	0.54	1.00	1.00	0.49	0.65
2d	0.79	0.60	0.45	0.75	0.92	0.72	0.62	0.84
3d	0.11	0.09	0.05	0.06	0.20	0.17	0.09	0.13
4d	0.20	0.19	0.08	0.15	0.31	0.30	0.16	0.27
5d	0.14	0.09	0.08	0.11	0.23	0.18	0.15	0.19
6d	0.10	0.10	0.05	0.05	0.22	0.19	0.08	0.15
1h	0.03	0.03	0.05	0.06	0.06	0.06	0.10	0.10
2h	0.03	0.10	0.03	0.04	0.11	0.15	0.10	0.10
3h	0.05	0.03	0.02	0.05	0.11	0.10	0.09	0.10
4h	0.03	0.02	0.01	0.02	0.05	0.06	0.06	0.04
1z	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2z	1.00	1.00	0.81	0.84	1.00	1.00	0.90	0.95

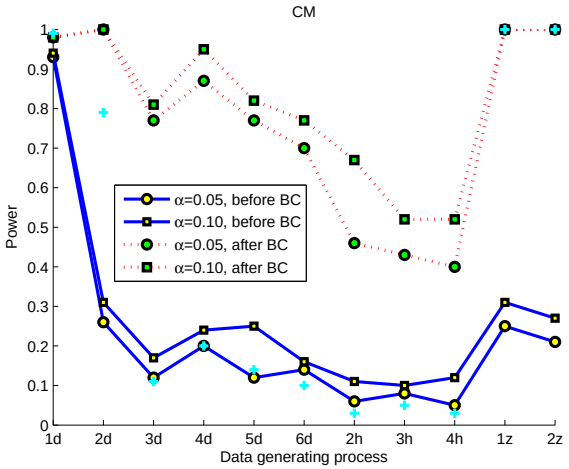

Figure 6.2: Power comparison, CM

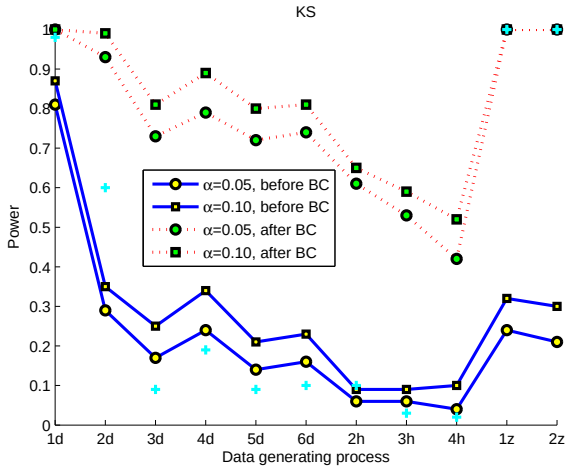

Figure 6.3: Power comparison, KS

Table 6.5: Rejection rate γ (before Box-Cox transformation, double kernel LLM)

	$T = 100, \alpha = 0.05$				$T = 100, \alpha = 0.1$			
	CM	KS	CT	RW	CM	KS	CT	RW
1i	0.05	0.05	0.06	0.04	0.12	0.10	0.10	0.09
2i	0.05	0.03	0.05	0.03	0.11	0.08	0.10	0.09
3i	0.05	0.06	0.05	0.05	0.11	0.09	0.10	0.08
4i	0.07	0.06	0.06	0.04	0.12	0.11	0.09	0.10
1d	0.98	0.90	0.82	0.76	1.00	0.96	0.89	0.85
2d	0.78	0.82	0.74	0.74	0.86	0.89	0.81	0.80
3d	0.07	0.08	0.07	0.09	0.11	0.15	0.10	0.14
4d	0.19	0.23	0.18	0.17	0.26	0.29	0.27	0.27
5d	0.10	0.10	0.08	0.05	0.14	0.15	0.15	0.14
6d	0.07	0.08	0.08	0.06	0.10	0.12	0.14	0.11
1h	0.06	0.07	0.06	0.06	0.12	0.13	0.09	0.10
2h	0.06	0.08	0.08	0.07	0.11	0.14	0.13	0.12
3h	0.06	0.07	0.07	0.06	0.12	0.11	0.10	0.09
4h	0.05	0.04	0.04	0.04	0.13	0.09	0.07	0.07
1z	0.93	0.91	0.82	0.73	0.96	0.96	0.86	0.85
2z	0.63	0.68	0.86	0.71	0.76	0.80	0.82	0.83

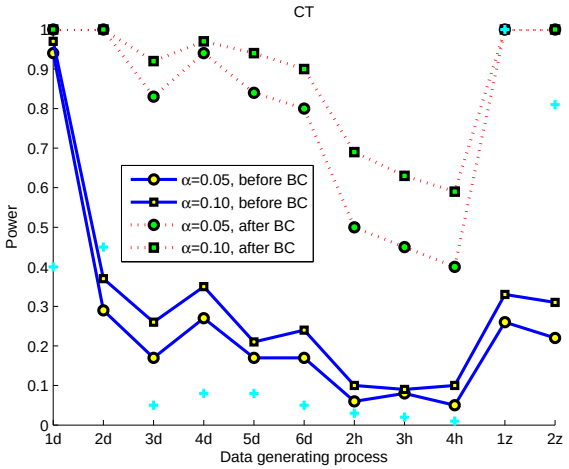

Figure 6.4: Power comparison, CT

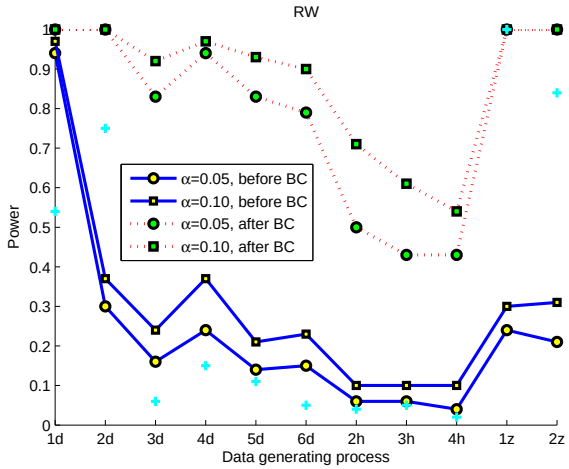

Figure 6.5: Power comparison, RW

Table 6.6: Rejection rate γ (before Box-Cox transformation, normal copula)

	$T = 100, \alpha = 0.05$				$T = 100, \alpha = 0.1$			
	CM	KS	CT	RW	CM	KS	CT	RW
1i	0.07	0.09	0.05	0.05	0.10	0.13	0.13	0.11
2i	0.03	0.04	0.04	0.04	0.09	0.10	0.10	0.09
3i	0.09	0.08	0.08	0.08	0.14	0.13	0.12	0.13
4i	0.04	0.05	0.04	0.04	0.09	0.11	0.09	0.09
1d	0.93	0.81	0.94	0.94	0.94	0.87	0.97	0.97
2d	0.26	0.30	0.29	0.30	0.31	0.35	0.37	0.37
3d	0.12	0.18	0.18	0.17	0.18	0.26	0.26	0.25
4d	0.20	0.27	0.28	0.25	0.24	0.36	0.36	0.37
5d	0.12	0.15	0.17	0.15	0.25	0.21	0.21	0.21
6d	0.14	0.16	0.17	0.16	0.16	0.23	0.24	0.24
1h	0.08	0.04	0.05	0.06	0.11	0.10	0.11	0.11
2h	0.06	0.06	0.06	0.06	0.11	0.09	0.10	0.10
3h	0.08	0.06	0.08	0.08	0.10	0.09	0.09	0.10
4h	0.05	0.04	0.05	0.04	0.12	0.10	0.10	0.10
1z	0.25	0.24	0.26	0.24	0.31	0.32	0.33	0.30
2z	0.21	0.22	0.23	0.22	0.28	0.31	0.32	0.31

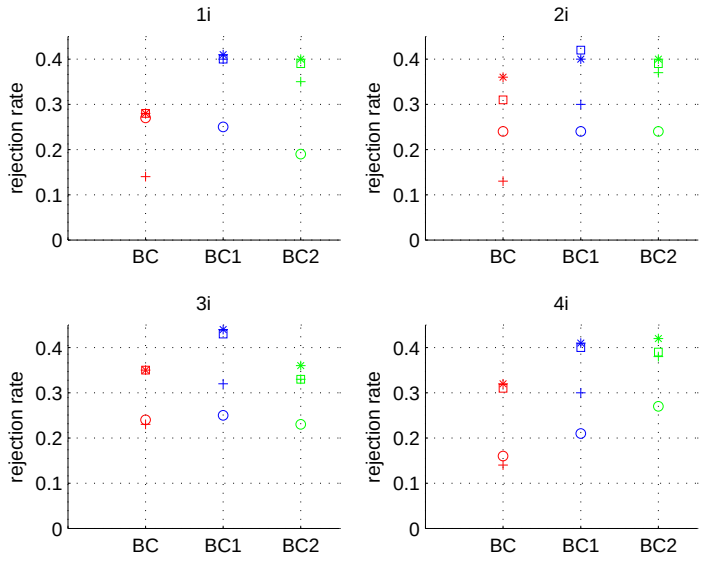

Figure 6.6: Rejection rate with respect to modified Box-Cox transformations, $*$i

Table 6.7: Rejection rate γ (before Box-Cox transformation, Archimedean copula)

	$T = 100, \alpha = 0.05$				$T = 100, \alpha = 0.1$			
	CM	KS	CT	RW	CM	KS	CT	RW
1i	0.02	0.03	0.02	0.02	0.06	0.05	0.06	0.06
2i	0.09	0.10	0.09	0.09	0.15	0.16	0.15	0.15
3i	0.04	0.04	0.04	0.02	0.05	0.07	0.06	0.08
4i	0.10	0.08	0.09	0.10	0.12	0.13	0.14	0.15
1d	0.88	1.00	0.98	1.0	0.91	1.00	0.99	1.00
2d	0.24	0.21	0.29	0.28	0.29	0.34	0.36	0.36
3d	0.14	0.15	0.14	0.17	0.21	0.19	0.21	0.21
4d	0.17	0.18	0.21	0.22	0.21	0.23	0.29	0.32
5d	0.06	0.07	0.10	0.11	0.12	0.18	0.15	0.16
6d	0.12	0.09	0.10	0.11	0.18	0.15	0.16	0.17
1h	0.04	0.03	0.04	0.04	0.10	0.07	0.09	0.06
2h	0.05	0.05	0.03	0.04	0.12	0.08	0.08	0.09
3h	0.06	0.05	0.05	0.06	0.08	0.13	0.10	0.13
4h	0.04	0.08	0.06	0.08	0.06	0.11	0.09	0.12
1z	0.18	0.19	0.19	0.20	0.26	0.25	0.27	0.28
2z	0.19	0.24	0.22	0.25	0.24	0.30	0.27	0.30

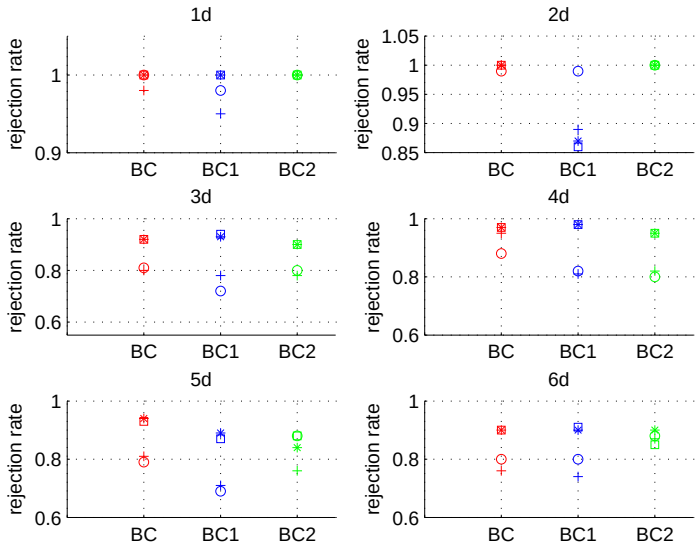

Figure 6.7: Rejection rate with respect to modified Box-Cox transformations, $*$d

Table 6.8: Rejection rate γ (after Box-Cox transformation, empirical distribution estimation)

	$T = 100, \alpha = 0.05$				$T = 100, \alpha = 0.1$			
	CM	KS	CT	RW	CM	KS	CT	RW
1i	0.11	0.11	0.07	0.07	0.18	0.20	0.18	0.11
2i	0.10	0.06	0.09	0.08	0.16	0.15	0.13	0.11
3i	0.09	0.09	0.09	0.08	0.20	0.21	0.15	0.12
4i	0.08	0.07	0.07	0.06	0.12	0.13	0.16	0.14
1d	1.00	1.00	0.57	0.60	1.00	1.00	0.68	0.72
2d	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3d	0.33	0.22	0.09	0.12	0.47	0.34	0.15	0.18
4d	0.42	0.35	0.21	0.40	0.52	0.46	0.31	0.50
5d	0.29	0.23	0.12	0.24	0.42	0.44	0.20	0.40
6d	0.20	0.17	0.13	0.18	0.33	0.24	0.19	0.40
1h	0.10	0.09	0.08	0.10	0.15	0.15	0.14	0.17
2h	0.10	0.12	0.07	0.06	0.19	0.19	0.15	0.14
3h	0.07	0.08	0.07	0.08	0.20	0.15	0.16	0.17
4h	0.09	0.09	0.06	0.05	0.16	0.17	0.09	0.10
1z	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2z	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

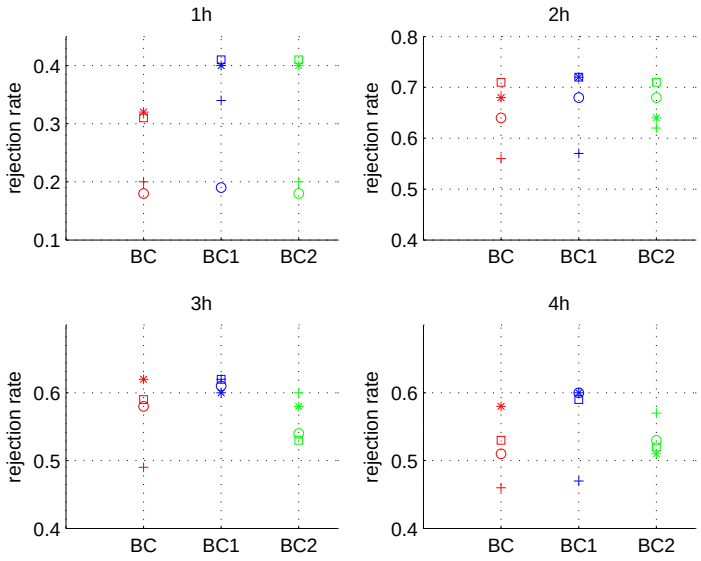

Figure 6.8: Rejection rate with respect to modified Box-Cox transformations, $*$h

Table 6.9: Rejection rate γ (after Box-Cox transformation, double kernel LLM)

	$T = 100, \alpha = 0.05$				$T = 100, \alpha = 0.1$			
	CM	KS	CT	RW	CM	KS	CT	RW
1i	0.31	0.28	0.27	0.26	0.37	0.31	0.35	0.32
2i	0.28	0.18	0.29	0.28	0.34	0.21	0.35	0.31
3i	0.32	0.19	0.31	0.30	0.37	0.23	0.32	0.33
4i	0.24	0.22	0.25	0.25	0.30	0.29	0.31	0.30
1d	1.00	0.98	0.96	0.96	1.00	1.00	0.99	0.99
2d	0.87	0.96	0.88	0.87	0.93	0.98	0.97	0.96
3d	0.50	0.47	0.41	0.41	0.57	0.54	0.50	0.50
4d	0.62	0.61	0.63	0.63	0.67	0.65	0.67	0.67
5d	0.53	0.58	0.59	0.57	0.58	0.63	0.64	0.64
6d	0.51	0.48	0.51	0.52	0.59	0.56	0.57	0.57
1h	0.30	0.32	0.31	0.31	0.35	0.35	0.34	0.33
2h	0.42	0.35	0.33	0.33	0.57	0.48	0.43	0.42
3h	0.41	0.33	0.32	0.33	0.53	0.41	0.40	0.41
4h	0.50	0.38	0.42	0.39	0.59	0.44	0.48	0.44
1z	1.00	1.00	0.95	1.00	1.00	1.00	0.97	1.00
2z	0.83	0.87	0.90	0.88	0.92	0.92	0.95	0.95

Table 6.10: Rejection rate γ (after Box-Cox transformation, normal copula)

	$T = 100, \alpha = 0.05$				$T = 100, \alpha = 0.1$			
	CM	KS	CT	RW	CM	KS	CT	RW
1i	0.11	0.21	0.15	0.15	0.14	0.27	0.28	0.28
2i	0.07	0.18	0.20	0.21	0.13	0.24	0.36	0.31
3i	0.17	0.18	0.22	0.22	0.23	0.23	0.35	0.35
4i	0.10	0.11	0.21	0.22	0.14	0.16	0.32	0.32
1d	0.98	1.00	1.00	1.00	0.98	1.00	1.00	1.00
2d	1.00	0.93	1.00	1.00	1.00	0.99	1.00	1.00
3d	0.77	0.72	0.83	0.83	0.80	0.81	0.92	0.92
4d	0.87	0.79	0.94	0.94	0.95	0.88	0.97	0.97
5d	0.77	0.71	0.84	0.83	0.81	0.79	0.94	0.93
6d	0.68	0.73	0.80	0.79	0.76	0.80	0.90	0.90
1h	0.14	0.13	0.20	0.19	0.20	0.18	0.32	0.31
2h	0.46	0.60	0.50	0.49	0.66	0.64	0.68	0.71
3h	0.42	0.53	0.44	0.42	0.52	0.58	0.62	0.59
4h	0.40	0.41	0.39	0.42	0.52	0.52	0.58	0.53
1z	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2z	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6.11: Rejection rate γ (after Box-Cox transformation, Archimedean copula)

	$T = 100, \alpha = 0.05$				$T = 100, \alpha = 0.1$			
	CM	KS	CT	RW	CM	KS	CT	RW
1i	0.04	0.04	0.03	0.04	0.08	0.09	0.08	0.09
2i	0.13	0.13	0.13	0.14	0.18	0.19	0.19	0.20
3i	0.05	0.04	0.05	0.05	0.11	0.09	0.12	0.12
4i	0.11	0.12	0.11	0.11	0.16	0.15	0.15	0.15
1d	0.91	0.83	1.00	1.00	0.94	0.86	1.00	1.00
2d	0.34	0.33	0.33	0.33	0.45	0.51	0.46	0.47
3d	0.25	0.25	0.25	0.26	0.34	0.34	0.33	0.33
4d	0.39	0.35	0.36	0.36	0.48	0.42	0.43	0.43
5d	0.32	0.32	0.32	0.30	0.34	0.39	0.38	0.38
6d	0.34	0.33	0.31	0.31	0.39	0.38	0.36	0.36
1h	0.32	0.33	0.34	0.34	0.37	0.39	0.39	0.37
2h	0.23	0.23	0.23	0.23	0.28	0.27	0.27	0.26
3h	0.29	0.29	0.28	0.29	0.35	0.35	0.35	0.34
4h	0.33	0.32	0.32	0.32	0.36	0.37	0.37	0.37
1z	0.21	0.23	0.20	0.21	0.36	0.36	0.37	0.36
2z	0.26	0.27	0.26	0.28	0.39	0.38	0.38	0.38

Table 6.12: Sample correlation coefficient before and after the Box-Cox transformation

	$\hat{r}_{org}$	$\hat{r}_{max}$		$\hat{r}_{org}$	$\hat{r}_{max}$
1i	0.0743	0.1534	1d	0.5248	0.5334
2i	0.0851	0.1629	2d	0.1507	0.3809
3i	0.0864	0.1636	3d	0.1057	0.2136
4i	0.0770	0.1597	4d	0.1304	0.2519
			5d	0.0982	0.2140
			6d	0.0911	0.1971
1h	0.0813	0.1546	1z	0.0553	0.5441
2h	0.0758	0.1742	2z	0.0521	0.4983
3h	0.0814	0.1761			
4h	0.0722	0.1655			

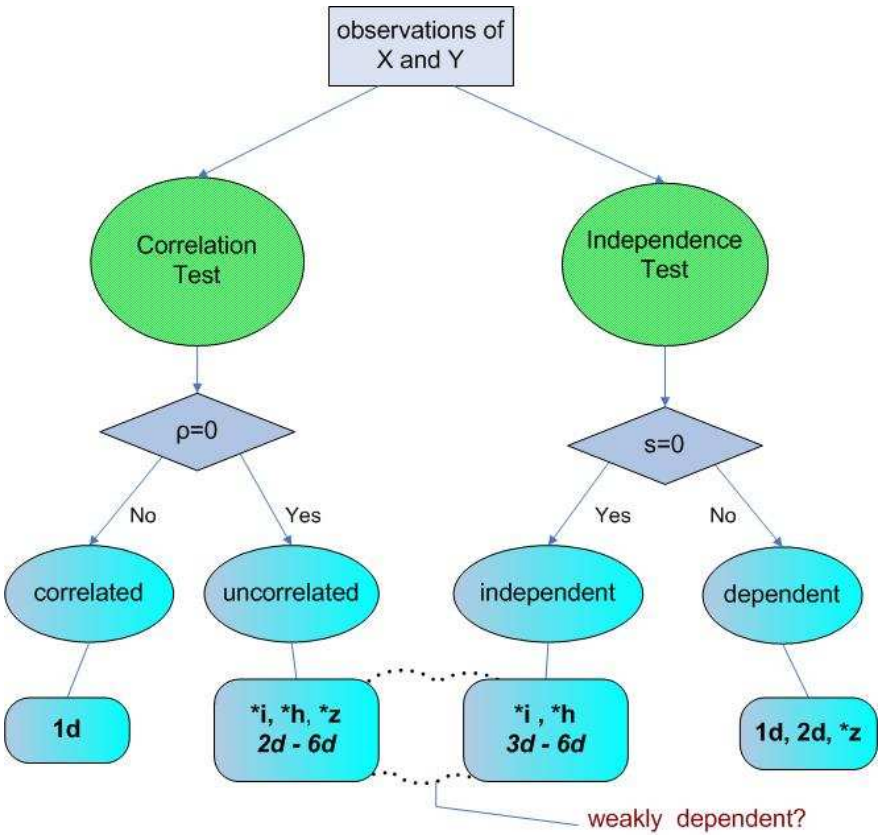

Figure 6.9: Test results before Box-Cox transformation

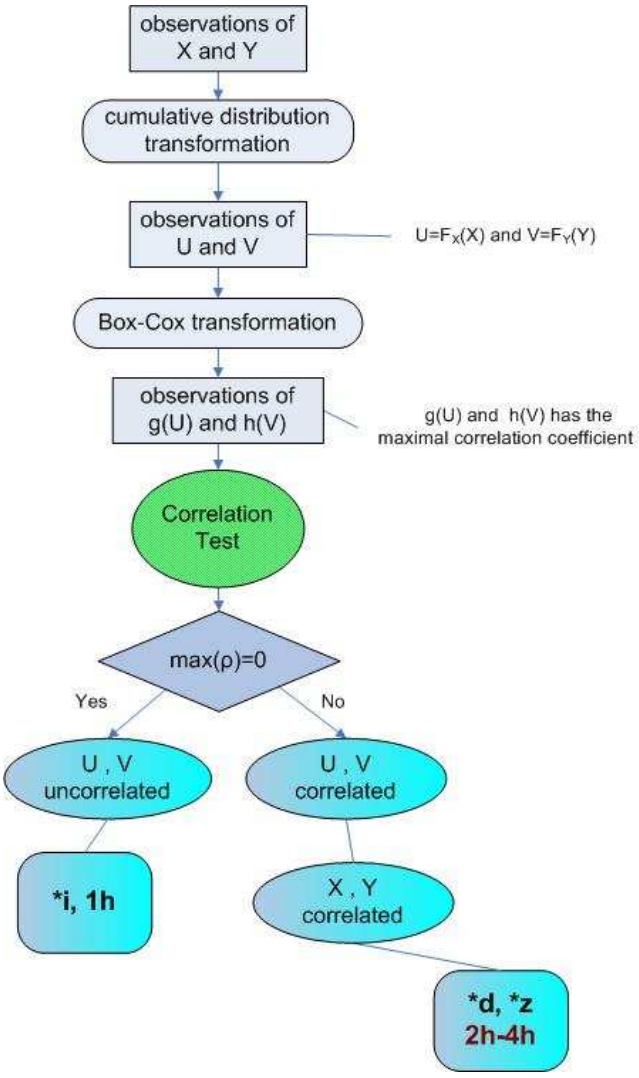

Figure 6.10: Correlation test results after Box-Cox transformation

CHAPTER 7. APPLICATIONS

For the time being, the independence tests and the correlation test we have developed in this study only address hypothetical data generating processes, for example, $*$i, $*$d, $*$h and $*$z, in attempting to distinguish weak dependence from genuine independence. In this section, the significance of after-transform correlation test in real instances is explored. Two possible applications, regression analysis and weak message identification, will be illustrated.

7.1 Regression Analysis

Regression analysis is a statistical tool for the investigation of relationship between variables, that is, to ascertain the causal effect of one variable upon another. For example, fitting a curve between a signal and its noise output in a circuit or finding the effect of the changes in the money supply upon the inflation rate is generally a regression problem. To explore such issues, we assemble data on the underlying variables of interest and employ regression to estimate the quantitative effect of the causal variables upon the variable that they influence. A variety of techniques are involved in regression analysis. One of the key steps to quantify the effect between variables, actually the starting point of regression analysis, is to identify the factors which may influence the variable of interest.

The responsibility of a network administrator is to maintain the network and keep it working normally. One of his routine tasks is monitoring the daily throughput of the network and using this parameter as part of the criteria to determine the network's health. Suppose y_t denotes the daily throughput fluctuation over a particular baseline. A mathematical model of y_t is certainly important in explaining and forecasting the workload of a network, and as a result is critical to determine whether or not the network works in a normal state. Since regression analysis is a technique for modeling the relationship between two or more variables, the statistical model of y_t can be built by

regressing y_t with some significant factors. With regard to regression illustration, the first thing we need to do is to identify and quantify the factors that determine y_t. It is intuitive that yesterday's throughput fluctuation y_{t-1} could be an important reference to predict today's throughput fluctuation y_t and might be used as one of the explanatory variables. Is there any other factor which may also influence y_t? One way is to find the factors which are correlated to y_t. Of course, the correlation between variables does not imply the causal relation. Correlation is not equal to causation, though correlation can still be a hint. Therefore, if the correlation between two random variables X and Y is measured to be zero, we will tend to think that there is no causation between them and thus one variable should not enter into the regression of the other variable. Consider the regression problem of the daily throughput fluctuation y_t. Suppose that y_t is monitored on a subnetwork which is an internal network of a campus network and does not connect to an outside network. However, note that the campus network has a few access points to the internet. Therefore, the campus network will be directly influenced by the computer viruses breaking out on the internet, while the subnetwork of interest will not be directly influenced by it. Assume the change of daily computer viruses breaking out on the internet is x_t. Since the network daily throughput fluctuation y_t of the subnetwork only has an indirect, weak and lagged correlation with x_t (for example, the students use the virused USB disks in the subnetwork afterwards), we assume $\mathrm{cov}(y_t, x_{t-1}) = 0$ without loss of generality. Actually, the daily throughput fluctuation y_t follows the rule of 3d, i.e.,

$$y_t = 0.5 y_{t-1} x_{t-1} + \varepsilon_{1,t}, \quad \text{and } x_t = 0.5 x_{t-1} + \varepsilon_{2,t}, \tag{7.1}$$

where y_{t-1} denotes the throughput fluctuation of the past day, x_{t-1} denotes the change of daily computer viruses breaking out on the internet and the noise terms $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d. $\mathcal{N}(0, I_2)$.

Using the collected observations from 3d, we wish to build a regression model for the response (dependent) variable y_t, where y_{t-1} and x_{t-1} are the possible explanatory (independent) variables and no knowledge of the relationship between y_t and $\{y_{t-1}, x_{t-1}\}$ is given. Usually, the regression analysis focuses more on the way to find an appropriate fitting curve between the dependent variable and the explanatory variables. In this procedure, the explanatory variables are always assumed to be known due to some obvious and plausible reasons. For example, the throughput fluctuation y_{t-1} can be thought of as a reasonable independent variable in explaining

the dependent variable y_t. As to the variable entry problem of regression analysis, there is no clear solution other than the stepwise method based on the current literature [92,93]. The stepwise method is a technique for choosing the independent variables to be included in a multiple regression model. Its aim is to obtain a parsimonious model which explains the most variance in the dependent variable containing the fewest independent variables. The method is a combination of a forward technique and a backward technique, adding variables when they are significant, and removing them when they are not significant. The forward technique starts with no model variables. At each step it adds the most statistically significant variable, the one with the lowest p-value, to the model until there are none left. On the other hand, the backward technique starts with all possible variables in the model and removes the least significant variable until all the remaining variables are statistically significant. According to [93], the stepwise technique has many problems associated with it, and should be used with extreme caution. However, the stepwise method is included in most statistical packages and thus is a convenient way to allow regression models to be built in a series of steps by adding or removing one independent variable at a time.

7.1.1 Variable Entry Method

In this section, the stepwise technique is employed to choose which variables should be entered into the regressing procedure, and thus a suitable regression model can be built to explain the variable of interest. The following results come from the Matlab stepwise function, which provides an interactive graphical interface to compare competing models.

Figure 7.1 shows the stepwise regression of the daily throughput fluctuation based on the 100 observations $\{y_t, y_{t-1}, x_{t-1}\}$ generated from (7.1). Two explanatory variables, y_{t-1} and x_{t-1}, are labeled as $X1$ and $X2$, respectively, in Figure 7.1. Consider the upper part of the plot. For each term on the y-axis, the plot shows the regression (least squares) coefficient as a dot with horizontal bars indicating confidence intervals. Blue dots represent terms that are in the model, while red dots indicate terms that are not currently in the model. In Figure 7.1, no variable has been added to the model yet, and no blue dot appears in the upper part of the plot.

A table listing the value of the regression coefficient for each term, along with its t-statistic and p-value, is on the right of each bar. The coefficient for a term that is not in the model is the coefficient that would result from adding that term to the current model. Therefore, if $X1$, i.e.,

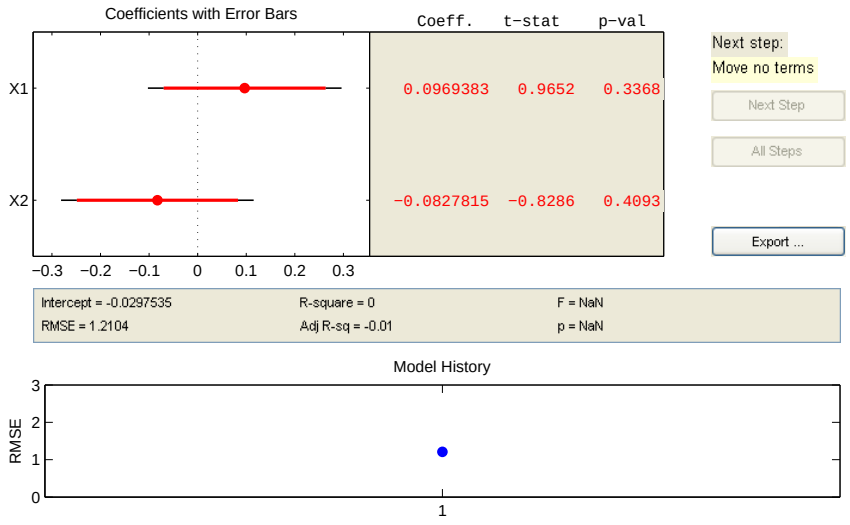

Figure 7.1: Stepwise regression of the daily throughput fluctuation

y_{t-1}, were added to the regression model, then $y_t = 0.09693y_{t-1} - 0.02975$. The bottom window of "Model History" , which shows the RMSE (root mean squared error) for every model generated during the current session, will be changed correspondingly.

The determination of whether or not to add a variable in a regression model depends on its significance. As for $X1$, i.e., y_{t-1}, in Figure 7.1, a p-value of 0.3368 is greater than the default significance level $\alpha = 0.05$ and therefore y_{t-1} should not be entered into the regression of y_t. Otherwise, if the p-value of y_{t-1} is less than 0.05, we know y_{t-1} should be classified as one of the predicators of y_t. More graphically and intuitively, such possible change to the model is represented by the recommended step shown under "Next Step" to the right of the table. In Figure 7.1, the recommended step is "Move no terms", which indicates that neither y_{t-1} nor x_{t-1} can be considered as a significant variable and therefore is not eligible to be added to the regressing model of y_t. In other words, no regression model is created.

This result is not surprising because the stepwise method is grounded on the correlation analysis and fails to find the corresponding independent variables of y_t when both $\text{cov}(y_t, y_{t-1})$ and $\text{cov}(y_t, x_{t-1})$ are zero. However, if y_{t-1} and x_{t-1} are considered dependent with y_t through the after-transform correlation test, it is sufficient to allow us counting them as the influence factors in curve fitting y_t, since dependence provides a stronger evidence of the relationship between variables than

correlation. Dependence with zero correlation only implies the nonlinear relationship between variables. For example, if a correct regressing model involving the independent variables y_{t-1} and x_{t-1} is built, i.e., $\hat{y}_t = \alpha \hat{y}_{t-1}\hat{x}_{t-1} + \varepsilon_t$, the estimated regression parameter $\alpha = 0.5205$ is close to the true parameter 0.5.

We now consider a telecommunication company which transfers the user data weekly from the working database to a raid disk array as a backup. Assume the normal transfer time is known given a base amount of user data. We care about the change of transfer time y_t since it is critical to the resource distribution. Assume x_t is the difference between the real amount of user data and the base amount of user data. When more user data is backed up, i.e., $x_t > 0$, more transfer time is needed and the corresponding change of the transfer time is greater than zero, i.e., $y_t > 0$, and vice versa. Let z_t be a global index representing the active status of the raid disk. If the raid disk works properly, z_t is a number far from zero. When the performance of the raid disk decreases, the magnitude of the index z_t becomes smaller. If the raid disk has some serious problems, for example, one or more disks of the cluster are not functioning, or the usage of the raid disk exceeds the threshold, z_t is close to zero. As we know, if the raid disk works normally, the change of transfer time y_t is mainly decided by the change of the amount of user data x_t. However, if the raid disk is in a bad condition, z_t may become the major factor that increases y_t. In practice, the change of transfer time y_t can be represented by the following rule,

$$y_t = 0.5x_t + 2\varphi(2z_t) + 0.5\varepsilon_t, \tag{7.2}$$

where $\{x_t, z_t, \varepsilon_t\}$ are i.i.d. $\mathcal{N}(0, \mathbf{I}_3)$ and φ is the standard normal density function. The difference of the amount of user data x_t has the same moving trend as the change of transfer time y_t. z_t's influence on y_t is much weaker than that of x_t on y_t when z_t is far from zero or the raid disk works normally. But when z_t is close to zero, i.e., the raid disk becomes dysfunctional, the impact that z_t puts on y_t cannot be ignored. Since y_t and z_t are nonlinearly related, the linear correlation between them is very small. In the particular case of the nonlinear normal density function φ, we have $\mathrm{cov}(y_t, z_t) = 0$.

Based on the observations collected from (7.2), the stepwise regression is performed. It includes two steps shown in Figures 7.2 and 7.3. The difference of the amount of user data x_t and

the index of the raid disk z_t are labeled as $X1$ and $X2$ in Figures 7.2 and 7.3. Figure 7.2 shows the first step of the regression procedure. The initial model contains no independent variables since there are only red dots in the upper left window. Because X1, i.e., x_t, has a smaller p-value, and therefore is statistically more significant compared to X2, i.e., z_t, the recommended step shown under "Next Step" to the right of the table is "Move X1 in". Figure 7.3 shows the model after $X1$ is moved in. In the upper left window of Figure 7.3, the blue dot of $X1$ indicates that $X1$ is in the model and the red dot of $X2$ indicates that $X2$ is not entered in the model. The recommended step shown under "Next Step" is "Move no terms". Therefore, there is no further steps, and the final regression model derived from the stepwise method is the one shown in Figure 7.3.

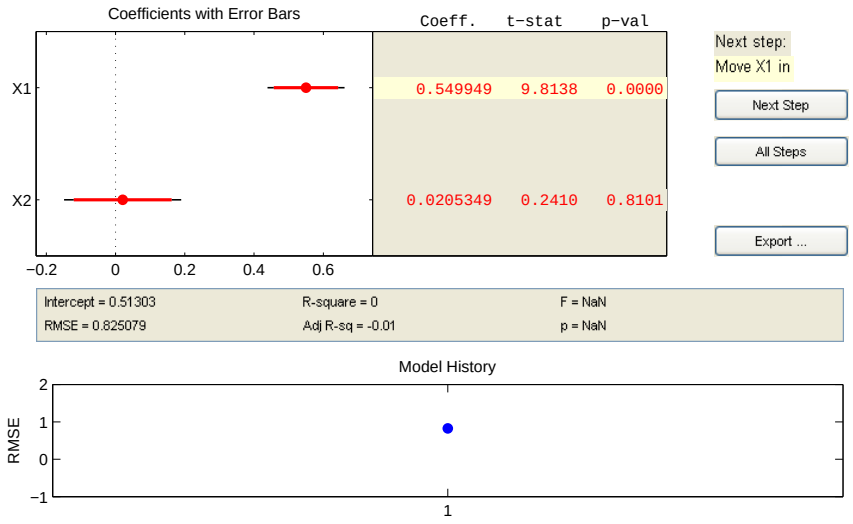

Figure 7.2: Stepwise regression of the total transfer time change, step 1

In the session of Figure 7.3, the tentative hypothesis is that higher levels of change in the amount of user data x_t causes higher levels of change in the transfer time y_t when other things are being equal, that is, the regression model is in the form of $\hat{y}_t = \alpha\hat{x}_t + \beta$, where the coefficient $\alpha = 0.5499$ and the intercept $\beta = 0.3782$. The corresponding RMSE is 0.5904. In Figure 7.4, the scatter plot of the observations (blue cross) indeed suggests that higher values of x_t tend to yield higher values of y_t, but the relationship is not perfect as represented by the solid regression line. It seems that the knowledge of x_t does not suffice for an entirely accurate predication about y_t. We can then deduce either that the effect of the change of user data amount upon the change of transfer

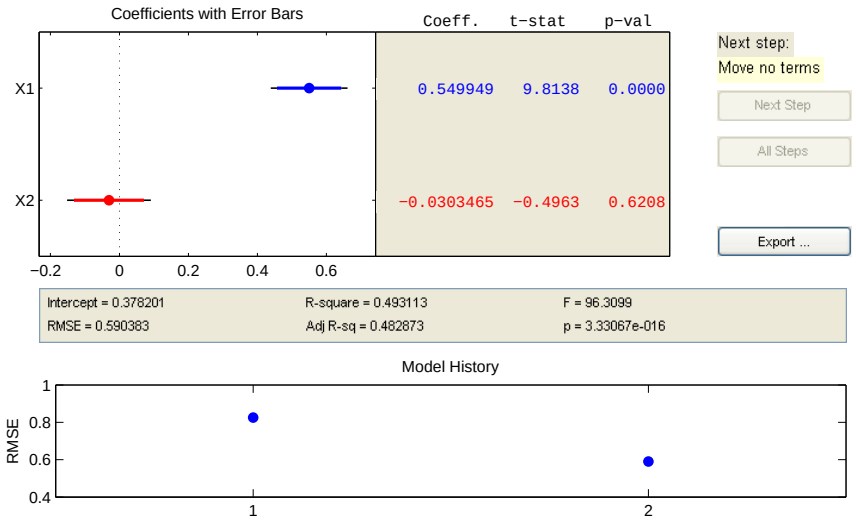

Figure 7.3: Stepwise regression of the total transfer time change, step 2

time varies with respect to time, or that factors other than the change of user data amount influence the change of transfer time. Regression analysis ordinarily embraces the latter explanation.

It should be noticed that the stepwise regression in Figures 7.2 and 7.3 only suggests x_t as the independent variable to explain y_t. z_t is omitted since its correlation with y_t is not as significant as that between x_t and y_t. In this way, some relevant factors whose change also causes the change of the response variable y_t are ignored.

Unlike the generally used stepwise technique based on the correlation, if we construct the regression model by taking into account the dependence relationship using the after-transform correlation test, the degree of confidence that we have in the accuracy of regression estimates is expected to increase. Formulate the regression model y_t as $\hat{y}_t = \alpha_1 \hat{x}_t + \alpha_2 \varphi(\alpha_2 \hat{z}_t)$, where φ is the standard normal density function. The estimated parameter $\alpha_1 = 0.5007$ and $\alpha_2 = 1.9860$. The coefficient of x_t alters from 0.5499 to 0.5007, which is closer to the true parameter 0.5. Note that the mean effect of the global index z_t on the change of transfer time y_t is positive. The coefficient of x_t is larger before the variable z_t is entered into the regression because it is erroneously capturing some of the positive influence of z_t as well as its own influence of x_t. Furthermore, the variable entry of z_t helps decrease RMSE from 0.5904 to 0.4902. As is expected, the accuracy of regression is improved.

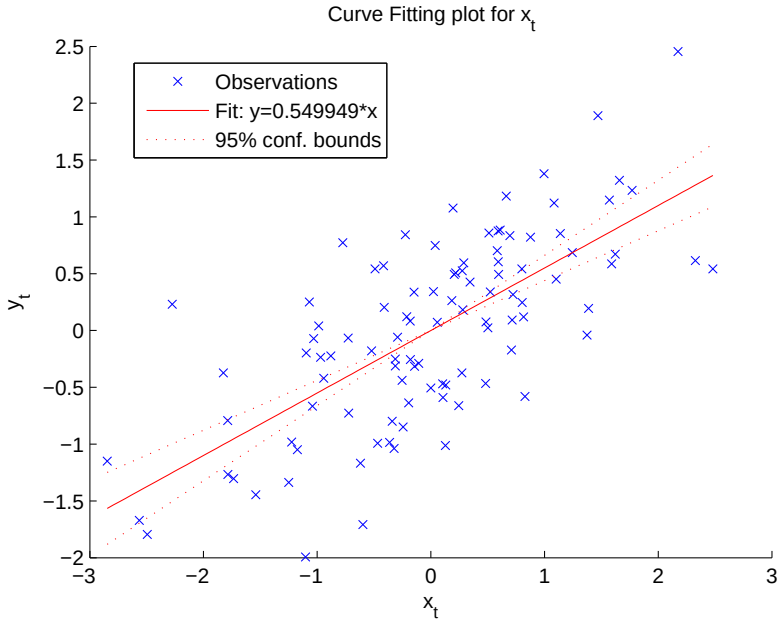

Figure 7.4: Regression of the total transfer time change y_t only by x_t

7.1.2 Variable Entry Discussion

Regression analysis is to produce an estimate of the regression parameters based upon the information contained in the data set, and more importantly upon some assumptions about the characteristics of the noise. For example, the noise term cannot always be negative or positive, i.e., it is on average equal to zero. This assumption suggests that the regressing curve lies roughly in the midst of the data, some observations below and some observations above. An effort to quantify the effect of some independent variables X upon the dependent variable Y without careful attention to the other factors that affect Y could create serious statistical difficulties, termed as omitted variables bias. In other words, the omission from a regression of some variables Z that affect the dependent variable Y may cause a bias. The problem arises because any omitted variable becomes part of the noise term, and the consequence could be a violation of the underlying assumptions required for the unbias, consistency or efficiency of the estimator. As is well known, the magnitude of the noise will affect the accuracy of the regression estimates with more noise leading to less accuracy on average. In the regression modeling of the change of transfer time y_t, the omission of the global index z_t means the average of the noise term increases to some positive value. The violation of the

zero mean noise assumption thus leads to the decrease in the explanatory power of the regression analysis and the statistical significance of its parameter estimates as described here: the coefficient of x_t diverges from 0.5007 to 0.5499, and the RMSE increases from 0.4902 to 0.5904.

As we know, correlation does not imply causality, but causality does mean there exists some kind of correlation, linear or nonlinear, between the variables. Since there is no easy-to-use equivalence of Pearson correlation that is capable of handling nonlinear relations, we usually use the Pearson correlation as the tool to identify the explanatory (independent) variables in regression analysis. However, this method is deficient when the explanatory variables and response variable are dependent but uncorrelated. As in our network throughput study, no explanatory variable, x_{t-1} or y_{t-1}, is found significant such that can be used to regress the network throughput fluctuation y_t by the stepwise method, since both $\mathrm{cov}(x_{t-1}, y_t)$ and $\mathrm{cov}(y_{t-1}, y_t)$ are zero. Furthermore, in the regression analysis of the change of transfer time y_t, a convenient and simplified regression model which has only one explanatory variable, x_t, is created. Because of the linear relationship between x_t and y_t, x_t is easily identified with statistical significance and hence enters into the regression model. But the other variable, the global index z_t, is not considered as significant in explaining y_t since no linear relation between them is detected, i.e., $\mathrm{cov}(y_t, z_t) = 0$. In short, at least one explanatory variables are skipped in the above two regression models, and therefore omitted variable bias occurs.

By detecting weak dependence and applying this dependence rather than the linear correlation as the criterion to determine whether an explanatory variable should enter a regression, our study is attempting to build a more appropriate regression model which not only fits better the observations but also reduces the estimation errors such as omitted variables bias. If two variables, the explanatory variable X and the response variable Y, are tested as dependent through the after-transform correlation test, we should take into account the role of X playing in the regression of Y. In other words, even if X is determined as an insignificant factor in explaining Y by a variable entry method that identifies the explanatory variables by the significance of the linear correlation, such as the stepwise technique, the nonlinear regressing relationship between X and Y should be considered given that X and Y are tested dependent. In particular, if X and Y are weakly dependent, only our after Box-Cox transformation test rather than other general indepen-

dence tests works efficiently in detecting such dependence relationship and consequently allows to build a right regression model.

7.2 Weak Signal Identification

Suppose x is a source signal coming from a known signal set $\{\xi_1, \cdots, \xi_m\}$. If x is not sent, we will receive a signal $y = n$, where n is a noise. If x is sent, it will be attenuated after transmission. Therefore, we will receive a signal $y = kx + n$, where k is the attenuation. We wish to determine whether or not $x = \xi_i$ is sent by analyzing the received signal y, i.e., signal identification. It is basically a problem of signal detection, but the source signal x is one of the known signals $\{\xi_1, \cdots, \xi_m\}$. When the attenuation k is very small, the received signal becomes a weak signal as the useful signal is vanishingly small compared to the noise disturbance [94] [95], and we call the problem of finding this weak signal as weak signal identification.

In nature, a fish possesses a fascinating ability to find its companions by recognizing their feeding sounds in an extremely noisy environment [96]. Suppose that groups of fish may appear in a few observation points and their feeding sounds at these points are recorded. A fish in a distant place is observed. If we can determine the connection between the sound that this fish hears and the feeding sounds recorded in the remote observation points, we can predict which direction this fish will go to join its group. Due to the long distance, we realize that the feeding sound of groups of fish is buried in the environmental noise. Therefore, it will be a problem to identify this weak but useful sound signal from the noisy sound that we receive.

Weak signal identification can be considered as a special case of weak signal detection. They both address the problem of finding a weak signal. In weak signal detection, the signal that needs to be detected is usually unobserved and unknown, while in weak signal identification, the signal that needs to be identified is provided. For example, in the identification process of fish-feeding activity, the feeding sounds of m observation points are measured. Based on this information, we can predicate which point the fish will head to by analyzing the noise-corrupted signal it hears.

In many real applications, systems are commonly designed to satisfy the minimum requirements of technique, costs, privacy, etc. These factors easily result in weak signal conditions. Therefore, it is important to solve the weak signal identification problems which are frequently en-

countered in practice and are more difficult to identify than moderately strong and strong signals. For example, it will be helpful for diagnosis of the disease if the link between some protein level in blood and the disease can be established at early stage.

In the following section, we use a weak signal identification example to show how the classic identification criteria and after-transform correlation test perform. The improvement of probability of identification when using the after-transform correlation test suggests that our method is a possible good candidate solution in identifying a weak signal.

7.2.1 Identification Criteria

A hypothesis test is conducted to identify weak signals. H_0 corresponds to "no signal" and H_1 corresponds to "signal found". That is,

$$\begin{aligned} H_0: &\quad y = n, \\ H_1: &\quad y = s + n, \end{aligned} \tag{7.3}$$

where y is the observation, s is the useful signal, and n is the environment noise. For example, y is the sound that the fish hears and s is the attenuated feeding sound from one of the observation points. When the measurement y, which is corrupted by the noise n, is observed, we wish to determine whether it contains the signal of interest.

In detection theory, specific criteria, such as maximum *a posteriori*, Bayes and Neyman-Pearson, are introduced to address the hypotheses in (7.3). We restate these criteria as follows.

1. Maximum *a posteriori*(MAP) criterion

   Let $P(H_i|y), i = 0, 1$, denote the *a posteriori* probability, i.e., the probability of each hypothesis holds given the observation y. If $P(H_0|y) > P(H_1|y)$, we say no signal is found. Otherwise, we decide a signal is identified.

   Assume the *a priori* probability $P(H_i)$ is known, that is, $P(H_0) = \pi_0$ and $P(H_1) = \pi_1$. Using the Bayes and total probability theorem, we have

$$\frac{p(y|H_1)}{p(y|H_0)} \underset{D_0}{\overset{D_1}{\gtrless}} \frac{\pi_0}{\pi_1}, \tag{7.4}$$

where D_0 is assumed to be the event associated with the decision of choosing H_0 and D_1 be the corresponding event associated with H_1, $p(y|H_1)/p(y|H_0) = L(y)$ is the likelihood ratio, and $\pi_0/\pi_1 = \tau_{MAP}$ is the threshold. D_0 occurs when y falls in the decision region R_0, i.e., $L(y) < \tau_{MAP}$, while D_1 occurs when y falls in the decision region R_1, i.e., $L(y) \geq \tau_{MAP}$. The corresponding MAP decision rule is summarized as

$$\delta_{MAP} = \begin{cases} 1, & \forall\, y \in R_1 \text{ or } L(y) \geq \tau_{MAP} \\ 0, & \forall\, y \in R_0 \text{ or } L(y) < \tau_{MAP}. \end{cases}$$

The error probability when making a specific decision is of particular interest in many applications. Suppose P_{ij} is the probability of deciding D_i when hypothesis H_j is correct, i.e.,

$$P_{ij} = \int_{R_i} p(y|H_j)\, dy. \tag{7.5}$$

Regarding to the hypotheses in (7.3), we have

- P_{00} is the probability of deciding D_0 when hypothesis H_0 is correct;
- P_{01} is the probability of "miss", P_m. That is, a signal is present, i.e., H_1 is correct, and we have missed it by deciding D_0;
- P_{10} is the probability of "false alarm", P_f. That is, a signal is not present, i.e., H_0 is correct, but we thought it showed up by deciding D_1;
- P_{11} is the probability of identification, P_i. That is, we decide that a signal is found when it is actually present.

Among them, P_{01} and P_{10} correspond to errors. Therefore, the average error probability is

$$P_E = \pi_0 P_{10} + \pi_1 P_{01}.$$

2. Bayes criterion

The MAP criterion does not take into account the fact that some decisions might be more important than others and some errors could be more costly than others. The Bayes criterion

solves this issue by choosing the decision which minimizes the Bayes risk r,

$$r = (P_{00}C_{00} + P_{10}C_{10})\pi_0 + (P_{01}C_{01} + P_{11}C_{11})\pi_1,$$

where C_{ij} and P_{ij} are the cost and the probability of deciding D_i when the hypothesis H_j is actually correct, and π_i is the *a priori* probability. The corresponding Bayes decision rule is:

$$\delta_B = \begin{cases} 1, & L(y) \geq \tau_B \\ 0, & L(y) < \tau_B \end{cases}$$

where $\tau_B = \frac{\pi_0(C_{10}-C_{00})}{\pi_1(C_{01}-C_{11})}$.

3. Neyman-Pearson criterion

When the costs and *a priori* probabilities are difficult or impossible to assign, the Neyman-Pearson (NP) criterion is applied. The idea of NP is to maximize the probability of identification P_i when the false-alarm probability P_f is constrained below some specified level α_f ($P_f = P(D_1|H_0) \leq \alpha_f$). The NP decision rule is:

$$\delta_{NP} = \begin{cases} 1, & L(y) > \tau_{NP} \\ \eta, & L(y) = \tau_{NP} \\ 0, & L(y) < \tau_{NP} \end{cases}$$

If $L(y)$ exceeds the threshold τ_{NP}, decide D_1; if $L(y)$ is less than τ_{NP}, decide D_0. However, if $L(y) = \tau_{NP}$, then we have η chance to decide D_1. The threshold τ_{NP} and η are chosen such that $P_f = \alpha_f$.

The MAP needs the *a priori* to form a decision, the Bayes asks for both *a priori* and costs to set the threshold, and the NP requires neither cost nor *a priori* for its formulation. These criteria are then applied in next section to investigate the performance in identifying weak signal.

7.2.2 Simulation Results

Regarding to the hypothesis test of (7.3), we assume the noise n is a white Gaussian noise with zero mean and σ^2 variance. We also assume that the source signal x and the noise n are statistically independent, and x is a Gaussian signal with zero mean and σ^2 variance. When y is received, the useful signal s is buried in the noise n. For simplicity, we let $s = kx$, where k is the attenuation representing the strength of the useful signal.

Based on (7.3), we have

$$p(y|H_0) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right),\ p(y|H_1) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{y^2}{2\sigma_1^2}\right),$$

where $\sigma_1^2 = (k^2+1)\sigma^2$, and the likelihood ratio

$$L(y) = \frac{p(y|H_1)}{p(y|H_0)} = \frac{\sigma}{\sigma_1}\exp\left(\frac{y^2}{2\sigma^2} - \frac{y^2}{2\sigma_1^2}\right). \tag{7.6}$$

Solve $L(y) \gtrless \tau$, where τ is the threshold τ_{MAP}, τ_B or τ_{NP} . The decision regions are:

$$\begin{aligned} R_0 &= \{y : -\sqrt{\upsilon} < y < \sqrt{\upsilon}\} \\ R_1 &= \{y : y \geq \sqrt{\upsilon} \text{ or } y \leq -\sqrt{\upsilon}\} \end{aligned} \quad \text{where } \upsilon = \frac{\ln\tau\frac{\sigma_1}{\sigma}}{\frac{1}{2\sigma^2} - \frac{1}{2\sigma_1^2}}.$$

The corresponding probability P_{ij} are

$$\begin{aligned} P_{00} &= \textstyle\int_{R_0} p(y|H_0)\,dy = \int_{-\sqrt{\upsilon}}^{\sqrt{\upsilon}} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right)dy = \mathrm{erf}(\frac{\sqrt{\upsilon}}{\sqrt{2}\sigma}) \\ P_{01} &= \textstyle\int_{R_0} p(y|H_1)\,dy = \int_{-\sqrt{\upsilon}}^{\sqrt{\upsilon}} \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{y^2}{2\sigma_1^2}\right)dy = \mathrm{erf}(\frac{\sqrt{\upsilon}}{\sqrt{2}\sigma_1}) \\ P_{10} &= \textstyle\int_{R_1} p(y|H_0)\,dy = \left(\int_{-\infty}^{-\sqrt{\upsilon}} + \int_{\sqrt{\upsilon}}^{\infty}\right)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right)dy = 1 - \mathrm{erf}(\frac{\sqrt{\upsilon}}{\sqrt{2}\sigma}) \\ P_{11} &= \textstyle\int_{R_1} p(y|H_1)\,dy = \left(\int_{-\infty}^{-\sqrt{\upsilon}} + \int_{\sqrt{\upsilon}}^{\infty}\right)\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{y^2}{2\sigma_1^2}\right)dy = 1 - \mathrm{erf}(\frac{\sqrt{\upsilon}}{\sqrt{2}\sigma_1}). \end{aligned}$$

Assume the *a priori* is $\pi_0 = \frac{2}{3}$ and $\pi_1 = \frac{1}{3}$. By MAP criterion, $\tau_{MAP} = \frac{\pi_0}{\pi_1} = 2$ and $\upsilon = \frac{\ln 2\frac{\sigma_1}{\sigma}}{\frac{1}{2\sigma^2} - \frac{1}{2\sigma_1^2}}$.

Assign the costs as $C_{00} = C_{11} = 0, C_{10} = 1$, and $C_{01} = 2$. That is, there is no cost to make a correct decision, and the cost of deciding D_0 when H_1 is correct is twice as large as the converse. Using Bayes criterion, $\tau_B = \frac{\pi_0(C_{10}-C_{00})}{\pi_1(C_{01}-C_{11})} = 1$ and $\upsilon = \frac{\ln\frac{\sigma_1}{\sigma}}{\frac{1}{2\sigma^2}-\frac{1}{2\sigma_1^2}}$.

Assume the false-alarm specification $\alpha_f = 0.1$. By letting $P_f = \alpha_f$, we know that $\upsilon = \sqrt{2}\sigma\text{erfinv}(1-\alpha_f)$ for the NP criterion.

The identification results by applying the decision criteria of MAP, Bayes and NP are displayed in Table 7.1, where the signal strength is adjusted by the attenuation k. Without loss of generality, the variance of noise is assumed to be $\sigma^2 = 1$.

The table shows the average probability of error P_e, the false-alarm probability P_f, and the probability of identification P_i that are achieved under each of the three criteria. The attenuation $k = 0.01, 0.05, 0.1, 1, 10$ corresponds to the signal to noise ratio SNR $= 20\log_{10} k$, that is, -40dB, -26dB, -20dB, 0dB, and 20dB, respectively. With the enhancement of the SNR, a stronger signal is observed and therefore it is more easily identified. For example, as the amplitude of the useful signal is 10 times than the amplitude of the noise, i.e., SNR $= 20$dB, a high probability of identification P_i (above 0.8) and low error probabilities P_e, P_f (most are below 0.1) are achieved. However, as the SNR decreases and the signal of interest becomes weaker, the performance of identification using classic criteria degrades. When the SNR is low, for example, SNR $= -40$dB, no criteria in Table 7.1 works. The highest $P_i = 0.3333$ when the Bayes criterion is used. Therefore, we conclude that weak signals cannot be efficiently identified through the traditional MAP, Bayes or NP criteria. In such cases, the after-transform correlation test provides an alternative to identify weak signals with more power.

Table 7.1: Results of classical identification criteria

	MAP			Bayes			NP		
	P_i	P_f	P_e	P_i	P_f	P_e	P_i	P_f	P_e
$k = 0.01$	0.0000	0.0000	0.3333	0.3173	0.3173	0.4391	0.1000	0.1000	0.3667
$k = 0.05$	0.0000	0.0000	0.3333	0.3176	0.3170	0.4388	0.1004	0.1000	0.3665
$k = 0.1$	0.0000	0.0000	0.3333	0.3185	0.3161	0.4379	0.1017	0.1000	0.3661
$k = 1$	0.1493	0.0414	0.3112	0.4051	0.2390	0.3577	0.2448	0.1000	0.3184
$k = 10$	0.8065	0.0138	0.0737	0.8299	0.0309	0.0773	0.8700	0.1000	0.1100

Suppose H_1 is correct, that is, "signal found". The received signal $y = s + n = kx + n$ where the source signal x and the noise obey a Gaussian distribution with zero mean and unit variance. x and y are observations of two random variables X and Y respectively. If k is small, X and Y are weakly dependent. Now assume that M signals are received, we are concerned with how many times the signal is believed to be present, that is, P_i. In terms of the terminology of our correlation test, it is equivalent to saying what the rejection rate γ is when the observations of X and Y are collected M times. Recall that γ is defined as the number of times that X and Y are decided to be dependent among these M times. If X and Y are considered as dependent, we conclude "signal found".

Figure 7.5 is the rejection rate comparison between before the Box-Cox transformation (dashdot line) and after the Box-Cox transformation (star line) when the significance level $\alpha = 0.1$. Since the signal is buried in the noise, its trace is hard to be found. This explains, in the dashdot line, why γ is small when k is small, especially when $k < 0.1$. The Box-Cox transformation enhances the linear relationship between the original signal x and the received signal y. The weak relationship is enlarged to some extent and therefore is more easily detected. The improvement is illustrated from the larger γ of the star line.

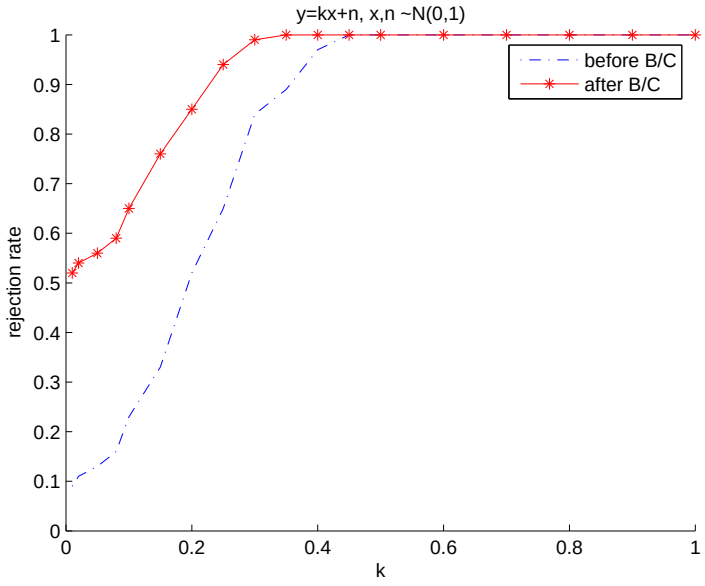

Figure 7.5: Rejection rate comparison before/after Box-Cox transformation

Consider the rejection rate γ, computed from the after-transform correlation test of $Y = kX + N$, as the probability of identification P_i. The comparison of P_i using the three classical identification criteria and the after-transform correlation test is shown in Figure 7.6. When the SNR is low, between -40dB to -20dB, none of the tests shows very good performance. Using the three classical identification criteria, the best chance to decide that the signal is present when it actually is is about 1/3. However, in such an enormous noise environment, the after-transform correlation test increases the identification probability to more than 0.5. As the signal strength gets stronger, that is, the SNR is between -20dB to -10dB, an apparent increase of P_i in the case of the correlation test emerges, while in other cases only tiny changes of P_i are observed. When the SNR is close to 0dB, the after-transform correlation test shows that $P_i = 1$. But the identification probabilities using the classical criteria slightly increase to 0.4 even if the signal and the noise are comparable. $P_i \approx 0.4$ does not meet the practical requirement.

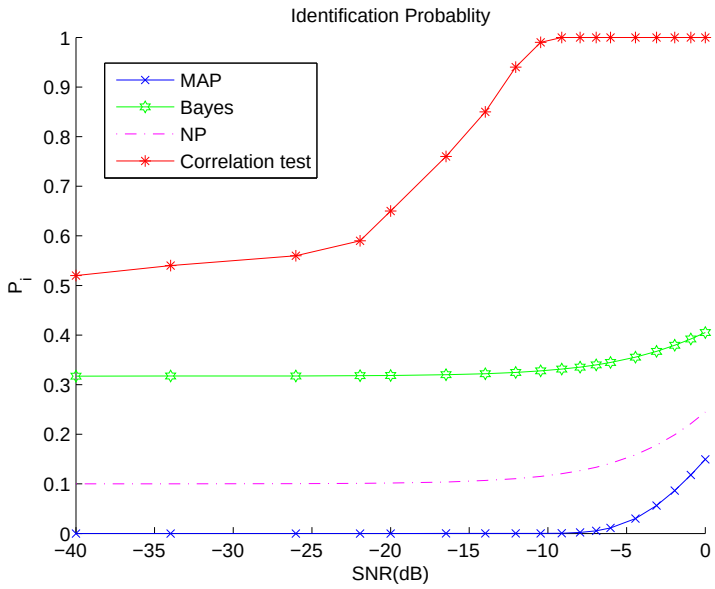

Figure 7.6: Comparison of identification probability

As we know, one single measure of performance is not sufficient to judge the identification criteria. Analogous to the above analysis, the false alarm probability P_f can be computed for the after-transform correlation test. Suppose H_0 is correct, that is, " no signal". We have $y = n$. An

after-transform correlation test is conducted. The resulting rejection rate $\gamma = 0.41$ is considered as the false alarm probability P_f since γ is the proportion of counting how many times X and Y are dependent, i.e., "signal found", among the total times. From the definition of (7.5), we see that the false alarm probability P_{10} and the identification probability P_{11} are calculated over the same decision region R_1, therefore it is always true that increasing P_i also tends to increase P_f. It explains the large P_f of the correlation test, in which P_i is also larger than that of other criteria. However, the false-alarm probability of 0.41 is large, especially when it is compared to the P_i when SNR is extremely low. For example, $P_i = 0.52$ when SNR$= -40$dB. The value of the correlation coefficient itself can provide more evidence in solving this problem.

Consider two models, where X and Y are variables of interest. Model 1: $Y = kX + N$, where $k = 0.01$ and N is the noise. Model 2: $Y = N$, X and Y are independent. Both X and N are Gaussian processes with zero mean and unit variance. $P_i = 0.52$ implies that the rejection rate is 0.52 when model 1 is tested. The result of $\gamma = 0.52$ shows that we only have sufficient evidence to say that the Pearson correlation coefficient between X and Y is nonzero 52 times among 100 experiments, when the correlation between X and Y is actually nonzero. On the other hand, $P_f = 0.41$ implies that the rejection rate is 0.41 when model 2 is tested. The result of $\gamma = 0.41$ shows that we actually have evidence to say that the Pearson correlation coefficient between X and Y is nonzero 41 times among 100 experiments, when the correlation between X and Y is in fact zero. In model 1, the larger γ, the better the result. In model 2, the smaller γ, the better the result. We wish the difference between the rejection rates in two models is equal to 1. However, it is practically impossible. Figure 7.7 gives us some clue for explaining the rejection rate with the actual correlation coefficient values recorded in the simulation. The left side plot of Figure 7.7 shows the sample r between X and Y after the Box-Cox transformation in 100 experiments. Sorting the correlation coefficients in an increasing order, we can clearly see the fluctuation of r over zero in the right side plot of Figure 7.7. The solid curve of model 2 shows that half of the r values are below 0 while another half values are above 0. It is a hint that the correlation between X and Y might be zero. On the other hand, the dashdot curve shows that about 75% of the correlation coefficients are above 0. Combining with the result that $\gamma = 0.52$, we are more likely to believe that, in model 1, X and Y are correlated and dependent. In other words, the decision of D_0 or D_1 is enforced by considering the fluctuation of sample r as an extra evidence in addition to the normal parameters P_i, P_f.

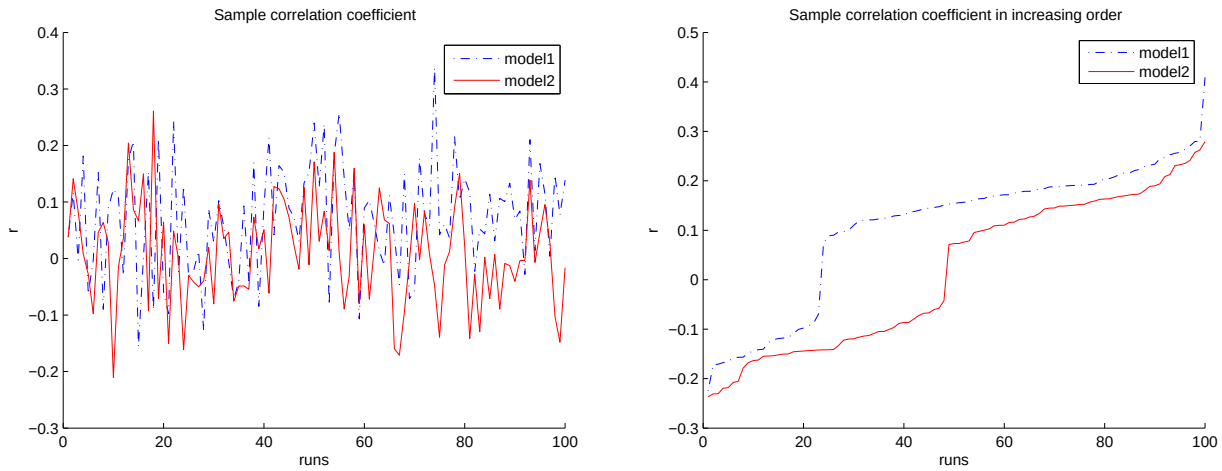

Figure 7.7: Sample correlation coefficient

In weak signal identification, the performance of the three classical detection criteria greatly degenerate. The simple example discussed in this section suggests that the after-transform correlation test might work as an efficient alternative in such cases to improve the identification probability.

CHAPTER 8. CONCLUSIONS

In this study, we present a dependence test using the Pearson correlation coefficient based on the Box-Cox transformation of random variables. This test represents a substantial performance improvement in weak dependence cases as compared to the classical independence tests using the Cramer-Von Mises distance, the Kolmogorov-Smirnov distance, or other criteria.

Our test statistic is conceptually straightforward and computationally simple. It is one of the essential features of probability and statistical theory that the dependence relationship between the after-transformation random variables may imply the dependence relationship between the original random variables. Moreover, our correlation test makes very few assumptions about the underlying distributions by using the bootstrap method. It is also fast since no time-consuming estimation of joint distribution or density is required.

We suggest using the semi-parametric estimation method when the joint distribution or density is needed to obtain a test statistic like CM or RW, i.e., estimating F_{XY} or f_{XY} by a copula together with the NW density estimator. This combination avoids the dimensionality curse of nonparametric estimation, and sustains fewer distribution restrictions of parametric estimation as well. Speaking of our simulation, the independence test using the normal copula after the Box-Cox transformation not only shows better performance but also saves 80% computation time compared to the same test using the double kernel local linear method.

To see its potential applications, we introduce two cases where our after-transformation correlation test would be helpful. In one case, by identifying the weak dependence through our test, we can avoid the omission of explanatory variables in regression analysis. In the other case using fish-feeding activity as the example, our method might explain how a useful message buried in noises can be found.

Although our test shows performance variance in different cases in this study, it presents a simple, convenient and reasonable method to solve a problem that is rather difficult but has not received sufficient attention by researchers in this field, that is, how to distinguish weak dependence from independence.

REFERENCES

[1] Tjøstheim, D., 1996. "Measures of dependence and tests for independence." *Statistics,* ***28***, pp. 249–284.

[2] Hoeffding, W., 1948. "A nonparametric test of independence." *The Annals of Mathematical Statistics,* ***19***, Dec, pp. 546–557.

[3] J. R. Blum, J. K., and Rosenblatt, M., 1961. "Distribution free tests of independence based on the sample distribution function." *Annals of Mathematical Statistics,* ***32***, pp. 485–498.

[4] Skaug, H. J., and Tjøstheim, D., 1993. "A nonparametric test of serial independence based on the empirical distribution function." *Biometrika,* ***80***, pp. 591–602.

[5] Rosenblatt, M., 1975. "A quadratic measure of deviation of two-dimensional density estimates and a test of independence." *Annals of Statistics,* ***3***, pp. 1–14.

[6] An, H., and Cheng, B., 1991. "A kolmogorov-smirnov type statistic with application to test for nonlinearity in time series." *International Statistical Review,* ***59***, pp. 287–307.

[7] Skaug, H. J., and Tjøstheim, D., 1996. "Measures of distance between densities with application to testing for serial independence." In *Time Series Analysis in Memory of E. J. Hannan*, P. Robinson and M. Rosenblatt, eds. Springer, New York.

[8] Robinson, P. M., 1991. "Consistent nonparametric entropy-based testing." *Review of Economic Studies,* ***58***, May, pp. 437–453.

[9] Skaug, H. J., and Tjøstheim, D., 1993. "Nonparametric tests of serial independence." In *Developments in Time Series Analysis*, M. B. Priestley and T. S. Rao, eds. Chapman and Hall, London, pp. 207–229.

[10] Hart, J. D., 1997. *Nonparametric Smoothing and Lack-of-fit Tests*. Springer & Verlag, New York, USA.

[11] Fernandes, M., 2001. "Nonparametric entropy-based tests of independence between stochastic processes.".

[12] Zhang, G., and Taniguchi, M., 1994. "Discrimination analysis for stationary time series." *Joural of Time Series Analysis,* ***15***, pp. 117–126.

[13] W. A. Brock, W. D. D., Scheinkman, J. A., and LeBaron, B., 1987. "A test for independence based on the correlation dimension.".

[14] M. Hallin, J. Jurečková, J. P., and Zahaf, T., 1999. "Nonparamtric tests of independence of two autoregressive time series based on autoregression rank scores." *Journal of Statitical Planing and Inference,* ***75***, Jan, pp. 319–330.

[15] Pinkse, J., 1998. "A consistent nonparametric test for serial independence." *Joural of Econometrics,* ***84***, pp. 205–231.

[16] W. A. Brock, W. D. Dechert, J. A. S., and LeBaron, B., 1996. "A test of independence based on the correlation dimension." *Econnometric Reviews,* ***15***, pp. 197–235.

[17] Su, L., and White, H., 2008. "A nonparametric hellinger metric test for conditional independence." *Econometric Theory,* ***24***, pp. 829–864.

[18] Box, G. E., and Cox, D. R., 1964. "An analysis of transformations." *Journal of the Royal Statistical Society,Series B,* ***26***, pp. 211–252.

[19] Box, G. E., and Cox, D. R., 1982. "An analysis of transformations revisited,rebutted." *Journal of American Statistical Association,* ***77***, pp. 209–210.

[20] Doukhan, P., and Louhichi, S., 1999. "A new weak dependence condition and application to moment inequalities." *Stochastic Processes and Their Applications,* ***84***, pp. 313–343.

[21] J. Dedecker, P. Doukhan, G. L. J. R. L. S. L., and Prieur, C., 2007. *Weak Dependence: With Examples and Applications*. Springer Science, New York, USA.

[22] Nze, P. A., and Doukhan, P., 2004. "Weak dependence: Models and applications to econometrics." *Econometric Theory,* ***20***, pp. 995–1045.

[23] Cohen, J., 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, New Jersey, USA.

[24] Sethuraman, J., 1990. "The asymptotic distribution of the renyi maximal correlation." *Communications in Statistics - Theory and Methods,* ***19***, pp. 4291–4298.

[25] Kumar, G., 2010. "Binary renyi correlation." *Information Theory*, July.

[26] Bollerslev, T., 1986. "Generalized autoregressive conditional heteroskedasticity." *Econometrics,* ***31***, pp. 307–327.

[27] Wand, M. P., and Jones, M. C., 1995. *Kernel Smoothing*. Chapman& Hall, London, UK.

[28] Silverman, B. W., 1986. *Density Estimation*. Chapman& Hall, London, UK.

[29] Tarter, M. E., and Lock, M. D., 1994. *Model-Free Cure Estiamtion*. Chapman& Hall, London, UK.

[30] Marron, J. S., and Ruppert, D., 1994. "Transformation to reduce boundary bias in kernel density estimation." *Royal Statistical Society,* ***56***(4), pp. 653–671.

[31] Schuster, E. F., 1985. "Incorporating support constraints into nonparametric estimators of densities." *Communication Statistics,* ***14***, pp. 1123–1136.

[32] Cowling, A., and Hall, P., 1996. "On pseudodata methods for removing boundary effects in kernel density estimation." *Royal Statistical Society,* ***58***, pp. 551–563.

[33] Jones, M. C., Linton, O., and Nielsen, J. P., 1995. "A simple bias reduction method for density estimation." *Biometrika,* ***82***, Jun, pp. 327–338.

[34] Fan, J., and Yao, Q., 2003. *Nonlinear Time Series:Nonparametric and Parametric Methods*. Springer & Verlag, New York, USA.

[35] J. Fan, Q. Y., and Tong, H., 1996. "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems." *Biometrika Trust,* ***83***, pp. 189–206.

[36] Fan, J., 1992. "Design-adaptive nonparametric regression." *Joural of the American Staitstical Association,* ***87***, Dec, pp. 998–1004.

[37] Chu, C., and Marron, J. S., 1991. "Choosing a kernel regression estimator." *Statistical Science,* ***6***, Nov, pp. 404–419.

[38] Gasser, T., and Muller, H. G., 1979. "Kerel estimation of regression functions." In *Smoothing Technuques for Curve Estimation*, T. Gasser and M. Rosenblatt, eds. Springer-Verlag, pp. 23–68.

[39] Moon, T., and Stirling, W. C., 2000. *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall, New Jersey, USA.

[40] Fan, J., and Yim, T. H., 2004. "A crossvalidation method for estimating conditional densities." *Biometrika,* ***91***, pp. 819–834.

[41] Bashtannyk, D. M., and Hyndman, R. J., 2001. "Bandwidth selection for kernel conditional density estimation." *Computational Statistics and Data Analysis,* ***36***, May, pp. 279–298.

[42] Hyndman, R. J., and Yao, Q., 2002. "Nonparametric estimation and symmetry tests for conditional density functions." *Nonparametric Statistics,* ***14***, pp. 259–278.

[43] Hardle, W., 1991. *Smoothing Techniques: With Implementation in S*. Springer Statistics, New York, USA.

[44] Embrechts, P., Hoing, A., and A.Juri, 2003. "Using copulae to bound the value-at-risk for functions of dependent risks." *Finance and Stochastics,* ***7***, Apr, pp. 145–167.

[45] Frees, E. W., and Wang, P., 2006. "Copula credibility for aggregate loss models." *Insurance: Mathematics and Economics,* ***38***(2), pp. 360–373.

[46] Frees, E. W., and Valdez, E. A., 1998. "Understanding relationships using copulas." *North American Actuarial Journal,* ***2***, Jan, pp. 1–25.

[47] Sklar, A., 1996. "Random variables, distribution functions, and copulas-a personal look backward and forward." *Institute of Mathematical Statistics Hayward,* ***31***, Sep, pp. 1–14.

[48] Nelsen, R. B., 2006. *An Introduction to Copulas*. Spring Science+Business Media, New York, NY, USA.

[49] Cherubini, U., Luciano, E., and Vecchiato, W., 2004. *Copula Methods in Finance*. John Wiley & Sons, West Sussex, England.

[50] Genest, C., and Mackay, J., 1986. "The joy of copulas:bivariate distributions with uniform marginals." *American Statistica,* ***40***, Nov, pp. 280–283.

[51] Genest, C., and Rivest, L., 1993. "Statistical inference procedures for bivariate archimedean copulas." *American Statistical Association,* ***88***, Sep, pp. 1034–1043.

[52] Melchiori, M. R., 2003. "Which archimedean copula is the right one?." *YieldCurve*, Sep.

[53] Joe, H., and Xu, J. J., 1996. "The estimation method of inference functions for margins for multivariate models." *Dept. of Statistics University of British Columbia,* ***166***.

[54] Joe, H., 1997. *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

[55] Hall, P., 1992. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York, USA.

[56] Efron, B., and Tibshirani, R., 1993. *An introduction to the bootstrap*. Chapman & Hall, New York, USA.

[57] Efron, B., 1979. "Bootstrap methods: Another look at the jackknife." *The Annals of Statistics,* ***7***, pp. 1–26.

[58] D. S. Moore, G. P. M., and Craig, B. A., 2009. *Introduction to the Practice of Statistics*. W.H.Freeman, New York, NY, USA.

[59] Rogers, J. L., and Nicewander, W. A., 1988. "Thirteen ways to look at the correlation coefficient." *The American Statistician,* ***42***, Feb, pp. 59–66.

[60] Stirling, W. C., 2008. *Notes for Probability and Probabilistic Reasoning for Electrical Engineering by Fine*. Unpublished.

[61] Fisher, R. A., 1915. "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population." *Biometrika,* ***10***, pp. 507–521.

[62] Fisher, R. A., 1921. "On the 'probable error' of a coefficient of correlation deduced from a small sample." *Metron,* ***1***, pp. 3–32.

[63] Olkin, I., and Pratt, J. W., 1958. "Unbiased estimation of certain correlation coefficients." *Annals of Mathematical Statistics,* ***29***, pp. 201–211.

[64] D. W. Zimmerman, B. D. Z., and Williams, R. H., 2003. "Bias in estimation and hypothesis testing of correlation." *Psicológica,* ***24***, pp. 133–158.

[65] Freeman, J., and R.Modarres, 2005. "Efficiency of test for independence after box-cox transformation." *Journal of Multivariate Analysis,* ***95***, pp. 107–118.

[66] SAKIA, R. M., 1992. "The box-cox transformation technique: a review." *The Statistician,* ***41***, pp. 169–178.

[67] Draper, N. R., and Cox, D. R., 1969. "On distributions and their transformation to normality." *Journal of the Royal Statistical Society,Series B,* ***31***, pp. 472–476.

[68] Chow, S. C., and Wang, S. G., 1993. *Advanced Linear Models: Theory and Applications*. CRC Press.

[69] "Engineering statistics handbook." In *National Institute of Standards and Technology*. http://www.itl.nist.gov/div898/handbook/.

[70] Shore, H., 2005. *Response Modeling Methodology: Empirical Modeling for Engineering and Science*. World Scientific Publishing Company.

[71] M. J. Gurka, L. J. Edwards, K. E. M., and Kupper, L. L., 2006. "Extending the box-cox transformation to the linear mixed model." *Jouranl of Royal Statistics,* ***169***, pp. 273–288.

[72] Freund, J. E., and Simon, G. A., 1992. *Statistics*. Prentice Hall, New Jersey, USA.

[73] F.Triola, M., 2004. *Elementary Statistics*. Pearson Education, New Jersey, USA.

[74] Kirk, R. E., 2007. *Statistics*. Cengage Learning, USA.

[75] Zhou, B., 1996. "High-frequency data and volatility in foreign-exchange rates." *Journal of Business and Economic Statistics,* ***14***, Jan, pp. 45–52.

[76] Lundbergh, S., and Terasvirta, T., 1999. "Modelling economic high-frequency time series with star-stagarch models." *Working Paper Series in Economics*.

[77] Martin, M. A., 2007. "Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties." *Computational Statistics and Data Analysis,* ***51***, pp. 6321–6342.

[78] Beran, R., 1986. "Simulated power functions." *The Annals of Statistics,* ***14***, pp. 151–173.

[79] Gulati, S., and Neus, J., 2003. "Goodness of fit statistics for the exponential distribution when the data are grouped." *Communications in Statistics,* ***32***, Sep., pp. 681–700.

[80] M. Zelditch, D. L. Swiderski, H. D. S., and Fink, W. L., 2004. *Geometric Morphometrics for Biologists: a Primer*. Elsevier Academic Press.

[81] Parzen, A., and Hall, A., 1993. "Stationary time series analysis using information and spectrial analysis." *In Development in Time Series Analysis,* ***The Priestly Birthday Volume***, pp. 139–148.

[82] Wolff, R. L., 1994. "Independence in time series: Another look at the bds test." *Philosophical Transactions of the Royal Society of London,* ***348***, pp. 383–395.

[83] Tong, H., 1995. "A personal review of nonlinear time series from a chaos perspective." *Journal of Statistics,* ***22***, pp. 399–446.

[84] Rosenblatt, M., and Wahlen, B., 1992. "A nonparametric measure of independence under a hypothesis of independent components." *Statistics and Probability Letters,* ***15***, pp. 245–252.

[85] Linton, O., and Gozalo, P., 1997. Conditional independence restrictions: Testing and estimation.

[86] Cover, T. M., and Thomas, J. A., 2006. *Elements of Information Theory*. Wiley & Sons, New Jersy, USA.

[87] Furuichi, S., 2006. "Information theoretical properties of tsallis entropies." *Journal of Mathematical Physics,* ***47***.

[88] Chan, N. H., and Tran, L. T., 1992. "Nonparametric test for serial dependence." *Joural of Time Series Analysis,* ***13***, pp. 19–28.

[89] Maasoumi, E., and Racine, J., 2002. "Entropy and predictability of stock market returns." *Journal of Econometrics,* ***107***, pp. 291–312.

[90] Manly, B., 1976. "Exponential data transformations." *Journal of the Royal Statistical Society,* ***25***, pp. 37–42.

[91] Yeo, I. K., and Johnson, R. A., 2000. "A new family of power transformations to improve normality or symmetry." *Biometrika Trust,* ***87***, pp. 954–959.

[92] L. H, K., 2008. *Regression Basics*. SAGE, London, United Kingdom.

[93] Miles, J., and Shevlin, M., 2001. *Applying Regression and Correlation*. SAGE, London,United Kingdom.

[94] I. Song, J. B., and Kim, S. Y., 2002. *Advanced Theory of Signal Dectection*. Springer-Verlag, Berlin, Germany.

[95] "http://www.combat-fishing.com/bass_sense.htm.".

[96] Myrberg, J., and Arthur, A., 1972. "Effectiveness of acoustic signals in attracting epipelagic sharks to an underwater sound source." *Bulletin of Marine Science,* ***22***, pp. 926–949.

[97] Park, B. U., and Marron, J. S., 1992. "On the use of pilot estimators in bandwidth selection." *Nonparameteric Statistics,* ***1***, pp. 231–240.

[98] Jones, M. C., and Foster, P. J., 1993. "Generalized jackknifing and higher order kernels." *Nonparametric Statistics,* ***3***, pp. 81–194.

[99] W. R Schucany, H. L. G., and Owen, D. B., 1971. "On bias reduction in estimation." *Statistics Association,* ***66***, pp. 524–533.

[100] Rice, J. A., 1984. "Boundary modification for kernel regression." *Communication Statistics,* ***13***, pp. 893–900.

[101] Jones, M. C., 1993. "Simple boundary correction for kernel density estimation." *Statistics and Computing,* ***3***, pp. 135–146.

[102] Ruppert, D., and Cline, D. B., 1994. "Bias reduction in kernel density estimation by smoothed empirical transformations." *Statistics,* ***22***, Mar, pp. 185–210.

[103] Stone, M., 1974. "Cross-validatory choice and assessment of statittical predictions." *Journal of the Royal Statistical Society,* ***36***, pp. 111–147.

[104] Fan, J., and Gijbels, I., 1995. "Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatical adaptation." *Journal of the Royal Statistical Society,* ***57***, pp. 371–394.

[105] D. Ruppert, S. J. S., and Wand, M. P., 1995. "An effective bandwidth selector for local least square regression." *Joural of the American Staitstical Association,* ***90***, pp. 1257–1269.

[106] Magnus, J. R., and Neudecker, H., 1988. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley & Sons, New Jersy, USA.

[107] Francq, C., and Zakoian, J. M., 2010. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley.

[108] Lindsey, J. K., 2004. *Statistical Ananlysis of Stochastic Processes in Time*. Cambridge University Press, New York, USA.

[109] McCabe, B., and Tremayne, A., 1993. *Elements of Modern Asymptotic Theory with Statistical Applications*. Manchester University Press.

[110] T. G. Andersen, R. A. Davis, J. K., and Mikosch, T., 2009. *Handbook of Financial Time Series*. Springer.

[111] Rachev, S. T., and Fabozzi, F. J., 2007. *Financial Econometrics: from basics to advanced modeling techniques*. Wiley.

[112] Bougerol, P., and Picard, N., 1992. "Stationarity of garch processes and of some nonnegarive time series." *Journal of Econometrics,* ***52***, pp. 115–127.

[113] Lutkepohl, H., 2009. *Introduction to Multiple Time Series Analysis*. Springer-Verlag, New York.

[114] A. Aue, S. Hormann, L. H., and Reimherr, M., 2009. "Break detection in the covariance structure of multivariate time series models." *Annals of Statistics,* ***37***, pp. 4046–4087.

[115] Tong, H., 1990. *Nonlinear Time Series*. Oxford University Press.

[116] Chan, K. S., and Tong, H., 1985. "On the use of the deterministic lyapunov function for the ergodicity of stochastic." *Advances in Applied Probability,* ***17***, Sep., pp. 666–678.

[117] Tjostheim, D., 1990. "Nonlinear time series and markov chains." *Advances in Applied Probability,* ***22***, Sep., pp. 587–611.

[118] Harris, C. J., and Valenca, J. M., 1983. *The Stability of Input-Output Dynamical Systems*. Academic Press.

[119] Silvenonomen, A., and Terasirta, T., 2009. "Multivariate garch models." In *Handbook of Financial Time Series*, J. K. T. G. Anderson, R. A. Davis and T. Mikosch, eds. Springer, New York, pp. 190–229.

[120] Jeatheau, T., 1998. "Strong consistency of estimators for multivariate arch models." *Econometric Theory,* ***14***, pp. 70–86.

APPENDIX A. TWO ISSUES OF NW DENSITY ESTIMATION

A.1 Bandwidth Selection Rules

In Nadaraya-Watson kernel density estimation, the selection of the bandwidth parameter h is crucial to the estimation performance. Recalling the bandwidth selection criterion of asymptotic mean integrated squared error(AMISE) discussed in Section 3.1,

$$\text{AMISE}\{\hat{f}(.;h)\} = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') \text{ as } n \to \infty,$$

where $R(g) = \int g(x)^2\, dx, \mu_2(K) = \int x^2K(x)\, dx$, n is the sample size, and K is the kernel function. The optimal bandwidth h minimizes AMISE. Let $\partial\text{AMISE}\{\hat{f}(.;h)\}/\partial h = 0$, then

$$h_{\text{AMISE}} = \left[\frac{R(K)}{\mu_2(K)^2R(f'')n}\right]^{1/5}. \tag{A.1}$$

The exact h_{AMISE} cannot be obtained since the real f is unknown, and therefore the density functional ψ_r is applied to estimate it.

Good estimation of $R(f'')$ ensures a good bandwidth $\hat{h}$ resulting from (A.1). Define $R(f^{(s)}) = \int f^{(s)}(x)^2\, dx = (-1)^s \int f^{(2s)}(x)f(x)\, dx$. Let $\psi_r = \int f^{(r)}(x)f(x)\, dx = E\{f^{(r)}(X)\}$ for r is even. When r is odd, $\psi_r = 0$. Since $\psi_r = E\{f^{(r)}(X)\}$, it is natural to use the average as its estimate. The estimator of ψ_r is:

$$\hat{\psi}_r(g) = n^{-1}\sum_{i=1}^{n}\hat{f}^{(r)}(X_i;g) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}L_g^{(r)}(X_i - X_j), \tag{A.2}$$

where g and L are, respectively, a bandwidth and kernel that are probably different from h and K. DPI, STE and SCV rules are based on this density functional.

A.1.1 Normal Scale Rule

A normal scale bandwidth selector simply involves using the bandwidth that is AMISE-optimal for the normal density with the same scale as that estimated for the underlying density f. If f is normal with variance σ^2, then from (A.1),

$$h_{\text{AMISE}} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2 n}\right]^{1/5}\sigma.$$

The NS bandwidth selector is thus the outcome by replacing σ by $\hat{\sigma}$

$$\hat{h}_{\text{NS}} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2 n}\right]^{1/5}\hat{\sigma},$$

where $\hat{\sigma}$ could be the sample standard deviation s or the standard interquartile range $\hat{\sigma}_{IQR}$.

In general, the normal scale bandwidth selector $\hat{h}_{NS}$ is used as a quick and FIRST guess of h, an initial h value.

A.1.2 Direct Plug-in Rule

Using the asymptotic mean squared error(MSE) properties of ψ_r in (A.2), the optimal g is:

$$g_{\text{AMSE}} = \left[\frac{k!L^{(r)}(0)}{-\mu_k(L)\psi_{r+k}n}\right]^{1/(r+k+1)}. \tag{A.3}$$

Since $R(f'') = \psi_4$, replacement of ψ_4 by the kernel estimator $\hat{\psi}_4(g)$ in (A.1) leads to the direct plug-in(DPI) rule:

$$\hat{h}_{\text{DPI}} = \left[\frac{R(K)}{\mu_2(K)^2\hat{\psi}_4(g)n}\right]^{1/5}. \tag{A.4}$$

But $\hat{h}_{\text{DPI}}$ depends on the choice of g. One way of choosing g is to appeal to the formula for the AMSE-optimal bandwidth for estimating $\hat{\psi}_4(g)$. If the same kernel K is used in $\hat{\psi}_4(g)$, then from(A.3) the AMSE-optimal bandwidth is

$$g_{\text{AMSE}} = \left[\frac{2K^{(4)}(0)}{-\mu_2(K)^2\psi_6(g)n}\right]^{1/7}.$$

However, $\hat{\psi}_6(g)$ is unknown. The optimal bandwidth g for estimating ψ_r depends on ψ_{r+2}. Again, like the normal scale rule, we get the first ψ_r estimate by assuming that f is a normal density with variance σ^2, then for r even,

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\pi^{1/2}}.$$

$\hat{h}_{\mathrm{DPI}}$ can be finally calculated by (A.4). The solve-the-equation rule(STE) and the smooth cross-validation(SCV) rule are also based on the density functional ψ_r. Instead of using the AMSE-optimal bandwidth for estimation of $\hat{\psi}_r(g)$ as the selection of new bandwidth g, the STE rule assumes that $g = \gamma(h)$, and estimate the function γ. The SCV rule obtains the optimal bandwidth g by minimizing the exact integrated squared bias instead of its asymptotic approximation. More information about the bandwidth selection rules, please see [27, 28, 97].

A.2 Boundary Effect Solution

Another important issue regarding the Nadaraya-Watson(NW) kernel density estimator is that it always has considerable bias near the support boundary due to the lack of the knowledge of these edges. This poor edge behavior is known as the boundary effect. Many correction approaches have been introduced to alleviate this problem. The following are some typical approaches.

A.2.1 Reflection Method

As shown in Section 3.1, the NW kernel density estimator is defined as:

$$\hat{f}(x;h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - x_{s_i}}{h}\right), \tag{A.5}$$

where K is the chosen kernel function, h is the bandwidth and $x_{s_1}, \cdots, x_{s_n}$ are observed samples taken from a continuous, univariate density f.

Suppose the support of f is $[a,b]$. Schuster [31] proposed a straightforward correction method 'reflection' or 'boundary folding'. The estimation $\breve{f}$ is represented as:

$$
\begin{aligned}
\check{f}(x) &= \hat{f}(x) + \hat{f}(2a-x) && \text{if } x \in [a, a+h) \\
&= \hat{f}(x) && \text{if } x \in [a+h, b-h] \\
&= \hat{f}(x) + \hat{f}(2b-x) && \text{if } x \in (b-h, b] \\
&= 0 && \text{if } x \notin [a, b].
\end{aligned}
$$

The reflection method is easy to compute and implement, but the theoretical analysis indicates that 'reflection' performs well only if the first derivative of the density f is 0 at the boundary, otherwise the estimation error in the boundary is still much worse than that in the interior. This correction does not improve the discontinuities in derivatives of the density f and therefore is not adaptable for the higher-order kernel estimator.

A.2.2 Generalized Jack-knifing

Jones [98] provided a generalized jack-knifing method based on the work of Schucany [99]. As to the NW kernel density estimator $\hat{f}$ in (A.5), assume s is the right boundary of the density f and let $[-B_K, B_K]$ denote the support of kernel K. Define

$$
m_p(s) = \int_{-B_K}^{\min(s, B_K)} u^p K(u)\, du.
$$

A slightly better $\bar{f}$ results from local renormalization of dividing $\hat{f}$ by $m_0(s)$. Choose another kernel function L with support $[-B_L, B_L]$, and define another normalization parameter

$$
n_p(s) = \int_{-B_L}^{\min(s, B_L)} u^p L(u)\, du.
$$

A similar $\tilde{f}$ is obtained by dividing $\hat{f}$ with $n_0(s)$. Jones's generalized Jack-knifing found a linear combination

$$
\check{f} = \alpha \bar{f} + \beta \tilde{f}
$$

with good asymptotic bias properties, where

$$\alpha = n_1(s)m_0(s)/(n_1(s)m_0(s) - m_1(s)n_0(s)),$$

$$\beta = -m_1(s)n_0(s)/(n_1(s)m_0(s) - m_1(s)n_0(s)).$$

The jack-knifing method provides a unified framework and generates a number of boundary kernel formulae which have been widely used in many fields [100, 101], but they have a common substantial disadvantage in that jack-knifing brings the negative values near the boundary.

A.2.3 Pseudodata Method

As the name indicates, this method generates pseudodata outside the support of estimation. These compensating data are linear functions of order statistics in the original sample $x_{s_1}, \cdots, x_{s_n}$ with weights calculated in terms of the kernel order. Then, such pseudodata are added to the original data samples to make a boundary-corrected estimation. One simple example in [32] is listed as below:

$$x_{-i} = 3(c-1)x_{s_i} + (1-3c)x_{s_{2i}} + cx_{s_{3i}},$$

where c is an arbitrary constant. x_{-i} are generated as the pseudodata outside the left boundary 0 which construct our knowledge of the left boundary and allow us to assign probability density values at the end points of left boundary. Technical details of pseudodata generation rules and the corresponding numerical analysis are referred to [32]. Compared to the common reflection method, it is more appropriate for higher-order kernel estimation. However, there is no clear clue for choosing some important parameters which might greatly influence its correction performance.

A.2.4 Transformation Method

There are different ways to transform the data so that the boundary effect could be eventually reduced, such as [30, 102] . We use Marron and Ruppert's method [30] as an illustration.

Since the reflection method plays poorly in removing the boundary effect when the first derivative of the density f does not equal zero in the boundary, Marron and Ruppert's idea is

to find a transformation so that such restriction could be avoided after the transformation. Their method contains three steps:

- Select a transformation g so that the first derivative of the density of $y_i = g(s_i)$ is close to zero in the boundary. The possible transformation g could be a quartic polynomial, a cumulative density function(cdf), or a linear combination of the polynomial and the cdf.
- Deploy the NW kernel density estimation of y_i with a reflection method.
- By the substitution of variables, convert the estimator of y_i to the boundary-corrected estimation of x_{s_i}.

The transformation method increases the computation complexity, but its performance is analogous to the jack-knifing method and it is even better in the sense that there are no negative values in the resulting density estimation.

In summary, all boundary effect correction methods which are listed here, and other methods such as variable kernel method [33], have both advantages and disadvantages and therefore should be applied differently according to different situations.

APPENDIX B. LLM BANDWIDTH SELECTION

In Section 3.2, the double-kernel local linear method (LLM) is introduced to estimate the conditional density $f_{Y|X}(y|x)$, i.e.,

$$E\{K_{h_2}(Y-y)|X=x\} \to f_{Y|X}(y|x) \text{ as } h_2 \to 0,$$

where the kernel K is a nonnegative density function and $K_h(y) = K(y/h)/h$, h is a bandwidth.

The specification of $E\{K_{h_2}(Y-y)|X=x\}$ can be considered as a nonparametric regression problem. A regression model relates a response variable Y to an explanatory variable X with a best prediction function $m(x) = E[Y|X=x]$ by embracing the minimum mean square error(MMSE) criterion. Therefore, $E\{K_{h_2}(Y-y)|X=x\}$ is the regression function of the random variable $K_{h_2}(Y-y)$ given $X=x$. According to the principle of local linear regression, it is equivalent to minimizing

$$\sum_{t=1}^{T}[K_{h_2}(y_t-y)-\alpha-\beta(x_t-x)]^2 W_{h_1}(x_t-x), \tag{B.1}$$

where the weight kernel W is a nonnegative density function. The minimization turns out:

$$\hat{f}_{Y|X}(y|x) = \frac{1}{nh_1h_2}\sum_{t=1}^{T} W_T\left(\frac{x_t-x}{h_1};x\right)K\left(\frac{y_t-y}{h_2}\right), \tag{B.2}$$

where

$$W_T(z;x) = W(z)\frac{s_{T,2}(x)-zh_1s_{T,1}(x)}{s_{T,0}(x)s_{T,2}(x)-s_{T,1}(x)^2} \text{ and } s_{T,j}(x) = \frac{1}{T}\sum_{t=1}^{T}(x_t-x)^j W_{h_1}(x_t-x), \text{ for } j=0,1,2.$$

Two bandwidths h_1 and h_2 represent the smoothness in both X direction and Y direction. Select the bandwidth h_2 first, for example, according to the normal scale rule [28],

$$\hat{h}_2 = \left[\frac{8\pi^{1/2}\int K^2(x)\,dx}{3(\int x^2 K(x)\,dx)^2}\right]^{1/5}\sigma T^{-1/5},$$

where σ could be the sample standard deviation s_y.

Given the bandwidth h_2, the estimation of conditional density in (B.1) is a regression function of $K_{h_2}(Y-y)$ given $X=x$. The methods to estimate h_1 based on this idea include: the crossvalidation rule [103], the residual-square rule [104], the plug-in rule [105], penalized average rule [41, 43], etc.

B.1 Crossvalidation Rule

Hyndman and Yao bring forward a simple tractable implementation of bandwidth selection [42]:

$$h_1 = 0.935(n\sigma^5/n|d_1|^5)^{1/6},\ h_2 = |d_1|h_1,$$

under the conditions: 1) W and K are normal kernels; 2) The conditional density is assumed to be $N(d_0+d_1x,\sigma^2)$ and the marginal density of X is assumed to be $N(\mu,\nu^2)$. Due to these restrictions, the above selection rule is usually applied as a useful tool for the initial bandwidth estimator. Other more effective method needs to be employed to refine the bandwidth estimator, for example, the crossvalidation rule introduced in [40, 103].

In order to obtain the estimator $\hat{f}_{Y|X}(y|x)$ of the conditional density $f_{Y|X}(y|x)$ on the interval $[a,b]$, the integrated squared error is defined as

$$\begin{aligned}\text{ISE} &= \int\{\hat{f}_{Y|X}(y|x)-f_{Y|X}(y|x)\}^2 f_X(x)\mathbf{1}_{[a,b]}(x)\,dxdy \\ &= \int \hat{f}^2_{Y|X}(y|x)f_X(x)\mathbf{1}_{[a,b]}(x)\,dxdy - 2\int \hat{f}_{Y|X}(y|x)f_{Y|X}(y|x)f_X(x)\mathbf{1}_{[a,b]}(x)\,dxdy \\ &+ \int f^2_{Y|X}(y|x)f_X(x)\mathbf{1}_{[a,b]}(x)\,dxdy,\end{aligned} \tag{B.3}$$

where $f_X(x)$ is the marginal density of X. Weighting the ISE by the marginal density $f_X(x)$ places more emphasis on the regions that have more data and it also eases the computational difficulty. $\mathbf{1}_{[a,b]}(x)$ is the indicator function, that is, $\mathbf{1}_{[a,b]}(x) = 1$ if $x \in [a,b]$ and $\mathbf{1}_{[a,b]}(x) = 0$ if $x \notin [a,b]$. Since the third term of (B.3) does not relate to the bandwidth h and therefore can be ignored in solving the problem of minimization of ISE. According to the crossvalidation rule, the minimization problem can be restated as [40, 103]:

$$\mathrm{cv}(h) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{[a,b]}(X_t) \int \hat{f}^2_{h,-t}(y|X_t)\, dy - \frac{2}{T}\sum_{t=1}^{T} \mathbf{1}_{[a,b]}(X_t)\hat{f}_{h,-t}(Y_t|X_t), \tag{B.4}$$

where $\hat{f}_{h,-t}(y|x)$ is the estimator of (B.2) based on the observations $\{X_j, Y_j\}_{j=1, j\neq t}^{T}$. The optimal bandwidth h is

$$\hat{h} = (\hat{h}_1, \hat{h}_2) = \min_{h_1,h_2} \mathrm{cv}(h).$$

B.2 Penalized Average Rule

The integrated squared error (ISE) is defined as [27]:

$$\mathrm{ISE} = \int \{\hat{f}_{Y|X}(y|x) - f_{Y|X}(y|x)\}^2 f_X(x)\, dxdy.$$

Since the above ISE is the expectation of $\int \{\hat{f}_{Y|X}(y|x) - f_{Y|X}(y|x)\}^2 dy$ with respect to X, we can estimate the ISE using the numerical samples by:

$$\mathrm{ISE} = \frac{\Delta}{T}\sum_{j=1}^{m}\sum_{t=1}^{T} \{f(y_j^*|x_t) - \hat{f}(y_j^*|x_t)\}^2, \tag{B.5}$$

where $\{x_t, y_t\}_{t=1}^{T}$ is the observations and $\mathbf{y}^* = \{y_1^*, \cdots, y_m^*\}$ is a vector of equally spaced values over the sample space of Y with $\Delta = y_{j+1}^* - y_j^*$. Since finding $f_{Y|X}(y|x)$ is a standard nonparametric problem of regressing $K_{h_2}(y_t - y)$ on x_t given the bandwidth h_2 and a specified value of y, Härdle proposes an alternative method in selecting the optimal bandwidth by minimizing the penalized average square prediction error based on (B.5) [41, 43].

Define the penalized average square prediction error as:

$$\begin{aligned}\mathrm{pa}(h) &= \frac{\Delta}{T}\sum_{j=1}^{m}\sum_{t=1}^{T}\{K_{h_2}(y_t - y_j^*) - \hat{f}(y_j^*|x_t)\}^2 p[w_t(x_t)] \\ &= \frac{\Delta}{T}\sum_{j=1}^{m}\sum_{t=1}^{T}\left\{K_{h_2}(y_t - y_j^*) - \sum_{k=1}^{T} w_k(x_t)K_{h_2}(y_k - y_j^*)\right\}^2 p[w_t(x_t)],\end{aligned} \tag{B.6}$$

where $\{y_1^*, \cdots, y_m^*\}$ are grid points over the range of y and the equally spaced interval $\Delta = y_{j+1}^* - y_j^*$. $p(\cdot)$ is the penalty function with the first order Taylor expansion $p(x) = 1 + 2x + O(x^2)$. In our numerical simulation, we choose the same penalty function $p(x) = (1+x)/(1-x)$ as used in [41].

Rewrite (B.6) in the matrix form for the convenience of computation as follows:

$$\mathrm{pa}(h) = \frac{\Delta}{T}\mathbf{p}^H(\mathbf{v} - \mathbf{w}^H\mathbf{v}) \odot (\mathbf{v} - \mathbf{w}^H\mathbf{v})\mathbf{1},$$

where the symbol $(.^H)$ denote the transpose. $\mathbf{v}$ is a $T \times m$ matrix with (i,j)th element $K_{h_2}(y_i - y_j^*)$, $\mathbf{w}$ is a $T \times T$ matrix with (i,j)th element $w_i(x_j)$, $\odot$ denotes the element-wise product, $\mathbf{1}$ is a $m \times 1$ vector of ones, and $\mathbf{p}$ is a $T \times 1$ vector with ith element $p(w_i(x_i))$.

The optimal bandwidth h is

$$\hat{h} = (\hat{h}_1, \hat{h}_2) = \min_{h_1,h_2} \mathrm{pa}(h).$$

Note that $\hat{f}(y_j^*|x_t)$ in (B.6) can be calculated from (B.2) of LLM, or, according to [41], is the modified Rosenblatt's estimator

$$\begin{aligned}\hat{f}_{Y|X}(y|x) &= \frac{\frac{1}{nh_1h_2}\sum_{t=1}^{T} K(\frac{y_t - y}{h_2})W(\frac{x_t - x}{h_1})}{\frac{1}{nh_1}\sum_{t=1}^{T} W(\frac{x_t - x}{h_1})} = \frac{\sum_{t=1}^{T} K_{h_2}(y_t - y)W_{h_1}(x_t - x)}{\sum_{t=1}^{T} W_{h_1}(x_t - x)} \\ &= \sum_{t=1}^{T} w_k(x)K_{h_2}(y_t - y),\end{aligned}$$

where $w_k(x) = \frac{W_{h_1}(x_k - x)}{\sum_{t=1}^{T} W_{h_1}(x_t - x)}$, the kernel K and W are nonnegative density functions.

APPENDIX C. COPULA PARAMETER ESTIMATION

A copula is characterized by its distinct functional form and specific parameters. The functional form comes from the copula family which mainly consists of Archimedean copulas and elliptical copulas, while the typical copula parameters measure the dependence among random variables and could be Kendall's correlation τ, Spearman's correlation ρ_s, or Pearson correlation ρ_{XY}. The general approaches to estimate parameters include the maximum likelihood (ML) method, the method of inference for the margins (IFM) and the canonical maximum likelihood (CML) method [49]. We introduce the ML method as an illustration.

Let $\{x_{1t},\cdots,x_{nt}\}$, $t=1,\cdots,T$ be known observations of n variables. According to (3.8), the log-likelihood function is expressed as:

$$l(\theta)=\sum_{t=1}^{T}\ln c(F_1(x_{1t}),\cdots,F_n(x_{nt}))+\sum_{t=1}^{T}\sum_{i=1}^{n}\ln f_i(x_{it}), \tag{C.1}$$

where θ is the set of all parameters of both the marginals and the copula. If the parameters of the marginal distributions are known (for example, the bandwidth of the NW kernel density estimation), θ only represents copula parameters.

The ML estimator $\hat{\theta}$ is the one which maximizes (C.1). Letting $\partial l/\partial\theta=0$, we obtain

$$\hat{\theta}_{ML}=\text{argmax } l(\theta).$$

The following shows the way to get the normal copula parameter of the correlation matrix $\mathbf{R}$ using the ML method. Given observations $\{x_{1t},\cdots,x_{nt}\}_{t=1}^{T}$, according to the definition of the normal copula density c^N in (3.10) and the expression of log-likelihood function in (C.1), the

log-likelihood function l^N of the normal copula is expressed as:

$$l^N = -\frac{T}{2}\log|\mathbf{R}| - \frac{1}{2}\sum_{t=1}^{T}\xi_t^H(\mathbf{R}^{-1}-\mathbf{I})\xi_t, \tag{C.2}$$

where $(\cdot)^H$ is the transpose. $\xi_t = (\Phi^{-1}(u_{1t}),\cdots,\Phi^{-1}(u_{nt}))^H$ and $u_{it} = F_i(x_{it}), i = 1,\cdots,n$.

From [39, 106], $\frac{\partial \log|\mathbf{A}|}{\partial \mathbf{A}} = \frac{1}{|\mathbf{A}|}\frac{\partial|\mathbf{A}|}{\partial \mathbf{A}}$, $\frac{\partial|\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}|\mathbf{A}^{-H}$ and $\frac{\partial \mathbf{a}^H\mathbf{A}^{-1}\mathbf{b}}{\partial \mathbf{A}} = -\mathbf{A}^{-H}\mathbf{a}\mathbf{b}^H\mathbf{A}^{-H}$, where $\mathbf{A}$ is a nonsingular square matrix, $\mathbf{a}$ and $\mathbf{b}$ are vectors. Let $\partial l^N/\partial \mathbf{R} = 0$, then

$$\begin{aligned}
&\Rightarrow -\frac{T}{2}\frac{1}{|\mathbf{R}|}|\mathbf{R}|\mathbf{R}^{-H} - \frac{1}{2}\sum_{t=1}^{T} -\mathbf{R}^{-H}\xi_t\xi_t^H\mathbf{R}^{-H} = 0\\
&\Rightarrow -\frac{T}{2}\mathbf{R}^{-H} + \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-H}\xi_t\xi_t^H\mathbf{R}^{-H} = 0\\
&\Rightarrow T\mathbf{I} - \sum_{t=1}^{T}\xi_t\xi_t^H\mathbf{R}^{-1} = 0\\
&\Rightarrow T\mathbf{R} = \sum_{t=1}^{T}\xi_t\xi_t^H.
\end{aligned}$$

The ML estimation of $\mathbf{R}$ is

$$\hat{\mathbf{R}}_{ML} = \frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t^H,$$

which is also the sample correlation matrix.

APPENDIX D. CORRELATION COMPUTATION

In this appendix, we compute the Pearson correlation coefficient between two random variables X and Y that are generated from data generating processes shown in Chapter 6. The Pearson correlation coefficient is defined as:

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}},$$

where $E(\cdot), \text{cov}(\cdot)$, and $\sigma_{(\cdot)}$ denote the expectation, the covariance, and the standard deviation, respectively.

Four types of data, $*$i, $*$d, $*$h, and $*$z, are applied in this study. The first three types of data come from the data generating processes introduced in Su and White's paper [17], which is published in 2008 and discusses the conditional independence. Suppose there are T observations $w_t = \{x_t, y_t\}$, $t = 1, \cdots, T$, where w_t denotes the realization of the random variables X and Y. The knowledge of covariance between X and Y is equivalent to calculating $\text{cov}(x_t, y_t)$.

D.1 Independent Case $*$i

Assume that the initial values, x_0 and y_0, are independent.

- 1i : $w_t = \{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d $\mathcal{N}(0, I_2)$.

  Here, $\text{cov}(X,Y) = \text{cov}(\varepsilon_{1,t}, \varepsilon_{2,t}) = 0$.

- 2i : $w_t = \{x_t, y_{t-1}\}$, $x_t = 0.5x_{t-1} + \varepsilon_{1,t}, y_t = 0.5y_{t-1} + \varepsilon_{2,t}$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d $\mathcal{N}(0, I_2)$.

$$
\begin{aligned}
\operatorname{cov}(x_1, y_0) &= E(x_1 y_0) - E(x_1)E(y_0) \\
&= E[(0.5x_0 + \varepsilon_{1,1})y_0] - E(0.5x_0 + \varepsilon_{1,1})E(y_0) \\
&= 0.5\operatorname{cov}(x_0, y_0) \\
&= 0. \\
\operatorname{cov}(x_2, y_1) &= E(x_2 y_1) - E(x_2)E(y_1) \\
&= E[(0.5x_1 + \varepsilon_{1,2})y_1] - E(0.5x_1 + \varepsilon_{1,2})E(y_1) \\
&= 0.5\operatorname{cov}(x_1, y_1) \\
&= 0.5[E(0.5x_0 + \varepsilon_{1,1})(0.5y_0 + \varepsilon_{2,1})] - 0.5E(0.5x_0 + \varepsilon_{1,1})E(0.5y_0 + \varepsilon_{2,1}) \\
&= 0.5^3\operatorname{cov}(x_0, y_0) \\
&= 0. \\
\operatorname{cov}(x_t, y_{t-1}) &= E(x_t y_{t-1}) - E(x_t)E(y_{t-1}) \\
&= E[(0.5x_{t-1} + \varepsilon_{1,t})y_{t-1}] - E(0.5x_{t-1} + \varepsilon_{1,t})E(y_{t-1}) \\
&= E(0.5x_{t-1}y_{t-1}) - E(0.5x_{t-1})E(y_{t-1}) \\
&= 0.5\operatorname{cov}(x_{t-1}, y_{t-1}) \\
&= 0.5E[x_{t-1}(0.5y_{t-2} + \varepsilon_{2,t-1})] - 0.5E(x_{t-1})E(0.5y_{t-2} + \varepsilon_{2,t-1}) \\
&= 0.5^2\operatorname{cov}(x_{t-1}, y_{t-2}) \\
&= 0.5^3\operatorname{cov}(x_{t-2}, y_{t-2}) \\
&\vdots \\
&= 0.5^{3(t-1)}\operatorname{cov}(x_0, y_0) \\
&= 0 \quad \text{for } t \geq 2.
\end{aligned}
$$

- 3i : $w_t = \{x_t, y_{t-1}\}$, $x_t = \sqrt{h_t}\varepsilon_{1,t}$, $h_t = 0.01 + 0.5x_{t-1}^2$, and $y_t = 0.5y_{t-1} + \varepsilon_{2,t}$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d $\mathcal{N}(0, I_2)$.

Since $E(y_t) = 0.5E(y_{t-1}) + E(\varepsilon_{2,t}) = 0$,

$$\text{cov}(x_t, y_{t-1}) = E(\sqrt{h_t}\varepsilon_{1,t}y_{t-1}) = E(\sqrt{0.01 + 0.5x_{t-1}^2})E(\varepsilon_{1,t})E(y_{t-1}) = 0.$$

- 4i : $w_t = \{x_t, y_{t-1}\}$, $x_t = \sqrt{h}_{1,t}\varepsilon_{1,t}$, $y_t = \sqrt{h}_{2,t}\varepsilon_{2,t}$, $h_{1,t} = 0.01 + 0.9h_{1,t-1} + 0.05x_{t-1}^2$, $h_{2,t} = 0.01 + 0.9h_{2,t-1} + 0.05y_{t-1}^2$.

  Since $E(y_t) = E(\sqrt{0.01 + 0.9h_{2,t-1} + 0.05y_{t-1}^2}\varepsilon_{2,t}) = 0$,

$$\begin{aligned}\text{cov}(x_t, y_{t-1}) &= E(\sqrt{h}_{1,t}\varepsilon_{1,t}y_{t-1}) \\ &= E(\sqrt{0.01 + 0.9h_{1,t-1} + 0.05x_{t-1}^2})E(\varepsilon_{1,t})E(y_{t-1}) \\ &= 0.\end{aligned}$$

Since data that are generated according to the processes $*$i possess independence relationship between X and Y, the corresponding covariance or correlation, which are calculated above, should equal zero.

D.2 Dependent Case $*$d

In the $*$d process, $w_t = \{x_t, y_{t-1}\}$ and $y_t = 0.5y_{t-1} + \varepsilon_{2,t}$, $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d $\mathcal{N}(0, I_2)$. Assume the initial values, x_0 and y_0, are independent.

- 1d : $x_t = 0.5x_{t-1} + 0.5y_{t-1} + \varepsilon_{1,t}$.

$$
\begin{aligned}
\operatorname{cov}(x_1, y_0) &= E[(0.5x_0 + 0.5y_0 + \varepsilon_{1,1})y_0] - E(0.5x_0 + 0.5y_0 + \varepsilon_{1,1})E(y_0) \\
&= 0.5\operatorname{cov}(x_0, y_0) + 0.5\operatorname{var}(y_0) \\
&= 0.5\operatorname{var}(y_0). \\
\operatorname{cov}(x_2, y_1) &= E[(0.5x_1 + 0.5y_1 + \varepsilon_{1,2})y_1] - E(0.5x_1 + 0.5y_1 + \varepsilon_{1,2})E(y_1) \\
&= 0.5\operatorname{cov}(x_1, y_1) + 0.5\operatorname{var}(y_1) \\
&= 0.5[E(x_1(0.5y_0 + \varepsilon_{2,1})] - 0.5E(x_1)E(0.5y_0 + \varepsilon_{2,1}) + 0.5\operatorname{var}(y_1) \\
&= 0.5^2\operatorname{cov}(x_1, y_0) + 0.5\operatorname{var}(y_1) \\
&= 0.5^3\operatorname{var}(y_0) + 0.5\operatorname{var}(y_1). \\
\operatorname{cov}(x_t, y_{t-1}) &= E(x_t y_{t-1}) - E(x_t)E(y_{t-1}) \\
&= E[(0.5x_{t-1} + 0.5y_{t-1} + \varepsilon_{1,t})y_{t-1}] - E(0.5x_{t-1} + 0.5y_{t-1} + \varepsilon_{1,t})E(y_{t-1}) \\
&= 0.5\operatorname{cov}(x_{t-1}, y_{t-1}) + 0.5\operatorname{var}(y_{t-1}) \\
&= 0.5E[x_{t-1}(0.5y_{t-2} + \varepsilon_{2,t-1})] - 0.5E(x_{t-1})E(0.5y_{t-2} + \varepsilon_{2,t-1}) + 0.5\operatorname{var}(y_{t-1}) \\
&= 0.5^2\operatorname{cov}(x_{t-1}, y_{t-2}) + 0.5\operatorname{var}(y_{t-1}) \\
&= 0.5E[x_{t-1}(0.5y_{t-1} + \varepsilon_{2,t-1})] - 0.5E(x_{t-1})E(0.5y_{t-2} + \varepsilon_{2,t-1}) + 0.5\operatorname{var}(y_{t-1}) \\
&= 0.5^2\operatorname{cov}(x_{t-1}, y_{t-2}) + 0.5\operatorname{var}(y_{t-1}) \\
&= 0.5^2[0.5\operatorname{cov}(x_{t-2}, y_{t-2}) + 0.5\operatorname{var}(y_{t-2})] + 0.5\operatorname{var}(y_{t-1}) \\
&\vdots \\
&= 0.5^{3(t-1)}\operatorname{cov}(x_0, y_0) + \sum_{i=1}^{t} 0.5^{2i-1}\operatorname{var}(y_{t-i}) \\
&= \sum_{i=1}^{t} 0.5^{2i-1}\operatorname{var}(y_{t-i}) \\
&\neq 0.
\end{aligned}
$$

Since the process y_t is an asymptotically strictly stationary process, we can calculate $\operatorname{var}(y_t) = \frac{4}{3}$, $t \to \infty$ as shown in (E.3). Therefore, the covariance between x_t and y_{t-1} will not tend to zero when t goes to infinity.

- 2d : $x_t = 0.5x_{t-1} + 0.5y_{t-1}^2 + \varepsilon_{1,t}$.

  Assume $E(y_0) = 0$, thus $E(y_t) = 0$ for all t.

$$
\begin{aligned}
\operatorname{cov}(x_1, y_0) &= E[(0.5x_0 + 0.5y_0^2 + \varepsilon_{1,1})y_0] - E(0.5x_0 + 0.5y_0^2 + \varepsilon_{1,1})E(y_0) \\
&= 0.5\operatorname{cov}(x_0, y_0) + 0.5E(y_0^3) \\
&= 0. \\
\operatorname{cov}(x_t, y_{t-1}) &= E[(0.5x_{t-1} + 0.5y_{t-1}^2 + \varepsilon_{1,t})y_{t-1}] - E(0.5x_{t-1} + 0.5y_{t-1}^2 + \varepsilon_{1,t})E(y_{t-1}) \\
&= E(0.5x_{t-1}y_{t-1}) + E(0.5y_{t-1}^3) - E(0.5x_{t-1})E(y_{t-1}) - 0.5E(y_{t-1}^2)E(y_{t-1}) \\
&= 0.5\operatorname{cov}(x_{t-1}, y_{t-1}) + 0.5E(y_{t-1}^3) \\
&= 0.5E[x_{t-1}(0.5y_{t-2} + \varepsilon_{2,t-1})] + 0.5E(y_{t-1}^3) \\
&= 0.5^2\operatorname{cov}(x_{t-1}, y_{t-2}) + 0.5E(y_{t-1}^3) \\
&= 0.5^3\operatorname{cov}(x_{t-2}, y_{t-2}) + 0.5E(y_{t-2}^3) + 0.5E(y_{t-1}^3) \\
&\vdots \\
&= 0.5^{3(t-1)}\operatorname{cov}(x_0, y_0) + \sum_{i=1}^{t} 0.5^{2i-1}E(y_{t-i}^3) \\
&= \sum_{i=1}^{t} 0.5^{2i-1}E(y_{t-i}^3) \\
&= 0.
\end{aligned}
$$

  As we know, if $X \sim \mathcal{N}(\mu, \sigma^2), Y \sim \mathcal{N}(\upsilon, \tau^2)$, and X and Y are independent, then $Z = X + Y \sim \mathcal{N}(\mu + \upsilon, \sigma^2 + \tau^2)$. Assume that $y_0 \sim \mathcal{N}(0,1)$, then $0.5y_0 \sim \mathcal{N}(0, 1/4)$. Since $\varepsilon_{2,t} \sim \mathcal{N}(0,1)$, we can derive that $y_1 \sim \mathcal{N}(0, 5/4), y_2 \sim \mathcal{N}(0, 21/16), y_3 \sim \mathcal{N}(0, 85/64), \cdots$. In other words, y_t are normally distributed for all t. Therefore $E(y_t^3) = 0$ due to the symmetry of the distribution.

- 3d : $x_t = 0.5x_{t-1}y_{t-1} + \varepsilon_{1,t}$.

Suppose $E(y_0) = 0$, thus $E(y_t) = 0$ for all t. Assume $E(x_0) = 0$.

$$
\begin{aligned}
\operatorname{cov}(x_1, y_0) &= E[(0.5x_0y_0 + \varepsilon_{1,1})y_0] - E(0.5x_0y_0 + \varepsilon 1, 1)E(y_0) \\
&= 0.5E(x_0y_0^2) \\
&= 0. \\
\operatorname{cov}(x_t, y_{t-1}) &= E[(0.5x_{t-1}y_{t-1} + \varepsilon_{1,t})y_{t-1}] - E(0.5x_{t-1}y_{t-1} + \varepsilon_{1,t})E(y_{t-1}) \\
&= 0.5E(x_{t-1}y_{t-1}^2) \\
&= 0.5E[x_{t-1}(0.5y_{t-2} + \varepsilon_{2,t-1})^2] \\
&= 0.5E[x_{t-1}(0.5^2y_{t-2}^2 + y_{t-2}\varepsilon_{2,t-1} + \varepsilon_{2,t-1}^2)] \\
&= 0.5E(0.5^2x_{t-1}y_{t-2}^2) \\
&= 0.5E[0.5^2(0.5x_{t-2}y_{t-2} + \varepsilon_{1,t-1})y_{t-2}^2] \\
&= 0.5^2E(0.5^2x_{t-2}y_{t-2}^3) \\
&= 0.5^4E[x_{t-2}(0.5y_{t-3} + \varepsilon_{2,t-2})^3] \\
&= 0.5^7E(x_{t-2}y_{t-3}^3) \\
&= 0.5^7E[(0.5x_{t-3}y_{t-3} + \varepsilon_{1,t-2})y_{t-3}^3] \\
&= 0.5^8E(x_{t-3}y_{t-3}^4) \\
&\vdots \\
&= a_tE(x_0y_0^{t+1}) \\
&= 0,
\end{aligned}
$$

where $a_t = a_{t-1}0.5^{t+1}$ and $a_1 = 0.5$.

- 4d : $x_t = 0.5x_{t-1} + 0.5y_{t-1}\varepsilon_{1,t}$.

$$\begin{aligned}
\text{cov}(x_t, y_{t-1}) &= E[(0.5x_{t-1} + 0.5y_{t-1}\varepsilon_{1,t})y_{t-1}] - E(0.5x_{t-1} + 0.5y_{t-1}\varepsilon_{1,t})E(y_{t-1}) \\
&= 0.5E(x_{t-1}y_{t-1}) - E(0.5x_{t-1})E(y_{t-1}) \\
&= 0.5\text{cov}(x_{t-1}, y_{t-1}) \\
&= 0.5E[x_{t-1}(0.5y_{t-2} + \varepsilon_{2,t-1})] - 0.5E(x_{t-1})E(0.5y_{t-2} + \varepsilon_{2,t-1}) \\
&= 0.5^2\text{cov}(x_{t-1}, y_{t-2}) \\
&= 0.5^3\text{cov}(x_{t-2}, y_{t-2}) \\
&= 0.5^4\text{cov}(x_{t-2}, y_{t-3}) \\
&\vdots \\
&= 0.5^{2t-1}\text{cov}(x_0, y_0) \\
&= 0.
\end{aligned}$$

- 5d : $x_t = \sqrt{h_t}\varepsilon_{1,t}$; $h_t = 0.01 + 0.5x_{t-1}^2 + 0.25y_{t-1}^2$.

  Since $E(x_t) = E(\sqrt{0.01 + 0.5x_{t-1}^2 + 0.25y_{t-1}^2}\varepsilon_{1,t}) = 0$,

$$\begin{aligned}
\text{cov}(x_t, y_{t-1}) &= E(x_t y_{t-1}) - E(x_t)E(y_{t-1}) \\
&= E(\sqrt{0.01 + 0.5x_{t-1}^2 + 0.25y_{t-1}^2}\varepsilon_{1,t}y_{t-1}) \\
&= E(\sqrt{0.01 + 0.5x_{t-1}^2 + 0.25y_{t-1}^2}y_{t-1})E(\varepsilon_{1,t}) \\
&= 0.
\end{aligned}$$

- 6d : $x_t = \sqrt{h_{1,t}}\varepsilon_{1,t}$, $y_t = \sqrt{h_{2,t}}\varepsilon_{2,t}$, $h_{1,t} = 0.01 + 0.1h_{1,t-1} + 0.4x_{t-1}^2 + 0.5y_{t-1}^2$, $h_{2,t} = 0.01 + 0.9h_{2,t-1} + 0.05y_{t-1}^2$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d $\mathcal{N}(0, I_2)$.

Since $E(y_t) = E(\sqrt{0.01 + 0.9h_{2,t-1} + 0.05y_{t-1}^2}\varepsilon_{2,t}) = 0$,

$$\begin{aligned}
\operatorname{cov}(x_t, y_{t-1}) &= E(x_t y_{t-1}) - E(x_t)E(y_{t-1}) \\
&= E(\sqrt{0.01 + 0.1h_{1,t-1} + 0.4x_{t-1}^2 + 0.5y_{t-1}^2}\varepsilon_{1,t}y_{t-1}) \\
&= E(\sqrt{0.01 + 0.1h_{1,t-1} + 0.4x_{t-1}^2 + 0.5y_{t-1}^2}y_{t-1})E(\varepsilon_{1,t}) \\
&= 0.
\end{aligned}$$

Data that are generated according to the processes $*$d relate with each other and therefore X and Y are dependent. However, except for 1d, the correlations between x_t and y_{t-1} equal 0 in 2d-6d processes when $\{x_0, y_0\}$ are i.i.d normally distributed.

D.3 High Frequency Case $*$h

$$*\text{h} : y_t = 0.5z_t + 4\alpha\varphi(x_t/\alpha) + 0.5\varepsilon_t,$$

where $\{x_t, z_t, \varepsilon_t\}$ are i.i.d $\mathcal{N}(0, \mathbf{I}_3)$ and φ is the standard normal density function.

For $*$h process, $w_t = \{x_t, y_t\}$. Considering $\alpha \in \{0, 0.5, 1, 2\}$, we obtain the corresponding processes 1h-4h. Note when $\alpha = 0$, 1h : $y_t = 0.5z_t + 0.5\varepsilon_t$.

For 1h, $\operatorname{cov}(x_t, y_t) = 0$ since X and Y are independent. As for 2h-4h,

$$\begin{aligned}
\operatorname{cov}(x_t, y_t) &= E(x_t y_t) - E(x_t)E(y_t) \\
&= E[x_t(0.5z_t + 4\alpha\varphi(x_t/\alpha) + 0.5\varepsilon_t)] \\
&= E(4\alpha x_t\varphi(x_t/\alpha)) \\
&= 0.
\end{aligned}$$

$\operatorname{cov}(x_t, y_t) = 0$ because $x_t\varphi(x_t/\alpha)$ is an odd function and the standard normal density is symmetric about 0. Therefore, only 1h represents independence relationship and all others (2h-4h) represent dependence relationship even if the correlation between x_t and y_t is 0 for all $*$h processes.

D.4 Zero Correlation Case ∗z

- 1z : $Y = X^2$,

  where X has a density function that is symmetric about 0 and the third moment of X exists. We have

$$\text{cov}(X,Y) = E(X^3) - E(X)E(X^2) = 0.$$

- 2z : $Y = ZX$,

  where $X \sim \mathcal{N}(0,1)$, Z and X are independent with $\text{P}(Z=1) = \text{P}(Z=-1) = \frac{1}{2}$. We have

$$\text{cov}(X,Y) = E(ZX^2) - E(Z)E(X)^2 = E(Z)\text{var}(X) = 0.$$

Therefore, X and Y are dependent with zero correlation in the process of ∗z.

Table D.1 summarizes the correlation relationship among random variables generated from the processes ∗i, ∗d, ∗h, and ∗z. In the simulations, the values of correlation are all zero by assuming the initial values x_0 and y_0 are i.i.d $\mathcal{N}(0,\mathbf{I}_2)$.

Table D.1: Correlation summary

Data type	**Description**
∗i	Independent case, $\text{corr}(X,Y) = 0$.
∗d	Dependent case, 1d: $\text{corr}(X,Y) \neq 0$. 2d: $\text{corr}(X,Y) = 0$, if $y_0 \sim \mathcal{N}(0,1)$. 3d: $\text{corr}(X,Y) = 0$, if $E(x_0) = E(y_0) = 0$. 4d-6d: $\text{corr}(X,Y) = 0$.
∗h	High frequency case, $\text{corr}(X,Y) = 0$. 1h: X and Y are independent. 2h-4h: X and Y are dependent.
∗z	Zero correlation case, $\text{corr}(X,Y) = 0$. 1z, 2z: X and Y are dependent.

APPENDIX E. PROCESS STATIONARITY

Regarding to the data generating processes $*$i, $*$d, and $*$h in this study, one of the most important properties is that they are asymptotically strictly stationary. Based on this property, x_t and y_t, which denote the realization of stochastic process X_t and Y_t at any time point, can be viewed as the realization of two random variables X and Y. The strict stationarity of processes will be discussed in this appendix.

A stochastic univariate process $\{X_t\}$, $t = 1, \cdots, T$ is said to be strictly, or strongly, stationary (SSS) if all sequences of consecutive responses of equal length in time have identical joint density functions, i.e.,

$$f_{X_{t_1},\cdots,X_{t_k}}(x_{t_1},\cdots,x_{t_k}) = f_{X_{t_1+\tau},\cdots,X_{t_k+\tau}}(x_{t_1},\cdots,x_{t_k}) \text{ for all } k,\tau. \tag{E.1}$$

In other words, shifting a fixed-width time observation window along a strictly stationary series always yields the same multivariate distributions.

A less restrictive assumption of stochastic process is weakly, or second-order, stationarity (WSS). Let the mean $E(X_t) = \mu_t$, the variance $\text{var}(X_t) = \gamma(t) = E(X_t - \mu_t)^2$, and the autocovariance $\text{cov}(X_t, X_{t-k}) = \gamma(t, t-k) = E[(X_t - \mu_t)(X_{t-k} - \mu_{t-k})]$, then the weak stationary condition means

- $\mu_t = \mu$ for all t,
- $\gamma(t, t-k) = \gamma_k$ for all t, k.

The first implies that the mean of observations over time is a constant. The second implies that the two observations have the same covariance as long as their relative locations, or the time distance, are fixed. Two special cases of (E.1) are:

$$f_{X_t}(x_t) = f_{X_{t+\tau}}(x_t) \text{ and } f_{X_{t_1},X_{t_2}}(x_{t_1},x_{t_2}) = f_{X_{t_1+\tau},X_{t_2+\tau}}(x_{t_1},x_{t_2}).$$

Therefore, if the mean and variance exist, SSS implies WSS.

As for a vector process $\{\mathbf{X}_t\}$ with m components, $\{\mathbf{X}_t\} = \{X_{1t}, \cdots, X_{mt}\}^H$, the strict stationarity definition (E.1) remains valid, while the process $\{\mathbf{X}_t\}$ is said to be weak stationary if [107]:

1. $EX_{it}^2 < \infty, \forall t \in \mathbb{Z}, i = 1, \cdots, m$;
2. $E\mathbf{X}_t = \mu, \forall t \in \mathbb{Z}$;
3. $\text{Cov}(\mathbf{X}_t, \mathbf{X}_{t+h}) = E[(\mathbf{X}_t - \mu)(\mathbf{X}_{t+h} - \mu)^H] = \Gamma(h), \forall t, h \in \mathbb{Z}$.

Furthermore, in time series analysis, an important concept called *equilibrium* is introduced to generalize the useful characteristic of stationarity such that the amount of information necessary to represent a stochastic process can be reduced enormously in practice. According to [108], the equilibrium is stated as:

Although a process may not be stationary when it starts, it may reach an equilibrium after a sufficiently long time, independent of the initial conditions. In other words, if an equilibrium has been reached, the probability that the process is in each given state, or the proportion of time spent in each state, has converged to a constant that does not depend on the initial conditions. This generally implies that eventually the process approaches closely to a stationary situation in the sense that, if it initially had the equilibrium distribution of states, it would be stationary.

In this study, we talk about the stationarity of the process in this generalized sense, that is, the process stationarity after the equilibrium is reached or asymptotically stationarity. We ignore the first couple of hundreds of samples to eliminate the influence of initial values and start from the samples which are collected after the process has already approaches the stationarity. Without causing the confusion, for simplicity, we speak of the process stationarity in the generalized sense and we say the process is stationary by not collecting the samples from the start.

All the processes used in this study, $*$i, $*$d, and $*$h, are financial series, i.e., they model some common economic phenomena. For example, the GARCH model, as shown in 3i or 4i, can be used to represent the returns of the Wilshire 5000 index. Therefore, instead of verifying the above stationarity conditions by computing the complicated multivariate distributions, we derive the strict stationarity of most of the processes on the basis of the inferences of stochastic regularity provided by a variety of econometric studies.

- 1i : $w_t = \{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d. $\mathcal{N}(0, I_2)$.

  $x_t = \varepsilon_{1,t}$ and $y_t = \varepsilon_{2,t}$ are i.i.d normally distributed, and therefore $\{x_t, y_t\}$ are both univariate SSS and joint SSS.

- 2i : $x_t = 0.5x_{t-1} + \varepsilon_{1,t}$, $y_t = 0.5y_{t-1} + \varepsilon_{2,t}$,

  where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, $t = 0, 1, \cdots$, are i.i.d $\mathcal{N}(0, I_2)$, and the initial values $\{x_0, y_0\}$ are also i.i.d $\mathcal{N}(0, I_2)$.

  We first investigate the strict stationarity of the process 2i by directly using definition (E.1). x_t is the summation of its lagged value x_{t-1} and a noise term $\varepsilon_{1,t}$, where x_{t-1} and $\varepsilon_{1,t}$ are independently normally distributed, and hence x_t is normally distributed. Considering the mean and variance of x_t, we have

$$\begin{aligned} E(x_1) &= 0.5E(x_0) + E(\varepsilon_{1,1}) = 0.5E(x_0) = 0; \\ E(x_2) &= 0.5E(x_1) + E(\varepsilon_{1,2}) = 0.5E(x_1) = 0.5^2E(x_0) = 0; \\ &\vdots \\ E(x_t) &= 0.5E(x_{t-1}) + E(\varepsilon_{1,t}) = 0.5E(x_{t-1}) = 0.5^tE(x_0) = 0; \end{aligned} \tag{E.2}$$

  and

$$\begin{aligned} E(x_1^2) &= E[(0.5x_0 + \varepsilon_{1,1})^2] = 0.5^2E(x_0^2) + E(\varepsilon_{1,1}^2); \\ E(x_2^2) &= E[(0.5x_1 + \varepsilon_{1,2})^2] = 0.5^2E(x_1^2) + E(\varepsilon_{1,2}^2) = 0.5^4E(x_0^2) + 0.5^2E(\varepsilon_{1,1}^2) + E(\varepsilon_{1,2}^2); \\ &\vdots \\ E(x_t^2) &= E[(0.5x_{t-1} + \varepsilon_{1,t})^2] = 0.5^{2t}E(x_0^2) + \sum_{k=1}^{t} 0.5^{2(t-k)}E(\varepsilon_{1,k}^2); \end{aligned}$$

  hence,

$$\gamma_0 = \text{var}(x_t) = Ex_t^2 - (Ex_t)^2 = 0.5^{2t} + \frac{4}{3}(1 - 0.5^{2t}). \tag{E.3}$$

  Equation (E.2) shows that the mean of x_t is a constant 0 and (E.3) shows that the variance γ_0 varies with respect to time. Therefore, x_t can not be generally considered as a station-

ary series. However, in (E.3), $\gamma_0 = 0.5^{2t} + \frac{4}{3}(1-0.5^{2t}) \to \frac{4}{3}$ as $t \to \infty$. After a sufficiently long time, the fact that the mean and the variance become constants implies that the normal distributions at every time point of x_t become invariant. Therefore, we can conclude that an equilibrium is reached and the process 2i, where x_t and y_t follow the same mathematical model, goes into stationarity. In other words, x_t and y_t are asymptotically stationary. Since x_t and y_t are independent, $\{x_t, y_t\}$ is jointly stationary as well.

As we know, x_t of the process 2i is an example of AR(1) model. We now consider the stationary conditions for autoregressive process(AR), more specifically, for AR(1) process defined as below:

$$x_t = \rho x_{t-1} + \varepsilon_t,$$

where ε_t is the error term, for example, a white noise. According to [109], the necessary and sufficient condition for strict stationarity of the AR(1) process is that $|\rho| < 1$. In 2i, $\rho = 0.5 < 1$, therefore x_t and y_t of 2i are strictly stationary.

- 3i : $x_t = \sqrt{h_t}\varepsilon_{1,t}$, $h_t = 0.01 + 0.5x_{t-1}^2$;

  4i : $x_t = \sqrt{h}_{1,t}\varepsilon_{1,t}$, $y_t = \sqrt{h}_{2,t}\varepsilon_{2,t}$, $h_{1,t} = 0.01 + 0.9h_{1,t-1} + 0.05x_{t-1}^2$, $h_{2,t} = 0.01 + 0.9h_{2,t-1} + 0.05y_{t-1}^2$;

  In both cases, $y_t = 0.5y_{t-1} + \varepsilon_{2,t}$. $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, $t = 0, 1, \cdots$, are i.i.d $\mathcal{N}(0, I_2)$, and the initial values $\{x_0, y_0\}$ are also i.i.d $\mathcal{N}(0, I_2)$.

  3i and 4i are two examples of GARCH model. In econometrics, AutoRegressive Conditional Heteroskedasticity (ARCH) models are used to characterize and model observed time series. A generalized ARCH (GARCH) model, GARCH(p,q) process for the time series r_t, is defined as [110, 111]

$$r_t = \sqrt{h_t}\eta_t,\ \eta_t \sim \mathcal{N}(0,1), \tag{E.4}$$

$$h_t = a_0 + \sum_{i=1}^{q} a_i r_{t-i}^2 + \sum_{i=1}^{p} b_i h_{t-i}. \tag{E.5}$$

In this study, we only consider the simple GARCH(1,1) model. It is most popular for modeling asset-return volatility and is written as:

$$r_t = \sqrt{h_t}\eta_t,\ \eta_t \sim \mathcal{N}(0,1),\ \text{ and }\ h_t = a_0 + a_1 r_{t-1}^2 + b_1 h_{t-1}. \tag{E.6}$$

To ensure the stationarity of r_t, the conditions on the parameters a's and b's need to be imposed. The necessary and sufficient condition for strict stationarity and ergodicity of the GARCH(1,1) model in (E.6) is obtained as follows [110, 111]:

$$E[\ln(b_1 + a_1\eta_t^2)] < 0. \tag{E.7}$$

The necessary and sufficient condition for strict stationarity and ergodicity of the general model (E.4) and (E.5) is established by Bougerol and Picard [111, 112] as follows

$$\sum_{i=1}^{q} a_i + \sum_{i=1}^{p} b_i \leq 1. \tag{E.8}$$

Applying the convenient condition (E.8) to x_t of 3i and 4i, which are analogical to r_t in (E.6), we see

3i: ARCH(1) or GARCH(0,1) model. $a_1 = 0.5, b_1 = 0$ and $a_1 + b_1 = 0.5 \leq 1$. The condition (E.8) is satisfied and x_t is strictly stationary.

4i: GARCH(1,1) model. $a_1 = 0.05, b_1 = 0.9$ and $a_1 + b_1 = 0.95 \leq 1$. The condition (E.8) is satisfied and x_t is strictly stationary.

The process x_t (GARCH(1,1)) will be stationary if $a_1 + b_1 \leq 1$. In econometrics, the sum $a_1 + b_1$ measures the persistence in volatility and, as is typical for financial return data, is very close to unity. The process y_t in 3i and 4i are same as that in 2i, and therefore has been proved to be strictly stationary. Since x_t and y_t are univariate stationary and independent, they are jointly stationary as well.

- 1d : $x_t = 0.5x_{t-1} + 0.5y_{t-1} + \varepsilon_{1,t},\ y_t = 0.5y_{t-1} + \varepsilon_{2,t}$;

where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, $t = 0, 1, \cdots$, are i.i.d $\mathcal{N}(0, I_2)$, and the initial values $\{x_0, y_0\}$ are also i.i.d $\mathcal{N}(0, I_2)$.

1d is an example of a vector autoregressive (VAR) model. A VAR model of order p is defined as [113]

$$\mathbf{y}_t = \mathbf{v} + \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \mathbf{u}_t, \, t \in \mathbb{Z}, \tag{E.9}$$

where $\mathbf{y}_t = (y_{1t}, \cdots, y_{mt})^H$ is a $m \times 1$ random vector, the coefficient matrices $A_i \in \mathbb{R}^m \times \mathbb{R}^m$, and $\mathbf{v}_t = (v_1, \cdots, v_m)^H$ is a fixed $m \times 1$ vector of intercept terms which may represent a nonzero $E(\mathbf{y}_t)$. Finally, $\mathbf{u}_t = (u_{1t}, \cdots, u_{mt})^H$ is a m-dimensional white noise, that is, $E(\mathbf{u}_t) = \mathbf{0}, E(\mathbf{u}_t \mathbf{u}_t^H) = \Sigma_u$ and $E(\mathbf{u}_t \mathbf{u}_s^H) = \mathbf{0}$ for $s \neq t$.

The process (E.9) is stable if its reverse characteristic polynomial has no roots in and on the complex unit circle, that is, $\mathbf{y}_t$ is stable if

$$\det(\mathbf{I}_m - \mathbf{A}_1 z - \cdots - \mathbf{A}_p z^p) \neq 0 \;\; |z| \leq 1, \tag{E.10}$$

where $\mathbf{I}_m$ is a $m \times m$ identity matrix. The strict stationarity condition of VAR(p) process in (E.9) is: a stable VAR(p) process $y_t, t \in \mathbb{Z}$ is stationary [113, 114]. The converse of this statement is not true, that is, an unstable process is not necessarily nonstationary.

The process 1d is stable, and hence stationary because

$$\det(\mathbf{I}_m - \mathbf{A}_1 z - \cdots - \mathbf{A}_p z^p) = 0 \Rightarrow \det\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0.5 & 0.5 \\ 0 & 0.5 \end{pmatrix} z\right] = 0$$

$$\Rightarrow \det\left[\begin{pmatrix} 1 - 0.5z & -0.5z \\ 0 & 1 - 0.5z \end{pmatrix}\right] = 0 \Rightarrow z_1 = 2 \text{ and } z_2 = 2.$$

$\det(\mathbf{I}_m - \mathbf{A}_1 z - \cdots - \mathbf{A}_p z^p) = 0$ has all its root outside the unit circle, i.e., (E.10) is satisfied. Therefore, the process 1d is stationary.

- 2d : $x_t = 0.5 x_{t-1} + 0.5 y_{t-1}^2 + \varepsilon_{1,t}$;

  4d : $x_t = 0.5 x_{t-1} + 0.5 y_{t-1} \varepsilon_{1,t}$;

and $y_t = 0.5y_{t-1} + \varepsilon_{2,t}$ for both 2d, 3d and 4d,

where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$, $t = 0, 1, \cdots$, are i.i.d $\mathcal{N}(0, I_2)$, and the initial values $\{x_0, y_0\}$ are also i.i.d $\mathcal{N}(0, I_2)$.

2d and 4d are examples of vector nonlinear autoregressive model (NLAR). A m-dimensional vector NLAR(1) model, or a Markov chain is defined as [115]:

$$\mathbf{y}_t = T(\mathbf{y}_{t-1}) + \varepsilon_t,\ t \geq 1,\ T : \mathbb{R}^m \to \mathbb{R}^m, \tag{E.11}$$

where ε_t are i.i.d, the marginal distribution of which is absolutely continuous and has an everywhere positive probability density function over $\mathbb{R}^m$. Also $E||\varepsilon_t|| < \infty$.

We say the process is geometrically ergodic if the rate of convergence is geometric, that is, there exists a $\rho \in (0,1)$ such that $\rho^{-t}||\mu_t(\mu_0) - \pi|| \to 0$ as $t \to \infty$, where π is a probability measure, $\mu_t(\mu_0)$ denotes the probability of $\mathbf{y}_t$ given that $\mathbf{y}_0$ has distribution μ_0, and $||\cdot||$ denotes a suitable norm, such as the total variation. *If we set $\mu_0 \equiv \pi$, then the Markov chain in (E.11) is strictly stationary. This practice will be adopted without further mention and π will be referred to as the stationary distribution and the Markov chain is said to be stationary* [115].

Suppose T is compact and can be decomposed into two parts, namely

$$T = T_h + T_d, \tag{E.12}$$

where T_h is homogeneous and continuous and T_d is bounded, then the process (E.11) is geometrically ergodic, and hence strictly stationary [116]. Regarding to 2d,

$$T(\mathbf{z}) = \begin{bmatrix} 0.5z_1 \\ 0.5z_2 \end{bmatrix} + \begin{bmatrix} 0.5z_2^2 \\ 0 \end{bmatrix},$$

where $\mathbf{z} = (z_1, z_2)^H$, $T_h(\mathbf{z}) = [0.5z_1, 0.5z_2]^H$ and $T_d(\mathbf{z}) = [0.5z_2, 0]^H$. T_h is a linear operator and therefore continuous and homogeneous, i.e., $T(c\mathbf{z}) = cT(\mathbf{z}), \forall c > 0, \mathbf{z} \in \mathbb{R}^m$. The condition for an operator T to be bounded is that there exists some constant M such that

$||T(\mathbf{z})|| \leq M||\mathbf{z}||$ for all $\mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^m$ and $||\cdot||$ denote a vector norm. Since

$$||T_d(\mathbf{z})|| = |0.5z_2| \leq 0.5(z_1^2 + z_2^2)^{1/2} = 0.5||\mathbf{z}||,$$

$M = 0.5$ and T_d is bounded. Equation (E.12) is satisfied and therefore the process 2d is strictly stationary.

$$\mathbf{y}_t = f_1(\mathbf{y}_{t-1}) + f_2(\mathbf{y}_{t-1})\varepsilon_t,\ t \geq 1, \tag{E.13}$$

In (E.13), the noise ε_t is same as that defined in (E.11), but is multiplicative. Suppose f_1 and f_2 are bounded on compact sets in $\mathbb{R}$, with $||f_1(\mathbf{z})|| \leq \alpha||\mathbf{z}||$, $0 < \alpha < 1$, for $||\mathbf{z}|| > r$ where $r > 0$, and f_2 is a positive measurable function, then the process in (E.13) is geometrically ergodic, and hence strictly stationary [117]. As for 4d,

$$f_1(\mathbf{z}) = [0.5z_1, 0.5z_2]^H \text{ and } f_2(\mathbf{z}) = [0.5z_2, 0]^H,$$

where $||f_1(\mathbf{z})|| = 0.5\sqrt{z_1^2 + z_2^2} \leq \alpha\sqrt{z_1^2 + z_2^2} = \alpha||z||$ if $0.5 \leq \alpha < 1$, $f_2(\mathbf{z}) > 0$ if $z_2 > 0$ and therefore is a positive measurable function [118]. The condition is satisfied and therefore the process 4d is strictly stationary.

- 6d : $x_t = \sqrt{h_{1,t}}\varepsilon_{1,t}$, $y_t = \sqrt{h_{2,t}}\varepsilon_{2,t}$, $h_{1,t} = 0.01 + 0.1h_{1,t-1} + 0.4x_{t-1}^2 + 0.5y_{t-1}^2$, $h_{2,t} = 0.01 + 0.9h_{2,t-1} + 0.05y_{t-1}^2$, where $\{\varepsilon_{1,t}, \varepsilon_{2,t}\}$ are i.i.d $\mathcal{N}(0, I_2)$.

  A vector sequence $\mathbf{r}_t$ with values in $\mathbb{R}^m$ follows a multivariate GARCH(p,q) process with constant correlation if [119, 120]

$$\mathbf{r}_t = \mathbf{H}_t^{1/2}\eta_t, \tag{E.14}$$

where η_t is an i.i.d vector error process such that $\mathrm{E}\eta_t\eta_t' = \mathbf{I}_m$ and $\mathbf{I}_m$ is a $m \times m$ identity matrix. $\mathbf{H}_t$ is a diagonal matrix and the elements of the diagonal $\mathbf{H}_{t,ii}$ satisfy, for all i, the

following relation:

$$\begin{pmatrix} H_{t,11} \\ \vdots \\ H_{t,mm} \end{pmatrix} = \mathbf{W} + \sum_{i=1}^{q} \mathbf{A}_i \begin{pmatrix} r_{t-i,1}^2 \\ \vdots \\ r_{t-i,m}^2 \end{pmatrix} + \sum_{i=1}^{p} \mathbf{B}_i \begin{pmatrix} H_{t-i,11} \\ \vdots \\ H_{t-i,mm} \end{pmatrix} \tag{E.15}$$

where $\mathbf{W} \in \mathbb{R}^m, \mathbf{A}_i$ and $\mathbf{B}_i \in \mathbb{R}^m \times \mathbb{R}^m$, and we assume that all coefficients of these matrices are positive. The process in (E.14) is called extended constant correlation GARCH (ECCC-GARCH) model since the conditional covariance matrix of $\mathbf{r}_t$, denoted by $\mathbf{H}_t$, is constant.

Jeantheau provides the sufficient condition for the stationarity of ECCC-GARCH model [120]. If $\det(\mathbf{I}_m - \sum_{i=1}^{n}(\mathbf{A}_i + \mathbf{B}_i)\lambda^i) = 0$ has its roots outside the unit circle, then the model of (E.14) is weakly stationary. Moreover, the process is also strictly stationary and ergodic.

Obviously, 6d is an example of ECCC-GARCH model shown in (E.14) and (E.15). As for 6d, $m = 2$.

$$\mathbf{I}_m - \sum_{i=1}^{n}(\mathbf{A}_i + \mathbf{B}_i)\lambda^i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \left[\begin{pmatrix} 0.4 & 0.5 \\ 0 & 0.05 \end{pmatrix} + \begin{pmatrix} 0.1 & 0 \\ 0 & 0.9 \end{pmatrix}\right]\lambda,$$

Solving $\det(\mathbf{I}_m - \sum_{i=1}^{n}(\mathbf{A}_i + \mathbf{B}_i)\lambda^i) = 0$, we get $\lambda_1 = 2$ and $\lambda_2 = 1.053$. Both λ are outside the unit circle, and therefore the vector process 6d is strictly stationary.

According to [17], the processes 3d, 5d and $*$h are strictly stationary β-mixing processes. Please refer to [34] to see the details of mixing conditions.