# Statistical Inference for High-dimensional Nonparametric Models

a thesis submitted to
The University of Kent at Canterbury
in the subject of statistics
for the degree
of doctor of philosophy by research

By
Jie Li
April 2021

University of Kent

DOCTORAL THESIS

# Statistical Inference for High-dimensional Nonparametric Models

*Author:*
Jie Li

*Supervisor:*
Prof. Jian Zhang

School of Mathematics, Statistics & Actuarial Science

April 2021

I would like to dedicate this thesis to my loving parents, wife and daughter.

谨以此文献给我深爱的父母，妻子和女儿。

# *Acknowledgements*

First, I appreciate the time and effort that the external examiner Professor Jianxin Pan and internal examiner Dr. Peng Liu have dedicated to during my PhD viva. Their valuable feedback indeed improve the level of my thesis. My deepest gratitude goes foremost to my supervisor, Professor Jian Zhang, for his continuous encouragement, support and guidance. I have learned from him not only an attitude of rigorous academic research but also a lot of practical skills to develop the theories of this thesis. Without his selfless and consistent help, this thesis could not reach the current level. I would also like to express my grateful thanks to my second supervisor, Dr. Alfred Kume, who gave me a lot of constructive suggestions in the completion of this thesis. I would like to thank Dr. Mark Burnley and Dr. Samantha Winter, School of Sport and Exercise Sciences, University of Kent, who offered the dataset of muscle contraction for this thesis.

Second, I would like to thank all the people of Statistics Group in the School of Mathematics, Statistics and Actuarial Sciences (SMSAS). Special thanks should go to Claire Carter who gave me much useful advice during my Ph.D. study and Sonnary Dearden who helped me through the submission of this thesis. High tribute shall be paid to Dr. Peng Liu who offered me a lot of support amid Covid-19 pandemic. I would like to express my huge thanks to all the Ph.D. students that I met in the past four years. Every moment that we spent together makes my Ph.D. life colourful and memorable.

Last, I express my sincere gratitude and high respect to my wife, Sifei Yang. She took the initiative to take care of our newborn daughter to save time and energy for me to complete my Ph.D. research. On behalf of my family, I would also like to express our grateful thanks to our beloved parents and parents-in-law. Their consecutive love, understanding and support are the most reliable source of our great confidence in life.

# Abstract

The thesis aims to develop methodologies for estimating high-dimensional nonparametric covariance models and the change-point detection in time series segments respectively.

With the development of statistical inference and the availabilities of big data, estimation of covariate-dependent conditional covariance matrix in a high-dimensional space poses a challenge to contemporary statistical research. The existing kernel estimators may not be adaptive to varying smoothness across different entries due to using a single bandwidth to explore the smoothness of the target matrix function. The nonparametric estimation of covariance matrix function may be degenerated or ill-conditioned when one confronts the curse of dimensionality. Furthermore, sparsity also has a significant effect on the bandwidth selection as zero entries have smoothness different from the non-zero entries. If the sparsity is high, then the zero entries will dominate the procedure of bandwidth selection and let the bandwidth go to infinity.

To address these issues, we have considered two possible methods. First, compared to the single bandwidth in the existing kernel estimators, we have adopted the multiple bandwidths for the different entries of covariance matrix. Meanwhile, we have also kept the covariance estimator positive definite. Second, one can detect the zero entries in advance and omit them temporarily. Next, the classical kernel estimation with single bandwidth can be applied to the rest of entries. Finally, considering the positive definiteness, the covariance estimator can be obtained by combining the estimators of non-zero entries and the detection of zero entries.

Based on the above analysis, we have proposed two novel frameworks in this thesis. One is the factorized estimation of high-dimensional nonparametric covariance model (NCM), the other is the so-called Divide-and-Combine estimation of high-dimensional nonparametric covariance model.

In the former, factorizing the target matrix into factors plays a significant role in improving the performance of NCM. These factors are in turn estimated by the kernel approach. The resulting estimator of covariance matrix is further regularized by thresholding and optimal shrinkage. Under certain mixing and

sparsity conditions, we show that the proposed estimator is well-conditioned and uniformly consistent with the underlying matrix even when the sample is dependent. A set of simulation studies show that the proposed estimator significantly outperforms its competitors in terms of integrated root-squared estimation error and computational speed. A real application of factorized NCM to a financial return dataset shows that factorized NCM can detect a number of interesting volatility and co-volatility patterns over different time periods.

In the latter, the key idea of Divide-and-Combine approach for the nonparametric covariance matrix is to divide it into three parts: the diagonal entries, the off-diagonal zero entries and the off-diagonal nonzero entries. We combine the three part estimations to form an estimator of the whole covariance matrix. We apply this model to seven scenarios, and the results show that the Divide-and-Combine NCM framework could also address the entries' smoothness problem under sparsity. The network analysis based on the historical return dataset shows that there exists a significant network change over the financial periods.

Besides the two methods of nonparametric covariance model, this thesis also aims to provide a method for the change-point detection in time series segments.

The change-point detection in time series segments thrives in many fields such as neurology, cardiology and sports science. The classical change-point detection methods can not be applied to the segments of time series directly because the change-point of time series segments are totally different from the change-point within a piece of time series. To coordinates with the existing methods, we need an appropriate statistic to summarize the segments into scalars or scores, then apply the change-point detection method to the scalars or scores.

We have proposed an innovative nonparametric relative entropy (RlEn) for the change-point detection in time series segments. It is a fact that the relative entropy is not only *transformation invariant* but also *background-noise-free*. More generally, we extend the relative entropy to the nonparametric settings. We have not only clarified the detailed steps of the nonparametric RlEn estimation, but also established a consistency theory of nonparametric RlEn. Under certain assumptions, the limiting distribution of nonparametric RlEn is Gaussian normal with of order $\sqrt{n}h^{(m+1)/2}$ where $m$ has an upper bound. Furthermore, we recommand using the BIC criterion to select the pre-determined parameter $m$. The consistency theory of BIC is developed to ensure that the estimator converges to the true lag order with probability 1. The results show that our algorithms of lag order selection and change-point detection using the RlEn are efficient in nonparametric settings. Finally, we apply our method to two real datasets: muscle contraction and Covid-19 dataset respectively to verify its performance in practice.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| ACC | **Acc**uracy of contingency table. |
| ACF | **A**uto-**c**orrelaton **F**unction. |
| AIC | **A**kaike **I**nformation **C**riterion. |
| ApEn | **Ap**proximate **En**tropy. |
| AR | **A**uto-**R**egression. |
| ARIMA | **A**uto-**R**egression **I**ntegrated **M**oving-**A**everage. |
| ARMA | **A**uto-**R**egression **M**oving-**A**everage. |
| ASE | **A**verage **S**quared **E**rror. |
| ATP | **A**denosine **Tri**phosphate. |
| BIC | **B**ayes **I**nformation **C**riterion. |
| CAPM | **C**apital **A**sset **P**ricing **M**odel. |
| CUSUM | **Cu**mulative **Sum**. |
| CV | **C**ross **V**alidation. |
| DAC | **D**ivide-**a**nd-**C**ombine. |
| DCM | **D**ynamic **C**ovariance **M**odel. |
| EDF | **E**ffective **D**egrees of **F**reedom. |
| EEG | **E**lectro**e**ncephalo**g**ram. |
| EWF | **E**ntry-**w**ise **F**unction. |
| FDR | **F**alse **D**iscovery **R**ate. |
| FzEn | **F**u**z**zy **En**tropy. |
| i.i.d. | **i**ndependent and **i**dentical **d**istribution. |
| IRSE | **I**ntegrated **R**oot-**s**quared **E**rror. |
| KDE | **K**ernel **D**ensity **E**stimation. |
| MA | **M**oving-**A**verage. |

| | |
|---|---|
| MAD | **M**ean **A**bsolute **D**istance. |
| MEG | **M**agneto**e**ncephalo**g**raphy. |
| MsEn | **M**ulti-**s**cale **En**tropy. |
| NCM | **N**onparametric **C**ovariance **M**odel. |
| OLS | **O**rdinary **L**east **S**quares. |
| PACF | **P**artial **A**uto-**c**orrelaton **F**unction. |
| PCA | **P**rincipal **C**omponent **A**nalysis. |
| PDF | **P**robability **D**ensity **F**unction. |
| PI | **P**lug-**i**n. |
| RlEn | **R**e**l**ative **En**tropy. |
| SARIMA | **S**easonal **A**uto-**R**egression **I**ntegrated **M**oving-**A**everage. |
| SEN | **Sen**sitivity or True Positive Rate. |
| SKD | **S**liding **kd** tree algorithm. |
| SPE | **Spe**cificity or True Negative Rate. |
| SpEn | **S**am**p**le **En**tropy. |
| SSVD | **S**parse **S**ingular **V**alue **D**ecomposition. |
| TSC | **T**ime **S**eries **C**lassification. |
| VWAP | **V**olume **W**eighted **A**verage **P**rice. |
| WLS | **W**eighted **L**east **S**quares. |

# List of Symbols

| | |
|---|---|
| $I_p$ | the $p$-dimensional identity matrix. |
| $I(\cdot)$ | the indicator function. |
| $\langle A, B \rangle$ | the inner product of matrices $A$ and $B$: $\mathrm{tr}(AB^T)/p$. |
| $\mathbb{I}$ | the interval $[0, 1]$. |
| $\lambda_{\max}(\cdot)$ | the maximum eigenvalue of a square matrix. |
| $\lambda_{\min}(\cdot)$ | the minimum eigenvalue of a square matrix. |
| $\|A\|_F$ | for a square matrix $A = (a_{ij})_{p\times p}$, its size-self-normalized Frobenius norm: $\sqrt{\mathrm{tr}(AA^T)/p}$. |
| $\|A\|_\infty$ | the infinity norm: $\max_{1\le i\le p} \sum_{i=1}^{n} \lvert a_{ij}\rvert$. |
| $\|A\|_{\max}$ | the maximum norm: $\max_{1\le i,j\le p} \lvert a_{ij}\rvert$. |
| $\|A\|$ | the spectral norm: $\lambda_{\max}^{1/2}(AA^T)$. |
| $\|x\|$ | the Euclidean norm of vector $x$. |
| $\mathrm{Diag}(A)$ | the diagonal matrix extracted from matrix $A$. |
| $\mathrm{diag}(x)$ | the diagonal matrix generated by vector $x$. |
| $\mathbb{1}(\cdot)$ | the indicator function. |
| $\mathcal{K}_h^{(m)}(\mathbf{x})$ | the multivariate Jackknife kernel function with dimension $m$. |
| $K_h^J(x)$ | the scaled Jackknife kernel function. |
| $K_h(x)$ | the scaled kernel function. |
| $K(x)$ | the kernel function. |
| $\mathbb{H}_0$ | the null hypothesis. |
| $\lceil x \rceil$ | the ceiling function of $x$. |
| $\lfloor x \rfloor$ | the floor function of $x$. |

| | |
|---|---|
| $\mathbb{N}$ | the integer set. |
| $A \succ 0$ | square matrix A is positive definite. |
| $tril(A)$ | the strictly lower triangular matrix of A. |
| $c \vee d$ | the maximum of two numbers $c$ and $d$. |
| $c \wedge d$ | the minimum of two numbers $c$ and $d$. |
| $\mathbb{R}$ | the real number set. |

# List of Algorithms

# Chapter 1

# Introduction

In contemporary statistical inference, covariance matrix estimation attracts lots of interests, and remarkable achievements have been made in the past two decades (Pourahmadi, 2013). For the estimation of covariance matrix in multiple regression under high-dimensional settings, the challenges of *high-dimensionality* and *positive definiteness* are extensively studied (Pourahmadi, 2013). There exist many classical methods in literature to make the covariance matrix positive definite in high-dimensional settings. For example, the sparse principal component analysis (PCA) (Shen & Huang, 2008; Johnstone & Lu, 2009), sparse singular value decomposition (SSVD) (Witten *et al.*, 2009; Lee *et al.*, 2010; Chen *et al.*, 2012; Yang *et al.*, 2014), sparse Gaussian graphic models (Huang *et al.*, 2006; Yuan & Lin, 2007; Friedman *et al.*, 2008; Lam & Fan, 2009; Peng *et al.*, 2009; Rothman, 2012) and their variants. The common underlying assumption in the above models is that the covariance matrix is fixed. However, the covariance matrix could be covariate-dependent in practice. In this circumstance, we want to estimate covariance matrix function.

In literature, there are a lot of approaches on the nonparametric covariance function estimation (e.g., see Hall *et al.*, 1994; Dette & Neumeyer, 2001; Yin *et al.*, 2010; Li, 2011; Chen *et al.*, 2013; Chen & Leng, 2016; Chen *et al.*, 2018; Wang *et al.*, 2020; Qiao *et al.*, 2020, and among others). Basically, there are two kinds of approaches on the nonparametric covariance function estimation. The first kind of approach directly estimates covariance matrix function or its inverse function. Covariance matrix and its precision matrix play a significant role in time-varying graphic model (e.g., Chen *et al.*, 2013). For the estimation of precision matrix, various penalties are imposed on the entries of precision matrix (the partial correlation coefficients) to form the networks changing with time (e.g., see Ahmed & Xing, 2009; Kolar *et al.*, 2010; Zhou *et al.*, 2010; Lu *et al.*, 2017; Hallac *et al.*, 2017; Yang & Peng, 2018, and among others). For the nonparametric covariance estimation, we refer to these existing works (e.g., see

Yin *et al.*, 2010; Chen & Leng, 2016)

The second kind of approach uses the factor model to estimate the covariance matrix (e.g., Chamberlain & Rothschild, 1983). Fan *et al.* (2013) focused on the static factor model and proposed *large covariance estimation by thresholding principal orthogonal complements*. Recently, Wang *et al.* (2020) proposed a *nonparametric estimation of large covariance matrices* which employed kernel smooth techniques to estimate the covariance matrix functions of both independent variables and noises under the framework proposed by Fan *et al.* (2013).

For the above two kinds of approaches, bandwidth selection plays an important role in covariance matrix estimation. However, the above models of nonparametric covariance matrix function did not clearly clarify the effect of sparsity on the bandwidth selection. In addition, to satisfy the positive definiteness (Yin *et al.*, 2010; Chen & Leng, 2016; Guo *et al.*, 2017), only one bandwidth is adopted in the procedure of covariance function estimation. This might be inappropriate because it is far-fetched to let the entries of covariance matrix share one common bandwidth. For example, when the covariance matrix function is sparse, zero entries have smoothness different from the non-zero entries.

Hence, compared to the challenges of covariance matrix estimation mentioned above, the sparsity effect on the estimation of covariance matrix function is also a challenge in nonparametric settings.

## 1.1 Nonparametric Covariance Model

To address the estimation of covariance matrix function under low-dimensional settings, Yin *et al.* (2010) proposed a general nonparametric covariance matrix estimation framework. The authors not only discussed the sampling properties of nonparametric covariance model (NCM) but also obtained the asymptotic normality of NCM. However, similar to the discussion in Fan *et al.* (2013), NCM also suffers from a curse of dimensionality. For example, in asset portfolio risk analysis, modelling market-dependent co-volatility of $p$ assets by use of historical return data over $n$ consecutive months involves estimating $p(p+1)/2$ nonparametric curves (Fama & French, 2004).

To extend the NCM to high-dimensional settings, Chen & Leng (2016) proposed the so-called dynamic covariance model (DCM). One advantage of DCM framework is the proposed *subset-y-variables* cross-validation procedure (Chen & Leng, 2016, p. 1200) which can eliminate the influence of high dimensionality in the kernel estimation of covariance matrix function. They also developed a uniform consistency theory of DCM in high-dimensional setting.

To satisfy the positive definiteness, Yin *et al.* (2010) and Chen & Leng (2016)

forced the entries of the covariance matrix to share a common bandwidth. However, it sacrifices the smoothness variability of entries. In more general case, the smoothness of covariance matrix entries could be different, which means that multiple bandwidths should be involved in the estimation of covariance matrix. At the same time, the desirable positive definite property should also be satisfied. Hence, it is a challenge to develop a framework that involves the multiple bandwidths and can satisfy the positive definiteness simultaneously.

The last challenge is the sparsity effect on covariance matrix function. The scientific researchers usually impose sparsity assumptions on the covariance matrix to ensure the consistency of covariance matrix with low-rank structure (Shen & Huang, 2008; Amini & Wainwright, 2008; Johnstone & Lu, 2009; Fan *et al.*, 2013, and among others). Similar to the covariance matrix estimation in high-dimensional settings, we also impose the sparsity assumption on the covariance matrix function. Throughout this thesis, we assume the covariance matrix function is sparse and the locations of zero entries in covariance matrix function are not dependent on the covariate.

We notice that Chen & Leng (2016) employed the threshold approach of covariance regularization (Bickel & Levina, 2008b) to make the covariance matrix estimator consistent under the sparsity assumption in high-dimensional regime. However, based on the framework of DCM, there is no discussion about the effect of sparsity on the bandwidth selection in the cross validation step. To achieve the sparsity, they used the threshold approach which immediately follows the cross validation step. In fact, the sparsity has a significant effect on the bandwidth selection. For example, if the sparsity of covariance matrix function is very high, say 95%, then the zero entries will dominate the bandwidth selection and let the bandwidth tend to infinity in cross validation step, see the pilot study in Section 3.2.4. Therefore, the effect of sparsity on covariance matrix function exists from the beginning of estimation rather than after the cross validation step.

Another goal of this research is to establish a novel estimation framework of covariance matrix function considering the sparsity effect from commence. It is not a simple adjustment of the estimation steps in DCM's framework. Actually, we have developed a novel framework to address the sparsity effect from the point view of the Divide-and-Combine approach.

The methods for solving the sparsity effect, tackling the conflict between positive definiteness and multiple bandwidths consist of the first task of this research. The second task that this research concentrates on in nonparametric high-dimensional settings is the change-point detection in time series segments.

## 1.2 Change-point Detection in Time Series Segments

In sports science, one type of time series segments can be described as follows: The time series consists of data from a series of consecutive experiments, each experiment corresponds to a different pattern, condition or category. Between two experiment records, there exists one pause designed by the researcher, see Figure 1.1. Figure 1.1 contains 659977 observations of muscle contractions[1]. Each

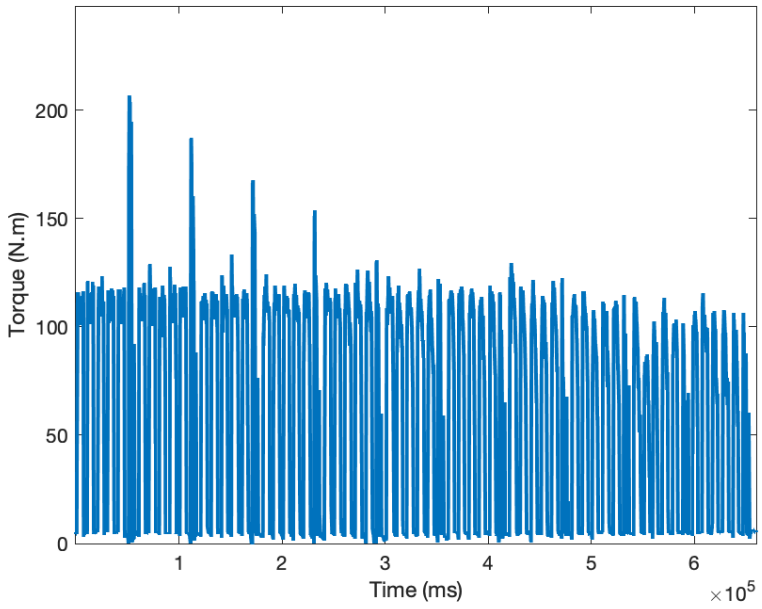



*Figure 1.1: The Data of Muscle Contractions*

contraction lasts six seconds, then the tester has a short break (four seconds). After the rest, another contraction starts. Sports scientists usually call it as the intermittent isometric experiment (Pethick *et al.*, 2016).

This type of signal is widely recorded in neurological field (Burioka *et al.*, 2005), heart rate analysis (Acharya U *et al.*, 2004) and sports science (Forrest *et al.*, 2014). The signal usually contains the information of various patterns or models. For instance, in Magnetoencephalography (MEG) or Electroencephalogram (EEG) experiments, neurologically healthy subjects respond differently in terms of EEG and MEG signal for different stimuli (faces v.s. scrambled faces, or familiar faces v.s. unfamiliar faces, see Wakeman & Henson, 2015); In cardiology, the pattern of heart rate varies along the states of human: sleeping, sitting, walking, jogging or running, see Acharya U. *et al.* (2005), Burioka *et al.* (2005), Shi *et al.* (2017) and among others; In sports science, the energy offered by Adenosine

[1]This dataset is offered by Dr. Mark Burnley and Dr. Samantha Winter, School of Sport and Exercise Sciences, University of Kent.

triphosphate (ATP) will gradually deduce along the time, the torque shows different models, see Figures 5.2(a) and 5.3(a). Scientists have a great interested in detecting the change-points of time series segments to help scientists to identify the patterns of brain activity, the heart disease or improve the performance of athlete.

Throughout this research, the terminology *change-point* refers to the "change-point" of time series segments rather than the "change-point" of a time series. The usual change-points are the break points within one piece of time series, see Killick *et al.* (2012), Fryzlewicz (2014), and Fryzlewicz (2020). However, the change-points in this thesis are the break points among the time series segments. For example, suppose we have extracted 55 time series segments from Figure 1.1. Sports scientists want to know the time of muscle fatigue occurrence. In this thesis, the time of muscle fatigue occurrence is called the change-point of time series segments. It is not the usual change-points within time series (Killick *et al.*, 2012; Fryzlewicz, 2014; Fryzlewicz, 2020).

The natural thoughts of this type of change-point detection is to find an appropriate statistic to compress each segment of time series into a scalar (score). This statistic should preferably have the following two properties: *transformation invariant* (Kullback & Leibler, 1951; Ihara, 1993) and *background-noise-free* (see, Propositions 5.1, 5.4 and 5.5). Then based on the observed scores of segments, one can use CUSUM (Page, 1954) or multiple change-points detection approach (Killick *et al.*, 2012) to find the change-points. A toy example in Section 5.1 shows that neither the mean nor the variance of time series are suitable as the statistic. Hence, the key goal is to find an appropriate statistic owning the above properties. Furthermore, we also need to discuss the merits of this statistic under the stationary ARMA process and nonparametric scenarios in high-dimensional settings. For the ARMA$(p,q)$ process, the high-dimensionality in time series here means that the $p$ and $q$ are infinite. For nonparametric settings, the high-dimensionality in time series here means that the lag order $m$ could be arbitrary large with an upper bound.

Finally, this thesis consists of three novel frameworks. The first two frameworks are estimations for high-dimensional nonparametric covariance models. The third framework is the change-point detection in time series segments under ARMA$(p,q)$ and nonparametric settings. Next, we will discuss the challenges we encountered in high-dimensional context.

## 1.3 Challenges in High-dimensional Context

Throughout the full thesis, we have encountered three challenges in high-dimensional context: sparsity effect, nonparametric correlation estimation with constraint and theory development.

### 1.3.1 Sparsity Effect

Sparsity has a significant effect on the bandwidth selection, see Section 3.2.4. The selected bandwidth of a very sparse covariance matrix could be far away from the true bandwidth, sometimes it will go to infinity. In this case, the errors of nonzero entries generated by the wrong bandwidth will become larger and larger. In low-dimensional setting, Yin *et al.* (2010) suggested using one single bandwidth for the entries of covariance. However, in high-dimensional settings with sparsity, the conflicts among sparsity, single bandwidth and positive definite property seem to be becoming more and more irreconcilable. Avoiding these conflicts is a big challenge in our research.

In fact, there exist two possible technical methods. Method 1 refers to the usage of multiple bandwidths in nonparametric covariance. There are two aspects worthy of attention, one is how to divide the elements of covariance and how many groups (or bandwidths) should we have? Another is how to keep positive definite covariance when multiple bandwidths are adopted. We will introduce a novel framework to address these issues in Chapter 3.

Method 2 needs to detect the zero entries in advance and delete them, then optimize the cross validation function only with respect to the nonzero entries. In this way, zero entries has less effect on the bandwidth selection. By contrast, the framework of Chen & Leng (2016) produces a sparse estimator via thresholding the kernel smoothed covariance after bandwidth selection step. In fact, zero entries still affect the bandwidth selection. Therefore, the swap of zero entries detection step and bandwidth selection step is the key of Method 2. It is also a main challenge to detect the zero entries. Moreover, Method 2 still concerns the positive definite property of covariance matrix. The details of Method 2 can be found in Chapter 4. Both Method 1 and Method 2 can reduce the sparsity effect on nonparametric covariance matrix in the high-dimensional settings.

### 1.3.2 Nonparametric Correlation Matrix Estimation

In Method 2, we divide the estimation procedure of nonparametric correlation matrix into three steps: estimation of diagonal entries, detection of zero entries and estimation of off-diagonal non-zero entries.

Within the estimation of off-diagonal non-zero entries, the nonparametric correlation estimators without constraints could be out of the range $[-1, 1]$, see Figure 4.1. This will bring in extra bias in terms of Frobenius norm loss. If the residuals are from Gaussian probability density function (PDF), we have developed a novel method to estimate the correlation coefficients with constraint via solving the nonparametric cubic equations. Furthermore, for each entry, each given bandwidth and each given explanatory variable without the $i$-th observation in cross validation function (4.16), we need to solve a cubic equation (4.15) and a nonlinear equation (4.13). In high-dimensional settings, the computational complexity increase much faster than the variable dimension. Therefore, it is a challenge to improve the algorithm to speed up the numerical computation of bandwidth selection via criterion (4.16).

### 1.3.3 Theory Development

The uniform consistency theory development in high-dimensional context is the biggest challenge in this research. For i.i.d. case, Yin *et al.* (2010) proved the uniform consistency of covariance matrix function in low-dimensional settings. Furthermore, Chen & Leng (2016) extended the uniform consistency theory to high-dimensional settings. However, for non i.i.d. case, the development of uniform consistency theory of covariance matrix function is difficult under sparsity assumption. For the factorization estimation of NCM, we hope to build a uniform consistency theory not only for the independent error terms but also for the dependent ones in high-dimensional context, see Section 3.4.

As for the change-point detection in time series segments, it is easy to develop the theory of relative entropy for time series in ARMA($p$,$q$) process, see Section 5.2.1. However, it needs to develop a relative entropy theory of time series for $m$-consecutive lags in high-dimensional context even though Hong & White (2005) had obtained an asymptotic distribution of relative entropy for pairwise variables. Furthermore, we have also proposed a criterion to determine the lag order $m$, and developed a corresponding theory for this criterion, see Section 5.2 and Section 5.3 respectively. Next, we briefly clarify the main contributions of the proposed frameworks in this thesis, then introduce the organization of the thesis.

## 1.4 Contributions

There are three contributions in this thesis: Factorized NCM, Divide-and-Combine NCM and the nonparametric relative entropy (RlEn) for the change-point detection. More specifically, two major features of factorized NCM esti-

mation are factorization of covariance matrix by a set of band matrices (3.12) and the criterion of cross-validation without the computation of precision matrix respectively. The former can reduce the effect of sparsity aforementioned in the previous section while the latter can speed up the computation. Most importantly, we have developed a consistency theory of NCM estimator for dependent samples.

In addition to the literal meaning of Divide-and-Combine NCM estimation, there are also two extra contributions to the nonparametric covariance model. One contribution is that the detection of zero entries occurs earlier than the selection of bandwidth. In this case, we can ignore the zero entries to reduce their effect on the bandwidth selection of off-diagonal nonzero entries. The other contribution is to propose a new nonparametric estimation approach of correlation coefficient with constraint, for instance, see Figure 4.1. It can guarantee that the correlation coefficient estimators exactly lie within the interval $[-1, 1]$. Next, we discuss the contributions of nonparametric relative entropy.

Firstly, we investigate the properties of nonparametric relative entropy when the time series is stationary. It concludes that the relative entropy has two properties: *transformation invariant* and *background-noise-free* for ARMA($p$,$q$) process. Furthermore, the relative entropy remains the same value if $m \geq p$ for AR($p$) or $m \geq q + 1$ for MA($q$) with finite $p$ and $q$. Secondly, in nonparametric settings, we not only propose a nonparametric relative entropy statistic for time series, but also develop a consistency theory of nonparametric relative entropy for i.i.d. samples. The limiting distribution of RlEn is Gaussian under the appropriate assumptions. Furthermore, we also construct a convergence theory of the Bayesian information criterion (BIC) for the lag order selection.

## 1.5 Organization of The Thesis

The statistical models we proposed in high-dimensional nonparametric settings constitute the following chapters in detail.

In Chapter 2, we briefly review the basic nonparametric models, bandwidth selection approaches as well as the complexity measures for time series in literature. These reviews and background are the cornerstones of the following chapters.

In Chapter 3, we have dwelt on the establishing procedures of factorized estimation of high nonparametric covariance models. This framework typically consists of five steps: standardization, factorization, bandwidth selection, thresholding and shrinkage.

In Chapter 4, we have proposed another framework to address the influence

of sparsity on covariance estimators. Inspired by the Divide-and-Conquer algorithm (Cormen, 2009), we employ this similar idea, called Divide-and-Combine, to reduce the influence of zero entries on the bandwidth selection.

In Chapter 5, we have discussed a new relative entropy for time series. This type of relative entropy is a two-stage procedure including lag order selection and relative entropy estimation.

Finally, Chapter 6 contains the conclusions and future works of this thesis. It not only highlights the conditions or constraints of the models we proposed in this thesis, but also suggests some potential extensions of our models in the future. The tables, figures and main proofs of Chapter 3, Chapter 4 and Chapter 5 are postponed to Appendix A, Appendix B and Appendix C respectively.

# Chapter 2

# Literature Review and Background

In this chapter, we give a literature overview of nonparametric covariance model and relative entropy. It is hard to review all the fields of *Nonparametric Statistics* since it is a broad research area. So, we limit our literature review to the areas of nonparametric covariance model, relative entropy, bandwidth selection and the relevant statistical methods used in this thesis.

## 2.1 Nonparametric Mean and Covariance Models

### 2.1.1 Nonparametric Mean Regression Model

To introduce the basic ideas of Nadaraya-Watson and local polynomial kernel estimators, we start from the nonparametric homoscedastic regression model:

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\{(x_i, y_i), i = 1, \ldots, n\}$ is a sample of random variable $(X, Y)$ satisfying Model (2.1). $\varepsilon_i, i = 1, \ldots, n$ are independent and identity distribution (i.i.d.) samples drawn from a distribution with mean zero and constant variance. $\varepsilon_i$ and $x_i, i = 1, \ldots, n$ are mutually independent. The *mean function* $m(\cdot)$ is a smooth unknown function of $X$. The support of $X$ could be $\mathbb{R}$ or an interval.

Let $K(u)$ be a kernel function and $K_h(u) = h^{-1}K(u/h)$ represent its scaled kernel function where $h \in \mathbb{R}^+$ is called the bandwidth or smoothing parameter. Nadaraya (1964) and Watson (1964) proposed the following estimator of

mean function, namely,

$$\hat{m}(x_i) = \sum_{j=1}^{n} \frac{K_h(x_j - x_i) y_i}{\sum_{s=1}^{n} K_h(x_s - x_i)} = \sum_{j=1}^{n} w_{ij} y_i. \tag{2.2}$$

In literature, people usually call (2.2) as Nadaraya-Watson type estimator.

Another important and extensively-used estimation of mean function in statistics is the local polynomial method, for example, see Section 5.2 in Wand & Jones (1995) or Chapter 3 in Fan & Gijbels (1996). We will briefly review the framework of local polynomial kernel estimator with degree $p$. Suppose $m(\cdot)$ has the $p$-th derivative at the given point $x_0$. For any $x$ in the neighbourhood of $x_0$, by the Taylor expansion, the mean function $m(x)$ can be approximated by

$$m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p.$$

Therefore, one can fit the above polynomial with respect to $x - x_0$ via minimizing $\sum_{i=1}^{n} [y_i - \sum_{s=0}^{p} \beta_s (x_i - x_0)^s]^2 K_h(x_i - x_0)$. Let the design matrix be

$$\mathbf{X}_{p,x_0} = \begin{bmatrix} 1 & x_1 - x_0 & \cdots & (x_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_0 & \cdots & (x_n - x_0)^p \end{bmatrix}.$$

Denote $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$, and $\mathbf{W}_{h,x_0}$ is an $n \times n$ diagonal matrix with $K_h(x_i - x_0), i = 1, \ldots, n$ on the main diagonal. The weighted least square (WLS) estimator of $\left[m(x_0), m'(x_0), \ldots, m^{(p)}(x_0)/(p!)\right]^T$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_{p,x_0}^T \mathbf{W}_{h,x_0} \mathbf{X}_{p,x_0}\right)^{-1} \mathbf{X}_{p,x_0}^T \mathbf{W}_{h,x_0} \mathbf{y}.$$

In practice, of great interest is the estimator of $m(x_0)$. Let $e_1$ be a vector in which the first entry is 1 and the other entries are zeros, then we have

$$\hat{m}(x_0) = e_1^T \left(\mathbf{X}_{p,x_0}^T \mathbf{W}_{h,x_0} \mathbf{X}_{p,x_0}\right)^{-1} \mathbf{X}_{p,x_0}^T \mathbf{W}_{h,x_0} \mathbf{y} = e_1^T \mathbf{S}_{p,h,x_0} \mathbf{y}, \tag{2.3}$$

where $\mathbf{S}_{p,h,x_0}$ is called the local polynomial smoother matrix. Especially, we note that estimator (2.3) degenerates to the Nadaraya-Watson estimator when $p = 0$. If $p = 1$, we call (2.3) a local linear kernel estimator, a simple form of $e_1^T \mathbf{S}_{p,h,x_0}$ can be easily obtained, e.g., the Equation (5.4) in Wand & Jones (1995).

Except for the kernel-based methods mentioned above, there are many other nonparametric methods for mean function estimation such as wavelet thresholding (Donoho & Johnstone, 1994; Donoho, 1994; Donoho & Johnstone, 1995;

Donoho, 1995; Donoho *et al.*, 1995; Donoho & Johnstone, 1998), spline smoothing (Kooperberg & Stone, 1991; Green & Silverman, 1994; Kooperberg *et al.*, 1995a; Kooperberg *et al.*, 1995b; Nychka, 1995), additive model and generalized additive model (Friedman & Stuetzle, 1981; Buja *et al.*, 1989; Hastie & Tibshirani, 1999).

Furthermore, there are also massive efficient approaches proposed to solve some specific issues in literature. For example, these issues include but not limited to: the optimal bandwidth selection, the boundary correction, the consistency theories, the combination with other regression models such as generalized linear regression, time series, multivariate regression, etc. These kernel-based models can be found in some classical nonparametric textbooks, e.g., Härdle (1990), Fan & Gijbels (1996), Wasserman (2006), and Li & Racine (2007).

These discussions are beyond the scope of our research because the first part of this thesis concentrates on the covariance function estimation. Specifically, it is no doubt that one can apply the local polynomial kernel regression method to the estimation of variance function. We will put off the review of general approaches with respect to the variance and covariance function in the following sections.

### 2.1.2 Nonparametric Variance Model

Hall & Marron (1990), Ruppert *et al.* (1997), Fan & Yao (1998), Yu & Jones (2004) and the references therein developed the univariate nonparametric variance model from different perspectives. Suppose that $(X, Y)$ are a pair of random variables, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ represent observations from model (2.4),

$$y_i = m(x_i) + \varepsilon_i, \qquad \mathrm{var}(\varepsilon_i|x_i) = v(x_i), \quad i = 1, \ldots, n, \tag{2.4}$$

where $\varepsilon_i, i = 1, \ldots, n$ are independent random variables with $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^4|x_i) < \infty$. Suppose both $m(\cdot)$ and $v(\cdot)$ are unknown smooth nonlinear functions of variable $X$.

Next, we discuss the classical estimation methods of variance function under the univariate scenario. In fact, if the mean function estimator (either parametric or nonparametric) can be absorbed by $y_i$, then Model (2.4) degenerates to

$$y_i = \varepsilon_i, \qquad \mathrm{var}(\varepsilon_i|x_i) = v(x_i), \quad i = 1, \ldots, n, \tag{2.5}$$

where in this case $E(y_i|x_i) = 0$. Model (2.5) represents a centralization of Model (2.4). Within the scope of this research, it is convenient to focus on the estimation of variance function regardless of the mean function temporarily. To keep the completeness, we still consider the mean function in the rest of this

chapter. But keeping in mind, even an explicit $m(\cdot)$ estimator is specified, it can be replaced by other existing nonparametric mean estimator in literature.

As for the variance function $v(\cdot)$, the current estimation approaches can be divided into two categories according to its homoscedasticity or heteroscedasticity. The rest of Section 2.1.2 introduces three classical variance function estimations. The first framework is based on the Nadaraya-Watson kernel under homoscedasticity assumption. The rest two frameworks address the boundary correction and positive definite problems under heteroscedasticity assumption respectively.

#### 2.1.2.1 Nadaraya-Watson Estimators

In nonparametric homoscedastic regression, there exist many methodologies in time series context, for example, see Rice (1984), Gasser *et al.* (1986), Hall & Marron (1990) and the references therein. Among these methodologies, we take the variance function estimation (Hall & Carroll, 1989) as an example to illustrate the framework in homoscedastic settings.

Suppose Model (2.4) holds with homoscedastic variance, i.e., $v(x_1) = \cdots = v(x_n) = \sigma^2$. Let $K(u)$ and $K_h(u)$ be the kernel and scaled kernel functions respectively, $h$ is a bandwidth. For simplicity, the mean function is estimated by the Nadaraya-Watson type of estimator in (2.2).

As mentioned above, one can use other types of mean function smoothers here, or just assume $m(\cdot)$ is zero by following the suggestion in Hall & Carroll (1989). Nevertheless, the kernel estimators of errors are $\hat{\varepsilon}_i = y_i - \hat{m}(x_i), \quad i = 1, \ldots, n$. The effective degrees of freedom (EDF) for errors (e.g., the definition in Hastie & Tibshirani, 1999, p. 54) has a simple form: $n_d = 2\sum_{i=1}^{n} w_{ii} - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$. Therefore, for the nonparametric homoscedastic regression, the Nadaraya-Watson type estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{1}{n - n_d}\sum\nolimits_{i=1}^{n} \hat{\varepsilon}_i^2. \tag{2.6}$$

As homoscedasticity, the error items in (2.6) share one common weight $1/(n-n_d)$. It is straightforward to extend estimator (2.6) to the heteroscedastic settings. Suppose that $K^*(x)$ and $K^*_{h^*}(x)$ represent another kernel and scaled kernel functions respectively, $h^*$ is another bandwidth. $v(\cdot)$ is an unknown smooth function of $x_i$. For any given $x_0$ in the support of $X$, the variance estimator is

$$v(x_0) = \sum\nolimits_{i=1}^{n} w_i^* \hat{\varepsilon}_i^2, \tag{2.7}$$

where $w_i^* = K^*_{h^*}(x_i - x_0)/\sum_{i=1}^{n} K^*_{h^*}(x_i - x_0)$. However, estimator (2.7) will encounter the boundary effect problem when $x_0$ is close to the boundaries of the support of $X$. There is a review on the boundary correction, see a list of methods

in Karunamuni & Alberts (2005). Among these existing methods, local linear smoother attracts many attentions due to its automatic correction (Cheng *et al.*, 1997) and asymptotic minimax efficiency properties (Fan *et al.*, 1997) in the context of nonparametric regression. A full and systematic introduction of local polynomial regression and its applications can be found in the textbooks (e.g., Fan & Gijbels, 1996; or Wand & Jones, 1995).

#### 2.1.2.2 Local Polynomial Estimators

Ruppert *et al.* (1997) applied the local $p$-th order polynomial smoother to the estimation of variance function (e.g., Härdle & Tsybakov, 1997). Suppose Model (2.4) holds and $h_1$, $h_2$ are two bandwidths distinguishing from $h$ and $h^*$. We notice that Ruppert *et al.* (1997) applied the local polynomial regression both to $m(\cdot)$ and $v(\cdot)$.

Given $h_1$ and the polynomial degree $p_1$, by estimator (2.3), we have

$$\hat{m}(x_0) = e_1^T \left(\mathbf{X}_{p_1,x_0}^T \mathbf{W}_{h_1,x_0} \mathbf{X}_{p_1,x_0}\right)^{-1} \mathbf{X}_{p_1,x_0}^T \mathbf{W}_{h_1,x_0} \mathbf{y} = e_1^T \mathbf{S}_{p_1,h_1,x_0} \mathbf{y}, \tag{2.8}$$

where $\mathbf{X}_{p_1,x_0}, \mathbf{W}_{h_1,x_0}, \mathbf{S}_{p_1,h_1,x_0}$ are the copies of $\mathbf{X}_{p,x_0}, \mathbf{W}_{h,x_0}, \mathbf{S}_{p,h,x_0}$ except that $(p,h)$ is replaced by $(p_1,h_1)$. When $x_0$ in (2.8) takes the values of $\{x_1,\ldots,x_n\}$, the estimator of $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^T$ can be written as $\hat{\boldsymbol{\varepsilon}} = (I_p - \mathbb{S}_{p_1,h_1})\,\mathbf{y}$, where the $i$-th row of $\mathbb{S}_{p_1,h_1}$ is $e_1^T \mathbf{S}_{p_1,h_1,x_i}$ and $I_p$ is an identity matrix.

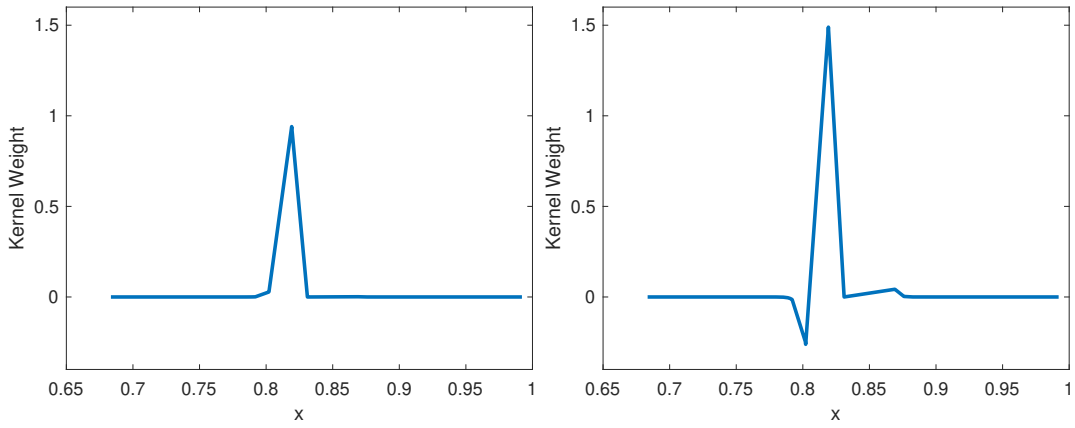
The contribution of Ruppert *et al.* (1997) is the local polynomial fit of $v(\cdot)$ with respect to the squared residuals $\hat{\boldsymbol{\varepsilon}}^2$. Given $p_2, h_2, x_0$, they proposed the following estimator

$$\hat{v}(x_0) = \hat{v}(x_0; p_1, h_1, p_2, h_2) = \frac{e_1^T \mathbf{S}_{p_2,h_2,x_0} \hat{\boldsymbol{\varepsilon}}^2}{1 + e_1^T \mathbf{S}_{p_2,h_2,x_0} \Delta}, \tag{2.9}$$

where $\Delta = \operatorname{diag}\left(\mathbb{S}_{p_1,h_1} \mathbb{S}_{p_1,h_1}^T - 2\mathbb{S}_{p_1,h_1}\right)$ represents an EDF vector. The denominator of (2.9) originates from the variance estimator in homoscedastic linear regression to reduce the bias. For example, in estimator (2.6), $\sum_{i=1}^n \hat{\varepsilon}_i^2$ is divided by $n - n_d$ rather than $n$ which makes (2.6) unbiased. The term $1 + e_1^T \mathbf{S}_{p_2,h_2,x_0} \Delta$ plays the similar role here. In nonparametric heteroscedastic regression, estimator (2.9) is biased even if it is adjusted by $1 + e_1^T \mathbf{S}_{p_2,h_2,x_0} \Delta$, see Theorem 1 in Ruppert *et al.* (1997).

It needs to point out that one can use other estimators of $m(\cdot)$ as long as the terms $\hat{\boldsymbol{\varepsilon}}^2$ and $\Delta$ in (2.9) change correspondingly. Another aspect needed to emphasize here is the equivalent kernel, see its definition and details in Section 3.2 in Fan & Gijbels (1996). The equivalent kernel $e_1^T \mathbf{S}_{p_2,h_2,x_0}$ might be negative such that $\hat{v}(x_0)$ is negative which violates the constraint $v(\cdot) \geq 0$. For instance,

let $p_2 = 0, 1$, Figure 2.1 clearly shows that the equivalent kernel of local linear ($p_2 = 1$) could be negative, the variance estimator at $x_{86} = 0.831$ is -0.005. In the following paragraph, we will review the methods that could solve this problem.




*(a)* $p_2 = 0$ *(b)* $p_2 = 1$

Note: this simulation is based on Model (2.5) with the following parameter settings: $n = 100$, $h = 0.01$, $v(x_i) = 1 - x_i^2$, $x_i$ is randomly generated from $U(0, 1)$. $y_i$, $i = 1, \ldots, n$ are randomly drawn from $N(0, v(x_i))$. The two figures show the equivalent kernels of local constant and linear regression in cross validation step.

*Figure 2.1: Equivalent Kernel Comparison*

#### 2.1.2.3 Maximum Locally Likelihood Estimators

The *maximum locally likelihood estimators* of variance functions (e.g., Fan & Yao, 1998; Yu & Jones, 2004) not only address the problem of negative variance, but also have the fully regression-adaptive merit, i.e., without knowing $m(\cdot)$, estimator $v(\cdot)$ performs as well as the variance estimator whence $m(\cdot)$ are known (Fan & Yao, 1998, p. 3).

The framework of maximum locally likelihood estimators requires that the distribution of $\varepsilon_i$ is known. Following Fan & Yao (1998) and Yu & Jones (2004), we demonstrate this framework assuming $\varepsilon_i$'s are independent and sampled randomly from a normal distribution. One can easily extend it to other error distributions.

The log-likelihood function can be expressed as

$$-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{\hat{\varepsilon}_i^2}{v(x_i)} + \log\left(v(x_i)\right)\right]. \tag{2.10}$$

Given $x_0$, Fan & Yao (1998) replaced $v(x_i)$ in (2.10) with its linear approximation $\alpha(x_0) + \beta(x_0)(x_i - x_0)$ at $x_0$, and proposed the following local log-likelihood

function

$$L(x_0,h)=-\frac{1}{2}\sum_{i=1}^{n}\Bigg[\frac{\hat{\varepsilon}_i^2}{\alpha(x_0)+\beta(x_0)(x_i-x_0)} + \log\left(\alpha(x_0)+\beta(x_0)(x_i-x_0)\right)\Bigg]K_h(x_i-x_0). \tag{2.11}$$

However, in (2.11), it needs special attention on the estimators of $\alpha(x_0), \beta(x_0)$ such that the logarithm item $\log\left(\alpha(x_0)+\beta(x_0)(x_i-x_0)\right)$ is well-defined. In contrast, Yu & Jones (2004) recommended replacing $v(x_i)$ with its log-linear expansion $\exp\left(\alpha_1(x_0)+\beta_1(x_0)(x_i-x_0)\right)$ which yields

$$L(x_0,h)=-\frac{1}{2}\sum_{i=1}^{n}\Bigg[\frac{\hat{\varepsilon}_i^2}{\exp\left(\alpha_1(x_0)+\beta_1(x_0)(x_i-x_0)\right)} + \alpha_1(x_0)+\beta_1(x_0)(x_i-x_0)\Bigg]K_h(x_i-x_0). \tag{2.12}$$

The advantage of (2.12) is that no extra constraints on $\alpha_1(x_0), \beta_1(x_0)$ are needed. The variance estimator at $x_0$ is $\hat{v}(x_0)=\exp(\hat{\alpha}_1(x_0))$. It is worth pointing out that there exist substantial approaches aforementioned to obtain $\hat{\varepsilon}_i$ regardless of whether $m(\cdot)$ is known or unknown. In Yu & Jones (2004), (2.12) is called the local log-linear estimator. We adopt the log-linear estimation of variance functions in Chapter 4 to evaluate the diagonal entries of covariance matrix.

So far, we have briefly reviewed the key works on variance function estimation under the univariate scenarios. These basic models and approaches are adopted and modified appropriately in Chapter 3 and Chapter 4 depending on the context of high-dimensional settings. Next, we discuss the existing models on estimation of covariance functions in literature.

### 2.1.3 Nonparametric Covariance Model

Yin *et al.* (2010) proposed the nonparametric covariance model under the low-dimensional settings which is the foundation of our research. Furthermore, Chen & Leng (2016) developed a novel dynamic covariance model to overcome the curse of dimensionality. Next, we provide a brief overview of these two models.

Let $\mathbf{Y}=(Y_1,\ldots,Y_p)^T\in\mathbb{R}^p$ be a $p$-dimensional random vector and $U\in\mathbb{R}$ be an independent random variable. Suppose that $(\mathbf{y}_i,u_i)_{i=1}^n$ with $\mathbf{y}_i=(y_{i1},\ldots,y_{ip})^T$ are random observations from the population $(\mathbf{Y},U)$, satisfying the equations $\mathbf{y}_i=\boldsymbol{\mu}(u_i)+\Sigma(u_i)^{1/2}\varepsilon_i,\quad i=1,\ldots,n$, where $\boldsymbol{\mu}(u_i)=(\mu_1(u_i),\ldots,\mu_p(u_i))^T$ and given $(u_i)_{i=1}^n$, $\varepsilon_i$'s are independent with zero means and unity covariance matrices (i.e., $E[\varepsilon_i|u_i]=0_p$, $\mathrm{cov}(\varepsilon_i|u_i)=I_p$). Let $K(u)$ and $K_h(u)=h^{-1}K(u/h)$ be the kernel function and its scaled kernel function respectively with bandwidth

$h > 0$. Yin *et al.* (2010) considered estimators $\hat{\boldsymbol{\mu}}(u) = \sum_{i=1}^{n} w_{ih^*}(u)\mathbf{y}_i$ and

$$\hat{\Sigma}(u) = \sum_{i=1}^{n} w_{ih}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))^T, \tag{2.13}$$

where $w_{ih}(u) = K_h(u_i - u)/\sum_{k=1}^{n} K_h(u_k - u)$, $h^*$ and $h$ are bandwidths of mean and covariance matrix functions respectively. For given $p$, they proved that $\hat{\Sigma}(u)$ is consistent if the convergence rate is of order $\sqrt{1/(nh)} + O(h^2)$.

However, when $p_0 = p(p+1)/2$ is close to or larger than $n$, the kernel covariance estimator proposed by Yin *et al.* (2010) can be degenerate or ill-conditioned with a high condition number. In existing researches for regularization, estimated covariance is usually done by banding, thresholding, or truncating the number of the leading eigenvalues (e.g., Bickel & Levina, 2008a; Cai & Liu, 2011; Fan *et al.*, 2013). Chen & Leng (2016) proposed a method: Dynamic Covariance Model to regularize the kernel covariance model by thresholding covariance entries.

In details, Chen & Leng (2016) used the mean function estimator $\hat{\boldsymbol{\mu}}(u)$ and a different covariance function estimator, namely,

$$\hat{\Sigma}_1(u) = \sum_{i=1}^{n} w_{ih}(u)\mathbf{y}_i\mathbf{y}_i^T - \hat{\boldsymbol{\mu}}(u)\hat{\boldsymbol{\mu}}(u)^T. \tag{2.14}$$

When $p$ is of order $O(\exp(n^{4/5}))$, they showed that both $\hat{\Sigma}(u)$ and $\hat{\Sigma}_1(u)$ are consistent as $n$ tends to infinity. Yin *et al.* (2010) used the following leave-one-out cross validation criterion

$$\begin{aligned} CV_\Sigma(h) = n^{-1}\sum_{i=1}^{n} \Big\{ &[\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i)]^T \hat{\Sigma}_{(-i)}^{-1}(u_i)\,[\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i)] \\ &- \log\left(\left|\hat{\Sigma}_{(-i)}^{-1}(u_i)\right|\right)\Big\}, \end{aligned} \tag{2.15}$$

in bandwidth selection, where $\hat{\Sigma}_{(-i)}(u_i)$ is the estimator computed according to $\hat{\Sigma}(u)$ or $\hat{\Sigma}_1(u)$ but without the $i$th observation. So when $p_0 \gg n$, it is impossible to estimate the precision matrix (e.g., inverse matrix) of covariance matrix accurately in equation (2.15). To overcome the effect of degeneration, Chen & Leng (2016) proposed a *subset-y-variables* cross-validation procedure.

In particular, they randomly choose $k (k < n)$ entries (scalar variables) from variable vector $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$, denoted as $\mathbf{Y}_s = (Y_{j_1}, \ldots, Y_{j_k})^T$ and repeat this $N$ times. Denote these $N$ subsets index as $s_1, \ldots, s_N$, then the cross validation (2.15) can be expressed as

$$\begin{aligned} CV(h) = &\frac{1}{N}\sum_{j=1}^{N}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big[\left\{\mathbf{y}_{i,s_j} - \hat{\boldsymbol{\mu}}_{s_j}(u_i)\right\}^T \hat{\Sigma}_{s_j(-i)}^{-1}(u_i)\left\{\mathbf{y}_{i,s_j} - \hat{\boldsymbol{\mu}}_{s_j}(u_i)\right\} \\ &+ \log\left(\left|\hat{\Sigma}_{s_j(-i)}(u_i)\right|\right)\Big]\right\}, \end{aligned} \tag{2.16}$$

where $\hat{\Sigma}_{s_j(-i)}(\cdot)$ is obtained by leaving out the $i$-th observation using responses $\mathbf{y}_{i,s_j} = (y_{i,s_{j_1}}, \ldots, y_{i,s_{j_k}})^T$ with the bandwidth $h^*$ and

$$\hat{\boldsymbol{\mu}}_{s_j}(u) = \left\{\sum_{i=1}^{n} K_{h^*}(u_i - u)\mathbf{y}_{i,s_j}\right\}\left\{\sum_{i=1}^{n} K_{h^*}(u_i - u)\right\}^{-1}.$$

In essence, they used the subset-$y$-variables to avoid the case of $p_0 > n$, however this method may omit the correlation among the variables in $(Y_1, \ldots, Y_p)^T$. Before turning to the next section, we need to remind that there also exist Bayes nonparametric covariance models, for example, see Fox & Dunson (2015) and the references therein. The following section is related to the bandwidth selection methods we used in the context of high-dimension.

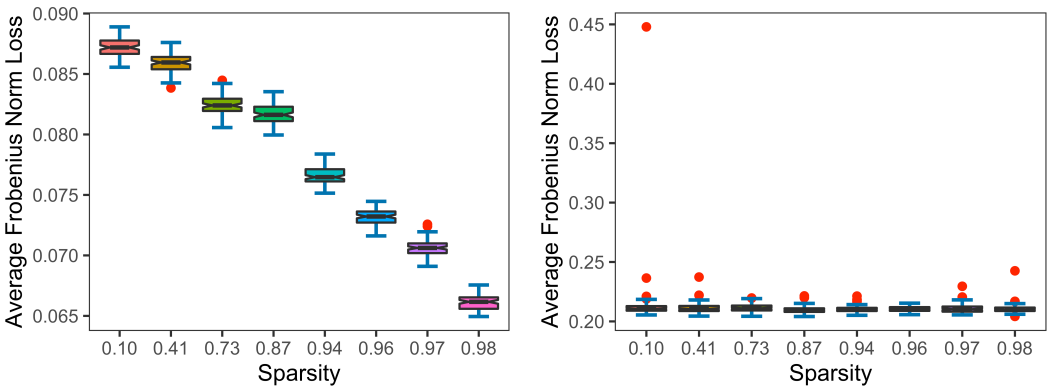
## 2.2 Bandwidth Selections

Both models of Yin *et al.* (2010) and Chen & Leng (2016) require the selection of the smoothing parameter $h$. Over the last 40 years, many bandwidth selection methods have been proposed under different circumstances.

For univariate regressor and single response case, there are four kinds of approaches to deal with bandwidth selection: ASE-based, Cross-validation-based, plug-in and bootstrap method. To the best of our knowledge, Rice (1984) firstly introduced the Average Squared Error (ASE) criterion. The cross-validation (CV) methods can be traced back to Clark (1977). The third refers to the plug-in approach. This approach minimizes the asymptotic mean integrated squared error and the unknown parts are usually replaced by the pilot estimators, for instance, Ruppert *et al.* (1995). Finally, there are also various methods based on bootstrap techniques including but not limited to Cao-abad & González-Manteiga (1993), González Manteiga *et al.* (2004) and references therein. Furthermore, a full overview of bandwidth selection methods can be found in many textbooks (e.g., see Chapter 5 in Härdle, 1990; Chapter 3 in Wand & Jones, 1995; Chapter 4 in Fan & Gijbels, 1996; Jones *et al.*, 1996). Besides, for the bandwidth selection in nonparametric time series, see Sections 5.4 and 6.3.5 in Fan & Yao (2003). For multiple regressors and single response case, the above 4 methods are also applicable. A full overview of bandwidth selection for multivariate kernel can be found from Chapter 3 in Chacón *et al.* (2018).

However, the covariance matrix of the response vector in nonparametric covariance model is a function of an explanatory variable. The bandwidth selection for nonparametric covariance matrix is not equivalent to that for univariate kernel regressor and single response case. If one applies the univariate kernel bandwidth selection methods to each element of covariance matrix and finally combine the

entries' estimators to form the covariance matrix estimator, then the covariance matrix could not be always positive definite, see Remark 1 in Yin *et al.* (2010). Figure 2.2 shows the comparison results of plug-in (PI) and cross validation (CV) methods in terms of Frobenius norm loss for the covariance model under different sparsities. Figure 2.2(a) shows the simulation results using cross validation




*(a) Cross Validation* *(b) Plug-in*

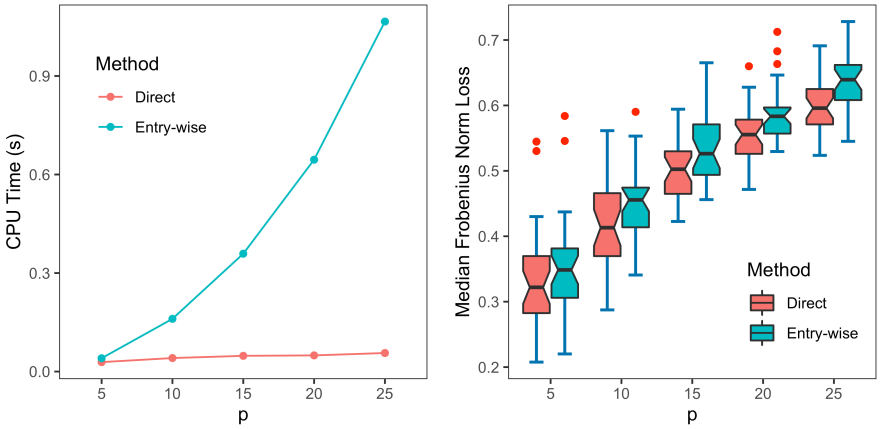
*Figure 2.2: Bandwidth Selection Comparison Between PI and CV for Different Sparsity*

method in Section 3.2.4. In contrast, for the same simulation as in Figure 2.2(a), we apply plug-in method to entries of covariance matrix function to obtain the entry-wise estimators and combine them to form the estimator of covariance matrix (without considering the positive definiteness). Figure 2.2(b) shows that the sparsity has tiny effect on the Frobenius norm loss if the plug-in bandwidth selection is adopted. Furthermore, the Frobenius norm losses of CV are below 0.09 while the Frobenius norm losses of plug-in are around 0.21.

In this case, the bandwidth selected by CV performs better than that selected by PI. In fact, we do not suggest using the entry-wise estimation because of the following two reasons: (1) the entry-wise estimation is not always positive definite even in low-dimensional settings; (2) the computational complexity will increase at a faster rate than the values of $p$, for example, see the CPU-time consumption study in Figure 2.3.

Figure 2.3 shows that the CPU-time consumption increases at a faster rate compared to the "Direct" method when $p$ increases. Furthermore, in this case there is no evidence that the entry-wisely estimation could be better than the "Direct" method in terms of Frobenius norm loss.

In high-dimensional settings, the covariance matrix estimator is not always positive definite even all the entries share one common bandwidth, see the discussions in Chen & Leng (2016). We notice that the cross-validation method at least can guarantee the positive definiteness in low-dimensional settings (Yin *et al.*, 2010). Following their proposal and based on the results in Figure 2.2, we




*(a) CPU time* *(b) Frobenius norm loss*

Note: "Direct" method represents that the bandwidth is chosen by equation (2.18). "Entry-wise" method represents that the bandwidth is chosen entry-wisely. We set $p = 5, 10, 15, 20, 25$, $n = 100$, $M = 30$. The simulation is based on the Setting 1 in Section 3.5.2.

*Figure 2.3: The CPU-time Consumptions of Two CV Methods*

employ the cross validation to choose the bandwidth in high-dimensional settings throughout this thesis.

Recall the mean function estimator (2.2) at $x_i$, the leave-one-out cross validation criterion can be defined as

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2, \tag{2.17}$$

where $\hat{m}_{-i}(x_i)$ represents the mean function estimator without the $i$-th observation. By minimizing (2.17), one can obtain the optimal bandwidth for $m(\cdot)$. Similarly, the criterion of bandwidth selection for nonparametric covariance can be written as the form of (2.15). This type of criterion is based on the log likelihood function which needs to compute the precision matrix $\hat{\Sigma}^{-1}_{(-i)}(u_i)$. It is a fact that for a $p \times p$ matrix, the computational complexity of inverse matrix is of order $O(p^3)$ in general[1]. It will be a computational burden when $(n, p)$ are sufficiently large. To avoid the computation of precision matrix, Biscay *et al.* (1997) suggested using the Frobenius norm loss, i.e.,

$$CV_{\Sigma}(h) = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\Sigma}_{(-i)}(u_i) - [\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i)]\,[\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i)]^T\right\|_F^2, \tag{2.18}$$

as the criterion of bandwidth selection. This research uses the criterion (2.18) to speed up the bandwidth selection, for example, see the comparison of CPU-time

[1]To the best of our knowledge, the optimal complexity of inverse matrix is up to $O(p^{2.373})$ (Davie & Stothers, 2013). However, algorithm design is beyond the scope of our research, we will not go further here.

consumption in Figure 3.1.

As for the bandwidth selection, it needs to highlight here that the entries of covariance matrix share a single bandwidth to keep it positive definite (see Remark 1 in Yin *et al.*, 2010, p. 471). The researchers, such as Chen & Leng (2015), Chen & Leng (2016), Xu *et al.* (2019), Jiang *et al.* (2020) adopt the common bandwidth by default. However, in high-dimensional settings with sparsity assumption, one single bandwidth could increase the Frobenius norm-based loss (see the table results in Appendix A). We develop a novel framework using multiple bandwidths for different entries of covariance matrix at the meanwhile the positive definite property also holds in theory (see the Factorized NCM in Chapter 3).

In the choice of bandwidth, another aspect that needs attention is the hybrid mean functions. For instance, suppose the mean functions are composed of linear functions and nonlinear functions. If one tends to use local linear smoother (with a common bandwidth) to estimate the mean functions, then it will bring in extra bias because the bandwidth will go to infinity for the linear functions (see, Fan & Gijbels, 1996, p.20 and Section 4.2.1). This indicates us to separate the linear mean functions if the local linear smoother is used. Similarly, if one uses Nadaraya-Watson smoother, then the constant functions should be identified as well. For example, in Chapter 4, we employ the *generalized likelihood ratio statistics* (Fan *et al.*, 2001) to detect the linear functions.

So far, we have reviewed the basic concepts of nonparametric covariance model and bandwidth selection in the context of kernel regression. There are also numerous literature relating to the kernel density estimation (KDE). Due to the space limitation of the thesis, we will not elaborate on the KDE further. Next, we discuss the existing complexity measures of times series.

## 2.3 Complexity Measures of Time Series

In the past few decades, various type of entropies such as Approximate Entropy (Pincus, 1991), Sample Entropy (Richman & Moorman, 2000), Multi-scale Entropy (Costa *et al.*, 2003), Fuzzy Entropy (Chen *et al.*, 2009) and Relative Entropy (Robinson, 1991; Hong & White, 2005) were proposed to evaluate the system complexity. Here, we briefly review the fundamental concepts of Approximate Entropy, Sample Entropy, Multi-scale Entropy, Fuzzy Entropy and Relative entropy. The topics related to algorithms, parameters determinant and the limiting distribution are also revised later in this section.

### 2.3.1 Approximate Entropy

Pincus (1991) proposed approximate entropy to measure the complexity of time series. To be specific, let $X_1, \ldots, X_N$ be the time series (equally spaced in time) variables and $x_1, x_2, \ldots, x_N$ represent their observations. Define $\mathbf{x}_i^{(m)} = [x_i, \ldots, x_{i+m-1}] \in \mathbb{R}^m$ where $m$ is the lag order. For simplicity, let $n = N - m + 1$, then for each $i, 1 \leq i \leq n$, denote

$$C_i^m(r) = \frac{\#\text{ of } \left\{ j : d\left(\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)}\right) \leq r, 1 \leq j \leq n \right\}}{n}, \tag{2.19}$$

where $d(\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)})$ is the Chebyshev distance between $\mathbf{x}_i^{(m)}$ and $\mathbf{x}_j^{(m)}$, namely, $d(\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)}) = \max_{k=1,\ldots,m}(|x_{i+k-1} - x_{j+k-1}|)$. Furthermore, let $\Phi^m(r) = n^{-1}\sum_{i=1}^n \log(C_i^m(r))$. The definition of approximate entropy (ApEn) can be described as:

$$\text{ApEn}\ (m, r, N) = \Phi^m(r) - \Phi^{m+1}(r). \tag{2.20}$$

In fact, Definition (2.20) is an approximation form of Eckmann-Ruelle (E-R) entropy, i.e., the average of logarithm of conditional probability that $d(\mathbf{x}_i^{(m+1)} - \mathbf{x}_j^{(m+1)}) \leq r$ given $d(\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)}) \leq r$, see Pincus (1991). There are three free parameters: $m, r, N$, each parameter has an effect on the validity of ApEn. However, Pincus (1991) did not discuss the choice of these parameters. In Pincus (1991), the value of $r$ lies in $[0.1 \times SD, 0.2 \times SD]$ for $m = 2, N = 1000$ where $SD$ is standard deviation of $x_i$.

### 2.3.2 Sample Entropy

Sample entropy has been proposed to overcome two disadvantages of ApEn: One disadvantage of ApEn is self-matched, see equation (2.19), where $j$ could be equal to $i$. Another disadvantage is that ApEn heavily depends on the data length $N$, for more discussions, see Richman & Moorman (2000). To be specific, let $d(\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)})$ be the distance between $\mathbf{x}_i^{(m)}$ and $\mathbf{x}_j^{(m)}$, then the sample entropy (SpEn) is defined as

$$\text{SpEn}(m, r, N) = -\log\left(\frac{\sum_{i=1}^{N-m} n_{j\neq i}(m+1)}{\sum_{i=1}^{N-m} n_{j\neq i}(m)}\right), \tag{2.21}$$

where $n_{j\neq i}(m+1)$ is the cardinality of $\{j : d(\mathbf{x}_i^{(m+1)} - \mathbf{x}_j^{(m+1)}) \leq r, j \neq i\}$ while $n_{j\neq i}(m)$ is the cardinality of $\{j : d(\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)}) \leq r, j \neq i\}$. $n_{j\neq i}(m+1)$ is always smaller than or equal to $n_{j\neq i}(m)$ so that equation (2.21) is well-defined. In practice, Richman & Moorman (2000) suggested that one can set $m = 2$ and $r = 0.2SD$ where $SD$ represents the standard deviation of data $x_1, \ldots, x_N$.

### 2.3.3 Multi-scale Entropy

Based on the sample entropy, Costa *et al.* (2003) proposed multi-scale entropy (MsEn) to account for the multiple timescales inherent in time series, i.e., how the sample entropy changes with different timescale. Given the scale factor $\tau$, they first divide the time series into $\lfloor N/\tau \rfloor$ windows, then take an average of the data in each window and obtain so-called coarse-grained time series $\{y_j^{(\tau)}\}$, namely,

$$y_j^{(\tau)} = \tau^{-1} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \quad 1 \le j \le \lfloor N/\tau \rfloor.$$

Finally, they apply the basic sample entropy algorithm to $\{y_j^{(\tau)}\}$ to obtain the multi-scale entropy. Specially, when $\tau = 1$, it degenerates to the sample entropy. A real example of multiple entropy can be found in Costa *et al.* (2005). Furthermore, Ahmed & Mandic (2012) extended the multi-scale entropy from univariate to multivariate case.

### 2.3.4 Fuzzy Entropy

Fuzzy entropy (Chen *et al.*, 2009) is an extension of ApEn. The significant difference between fuzzy entropy and ApEn or SpEn is that cardinality is replaced by similarity degree, see equations (2.19), (2.21) and (2.22). To be specific, let $\bar{\mathbf{x}}_i^{(m)}$ be the centralization of $\mathbf{x}_i^{(m)}$, i.e., $\bar{\mathbf{x}}_i^{(m)} = (x_i - \bar{x}_i(m), \ldots, x_{i+m-1} - \bar{x}_i(m))$, where $\bar{x}_i(m) = 1/m \sum_{j=0}^{m-1} x_{i+j}$. Let $d_{ij}(m)$ be the distance between $\bar{\mathbf{x}}_i^{(m)}$ and $\bar{\mathbf{x}}_j^{(m)}$, i.e.,

$$d_{ij}(m) = d\left(\bar{\mathbf{x}}_i^{(m)} - \bar{\mathbf{x}}_j^{(m)}\right). \tag{2.22}$$

Given $q$ and $v$, let $D_{ij}(m, q, v)$ denote the fuzzy function[2] $\exp\left(-(d_{ij}(m))^q/v\right)$, define

$$\Phi^m(q, v) = \frac{1}{N-m} \sum_{i=1}^{N-m} \left( \frac{1}{N-m} \sum_{j=1, j\neq i}^{N-m+1} D_{ij}(m, q, v) \right),$$

and

$$\Phi^{m+1}(q, v) = \frac{1}{N-m} \sum_{i=1}^{N-m} \left( \frac{1}{N-m-1} \sum_{j=1, j\neq i}^{N-m} D_{ij}(m, q, v) \right).$$

Finally, the fuzzy entropy (FzEn) is defined as

$$\text{FzEn}\ (m, q, v, N) = \log \Phi^m(q, v) - \log \Phi^{m+1}(q, v).$$

We want to mention that fuzzy entropy also brings in a new parameter $q$ which increases the complexity of computation.

[2] In fact, Chen *et al.* (2009) did not explicitly specify the fuzzy function. For simplicity, we take the Gaussian similarity function as an example.

### 2.3.5 Nonparametric Relative Entropy

Robinson (1991) proposed a consistent nonparametric entropy-based test for independence in time series. To be specific, let $\{X_t\}$ be a real-valued strictly stationary time series. It is assumed that $X_t$ has a common probability density function $h(x)$, $X_1$ and $X_2$ have a joint probability density function $f(x, y)$. Suppose that $X_t$ and $X_{t+1}$ are independent, then the null hypothesis is

$$\mathbb{H}_0 : f(x, y) = h(x)h(y), \quad \text{for all } x, y. \tag{2.23}$$

Robinson (1991) constructed a nonparametric test statistic for hypothesis (2.23) based on Kullback and Leibler entropy. Let $\hat{f}(x, y)$ and $\hat{h}(x)$ be the nonparametric density estimators of $f(x, y)$ and $h(x)$. The entropy-based test statistic is

$$I\left(\hat{f}, \hat{h}\hat{h}\right) = \int\int \hat{f}(x, y) \log \hat{f}(x, y)\, \mathrm{d}x\mathrm{d}y - 2\int \hat{h}(x) \log \hat{h}(x)\, \mathrm{d}x. \tag{2.24}$$

The integrals in (2.24) bring difficulty to calculation. It is a common technique to replace the integrals with summands (Robinson, 1991), then an alternative estimator is

$$\hat{I}\left(\hat{f}, \hat{h}\hat{h}\right) = \frac{1}{N}\sum_{t\in S} \log\left(\frac{\hat{f}(X_t, X_{t+1})}{\hat{h}(X_t)\hat{h}(X_{t+1})}\right), \tag{2.25}$$

where $S = \{t \in \mathbb{N} : \hat{f}(X_t, X_{t+1}) \geq 0, \hat{h}(X_t) \geq 0, 1 \leq t \leq N-1\}$ and $N$ is the cardinality of $S$. By *sample-splitting device*, the author generalized (2.25) to

$$\hat{I}_\gamma\left(\hat{f}, \hat{h}\hat{h}\right) = \frac{1}{N_\gamma}\sum_{t\in S} c_t(\gamma) \log\left(\frac{\hat{f}(X_t, X_{t+1})}{\hat{h}(X_t)\hat{h}(X_{t+1})}\right), \tag{2.26}$$

where $c_t(\gamma) = 1 + \gamma$ if $t$ is odd; otherwise $c_t(\gamma) = 1 - \gamma$, $\gamma \geq 0$ and $N_\gamma = N$ for $N$ even and $N + \gamma$ for $N$ odd. Under the appropriate conditions (see, Robinson, 1991, p. 441), estimator (2.26) is consistent.

As Robinson (1991) pointed out, the choice of $\gamma$ is an open problem. To avoid the selection of tuning parameter, Hong & White (2005) developed an *asymptotic distribution theory for nonparametric entropy of serial dependence*, which does not involve the sample splitting device. To be clearly, denote $Z_{jt} = (X_t, X_{t+j})^T$ where $j = 1, 2, \ldots, N-1$ is a given lag order, similar to estimator (2.25), Hong & White (2005) proposed

$$\hat{I}(\hat{f}, \hat{h}\hat{h}) = \frac{1}{N-j}\sum_{t\in S_j} \log\left(\frac{\hat{f}(X_t, X_{t+j})}{\hat{h}(X_t)\hat{h}(X_{t+j})}\right), \tag{2.27}$$

as a test statistic for the null hypothesis: $X_t$ and $X_{t+j}$ are independent, where

$S_j = \{t \in \mathbb{N} : \hat{f}(X_t, X_{t+j}) \geq 0, \hat{h}(X_t) \geq 0, \hat{h}(X_{t+j}) \geq 0, 1 \leq t \leq N - j\}$. One appealing feature of theory in Hong & White (2005) is that estimator (2.27) is consistent and has a limiting distribution even $j \to \infty$. We extend estimator (2.27) to the case of $m$ consecutive lag order variables and develop a consistent theory when $m$ has upper bound, see Section 5.3.

### 2.3.6 Limiting Distribution

Pincus & Huang (1992) discussed statistical properties and applications of approximate entropy. The smaller value of approximate entropy implies regularity and predicability while larger approximate entropy means the substantial fluctuations and irregularity in the time series. One can use ApEn to evaluate the randomness. However, some problems, such as parameter choice, limiting distribution of ApEn, remained unknown at that time.

Rukhin (2000) proved that the limiting distribution of ApEn converges in distribution to $\chi^2$-distribution when $m$ is given. When $m$ increases to infinity, Rukhin (2000) also obtained the normal limiting distribution with the parameters estimated by Poisson approximation, see Section 3 in Rukhin (2000). Robinson (1991) developed a theory for nonparametric relative entropy (RlEn) using the sample-splitting device. To overcome the difficulty of tuning parameter selection in Robinson (1991)'s theory, Hong & White (2005) proposed a nonparametric relative entropy test statistic to pairwise variables $(X_t, X_{t+j})$. The limiting distribution is still Gaussian even when $j \to \infty$.

### 2.3.7 Parameter Selection and Algorithm

As for the tuning parameters in ApEn, SpEn and FzEn, many approaches in literature can address the selection of parameters. For example, Lu *et al.* (2008) proposed using maximum ApEn to automatically select the optimal $r$. Based on Monte Carlo simulations, they also obtained general equations for computing the parameter $r$ given $m$. There is a discussion about the $r$ selection, see Chon *et al.* (2009) and Section 2.2 in Udhayakumar *et al.* (2017).

As far as we know, there is no theory that describes how to select the embedded parameter $m$ and how to set up the length of time series $N$ for ApEn, SpEn, FzEn and RlEn. Nevertheless, Kaffashi *et al.* (2008) studied the effect of time delay on ApEn and SpEn, they added the time delay parameter $\tau$ in the entropy and discussed the $\tau$ selection, but no general criterion was proposed.

Besides the choice of tuning parameters, algorithms have been developed to speed up the computation of entropy. Manis (2008) proposed two algorithms: *improved basic algorithm* and *bucket-assisted algorithm* to compute the approx-

imate entropy. However, the author did not mention whether these two algorithms can be applied to the sample entropy. Pan *et al.* (2011) proposed *kd tree algorithm*, *sliding kd tree algorithm* (SKD) and *adaptive kd tree algorithm* (adaptive SKD) to compute both ApEn and SpEn. The complexity of computation reduces from $O(N^2)$ to $O(N^{3/2})$ with a price of $O(N)$ memory storage. Zurek *et al.* (2012) proposed *norm-components matrix* to accelerate the computation of *correlation dimension*[3], the algorithm can be found in `https://github.com/sebzur/NCM-algorithm` and Appendix in Zurek *et al.* (2012). Manis *et al.* (2018) proposed three algorithms to obtain the sample entropy. These three algorithms are *kd-tree, bucket assisted* and *lightweight* which is the fastest among them according to the authors' results.

There are also some overviews which sum up the above entropies as well, e.g., for a review of approximate entropy, see Chen *et al.* (2009); the appropriate use of ApEn and SpEn with the short datasets, see Yentes *et al.* (2013); the application of FzEn, see Hsu (2015). We will not go further here and begin to introduce the other statistical techniques used in this thesis.

## 2.4 Other Techniques Used in Thesis

In this section, we provide a brief overviews of covariance shrinkage, false discovery rate (FDR), four basic concepts in graphic model and Jackknife kernel which are adopted in this thesis.

### 2.4.1 Covariance Shrinkage

Ledoit & Wolf (2004) proposed a framework of large covariance shrinkage such that the covariance is well-conditioned to obtain the precision matrix. This framework can be described as follows.

Let $X$ be a $p \times n$ matrix with columns representing $n$ i.i.d. observations from a distribution with zero mean and covariance matrix $\Sigma$. Suppose $X$ has a finite fourth moment. The sample covariance matrix estimator is $S = n^{-1}XX^T$. When $p$ is larger than $n$, the estimator $S$ is usually ill-conditioned which hinders the statistical inference of covariance matrix. In numerical calculations, one common knowledge of dealing with ill-conditioned is by adding an identity matrix $I$ to $S$, i.e., $\Sigma^* = \rho_1 I + \rho_2 S$, where $\rho_1, \rho_2$ are the coefficients to be determined. Define the inner product of two $p \times p$ matrices $A_1, A_2$ by $\langle A_1, A_2 \rangle = \mathrm{tr}(A_1 A_2^T)/p$. Denote $\|\cdot\|_F^2$ as the squared Frobenius norm loss, we have $\|A\|_F^2 = \mathrm{tr}(AA^T)/p$. By using these notations, let $\mu = \langle \Sigma, I \rangle$, $\alpha^2 = \|\Sigma - \mu I\|_F^2$, $\beta^2 = \mathrm{E}\|S - \Sigma\|_F^2$, $\delta^2 = \mathrm{E}\|S - \mu I\|_F^2$.

[3]The terminology *correlation dimension* is a concept used in the computation of ApEn and SpEn, see Zurek *et al.* (2012) for more details.

Then we have $\alpha^2 + \beta^2 = \delta^2$, see Lemma 2.1 in Ledoit & Wolf (2004). Next, define the objective function as

$$\mathrm{E} = \mathrm{E}\|\Sigma^* - \Sigma\|_F^2. \tag{2.28}$$

The aim is to find the optimal $\rho_1, \rho_2$ such that (2.28) is minimal. The solutions verify that

$$\Sigma^* = \frac{\beta^2}{\delta^2}\mu I + \frac{\alpha^2}{\delta^2}S, \quad \rho_1 = \frac{\beta^2}{\delta^2}\mu, \quad \rho_2 = \frac{\alpha^2}{\delta^2},$$

for more details, see Theorem 1 in Ledoit & Wolf (2004).

It is worth pointing out that the above results hold if the $X$ is composed of the i.i.d. observations. We need to modify this framework to adapt the non i.i.d. case, for example, see Appendix A.1.

### 2.4.2 False Discovery Rate

We briefly review the main content of FDR. Let $m_{1\cdot}$ represent the number of true hypothesis, $m_{2\cdot}$ represents the number of false hypothesis, so the total number of test is $m_{1\cdot} + m_{2\cdot} = m$. Furthermore, let $m_{\cdot 1}$ represent the number of rejecting null hypothesis, $m_{\cdot 2}$ represents the number of accepting null hypothesis, also $m_{\cdot 1} + m_{\cdot 2} = m$. These can be summarized by Table 2.1. $m_{11}$ is the number

*Table 2.1: Classification of Multiple Hypothesis Tests*

| | $H_0$ is true | $H_0$ is false | |
|---|---|---|---|
| Reject $H_0$ | $m_{11}$ | $m_{21}$ | $m_{\cdot 1}$ |
| Accept $H_0$ | $m_{12}$ | $m_{22}$ | $m_{\cdot 2}$ |
| | $m_{1\cdot}$ | $m_{2\cdot}$ | $m$ |

of false discoveries, $m_{21}$ is the number of true discoveries, $m_{12}$ is the number of true negatives, $m_{22}$ is the number of false negatives. $m_{\cdot 1}$ is an observable random variable while $m_{11}, m_{12}, m_{21}, m_{22}$ are unobservable random variables. The false discovery rate is defined as

$$FDR = \frac{m_{11}}{m_{\cdot 1}}.$$

We use Benjamini-Hochberg procedure (Benjamini *et al.*, 2006) to control the FDR at level $\alpha$ in Chapter 4.

### 2.4.3 Basic Concepts in Graphical Network

In this subsection, we will review four basic concepts in graphical network: Edge Density, Vertex Strength, Clustering Coefficient and Centrality. We use these definitions to evaluate the changes of network with respect to different financial periods in Chapter 4.

Let $V$ and $E$ represent the vertex and edge, define the network graph as $\mathcal{G} = (V, E)$. The definition of Edge Density is

$$\text{den}(\mathcal{G}) = \frac{|E|}{|V|(|V|-1)/2},$$

where $|\cdot|$ represents the cardinality operator. For a weighted network, the Vertex Strength or Weighted Vertex Degree is simply to add the weights of edges which are directly connected to a given vertex. Let $A = (a_{ij})_{1\le i,j\le p}$ and $W = (w_{ij})_{1\le i,j\le p}$ be the adjacency matrix and weight matrix, then we define the vertex strength of $\mathcal{G}$ as

$$\text{stren}(\mathcal{G}) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}w_{ij}.$$

For any vertex $u \in V$, the vertex centrality (Sabidussi, 1966) is

$$c(u) = \frac{1}{\sum_{v\in V} d(u,v)}, \tag{2.29}$$

where $d(u, v)$ is the distance between vertices $u$ and $v$. This equation measures the vertex centrality for Graph $\mathcal{G}$. Freeman (1978) gave a measure of graph level centrality, let $c(u^*)$ be the maximum centrality among the $q$ vertices, the graph level centrality is

$$c(\mathcal{G}) = \frac{\sum_{u\in V}[c(u^*) - c(u)]}{(q^2-3q+2)/(2q-3)}.$$

Clustering coefficient is defined as

$$\text{clust}(\mathcal{G}) = \frac{3\tau_\Delta(\mathcal{G})}{\tau(\mathcal{G})},$$

where $\tau_\Delta(\mathcal{G})$ represents the number of triangles in $\mathcal{G}$ and $\tau(\mathcal{G})$ is the connected triples, i.e., a sub-graph of three vertices connected by two edges, for more detail, see Kolaczyk (2009).

### 2.4.4 Jackknife Kernel

For bounded support of $x$, researchers prefer to use the Jackknife kernel to correct the boundary effect. Denote $K_h(x) = K(x/h)/h$ as the scaled kernel. Let $f_0(\cdot)$ be the density function of univariate $x$ with support $[0, 1]$. The nonparametric density estimator of $f_0(x)$ is $\hat{f}_0(x) = n^{-1}\sum_{i=1}^{n} K_h(x - x_i)$. Using change of variable and second order Taylor expansion, the expectation of $\hat{f}_0(x)$ can be

obtained as

$$\begin{aligned}\mathrm{E}\left(\hat{f}_0(x)\right) \approx f_0(x)\int_{(x-1)/h}^{x/h} K(u)\,\mathrm{d}u - hf_0'(x)\int_{(x-1)/h}^{x/h} uK(u)\,\mathrm{d}u \\ + \frac{1}{2}h^2 f_0''(x)\int_{(x-1)/h}^{x/h} u^2K(u)\,\mathrm{d}u.\end{aligned} \tag{2.30}$$

If $x \leq 1-h$, then $(x-1)/h \leq -1$. Hence, $(x-1)/h$ in equation (2.30) can be written as -1 because $K(\cdot)$ has the support $[-1,1]$. Furthermore, let $x = \rho h$, $\rho \geq 0$, then equation (2.30) can be simply expressed as

$$\mathrm{E}\left(\hat{f}_0(x)\right) \approx f_0(x)\omega_0(\rho) - hf_0'(x)\omega_1(\rho) + \frac{1}{2}h^2 f_0''(x)\omega_2(\rho), \tag{2.31}$$

where $\omega_l(\rho) = \int_{-1}^{\rho} u^l K(u)\,\mathrm{d}u$, $l = 0,1,2$. Note that if $\rho \geq 1$ (equivalently $x \geq h$), then according to Assumption 1 (on page 108), $\omega_0(\rho) = 1$ and $\omega_1(\rho) = 0$. Therefore, we have

$$\mathrm{E}\left(\hat{f}_0(x)\right) \approx f_0(x) + O(h^2), \tag{2.32}$$

which indicates $\hat{f}_0(x)$ is asymptotically unbiased with of order $O(h^2)$ if $x \in [h, 1-h]$. However, when $x \in [0,h)$ (or $0 \leq \rho < 1$), we can see $1/2 \leq \omega_0(\rho) < 1$. This means estimator $\hat{f}_0(x)$ is biased. To ensure the asymptotically unbiased property, one can construct a 'self-normalized' estimator by $\hat{f}_N(x) = \hat{f}_0(x)/\omega_0(\rho)$. The expectation of $\hat{f}_N(x)$ is

$$\mathrm{E}\left(\hat{f}_N(x)\right) \approx f_0(x) - hf_0'(x)R_1(\rho) + \frac{1}{2}h^2 f_0''(x)R_2(\rho), \tag{2.33}$$

where $R_l(\rho) = \omega_l(\rho)/\omega_0(\rho)$, $l = 1,2$. Apparently the leading bias term in equation (2.33) is of order $O(h)$ rather than $O(h^2)$ except $f_0'(x) = 0$. In order to let $\hat{f}_N(x)$ be of the same order $O(h^2)$, John (1984) proposed Jackknife kernel to eliminate the $O(h)$ term. Only in the computation of RlEn in Chapter 5, we adopt John (1984)'s Jackknife kernel method, more generalized Jackknife kernels and discussions can be found in Jones (1993). Given bandwidth $h$, let $\hat{f}_N(x;h)$ be the estimator of $f_0(x)$ if $x = \rho h$ and $0 \leq \rho < 1$. Similarly, denote $\hat{f}_N(x;h_1)$ as the estimator of $f_0(x)$ based on another bandwidth $h_1$ if $x = \rho_1 h_1$ and $0 \leq \rho_1 < 1$. The essence of John (1984)'s Jackknife kernel method is the linear combination of two normalized kernel estimators with bandwidths $h$ and $h_1$ respectively, i.e.,

$$\bar{g}_N(x) = (1+\beta)\hat{f}_N(x;h) - \beta\hat{f}_N(x;h_1),$$

where $\beta$ is the parameter to be determined later. One can easily verify that

$$\mathrm{E}\left(\bar{g}_N(x)\right) \approx f_0(x) + f_0'(x)\left[-h(1+\beta)R_1(\rho) + h_1\beta R_1(\rho_1)\right] + O(h^2 + {h_1}^2). \tag{2.34}$$

Let $h_1 = \alpha h$, according to $\rho_1 h_1 = \rho h$, then $\rho_1 = \rho/\alpha$. The leading bias term of equation (2.34) is

$$h f_0'(x) \left[-(1+\beta)R_1(\rho) + \alpha\beta R_1(\rho/\alpha)\right]. \tag{2.35}$$

With the appropriate choice of $\beta$:

$$\beta(\rho) = \frac{R_1(\rho)}{\alpha R_1(\rho/\alpha) - R_1(\rho)},$$

the leading bias term (2.35) vanishes so that $\mathrm{E}\left(\bar{g}_N(x)\right) - f_0(x)$ is of order $O(h^2)$ as interior interval $[h, 1-h]$. It is easy to see that the Jackknife kernel is $K_\rho(u) = h^{-1}k_\rho(u)$, where

$$k_\rho(u) = (1+\beta)\frac{K(u)}{\omega_0(\rho)} - \frac{\beta}{\alpha}\frac{K(u/\alpha)}{\omega_0(\rho/\alpha)}. \tag{2.36}$$

Now, we have obtained the Jackknife kernel for interval $[0, h)$, using the same way, we can get the Jackknife kernel for interval $(1-h, 1]$ as well. If $1-h < x \leq 1$ and let $1 - x = \rho h$ then equation (2.31) is

$$\mathrm{E}\left(\hat{f}_0(x)\right) \approx f_0(x)\omega_0^*(\rho) - h f_0'(x)\omega_1^*(\rho) + \frac{1}{2}h^2 f_0''(x)\omega_2^*(\rho),$$

where $\omega_l^*(\rho) = \int_{-\rho}^{1} u^l K(u)\,\mathrm{d}u$, $l = 0, 1, 2$. Using the same discussions as equations (2.32)-(2.35), one has $\beta^*(\rho) = \frac{R_1^*(\rho)}{\alpha R_1^*(\rho/\alpha) - R_1^*(\rho)}$, where $R_1^*(\cdot) = \omega_1^*(\cdot)/\omega_0^*(\cdot)$. Since $K(\cdot)$ is symmetric, we have $\omega_0^*(\cdot) = \omega_0(\cdot)$, $\omega_1^*(\cdot) = -\omega_1(\cdot)$, $R_1^*(\cdot) = -R_1(\cdot)$ and $\beta^*(\rho) = \beta(\rho)$. Therefore, for $x \in (1-h, 1]$, $\rho = (1-x)/h$, the Jackknife kernel has the same form of equation (2.36). Recalling that $\alpha$ is not yet determined, in this thesis, we follow the choice of $\alpha$ in John (1984) and let $\alpha = 2 - \rho$. Finally, for univariate $x$, the Jackknife kernel is

$$K_h^J(x-y) = \begin{cases} h^{-1}k_{(x/h)}\left(\frac{x-y}{h}\right), & \text{if } x \in [0, h). \\ h^{-1}K\left(\frac{x-y}{h}\right), & \text{if } x \in [h, 1-h]. \\ h^{-1}k_{[(1-x)/h]}\left(\frac{x-y}{h}\right), & \text{if } x \in (1-h, 1]. \end{cases} \tag{2.37}$$

For more details, see John (1984) and Hong & White (2005).

We try our best to review the references relevant to this research. We hope it outlines a full background from a bird's eye view. Before we turn to the next chapter, the remark on notations needs to be emphasized here. In this chapter, we keep the symbol notations consistent with those in the references as possible as we can. Because we quote many statistical methods in this chapter, it is hard to coordinate the models using one set of notations. So whence we introduce the statistical methods in one specific section, the notations used in this section will

be emphasized again. But, from now on, the symbol notations in each chapter are consistent to avoid the confusion. If the same notation appears in different chapters, we will emphasize or re-define it in where it needs to. The general notations for the following chapters can be found in List of Symbols.

As for the kernel function, we adopt the Gaussian kernel in Chapter 3 and Chapter 4. In Chapter 5, we use the Jackknife kernel. No doubt, the frameworks and relative entropy can be extended to the other kernels, however, this extension is beyond this research, we will not discuss it any more.

# Chapter 3

# Factorized Estimation of High-dimensional Nonparametric Covariance Models

## 3.1 Introduction

Nonparametric estimation of the covariate-dependent conditional covariance matrix $\Sigma(u)$ in covariance models is fundamental to contemporary scientific research including neuroimaging in neuroscience, disease mapping in health science, daily ozone concentration analysis in environmental science, asset portfolio risk analysis in finance and among others (Ledoit & Wolf, 2004; Yin *et al.*, 2010; Reich *et al.*, 2011; Lamus *et al.*, 2012; Fan *et al.*, 2013; Fox & Dunson, 2015; Zhang & Liu, 2015; Zhang & Su, 2015; Chen & Leng, 2016). However, most efforts in nonparametric covariance estimation suffer from a curse of dimensionality (Fan *et al.*, 2013). For example, the dataset we are studying in this chapter contains historical returns of 75 assets over three time-periods, namely before-financial-crisis, in-financial-crisis and after-financial-crisis with $n$ equal to 84, 36 and 95 months respectively. Note that many more assets can be collected for investigation whereas the number of months $n$ in a period is sometimes quite limited (Engle *et al.*, 2017).

These authors pointed out that the resulting covariance estimator can still be ill-conditioned for finite samples, where an ad-hoc and small constant is required to add to its eigenvalues. These authors also established a consistency theory for their estimators when the sample is from an independent identical distribution. There are three main issues when we use these existing methods.

Firstly, the performance of these methods can be compromised by employing the same smoothing bandwidth for the entries which have varying degrees of smoothness. In particular, under the sparsity assumption, the covariance matrix

function contains many zero entries which are in favour with infinite bandwidth and thus affect the estimation of other nonzero entries if we use a single bandwidth for the entries. On the other hand, letting each entry having its own bandwidth will generate $p(p+1)/2$ tuning parameters to choose. The resulting estimator may not be an appropriate covariance matrix estimator as it can be negative definite for finite samples. Secondly, as the ad-hoc eigenvalues adjustment of Chen & Leng (2016) to the estimated matrix is not principle-guided, it is desirable to explore an optimal shrinkage procedure. Finally, the existing asymptotic theory holds only for i.i.d. samples, in most of the applications, the samples are dependent. For instance, in the above asset portfolio risk analysis, both market returns and asset returns are serially correlated time series.

In this chapter, we have proposed a novel framework to address these issues. It is based on a variance-correlation factorization of $\Sigma(u)$ in the form of $\Sigma(u) = Q_0(u)C_0(u)Q_0(u)^T$, where $Q_0(u)$ is a diagonal matrix function composed by the square roots of the diagonal entries of $\Sigma(u)$ and $C_0(u)$ is the correlation matrix function. We further factorize $C_0(u)$ into the product of invertible band matrix factors of $\Sigma(u)$. In general, we choose band matrices which are less complex than $\Sigma(u)$. In the proposal, we first estimate these band matrices in turn with separate kernel bandwidths, followed by entry-wise thresholding on the resulting estimator of $C_0(u)$. Estimation of these band matrices with different bandwidths is expected to improve the flexibility of the proposal and thus to provide a more accurate estimator for $\Sigma(u)$. Intuitively, performing thresholding on estimated correlations is better than on covariances, since the variation of the estimated correlations is likely to be smaller and more homogenous than that of the estimated covariances. The reason is that our model is *heteroscedastic*, so thresholding correlation coefficients is more appropriate than thresholding the covariance matrix. In fact, thresholding correlations has been proved adaptive to the variability of individual entries of covariance matrix (Cai & Liu, 2011). Finally, a well-conditioned and optimal shrinkage estimator of $\Sigma(u)$ is derived by minimizing the Frobenius norm loss. In summary, the proposed framework differs from the DCM in using multiple factorization-based bandwidths, thresholding correlations and taking into account the shrinkage effect. The proposal can be viewed as a nonparametric extension of the so-called DCC-GARCH approach (Engle *et al.*, 2017), a popular technique for estimating a multivariate time series model.

To evaluate the performance of the new proposal, a set of simulation studies are conducted. The results demonstrate that the new proposal substantially outperforms its counterparts in terms of the Frobenius norm loss and other criteria. The proposed method is illustrated through an application to the analysis

of monthly return data for a group of risky assets mentioned above. The analysis reports the following findings: (1) Some asset returns present a striking nonlinear departure from the Capital Asset Pricing Model (CAPM) (Fama & French, 2004). (2) Both volatility and co-volatility of these asset returns are market-dependent, see Figures A.1–A.3 for more details. These two findings provide empirical support for building a nonparametric CAPM for risk assessment and portfolio selection. We also establish an asymptotic theory for the new proposal: under some mixing and regularity conditions, the proposed estimator is asymptotically consistent with the underlying covariance matrix function even when the samples are dependent. In the procedure, the thresholding step ensures that the resulting estimator converges to the true covariance matrix with a good rate while the shrinkage step makes the resulting estimator not ill-conditioned even infinite. To prove the above theory, a dedicated concentration inequality different from Chen & Leng (2016) is employed for dependent samples. In particular, the proof for the convergence rate of the proposed shrinkage is non-trivial. Note that without the extra thresholding step, a standard shrinkage estimator is expected to have convergence rate of $O_p(\sqrt{p/(nh)})$, where $h$ is the bandwidth in the kernel estimation (Ledoit & Wolf, 2004). After adding the extra thresholding step in the shrinkage procedure, we show that the resulting estimator has a faster convergence rate $O_p(\sqrt{\log(p/h)/(nh)})$ than the standard shrinkage if the underlying covariance matrix is sparse.

The rest of this chapter is organized as follows. First, we introduce our motivations behind the framework in Section 3.2, then the proposed factorized estimators are introduced in Section 3.3. The corresponding algorithms are developed to determine the bandwidths in the related kernel smoothing as well as the levels of thresholding and shrinkage. The uniform consistency and the convergence rate of the proposed estimator are established with dependent samples in Section 3.4. In Section 3.5, simulation studies are conducted to evaluate the performance of the proposed method and compare it to the existing method. The proposed procedure is employed to analyse financial returns for a group of assets. We conclude with a discussion in Section 3.6. The numerical results are postponed to Appendix A.

Throughout this chapter, we let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a square matrix. For a vector $x$, let $||x||$ denote its Euclidean norm. For a square matrix $A = (a_{ik})_{p\times p}$, let $||A||_F = \sqrt{\mathrm{tr}(AA^T)/p}$, $||A|| = \lambda^{1/2}_{\max}(AA^T)$, $||A||_{\max} = \max_{ik}|a_{ik}|$ and $||A||_\infty = \max_{1\le i\le p}\sum_{k=1}^{n}|a_{ik}|$ denote its (size-normalized) Frobenius, spectral, max and $\infty$-norms. Let $\langle A, B\rangle = \mathrm{tr}(AB^T)/p$ be the inner product of square matrices $A$ and $B$. Note that these norms satisfy $||A||_F \le ||A|| \le ||A||_\infty \le \max_{1\le i\le p}\sum_{j=1}^{p} I(|a_{ik}| > 0)||A||_{\max}$. Let

$c \wedge d$ and $c \vee d$ denote the minimum and maximum of numbers $c$ and $d$. Let $I_p$ be a $p-$dimensional identity matrix. Next, we introduce the motivations of Factorized NCM.

## 3.2 Motivation

In this section, we will analyse the current issues in dynamic covariance model such as the choice of covariance estimator, the cross validation criterion, the adhocly adjusted estimator and the effect of sparsity.

### 3.2.1 The Choice of Covariance Estimator

There are two covariance estimators in literature, see Section 2.1.3. One is estimator (2.13) (Yin *et al.*, 2010) and the other is estimator (2.14) (Chen & Leng, 2016). Estimators (2.13) and (2.14) are

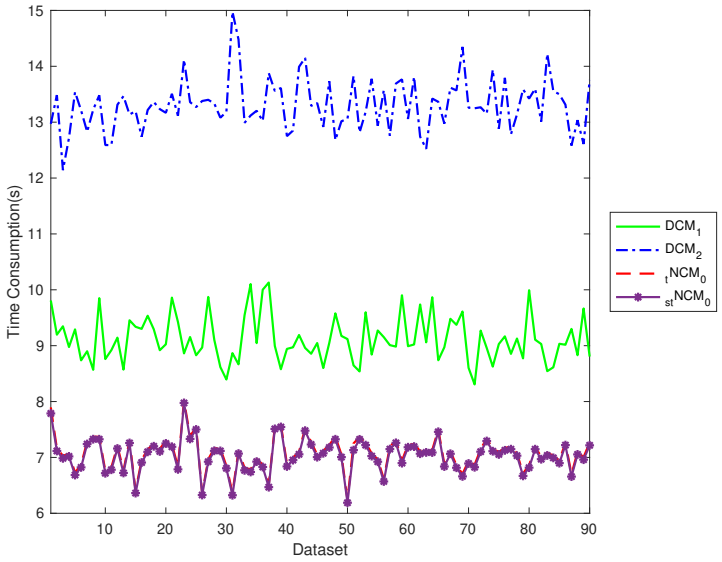
$$\hat{\Sigma}(u) = \sum_{i=1}^{n} w_{ih}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))^T,$$
$$\hat{\Sigma}_1(u) = \sum_{i=1}^{n} w_{ih}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u))^T.$$

If the mean function $\hat{\boldsymbol{\mu}}(u_i)$ is constant with respect to $u_i$, then $\hat{\Sigma}(u)$ and $\hat{\Sigma}_1(u)$ are equivalent especially when the mean function is zero. Note that $\hat{\Sigma}(u)$ differs from $\hat{\Sigma}_1(u)$ in terms of estimating the residuals: The former uses estimators $\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i) = \Sigma(u_i)^{1/2}\boldsymbol{\varepsilon}_i + \boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u_i)$, $1 \leq i \leq n$ while the latter adopts estimators $\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u) = \Sigma(u_i)^{1/2}\boldsymbol{\varepsilon}_i + \boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u)$, $1 \leq i \leq n$. Here, compared to $\boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u_i)$, $\boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u) = \boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u_i) + \hat{\boldsymbol{\mu}}(u_i) - \hat{\boldsymbol{\mu}}(u)$ has an extra bias $\hat{\boldsymbol{\mu}}(u_i) - \hat{\boldsymbol{\mu}}(u)$. So $\hat{\Sigma}(u)$ is expected to perform better than $\hat{\Sigma}_1(u)$, see Remark 2 in Yin *et al.* (2010) and table results in Appendix A. In Chen & Leng (2016)'s work, they assumed the mean function is zero so that these two estimators are equivalent. In general, we consider both mean function and covariance function. Without loss of generality, we adopt $\hat{\Sigma}(u)$ as the covariance matrix estimator in Chapter 3 and Chapter 4.

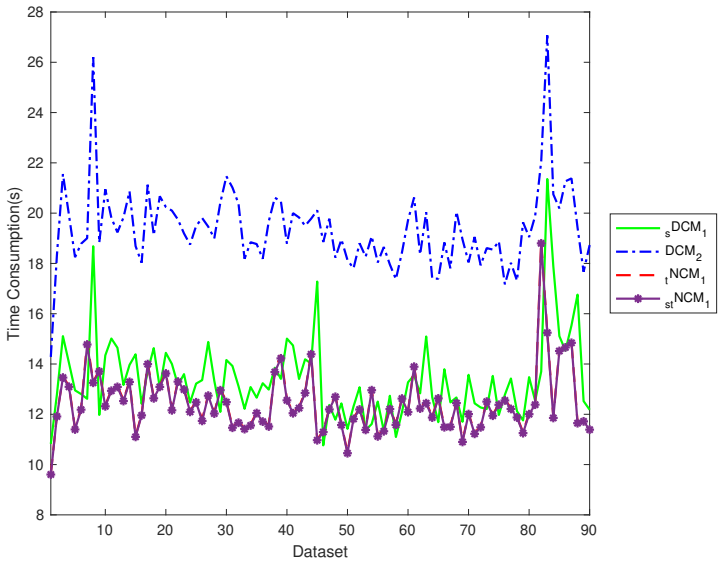
### 3.2.2 The Criterion of Cross Validation

The criterion of bandwidth selection is crucial in nonparametric covariance estimation. It not only determines the performance of nonparametric covariance estimator but also has a big influence on computational complexity. Appropriate criterion can speed up the selection of bandwidth, see Figure 3.1. Both cross validation criteria: (2.15) and (2.16) include the computation of inverse of covariance estimator. As aforementioned in Section 2.2, for a $p \times p$ square matrix $\hat{\Sigma}(u)$

in high-dimensional setting, the matrix inversion in (2.15) and (2.16) is still a computational burden. Hence, we employ the criterion (2.18) to avoid the computation of precision matrix. Criterion (2.18) avoids the computation of matrix inversion, so our method performs efficiently in terms of time consumption as illustrated in Figures 3.1(a) and 3.1(b).




*(a) Comparison 1*




*(b) Comparison 2*

Note: In Figure (a), the green line, blue dashed line, red dashed line and purple line with $*$ represent $\mathrm{DCM}_1$, $\mathrm{DCM}_2$, ${}_{\mathrm{t}}\mathrm{NCM}_0$ and ${}_{\mathrm{st}}\mathrm{NCM}_0$ for 90 datasets respectively. In Figure (b), the green line, blue dashed line, red dashed line and purple line with $*$ represent ${}_{\mathrm{s}}\mathrm{DCM}_1$, $\mathrm{DCM}_2$, ${}_{\mathrm{t}}\mathrm{NCM}_1$ and ${}_{\mathrm{st}}\mathrm{NCM}_1$ for 90 datasets respectively. There are 90 datasets, $n = 100$ and $p = 100$.

*Figure 3.1: Comparison of The CPU-time Consumptions*

Figures 3.1(a) and 3.1(b) show the comparison of time consumptions in sec-

onds when we apply $\text{DCM}_1$, $_{\text{s}}\text{DCM}_1$, $\text{DCM}_2$, $_{\text{t}}\text{NCM}0$, $_{\text{st}}\text{NCM}_0$, $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ to each of 90 datasets in Macbook Pro (CPU: 2.6 GHz 6-Core Intel Core i7, macOS: Big Sur). These datasets are simulated from the Setting 1 with $\rho = 0$. There are four different methods in Figure 3.1(a) and Figure 3.1(b), see Section 3.5 for more details. We can see that DCM estimators using criterion (2.16) consume more time than our nonparametric covariance estimators.

### 3.2.3 The Modified Covariance Estimators

In practice, the dimension $p$ is frequently much larger than the effective local sample size $nh$. This results in a degenerate covariance estimator. To mitigate the high-dimensional effect, researchers modify the covariance estimator by trimming its entries (Bickel & Levina, 2008a). Chen & Leng (2016) also adopted Bickel & Levina (2008a)'s threshold method in the DCM framework. Furthermore, under sparsity and regularity conditions on $(\mathbf{y}_i, u_i)_{i=1}^n$, the threshold covariance estimator is shown to be consistent for an i.i.d. sample with a convergence rate $\sqrt{\log(p)}(O_p(\sqrt{1/(nh)}) + h^2)$ (Chen & Leng, 2016).

However, for a finite sample, Chen & Leng (2016) pointed out that the proposed estimator may be degenerated. To obtain positive definiteness, they adhocly adjusted the estimator by adding the smallest eigenvalue-dependent number to its diagonals, i.e., the modified estimator is

$$\hat{\Sigma}_c(u) = s_\lambda(\hat{\Sigma}(u)) + \{-\hat{a}(u) + c_n\}I_p, \tag{3.1}$$

where $\hat{a}(u)$ is the smallest eigenvalue of $s_\lambda(\hat{\Sigma}(u))$ and $s_\lambda(z) = zI(|z| \geq \lambda)$ represents the hard threshold function. They suggested using

$$c_n = O\left(c_0(p)\left(\sqrt{\log p/(nh)} + h^2\sqrt{\log p}\right)^{1-q}\right), \tag{3.2}$$

as a small positive number in the adjusted estimator (3.1). As far as we know, there is no available way in literature to determine the constant $c_0(p)$.

As mentioned above, Chen & Leng (2016) employed (3.1) to guarantee the positive definiteness of threshold nonparametric covariance estimator. However, the authors did not point out how to determine the constant (3.2). Hence, we need to find another method. We also notice that Ledoit & Wolf (2004) proposed a shrinkage method for the i.i.d. sample, see the review in Section 2.4.

The idea behind the modified covariance matrix estimator (Ledoit & Wolf, 2004) is to find the optimal linear combination of the sample covariance matrix $S$ and the identity matrix $I_p$ by minimizing the expected squared Frobenius norm, see Section 2.4.1. Compared with adding $c_n I_p$ in equation (3.1), Ledoit & Wolf

(2004)'s shrinkage coefficient $\rho$ is determined by the observations. This is a more desirable property if we can extend Ledoit & Wolf (2004)'s shrinkage method to nonparametric covariance estimation. Both Chen & Leng (2016) and Ledoit & Wolf (2004)'s methods can make the covariance matrix estimators positive-definite, however there is no standard criterion to choose the constant (3.2). For this purpose, we have derived the optimal shrinkage coefficient in nonparametric covariance circumstance in Appendix A.1 and the results in Appendix A.2 show that the shrinkage estimator performs consistently better than the non-shrinkage estimator.

### 3.2.4 The Effect of Sparsity

In high-dimensional settings, sparsity assumption is frequently required to make sure the statistical estimator is consistent, e.g., Chaudhuri *et al.* (2007), Cai & Liu (2011), Bien & Tibshirani (2011), Rothman *et al.* (2009), and Rothman (2012), and the references therein. We explore the smoothness of $\Sigma(u)$ with a common shared bandwidth as Yin *et al.* (2010)'s suggestion in high-dimensional settings. The conclusion reports that it may introduce a large bias to estimate nonzero entries when $\Sigma(u)$ contains many unknown zero entries as illustrated by the following pilot study. In this study, we clarify the *zero entries problem*, then show how factorization can reduce the effect of zero entries in terms of Frobenius norm loss. Note that our algorithm includes mean function estimation, standardization, factorization, threshold and shrinkage, see Section 3.3. To gain straightforward insight into factorization, we suppose the mean function is zero, the underlying variance-covariance matrix is a sparse correlation coefficient matrix so that we do not need to estimate the mean function and standardize covariance matrix any more.

First, we introduce how to generate sparse symmetric correlation coefficient matrix. We need to introduce two definitions: *sparsity of matrix* and *sparsity of strictly lower triangular matrix*. Throughout this chapter, we define the *sparsity* of a $p$-dimensional matrix as *the number of zero entries divided by* $p^2$, denoted as $\mathcal{S}$. Next, we define the strictly lower triangular matrix $L = (l_{ij})_{p\times p}$, where $l_{ij} = a_{ij}$ if $i > j$; otherwise 0. Denote $p_0 = p(p-1)/2$, $n_z = \#\{a_{ij} = 0, 1 \le j < i \le p\}$, then the sparsity of $L$ is defined as $\mathcal{S}_L^* = n_z/p_0$. The sparse symmetric correlation coefficient matrix is generated by three steps: (1) Let $R(u) = (r_{ij}(u))_{p\times p}$, where $r_{ij}(u) = \exp(100u\sin(ij))\sin(\pi u)$, then initialize the indicator matrix $E$ by letting the entries be 1. (2) Let $L_E$ represent the strictly lower triangular matrix of $E$, then there are $p_0$ 1-entries in $L_E$. Given $\mathcal{S}_L^*$, we randomly draw $p^* = \lfloor p_0\mathcal{S}_L^* \rfloor$ entries from $L_E$ and change the values at these $p^*$ entries to 0, update the indicator matrix $E$ by $L_E + I_p + L_E^T$, where $I_p$ represents the identity matrix and

superscript $T$ denotes transpose. (3) Renew $R(u) = R(u) \odot E$, where $\odot$ represents matrix component-wise multiplication. Now $R(u)$ is sparse and symmetric but may be negative definite. To satisfy the positive definite requirement of variance-covariance matrix, we let $\mathcal{C}(u) = R(u) \times R(u)$. Finally, standardize $\mathcal{C}(u)$ to obtain $\Sigma(u) = [\text{Diag}(\mathcal{C}(u))]^{-1/2}\mathcal{C}(u)[\text{Diag}(\mathcal{C}(u))]^{-1/2}$.

For simulation, we let $n = 250$, $p = 100$. Covariates $u_i, i = 1, \ldots, n$ are randomly drawn from interval $[-0.95, 0.95]$. Without loss of generality, we let $\Sigma(u_i), i = 1, \ldots, n$ be correlation coefficient matrices. Given $\mathcal{S}_L^*$, for each $u_i$, we obtain the $\Sigma(u_i)$ using the above procedure, then draw one observation from multivariate norm distribution with zero mean and covariance $\Sigma(u_i)$. In this example, we specify $\mathcal{S}_L^*$ as 0.855, 0.91, 0.95, 0.97, 0.98, 0.985, 0.99, 0.995, the corresponding final sparsity of variance-covariance matrix, i.e., $\mathcal{S}_\Sigma$, are 0.10, 0.40, 0.74, 0.88, 0.93, 0.95, 0.97, 0.98 ($\mathcal{S}_\Sigma$ maybe fluctuate with different random seed).

Based on the sample, we calculate the curves of cross-validation values against bandwidth for each $\mathcal{S}_\Sigma = 0.10, 0.40, 0.74, 0.88, 0.93, 0.95, 0.97, 0.98$. These curves have different scales, so we rescale these curves using *each curve divided by its minimum value*, these results are illustrated in Figure 3.2. The numbers behind

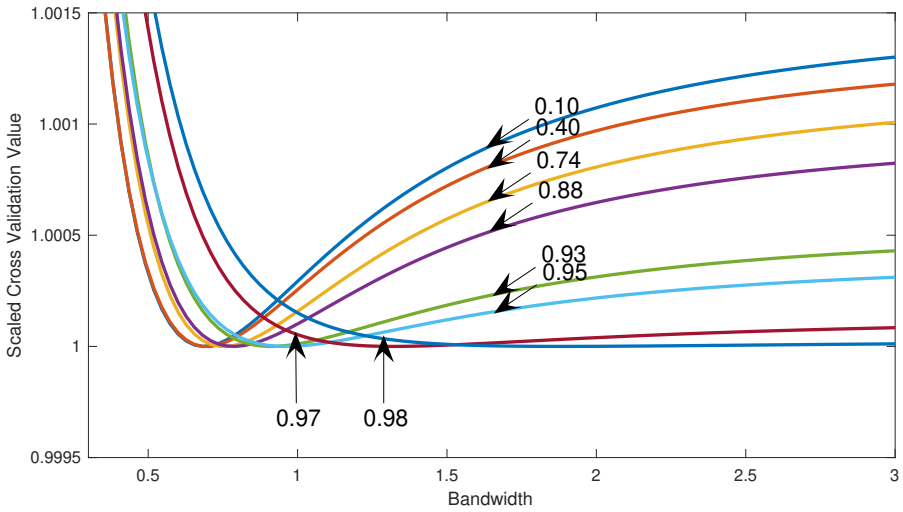
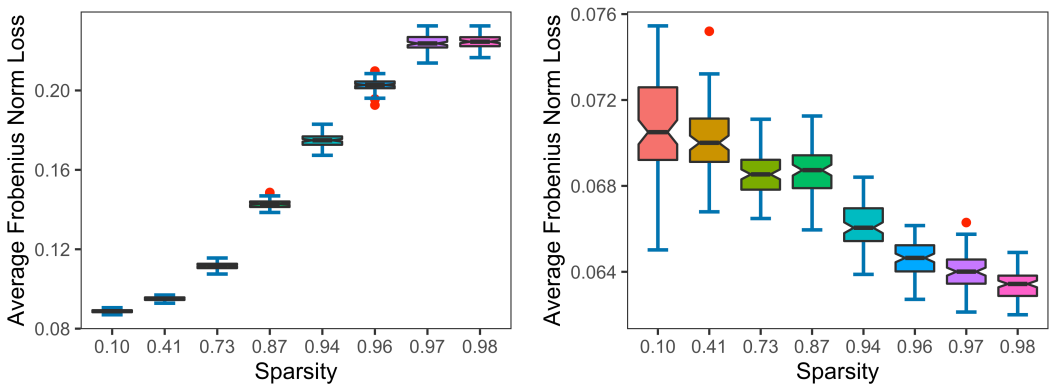



*Figure 3.2: Cross Validation Curves with Different Sparsity*

arrows are the sparsity of variance-covariance matrix, i.e., $\mathcal{S}_\Sigma$. As sparsity $\mathcal{S}_\Sigma$ increases from 0.10 to 0.98, the curves become flatter, and the optimal bandwidth tends to infinity. Furthermore, we repeat the previous process 90 times and compute the Frobenius norm loss for zero and nonzero off-diagonal entries, the box-plots in Figure 3.3 represent the 90 average Frobenius norm losses. The optimal bandwidth goes to infinity because zero entries gradually dominate the bandwidth selection whence the sparsity increases. That means one will use the global arithmetic average of $\mathbf{y}_i\mathbf{y}_i^T$ to estimate the local value of correlation coefficient function, or simply, using constant line to estimate correlation coefficient



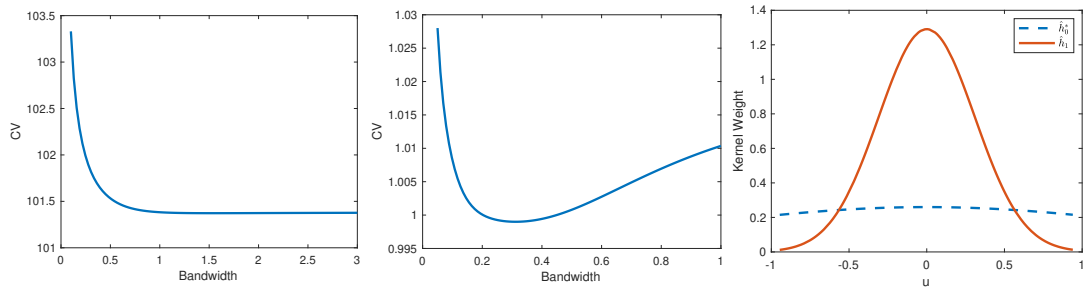

*(a) off-diagonal, nonzero part* *(b) off-diagonal, zero part*

*Figure 3.3: Frobenius Norm Loss for Different Sparsity*

function for nonzero off-diagonal entries when the bandwidth $h \to +\infty$. Hence, the bias comes from the wrong selected bandwidth for nonzero off-diagonal entries. Therefore, the average Frobenius norm losses in nonzero part become larger and larger while the losses in zero part become smaller and smaller. In Chapter 3 and Chapter 4, we call it *zero entries problem*. This phenomenon exists in the nonparametric mean function estimation as well, there still exists *zero entries problem*.

We develop a factorized nonparametric covariance model to solve the *zero entries problem*, for more technical details, see Section 3.3. Next, we will show how factorized nonparametric covariance model could reduce the bias of nonzero entries.

In the procedure of covariance matrix estimation, if we only adopt the factor matrix $Q_0$, then the off-diagonal entries share one common bandwidth. If there are many zeroes among the off-diagonal entries, then we can encounter the *zero entries problem*. To illustrate how the factorization matrices $Q_0, Q_1, \ldots$ can reduce the effect of the zero entries, we simply compare 2 methods here. $\mathbf{Q_1}$ represents the method using $\hat{Q}_1(u_i)$ factor while $\mathbf{Q_0}$ represents method without using factorization. In fact, $\hat{Q}_0(u_i)$ factor belongs to the standardization step which in this simulation we need not estimate. Figure 3.4(a) shows the bandwidth selec-

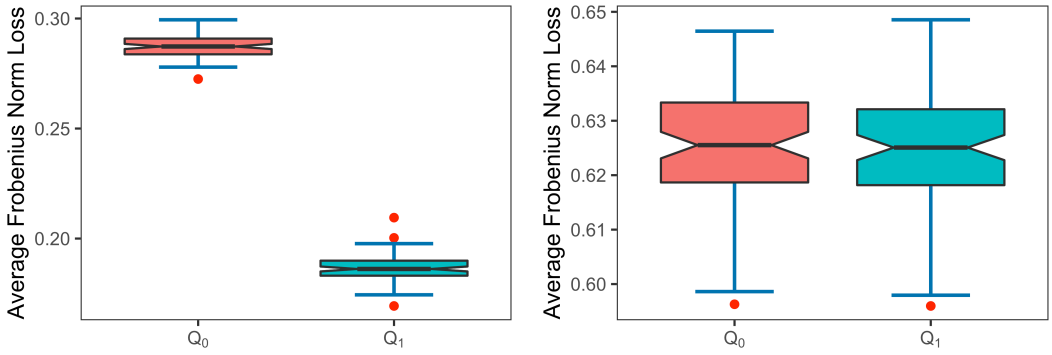



*(a) Method* $\mathbf{Q_0}$ *(b) Method* $\mathbf{Q_1}$ *(c) Kernel Weight at* $u_0 = 0$

*Figure 3.4: Bandwidth Selection for Method* $\mathbf{Q_0}$ *and* $\mathbf{Q_1}$ *when* $\mathcal{S}_\Sigma = 0.97$

tion in method $\mathbf{Q_0}$. Since there is no factorization and the sparsity $\mathcal{S}_\Sigma = 0.97$, zero entries contribute a major effect on bandwidth selection and let the CV against bandwidth $h_0^*$ curve be flatter. However, in Figure 3.4(b), the optimal bandwidth $h_1$ can be obtained at around 0.3091. In fact, the optimal bandwidth in Figure 3.4(a) is 1.5354 which is larger than 1. To investigate the kernel weight performance, let $u_0 = 0$, Figure 3.4(c) shows the kernel weight $K_h(u - u_0)$ for $\hat{h}_0^* = 1.5354$ and $\hat{h}_1 = 0.3091$ respectively. Clearly, the bandwidth $\hat{h}_0^* = 1.5354$ means that each observation has almost the same weight. When the sparsity continues to increase, the bandwidth will go to infinity as Figure 3.2 shows.

Note that, in this simulation, the locations of zero entries are randomly drawn and the elements of $\tilde{\mathbf{y}}_i$'s are re-ranked, we can guarantee that at least the first off-diagonal entries of $\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T$ can not be estimated by constant. Figure 3.5 shows the average Frobenius loss. At each condition $u_i, i = 1, \ldots, n$, we evaluate the Frobenius norm loss of correlation coefficient matrix, then take average of $n$ Frobenius norm losses respectively. Finally, we repeat this procedure 90 times. To avoid the effect of threshold and shrinkage, we do not implement these two steps in this simulation. Because the off-diagonal entries contain zero and nonzero entries, and their locations are already known in the data generation process. We compute the Frobenius norm loss for both zero and nonzero parts respectively, see Figures 3.5(a) and 3.5(b). The losses of zero part for method $\mathbf{Q_0}$ and $\mathbf{Q_1}$ are almost equivalent, but the Frobenius loss of method $\mathbf{Q_1}$ in nonzero part is significantly lower than the loss of method $\mathbf{Q_0}$. Similar to Chen & Leng (2016), they used




*(a) off-diagonal nonzero part* *(b) off-diagonal zero part*

*Figure 3.5: The Frobenius Norm Loss of Method* $\mathbf{Q_0}$ *and* $\mathbf{Q_1}$

the median Frobenius norm loss to evaluate the performance of different method, here we want to mention that a smaller loss will imply a better performance. In this simulation, the median Frobenius norm loss (3.16) of both methods $\mathbf{Q_0}$ and $\mathbf{Q_1}$ are 0.287 and 0.186 respectively. The relative improvement of method $\mathbf{Q_1}$ to method $\mathbf{Q_0}$ under this sparsity is 35.19%. This means that the NCM with one factor band matrix ($m = 1$) performs better than that without factor band

matrix ($m = 0$). Note that we only concentrate on the factorization step in this case and just one factor $\hat{Q}_1(u_i)$ involved under $\mathcal{S}_\Sigma = 0.97$. When more factors are added in, the improvement will be much larger. However, each added-in factor will increase the computational complexity in terms of computing inverse matrix $\hat{Q}_m^{-1}(u_i)$ and bandwidth selection. An *et al.* (2014) proposed a hypothesis testing to detect the band size under high-dimensional banded precision matrices circumstance. However, we can not directly extend this parametric hypothesis test to our nonparametric case, because we did not assume that the precision matrix is a band matrix, e.g., setting 2 in Section 3.5. In this simulation, the zero entries locations are randomized which means our variance-covariance matrix is more flexible than just the band matrix.

To the best of our knowledge, the usual criteria to select model are AIC, BIC or cross validation method. However, they fail to select the number of factors since it is hard to establish the criteria in terms of the Frobenius norm loss. In practise, I suggest two alternative methods to select $m$. (1) Let $m$ increase 1 by each step, repeat the whole NCM algorithm until the difference between Frobenius norm loss in $m$ and $m+1$ step is less than user-specified threshold value, say $10^{-4}$. (2) Let $m$ increase 1 by each step, if the bandwidth $\hat{h}_m$ is very large, say 100, then it means the entries share the same weight, the improvement seems to be insignificant. In this sense, one can stop here and return $m$. However, these two methods are not verified by simulation since for each fixed $m$, one implementation of NCM will cost about one week in high performance computing cluster with 96 CPU cores. Therefore, throughout this chapter we let $m = 1$ due to the trade-off between loss improvement and computational complexity.

Furthermore, we also extend our model to non i.i.d. case, see the settings in Section 3.5. The results in Appendix A show that our factorized estimation of nonparametric covariance model uniformly performs better than DCM method for non i.i.d. as well.

## 3.3 Methodology

Let $\mathbf{Y} = (Y_1, \ldots, Y_p)^T \in \mathbb{R}^p$ be a $p-$dimensional random vector and $U \in \mathbb{R}$ be the associated index random variable. We model the conditional mean and covariance matrix of $\mathbf{Y}$ given $U = u$ as $\boldsymbol{\mu}(u) = \mathrm{E}[\mathbf{Y}|U = u]$ and $\mathrm{cov}(\mathbf{Y}|U = u) = \Sigma(u)$ respectively whose entries are assumed to be an unknown smooth function of $u$. Suppose that $(\mathbf{y}_i, u_i)_{i=1}^n$ with $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^T$, are random observations from the population $(\mathbf{Y}, U)$, satisfying the equations

$$\mathbf{y}_i = \boldsymbol{\mu}(u_i) + \Sigma(u_i)^{1/2}\varepsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\mu}(u_i) = (\mu_1(u_i), \ldots, \mu_p(u_i))^T$, $(u_i)_{i=1}^n$ is a dependent random sample of $U$. Also, given $(u_i)_{i=1}^n$, $\varepsilon_i$'s are dependent with zero means and identity matrices (i.e., $\mathrm{E}[\varepsilon_i|u_i] = 0_p$, $\mathrm{cov}(\varepsilon_i|u_i) = I_p$ and $\mathrm{E}[\varepsilon_i\varepsilon_j^T] \neq 0, i \neq j$).

Let $K(u)$ be a kernel density function, $K_h(u) = h^{-1}K(u/h)$ (the scaled kernel function with a bandwidth $h > 0$) and $w_{ih}(u) = K_h(u_i - u)/\sum_{k=1}^n K_h(u_k - u)$ (the weighting function). Yin *et al.* (2010) considered the following kernel estimators for $\boldsymbol{\mu}(\cdot)$ and $\Sigma(\cdot)$:

$$\begin{aligned} \hat{\boldsymbol{\mu}}(u) &= \sum_{i=1}^n w_{ih_\alpha}(u)\mathbf{y}_i, \\ \hat{\Sigma}(u) &= \sum_{i=1}^n w_{ih_\beta}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))^T \triangleq (\hat{\sigma}_{kj}(u))_{1\leq k,j\leq p}, \end{aligned} \tag{3.3}$$

where $h_\alpha$ and $h_\beta$ are bandwidths for mean and covariance matrix functions. As mentioned in Section 3.2, estimator (3.3) will encounter the problems of zero-entry effect and non-positive definiteness in high-dimensional settings. We will develop a framework to overcome these problems.

### 3.3.1 Standardization

To improve the above covariance estimator, we consider a variance-correlation factorization in the form $\Sigma(u) = Q_0(u)C_0(u)Q_0(u)$, where $Q_0(u) = \mathrm{Diag}(\Sigma(u))^{1/2}$ and $C_0(u) = Q_0(u)^{-1}\Sigma(u)Q_0(u)^{-1}$.

Firstly, we need to estimate the diagonal entries $Q_0(u)$ with a $Q_0(u)$-specified bandwidth $h = h_0$. Given $u$, the nonparametric variance estimator of each entry of $\mathbf{y}_i$ can be written as

$$\begin{aligned} \hat{\sigma}_{11}(u) &= \sum_{j=1}^n \frac{K_{h_0}(u_j - u)}{\sum_{j=1}^n K_{h_0}(u_j - u)}(y_{1j} - \hat{\mu}_1(u_j))^2, \\ \vdots \quad &= \qquad\qquad \vdots \\ \hat{\sigma}_{pp}(u) &= \sum_{j=1}^n \frac{K_{h_0}(u_j - u)}{\sum_{j=1}^n K_{h_0}(u_j - u)}(y_{pj} - \hat{\mu}_p(u_j))^2. \end{aligned} \tag{3.4}$$

Let $\hat{Q}_0(u) = \mathrm{diag}\left(\sqrt{\hat{\sigma}_{kk}(u)} : 1 \leq k \leq p\right)$, for each $u = u_i, i = 1, \ldots, n$, we can compute the diagonal entries' estimators $\hat{Q}_0(u_1), \ldots, \hat{Q}_0(u_n)$. Then, we standardize $\mathbf{y}_i, 1 \leq i \leq n$ by using $\hat{\boldsymbol{\mu}}(u_i)$ and $\hat{Q}_0(u_i)$, i.e., $\tilde{\mathbf{y}}_i = \hat{Q}_0^{-1}(u_i)\,(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))$.

There are two bandwidths need to be selected here.

*Bandwidth selection for* $\hat{\boldsymbol{\mu}}(u)$. We minimize the following function

$$CV_{\boldsymbol{\mu}}(h_\alpha) = n^{-1}\sum\nolimits_{i=1}^n [\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{h_\alpha,-i}(u_i)]^2\omega(u_i),$$

to choose the optimal bandwidth $\hat{h}_\alpha$, where $\hat{\boldsymbol{\mu}}_{h_\alpha,-i}(u_i)$ is the leave-one-out kernel mean function estimator, which is estimated by the local linear smoother based on the data without the $i$th observation. In analogy to the ASE function (Gasser & Müller, 1979), we apply the trimming function $\omega(u_i) = \mathbb{1}(u_{(l+1)} \leq u_i \leq u_{(n-l)})$ to $CV_{\boldsymbol{\mu}}(h_\alpha)$ to get rid of the boundary effects, where $u_{(l)}$ is the $l$th order statistic of $(u_i)_{i=1}^n$ and in our simulation, we set $l = \lfloor 0.05n \rfloor$. The optimal bandwidth for $\hat{\boldsymbol{\mu}}(u)$ is $\hat{h}_\alpha = \arg\min_{h_\alpha} CV_{\boldsymbol{\mu}}(h_\alpha)$.

*Bandwidth selection for* $\hat{Q}_0(u_i)$. Regarding the selection of smoothing parameter for $\hat{\sigma}_{kk}(u_i), i = 1, \ldots, n, k = 1, \ldots, p$, we use the following cross validation criterion to select the optimal bandwidth:

$$CV_0(h_0) = \sum_{i=1}^{n} \sum_{k=1}^{p} \left\{ \frac{(y_{ki} - \hat{\mu}_k(u_i))^2}{\hat{\sigma}_{kk(-i)}(u_i)} + \log(\hat{\sigma}_{kk(-i)}(u_i)) \right\}, \tag{3.5}$$

where $\hat{\sigma}_{kk(-i)}(u_i)$ represents the $k$-entry variance estimated by equation (3.4) without the $i$-th observation at given $u_i$. The optimal bandwidth for $\hat{Q}_0(u_i)$ is $\hat{h}_0 = \arg\min_{h_0} CV_0(h_0)$.

### 3.3.2 Factorization

There are a few matrix factorization algorithms for estimating a covariance matrix in literature, for example, the Cholesky algorithm (Rothman *et al.*, 2010). In this chapter, we introduce a factorization method based on a series of pre-selected invertible band matrices.

In the previous section, we have obtained the standardization of $\mathbf{y}_i$, hence the correlation coefficient matrix has the plug-in estimator

$$\hat{C}_0(u) = \sum\nolimits_{i=1}^{n} w_{ih}(u) \tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^T, \tag{3.6}$$

with bandwidth $h$ to be discussed later in this section. To reduce the zero entries effect introduced in Section 3.2.4, we further factorize $C_0(u)$ into $Q_1(u)C_1(u)Q_1(u)^T$ with

$$Q_1(u) = \begin{bmatrix} 1 & \rho_{12}(u) & 0 & 0 & \cdots & 0 \\ 0 & 1 & \rho_{23}(u) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & \rho_{(p-1)p}(u) \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}. \tag{3.7}$$

Note that (3.7) has the explicit iterated inverse matrix (e.g., see Kiliç & Stanica, 2013). To avoid $\rho_{12}(u), \ldots, \rho_{(p-1)p}(u)$ including too many zero entries, we re-

rank the coordinates of $\mathbf{Y}$ by maximizing the marginal correlations of consecutive coordinates as follows:

Let $\tilde{\mathbf{y}}^{(s)}$ denote the $s$th row of the standardized data matrix $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n)$. Let $s_1 = 1$ and $S_1 = \{2, \ldots, p\}$. For $k = 2, \ldots, p$, define

$$s_k = \arg\max_{s \in S_{k-1}} \left| \text{corr}\left(\tilde{\mathbf{y}}^{(s)}, \tilde{\mathbf{y}}^{(s_{k-1})}\right) \right|,$$

and $S_k = S_{k-1} \backslash \{s_k\}$, where $\text{corr}(\cdot, \cdot)$ denotes the operator for calculating the sample correlation between two random vectors. Let $\mathbf{Y}^* = \left(Y_{s_1}, \ldots, Y_{s_p}\right)$ be the re-ranked $\mathbf{Y}$. Then the largest absolute correlation coefficient entries (except the diagonal) in $C_0(u)$ are likely re-arranged to be close to the diagonal band. To simplify the notation, in the following we assume that the underlying coordinates have already been ranked. In many applications, the above assumption may hold when coordinates in $\mathbf{Y}$ have a natural ordering.

The estimator of $Q_1(u)$ can be obtained by nonparametric kernel estimation based on the standardized observations $\tilde{\mathbf{y}}_i$, namely, $\hat{Q}_1(u) = (\hat{q}_{kj}^{(1)})_{1 \le k,j \le p}$, where

$$\hat{q}_{kj}^{(1)} = \begin{cases} 1, & 1 \le k = j \le p. \\ \sum_{i=1}^n w_{ih_1}(u) \tilde{y}_{ki} \tilde{y}_{(k+1)i}, & j = k+1, 1 \le k \le p-1. \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

where $h_1$ is the bandwidth only for the first diagonal entries in factor (3.7). Hence, using the factors $\hat{Q}_1(u_i)$, $i = 1, \ldots, n$, we have the following transformation:

$$\breve{\mathbf{y}}_i = \hat{Q}_1(u_i) \tilde{\mathbf{y}}_i, \quad i = 1, \ldots, n. \tag{3.9}$$

Based on this transformation, the correlation coefficient matrix estimator is

$$\hat{C}_0(u) = \hat{Q}_1^{-1}(u) \hat{C}_1(u) \hat{Q}_1^{-1}(u)^T, \tag{3.10}$$

where

$$\hat{C}_1(u) = \sum_{i=1}^n w_{ih}(u) \breve{\mathbf{y}}_i \breve{\mathbf{y}}_i^T. \tag{3.11}$$

Now, we compare the bandwidths in two estimators (3.6) and (3.10). Without factorization, the bandwidth in former estimator (3.6) is shared by the off-diagonal entries, the bandwidth $h_0$ of main diagonal entries is selected via criterion (3.5). Considering $Q_1(u)$-factorization, the bandwidth $h_1$ in $\hat{Q}_1(u)$ is shared by the first off-diagonal entries in the latter estimator (3.10). The rest of entries (except the main and first off-diagonal entries) in the latter estimator (3.10) share one common bandwidth as illustrated in estimator (3.11).

Clearly, we can see that the transformation (3.9) can guarantee at least the first diagonal entries of $\hat{C}_0(u)$ not to be affected by the zero entries problem. More generally, we can repeat the same factorization step for the second off-diagonal, the third off-diagonal, ..., the $m$th off-diagonal as follows:

$$\begin{aligned}
\hat{Q}_2(u) &= \left(\hat{q}_{kj}^{(2)}\right)_{1\le k,j\le p}, \quad \hat{q}_{kj}^{(2)} = \begin{cases} 1, & 1 \le k = j \le p. \\ \sum_{i=1}^{n} w_{ih_2}(u)\tilde{y}_{ki}\tilde{y}_{(k+1)i}, & j = k+2, \\ & 1 \le k \le p-2. \\ 0, & \text{otherwise.} \end{cases} \\
\hat{Q}_3(u) &= \left(\hat{q}_{kj}^{(3)}\right)_{1\le k,j\le p}, \quad \hat{q}_{kj}^{(3)} = \begin{cases} 1, & 1 \le k = j \le p. \\ \sum_{i=1}^{n} w_{ih_3}(u)\tilde{y}_{ki}\tilde{y}_{(k+1)i}, & j = k+3, \\ & 1 \le k \le p-3. \\ 0, & \text{otherwise.} \end{cases} \\
&\vdots \\
\hat{Q}_m(u) &= \left(\hat{q}_{kj}^{(m)}\right)_{1\le k,j\le p}, \quad \hat{q}_{kj}^{(m)} = \begin{cases} 1, & 1 \le k = j \le p. \\ \sum_{i=1}^{n} w_{ih_m}(u)\tilde{y}_{ki}\tilde{y}_{(k+1)i}, & j = k+m, \\ & 1 \le k \le p-m. \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.12}$$

with bandwidths $h_2, \ldots, h_m$. Using the above band matrices, we make the transformation: $\breve{\mathbf{y}}_i = \hat{Q}_m^{-1}(u_i)\cdots\hat{Q}_1^{-1}(u_i)\tilde{\mathbf{y}}_i, 1 \le i \le n$. For $m \ge 1$, we estimate $C_m(u)$ by

$$\hat{C}_m(u) = \sum_{i=1}^{n} w_{ih}(u)\breve{\mathbf{y}}_i\breve{\mathbf{y}}_i^T. \tag{3.13}$$

Let $\hat{Q}(u) = \hat{Q}_1(u)\hat{Q}_2(u)\cdots\hat{Q}_m(u)$, $\hat{P}(u) = \hat{Q}^{-1}(u)$, then we have

$$\hat{C}_0(u) = \hat{P}(u)\hat{C}_m(u)\hat{P}(u)^T. \tag{3.14}$$

For pre-fixed $m$, correlation coefficient matrix estimator (3.14) includes $m+1$ separate bandwidths which can overcome the zero entries problem stated in Section 3.2.4. Because estimator (3.14) has multiple bandwidths which implies that the zero entries will not affect the bandwidth selection for the first $m$ off-diagonal entries if $m$ is appropriately pre-specified.

*Bandwidth for estimating $Q_k(u)$, $1 \le k \le m$.* We choose the bandwidth $\hat{h}_k = \arg\min \mathrm{CV}_k(h_k)$ at which the following criterion attains the minimum:

$$\mathrm{CV}_k(h) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p-k}\left(\hat{\rho}_{j(j+k)(-i)}(u_i) - \tilde{y}_{ij}\tilde{y}_{i(j+k)}\right)^2,$$

where $\hat{\rho}_{j(j+k)(-i)}(u_i)$ is the kernel estimator of the $j(j+k)$th correlation $\rho_{j(j+k)}(u_i)$ based on the leave-one-out dataset $(\tilde{\mathbf{y}}_t, u_t)_{t\ne i}$.

*Bandwidth for estimating $\hat{C}_0(u)$.* There are two existing cross-validation methods for selecting the bandwidth $h$ for $C_m(u)$. One is a Stein-loss-based

approach (Yin *et al.*, 2010) which is applicable only to low-dimensional data. The other is a subset-based approach (Chen & Leng, 2016) for high-dimensional data. As discussed in Section 3.2.2, for simplicity, we choose the optimal bandwidth based on the Frobenius norm to avoid the computation of matrix inversion. Without factorization ($m = 0$), the criterion of bandwidth selection is

$$\mathrm{CV}_{C_0}(h) = n^{-1} \sum\nolimits_{i=1}^{n} \left\| \hat{C}_{0(-i)}(u_i) - \tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^T \right\|_F^2,$$

where $\hat{C}_{0(-i)}(u_i)$ is the kernel estimator of (3.6) without the $i$-th observation. The optimal bandwidth is $\hat{h} = \arg\min_h \mathrm{CV}_{C_0}(h)$.

*Bandwidth for estimating* $\hat{C}_m(u)$. The cross validation criterion is

$$\mathrm{CV}_{C}(h) = n^{-1} \sum\nolimits_{i=1}^{n} \left\| \hat{C}_{m(-i)}(u_i) - \breve{\mathbf{y}}_i \breve{\mathbf{y}}_i^T \right\|_F^2,$$

where $\hat{C}_{m(-i)}(u_i)$ is the kernel estimator of $C_m(u)$ based on the leave-one-out dataset $(\breve{\mathbf{y}}_j, u_j)_{j \neq i}$ like estimator (3.13). The optimal bandwidth for estimating $\hat{C}_m(u)$ is $\hat{h} = \arg\min_h \mathrm{CV}_C(h)$.

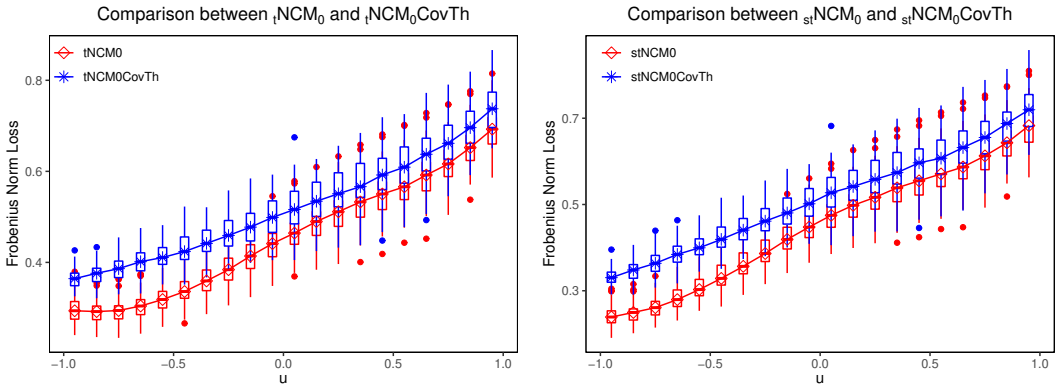
### 3.3.3 Threshold

Note that the dimension $p$ is larger than the local sample size $nh$. This results in a degenerate estimator $\hat{C}_0(u)$. Following Bickel & Levina (2008a), we regularize the above correlation matrix estimator by thresholding its entries as follows:

$$\hat{C}_0^{(t)}(u) = \left( \hat{c}_{jk}(u) \mathbb{1} \left( |\hat{c}_{jk}(u)| > t_0(u) \sqrt{\log(p/h)/(nh)} \right) \right)_{1 \leq j,k \leq p},$$

where $\hat{c}_{jk}(u)$ is the $(j,k)$-th entry of $\hat{C}_0(u)$ and $\mathbb{1}(\cdot)$ is an indicator function and $t_0(u)$ is a positive function of $u$. The above rate of thresholding is suggested by Theorem 3.2 in Section 3.4 below. Here, $p/h$ is related to the dimension of an approximate parametric model to the original model: $[a, b]$ is partitioned into $(b-a)/h$ intervals in which the $p$-dimensional nonparametric model are approximated by a $p(b-a)/h$-dimensional step model. Note that unlike the covariance matrix, the correlation matrix is scale-invariant and with homogenous diagonals. Therefore, the thresholding correlation matrix is expected to make fewer errors than thresholding covariance matrix, see the comparisons in Figures 3.6(a) and 3.6(b). Note that in ${}_{\mathrm{t}}\mathrm{NCM}_0$, we applied threshold to the estimated correlation matrix. Here, to investigate the advantage of ${}_{\mathrm{t}}\mathrm{NCM}_0$ over its variation, we consider thresholding the estimated covariance matrix instead of the estimated correlation matrix. The following figures demonstrate that thresholding correlations is sub-

stantially better than thresholding covariances.




*(a)* ${}_{t}NCM_0(u)$ *and* ${}_{t}NCM_0 CovTh(u)$ *(b)* ${}_{st}NCM_0(u)$ *and* ${}_{st}NCM_0 CovTh(u)$

Note: ${}_{t}\mathrm{NCM}_0 CovTh(u)$ and ${}_{st}\mathrm{NCM}_0 CovTh(u)$ are ${}_{t}\mathrm{NCM}_0(u)$ and ${}_{st}\mathrm{NCM}_0(u)$ with covariance threshold respectively. These four estimators are applied to 90 datasets generated from the Setting 1 with $n = 100$ and $p = 100$, where $u = -1.05 + k/10, k = 1, \ldots, 20$.

*Figure 3.6: Box-plots of The Frobenius Norm Loss Comparison*

In particular, when individual variances $\sigma_{kk}(u), 1 \leq k \leq p$ differ from each other (Cai & Liu, 2011). Using the above estimators, we construct a plug-in estimator of $\Sigma(u)$ in form of $\hat{\Sigma}^{(t)}(u) = \hat{Q}_0(u)\hat{C}_0^{(t)}\hat{Q}_0(u)$.

*Thresholding level for* $\hat{C}^{(t)}(u)$. Following Bickel & Levina (2008a), we split the sample into two sub-samples called trial and testing samples and select the threshold by minimizing the Frobenius norm of the difference between the trial-sample-based thresholding estimator and the testing-sample-based covariance matrix. Specifically, we divide the original sample into two samples at random of size $n_1$ and $n_2$, where $n_1 = n(1-1/\log(n))$ and $n_2 = n/\log(n)$, and repeat this $N_1$ times. Here, we set $N_1 = 100$ as the default value according to our numerical experience. Let $\hat{C}_{1,s}(u)$ and $\hat{C}_{2,s}(u)$ be the plug-in estimators based on $n_1$ and $n_2$ observations respectively with the bandwidth selected by the leave-one-out cross validation. Let $\hat{C}_{1,s}^{(t)}$ be the thresholding estimator derived from $\hat{C}_{1,s}(u)$ with the thresholding level $t_0(u)$. Given $u$, we select $t_0(u)$ by minimizing $N_1^{-1}\sum_{s=1}^{N_1}\|\hat{C}_{1,s}^{(t)} - \hat{C}_{2,s}\|_F$.

### 3.3.4 Shrinkage

In Section 3.4, under sparsity and regularity conditions, we show that under certain regularity conditions the above thresholding covariance estimator is consistent with the underlying covariance matrix function as $n$ and $p$ tend to infinity. However, for a finite sample, the proposed estimator may still be ill-conditioned. To ameliorate it, we propose to shrink $\hat{\Sigma}^{(t)}(u)$ to the identity matrix $I_p$, where the amount of shrinkage is optimized in terms of the Frobenius loss. There are other covariance shrinkage methods in literature, but most of them are developed for

estimating covariance models without covariates, see Jolliffe (2002), Bai & Silverstein (2010) and references therein. To find the optimal amount of shrinkage, we first consider a population version, namely a linear combination of $I_p$ and $\hat{\Sigma}^{(t)}(u)$, $\Sigma^*(u) = \rho a I_p + (1-\rho)\hat{\Sigma}^{(t)}(u)$, whose expected Frobenius loss $\mathrm{E}\|\Sigma^*(u) - \Sigma(u)\|_F^2$ attains the minimum with respect to $0 \leq \rho \leq 1$ and $a \in \mathbb{R}$. The resulting solutions depend on $\Sigma(u)$ as well as variability of $\hat{\Sigma}^{(t)}(u)$. Replacing these unknown quantities by their estimators, we obtain the following plug-in estimator of $\Sigma(u)$ with a data-driven optimal amount of shrinkage:

$$\hat{\Sigma}^{(st)}(u) = \frac{\hat{\beta}_p^2(u)}{\hat{\alpha}_p^2(u) + \hat{\beta}_p^2(u)} p^{-1}\mathrm{tr}\left(\hat{\Sigma}^{(t)}(u)\right) I_p + \frac{\hat{\alpha}_p^2(u)}{\hat{\alpha}_p^2(u) + \hat{\beta}_p^2(u)}\hat{\Sigma}^{(t)}(u). \tag{3.15}$$

where

$$\hat{\alpha}_p^2(u) = \left\|\hat{\Sigma}^{(t)}(u) - p^{-1}\mathrm{tr}\left(\hat{\Sigma}^{(t)}(u)\right) I_p\right\|_F^2,$$

$$\begin{aligned}\hat{\beta}_p^2(u) &= \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{i=1}^{n} w_{ih}^2(u)((y_{ij} - \hat{\mu}_j(u_i))(y_{ik} - \hat{\mu}_k(u_i)) - \hat{\sigma}_{jk}(u))^2 \\ &\quad \times \mathbb{1}(|\hat{\sigma}_{jk}(u)| > t_0(u)\sqrt{\log(p/h)/(nh)}).\end{aligned}$$

Note that $\hat{\alpha}_p^2(u)$ is a plug-in bias when we use $p^{-1}\mathrm{tr}(\hat{\Sigma}^{(t)}(u))I_p$ to estimate $\Sigma(u)$ while $\hat{\beta}_p^2(u)$ gauges the variability of $\hat{\Sigma}^{(t)}(u)$ as an estimator of $\Sigma(u)$. So estimator (3.15) is intended to strike a balance between variability and bias of covariance estimators. Our idea is general, which can be directly used to improve other nonparametric covariance matrix estimators including the DCM, see Appendix A for the detailed derivation.

Finally, we end up with a general procedure for estimating the covariance matrix.

## 3.4 Theory

In this section, we develop an asymptotic theory for the proposed estimators which covers both i.i.d. and non i.i.d. cases and thus is more general than Chen & Leng (2016). Under certain regularity conditions, the proposed estimators are shown to be consistent with the underlying matrix function if we let the related bandwidths be different from each other but have the same convergence rate to zero.

Let $\mathcal{F}_{k_0}$ and $\mathcal{F}_{k_0+k}^{\infty}$ be the $\sigma$-algebras generated by $\{(\mathbf{y}_i, u_i) : 1 \leq i \leq k_0\}$ and

$\{(\mathbf{y}_i, u_i) : k_0 + 1 \leq k < \infty\}$. Define

$$\alpha(k) = \max_{k_0 \geq 1} \sup_{A \in \mathcal{F}_{k_0}, B \in \mathcal{F}_{k_0+k}^{\infty}} |P(A)P(B) - P(A \cap B)|.$$

We assume the following regularity conditions:

(C1) The symmetric kernel function $K(\cdot)$ on $\mathbb{R}$ with derivative $K'(\cdot)$ satisfies

$$K_0 = \sup_z K(z) < +\infty, \quad K_1 = \sup_z |K'(z)| < +\infty,$$
$$\int K(z)dz = 1, \quad \int zK(z)dz = 0,$$
$$\int z^2 K(z)dz < +\infty, \quad \int |z|^3 k(z)dz < \infty.$$

(C2) The density function of $U$, $g(u)$, has the second order continuous derivative $g''(\cdot)$ over a compact support $[a, b]$ and $\inf_{u \in [a,b]} g(u) > 0$. For any $i \neq i_1$, the joint density of $u_i$ and $u_{i_1}$, $\max_{i \neq i_1} \sup_{z, z_1 \in [a,b]} g_{ii_1}(z, z_1)$ is bounded.

(C3) There exist positive constants $\tau_2$ and $\kappa_2 < 1$ such that for $k \geq 1$, $\alpha(k) \leq \exp(-\tau_2 k^{\kappa_2})$.

(C4) There exist constants $0 < \kappa_1 \leq 1, \tau_1 > 0$ such that

$$\max_{1 \leq j \leq p} P(|y_{ij}| > v) \leq \exp(1 - \tau_1 v^{\kappa_1}).$$

(C5) The second derivatives of $\mu_j(u) = E[y_{1j}|U = u]$, $1 \leq j \leq p$ are uniformly bounded in the sense that $\max_{1 \leq j \leq p} \sup_{u \in [a,b]} |\mu_j''(u)| < \infty$.

(C6) The first-order derivatives of $\sigma_j^2(u) = E[(y_{ij} - \mu_j(u_i))^2 | u_i = u]$, $1 \leq j \leq p$, are bounded below from zero uniformly for $1 \leq j \leq p$ and $u \in [a, b]$. Their first-order derivatives are also uniformly bounded. The conditional expectations $E[(y_{ij} - \mu_j(u_i))(y_{(i+t)j} - \mu_j(u_{i+t}))|u_i = z, u_{i+t} = z_1]$ with $z, z_1 \in [a, b]$, $1 \leq i < \infty$, $1 \leq t \leq \infty$, $1 \leq j \leq p$, are uniformly bounded in $i, t, z$ and $z_1$.

The above conditions are routinely used in the literature of nonlinear time series analysis, see Fan & Yao (2003), Zhang & Liu (2015), and Lam & Yao (2012). It follows from (C5) that $b_2 \triangleq \max_{1 \leq j \leq p} \sup_{u \in [a,b]} |\mu_j(u)| < \infty$. (C3) and (C4) assume that the response observations have an exponentially fast mixing rate and sub-exponential tails. Note that these conditions are imposed to facilitate the proofs and thus may not be the weakest possible for establishing the theory below.

Let $\hat{g}_{h_a}(u) = 1/n \sum_{i=1}^n K_{h_a}(u_i - u)$ be a kernel density estimator of $g(u)$. It follows from Proposition 0.1 in Supplementary Material in Zhang & Li (2021)

that $\hat{g}_{h_a}(u)$ is uniformly consistent with $g(u)$.

Letting $1/\gamma_1 = 1/\kappa_1 + 1/\kappa_2$, we state a uniform consistency result for estimator $\hat{\mu}_j(u)$ in the following theorem.

**Theorem 3.1.** *Under Conditions (C1) ∼ (C6), if as $n, p \to \infty$ and $h_a \to 0$,*

$$(\log(p))^{2/\gamma_1 - 1}/n = O(1), \quad \frac{\log(h_a^{-4} np)}{(nh_a \log(p/h_a))^{\gamma_1/2}} = O(1),$$
$$\frac{(\log(nh_a \log(p/h_a)))^{\gamma_1} \log(1/h_a)}{(nh_a \log(p/h_a))^{\gamma_1(1-\gamma_1)/2}} = O(1),$$

*then*

$$\max_{1 \le j \le p} \sup_{u \in [a,b]} |\hat{\mu}_j(u) - \mu_j(u)| = O_p\left(\sqrt{\frac{\log(p/h_a)}{nh_a}}\right) + O(h_a^2).$$

Note that $0 < \gamma_1 < 1/2$ as $\kappa_1 \le 1$ and $\kappa_2 < 1$. The above bandwidth condition imposed on $h_a$ holds and $\sqrt{\log(p/h_a)/(nh_a)} = o(1)$ if $h_a = c_0 n^{-1/5}$ and $(\log(p))^d/n = o(1)$ for a constant $c_0$ and $d = \max\{1/(2\gamma_1), 2/\gamma_1 - 1\}$.

Let $1/\gamma_2 = 2/\kappa_1 + 1/\kappa_2$. In the next theorem, we show that the entries of the proposed covariance matrix estimator are consistent with the underlying ones uniformly in $u$ and indices $1 \le j, k \le p$. We say $h_a, h_v, h_r, h \to 0$ with the same convergence rate if $h/h_a + h_a/h = O(1)$, $h/h_r + h_r/h = O(1)$, $h/h_v + h_v/h = O(1)$, $0 \le v \le m$. $h_a, h_v, h_r, h$ are different bandwidths of $g(u)$ for different terms in the proof of Theorem 3.2, see details in Zhang & Li (2021).

**Theorem 3.2.** *Under Conditions (C1) ∼ (C6), if as $n, p \to \infty$, $h_a, h_v, h_r, h \to 0$ with the same rate, for $w = 1, 2$, $\log(p)^{2/\gamma_w - 1}/n = O(1)$ and*

$$\frac{\log(nph^{-4})}{(nh \log(p/h))^{\gamma_w/2}} = O(1), \quad \frac{(\log(nh \log(p/h)))^{\gamma_w} \log(1/h)}{(nh \log(p/h))^{\gamma_w(1-\gamma_w)/2}} = O(1),$$

*then*

$$\max_{1 \le j,k \le p} \sup_{u \in [a,b]} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| = O_p\left(\sqrt{\frac{\log(p/h)}{nh}} + h^2\right),$$
$$\max_{1 \le j,k \le p} \sup_{u \in [a,b]} |\hat{c}_{jk}(u) - c_{jk}(u)| = O_p\left(\sqrt{\frac{\log(p/h)}{nh}} + h^2\right).$$

Note that $0 < \gamma_2 < 1/3$ as $\kappa_1 \le 1$ and $\kappa_2 < 1$. The bandwidth condition imposed on $h_a, h_v, h_r, h$ holds and $\sqrt{\log(p/h)/(nh)} = o(1)$ if $h_a = c_0 n^{-1/5}$ (which is the optimal bandwidth for the univariate nonparametric regression estimator with $c_0$ a constant) and $(\log(p))^d/n = o(1)$ for $d = \max\{1/(2\gamma_1), 2/\gamma_1 - 1, 1/(2\gamma_2), 2/\gamma_2 - 1\}$.

Put $\alpha_p(u) = \|\Sigma(u) - \langle \Sigma(u), I_p \rangle I_p\|_F$ and $\tau_{np} = \sqrt{\log(p/h)/(nh)}$. Let $\hat{t}_0(u)$ be an estimator of the thresholding function $t_0(u)$ used in $\hat{\Sigma}^{(t)}(u)$ and $\hat{\Sigma}^{(st)}(u)$. $m_p(u) = \max_{1 \le k \le p} \sum_{j=1}^{p} I(\sigma_{kj}(u) > 0)$ denotes a sparsity index of $\Sigma(u)$. The smaller $m_p(u)$, the sparser $\Sigma(u)$ is. To state the next theorem, we introduce the following conditions on separability between $\Sigma(u)$ and $I_p$, sparsity and bounds of $\Sigma(u)$ respectively.

(C7) $\tau_{np}/(\log(p/h) \inf_{u \in [a,b]} \alpha_p^2(u)) = O(1)$, $\sup_{u \in [a,b]} m_p(u)\tau_{np}/\alpha_p(u) = o(1)$.

(C8) There exists a positive constant $s_1$ such that $\sup_{u \in [a,b]} \|\Sigma(u)\| \le s_1$.

(C9) There exists a positive constant $s_{0p}$ such that as $p \to \infty$

$$\frac{s_{0p}}{\sqrt{\sup\limits_{u \in [a,b]} m_p(u)\tau_{np}}} \to \infty, \qquad \inf_{u \in [a,b]} \|\Sigma(u)\| \ge s_{0p}.$$

(C10) $\sup_{u \in [a,b]} \left|\hat{t}_0(u) - t_0(u)\right| = o(1)$ and there exist positive constants $t_a < t_b$ such that for $t_a < \inf_{u \in [a,b]} t_0(u) \le \sup_{u \in [a,b]} t_0(u) < t_b$.

Note that Condition (C7) implies that $\Sigma(u)$ is not close to $cI_p$ in a distance less than the product of the sparsity index and the rate $\tau_{np}/\log(p/h)$, where $c$ is any arbitrary constant. Conditions (C8) and (C9) are about the uniform boundedness of $\|\Sigma(u)\|$ from above and away from zero in an order of $\tau_{np}$ multiplied by the sparsity index. Finally, we can see from Theorem 3.3 that although (C10) requires the tuning constant $\hat{t}_0(u)$ has a finite limit as $n$ tends to infinity, the order of the convergence rate of the corresponding estimator $\hat{\Sigma}^{(st)}(u)$ is independent of such a limit.

Under these conditions, we state the uniform consistent result for $\hat{\Sigma}^{(st)}(u)$ as follows.

**Theorem 3.3.** *Under Conditions (C1) ∼ (C8), if as* $n, p \to \infty$, $h_a, h_v, h_r, h \to 0$ *with the same rate, and for* $w = 1, 2$, $(\log(p))^{2/\gamma_w - 1}/n = O(1)$, $nh^5/\log(p/h) = O(1)$ *and*

$$\frac{\log(nph^{-4})}{(nh\log(p/h))^{\gamma_w/2}} = O(1), \quad \frac{(\log(nh\log(p/h)))^{\gamma_w}\log(1/h)}{(nh\log(p/h))^{\gamma_w(1-\gamma_w)/2}} = O(1),$$

*and if* $\sup_{u \in [a,b]} m_p(u)\tau_{np} = o(1)$, *then uniformly in* $u \in [a, b]$,

$$\left\|\hat{\Sigma}^{(st)}(u) - \Sigma(u)\right\| = O_p(m_p(u)\tau_{np}).$$

*In addition to the above conditions, if Condition (C9) holds, then uniformly in*

$u \in [a, b]$,

$$\begin{aligned}
\left\|\hat{\Sigma}^{(st)}(u)\Sigma^{-1}(u) - I_p\right\| &= O_p\left(m_p(u)\tau_{np}s_{0p}^{-1}\right) = o_p\left(\sqrt{m_p(u)\tau_{np}}\right), \\
\left\|\Sigma(u)\left(\hat{\Sigma}^{(st)}(u)\right)^{-1} - I_p\right\| &= O_p\left(m_p(u)\tau_{np}s_{0p}^{-1}\right) = o_p\left(\sqrt{m_p(u)\tau_{np}}\right), \\
\left\|\left(\hat{\Sigma}^{(st)}(u)\right)^{-1} - \Sigma^{-1}(u)\right\| &= O_p\left(m_p(u)\tau_{np}s_{0p}^{-2}\right) = o_p(1).
\end{aligned}$$

*Finally, in addition to the above conditions, if Condition (C10) holds, then the above results continue to hold after replacing* $t_0(u)$ *by* $\hat{t}_0(u)$ *in* $\hat{\Sigma}^{(t)}(u)$ *and* $\hat{\Sigma}^{(st)}(u)$.

Note that if $h = c_0 n^{-1/5}$ ($c_0$ is a constant) and for $d = \max\{1/(2\gamma_1), 2/\gamma_1 - 1, 1/(2\gamma_2), 2/\gamma_2 - 1\}$, $(\log(p))^d/n = o(1)$, the above condition imposed on $h$ holds. Note that the above bandwidth assumption that they have the same convergence rate to zero does not rule out these bandwidths are different. However, the cross-validation (or the so-called subset) selected bandwidths may not tend to zero. In particular, some of these bandwidths may tend to infinity when there are many zeros and a few non-zeros in the underlying covariance matrix. In this situation, simulation studies in the next section show that the proposed estimators could reduce the bias and outperformed the DCM in terms of integrated mean squared errors. The theoretical development along this aspect will be spelled out in a future paper. All the details of proofs of the above the theorems can be found in Zhang & Li (2021) and the online Supplementary Material.

## 3.5 Numerical Studies

In this section, to demonstrate the merits of the proposed estimators in finite sample settings, we apply the proposed procedure to both synthetic and real dataset. We present the numerical results for the proposed estimators using $m$ band matrix factors.

To facilitate the presentation, let ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ denote the proposed estimators $\hat{\Sigma}^{(t)}(u)$ and $\hat{\Sigma}^{(st)}(u)$ respectively with $m = 0$. Let ${}_{\text{t}}\text{NCM}_1$ and ${}_{\text{st}}\text{NCM}_1$ denote the proposed estimators $\hat{\Sigma}^{(t)}(u)$ and $\hat{\Sigma}^{(st)}(u)$ respectively with $m = 1$. Let $\text{DCM}_1$ and $\text{DCM}_2$ denote two DCM estimators defined by

$$\begin{aligned}
\text{DCM}_1(u) &= (\hat{\sigma}_{1jk}(u)I(\hat{\sigma}_{1jk}(u) \geq d(u)))_{1\leq j,k\leq p}, \\
\text{DCM}_2(u) &= (\hat{\sigma}_{2jk}(u)I(\hat{\sigma}_{2jk}(u) \geq d(u)))_{1\leq j,k\leq p},
\end{aligned}$$

where $d(u)$ is the level of thresholding and

$$\begin{aligned}\tilde{\Sigma}_1(u) &= \sum_{i=1}^{n} w_{ih}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))^T \triangleq (\hat{\sigma}_{1jk}(u))_{1\le j,k\le p},\\ \tilde{\Sigma}_2(u) &= \sum_{i=1}^{n} w_{ih}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u))^T \triangleq (\hat{\sigma}_{2jk}(u))_{1\le j,k\le p}.\end{aligned}$$

Note that DCM$_1$ differs from DCM$_2$ in the way of estimating the residuals, see the discussion in Section 3.2.1. So, DCM$_1$ is expected to perform better than DCM$_2$. Following the same procedure as in $_{\text{st}}$NCM$_0$, we improve DCM$_1$ by incorporating the effects of shrinkage on it. Let $_{\text{s}}$DCM$_1$ denote the optimal shrinkage estimator after replacing $_{\text{t}}$NCM$_0$ by DCM$_1$ in the definition of $_{\text{st}}$NCM$_0$. As for the tuning parameters for estimating DCM, we follow the method in Chen & Leng (2016), i.e., the bandwidth $h$ and the level of thresholding of the DCM estimators in (3.3) are determined by the so-called subset and sample-splitting approaches respectively.

### 3.5.1 Criteria for Performance Assessment

We need a criterion to evaluate the performance of a nonparametric covariance matrix estimator. There are multiple possible criteria, but one particularly convenient choice is integrated root-squared error (IRSE). For any estimator $\hat{\Psi}(u)$ of $\Sigma(u)$, $u \in [a, b]$, the IRSE is defined as

$$\text{IRSE}(\hat{\Psi}) = \int_a^b \left\|\hat{\Psi}(u) - \Sigma(u)\right\|_F \mathrm{d}u \approx \frac{1}{K_0}\sum_{k=1}^{K_0}\left\|\hat{\Psi}(v_k) - \Sigma(v_k)\right\|_F, \tag{3.16}$$

where $v_k, 1 \le k \le K_0$ are grids evenly distributed over the interval $(a, b)$. In the following, we set $K_0 = 20$ for $(a, b) = (-0.95, 0.95)$. In our study, we also consider a spectral-norm based IRSE . As they are similar to the Frobenius version, the results are put off in Appendix A.2.

We also evaluate the performance of the proposed procedure in discovering zero entries in the covariance matrix. Let $p_1$ ($p_2$) be the number of nonzero (zero) entries in $\Sigma(u)$. For any estimator $\hat{\Psi}(u)$ of $\Sigma(u)$, let $n_{11}$ be the number of true discoveries of nonzero entries in $\Sigma(u)$ by $\hat{\Psi}(u)$. Similarly, let $n_{22}$ denote the number of true discoveries of zero entries in $\Sigma(u)$ by $\hat{\Psi}(u)$. Let SEN, SPE and ACC denote sensitivity, specificity and accuracy in the above tests, namely,

$$\text{SEN} = \frac{n_{11}}{p_1}, \quad \text{SPE} = \frac{n_{22}}{p_2}, \quad \text{ACC} = \frac{n_{11} + n_{22}}{p_1 + p_2}.$$

### 3.5.2 Synthetic Data

In this subsection, we carry out a set of simulation studies. We consider three settings for $\boldsymbol{\mu}(u)$ and $\Sigma(u)$ in our simulations.

**Setting 1:** Following Yuan & Cai (2010) and Chen & Leng (2016), we set $\boldsymbol{\mu}(u)$ and $\Sigma(u)$ as follows. Let $\boldsymbol{\mu}(u) = (\mu_1(u), \ldots, \mu_p(u))^T$ with

$$\mu_j(u) = \sum_{k=1}^{50} \frac{(-1)^{k+1}}{k^2} Z_{jk} \cos(k\pi u), \quad 1 \le j \le p,$$

where $\{Z_{jk} : 1 \le j \le p, 1 \le k \le 50\}$ is an independent sample drawn from the uniform distribution over $[-5, 5]$. Let $\Sigma(u) = \{\sigma_{ij}(u)\}_{1 \le i,j \le p}$ with $\sigma_{ij}(u) = \exp(u/2)[\{\phi(u) + 0.1\}I(|i-j| = 1) + \phi(u)I(|i-j| = 2) + I(i = j)]$ and $\phi(u)$ is the standard normal density. Note that $\mathrm{Diag}(\Sigma(u)) = \exp(u/2)I_p$ is spherical and the correlation matrix $C_0(u) = (c_{ij}(u))_{1 \le i,j \le p}$ with $c_{ij}(u) = I(|i-j| = 1) + \phi(u)I(|i-j| = 2) + I(i = j)$ which is equal to zero when $|i-j| \ge 3$. Therefore, $C_0(u)$ is sparse as it is a band matrix with bandwidth 2.

**Setting 2:** Following Zhang & Liu (2015), let $\boldsymbol{\mu}(u) = (\mu_1(u), \ldots, \mu_p(u))^T$ with

$$\mu_j(u) = Z_j \exp\left(\frac{(u-\tau_j)^2}{4}\right) \sin(2\pi(u-\tau_j)), \qquad 1 \le j \le p,$$

where $Z_j, j = 1, \ldots, p$ are independently drawn from the uniform distribution $U(-5, 5)$, $\tau = (\tau_1, \ldots, \tau_p)$ is a row vector of $p$ evenly spaced points between $-1$ and 1. Set $\Sigma(u) = \{\sigma_{ij}(u)\}_{1 \le i,j \le p}$ with $\sigma_{ij}(u) = \exp(u/2)\phi(u)^{|i-j|}$. Note that $\mathrm{Diag}(\Sigma(u)) = \exp(u/2)I_p$ is spherical and the correlation matrix $C_0(u) = (c_{ij}(u))_{1 \le i,j \le p}$ with $c_{ij}(u) = \phi(u)^{|i-j|}$. Therefore, $c_{ij}(u)$ is decreasing exponentially fast but is not sparse.

**Setting 3:** Let $\boldsymbol{\mu}(u)$ be the same as that in **Setting 1**. Let $\Sigma(u) = A^T(u)A(u)$, where the $(i,j)$-th entry of $A(u)$ equals

$$\begin{aligned} a_{ij}(u) \;=\; & \exp\left(\frac{u\sin(ij)}{2}\right) \Big\{ [\sin(\pi u) + 0.1]\, I(|i-j| = 1) \\ & + \sin(\pi u) I(|i-j| = 2) + I(i = j) \Big\}. \end{aligned}$$

Note that $\mathrm{Diag}(\Sigma(u)) = \mathrm{diag}(\sum_{j=1}^{p} a_{ij}^2(u) : 1 \le i \le p)$ is not spherical. $C_0(u)$ is sparse as it is a band matrix with bandwidth 4.

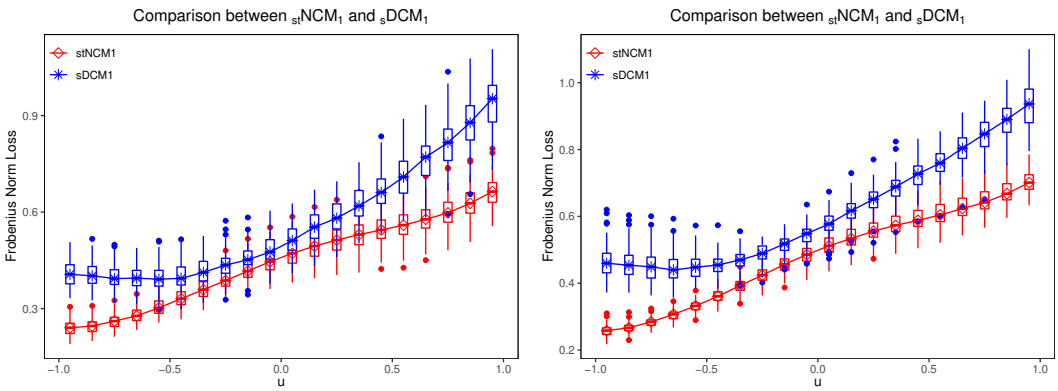
For each combination of $(n, p)$ with $n = 100, 200, 500$ and $p = 50, 100, 150, 300, 500$, we repeat the experiment 90 times, generating 90 datasets of $(\mathbf{y}_i, u_i)$, $1 \le i \le n$. Each dataset is obtained in two steps. In Step 1, we independently

draw $u_i, i = 1, \ldots, n$ from the uniform distribution $U(-1, 1)$. In Step 2, for each given $u_i$, we draw $\mathbf{y}_i$ from the covariance model $\mathbf{y}_i = \boldsymbol{\mu}(u_i) + \Sigma(u_i)^{1/2}\boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$ are iteratively drawn from the vector VAR(1) model

$$\boldsymbol{\varepsilon}_0 = \xi_0, \quad \boldsymbol{\varepsilon}_i = \rho\boldsymbol{\varepsilon}_{i-1} + \xi_i, \quad i = 1, \ldots, n,$$

with $0 \leq \rho < 1$ and $\xi_k, k = 0, 1, \ldots$ are independently sampled from the standard $p$-dimensional normal $N(0, I_p)$. We consider $\rho = 0, 0.3, 0.8$.

For each combination of $(n, p, \rho)$, we apply the ${}_{\text{t}}\text{NCM}_m$, ${}_{\text{st}}\text{NCM}_m$ $(m = 0, 1)$, $\text{DCM}_1$, ${}_{\text{s}}\text{DCM}_1$ and $\text{DCM}_2$ to each of 90 datasets and calculate their IRSE values and (SEN, SPE, ACC) values. The mean and standard error of these values are displayed in Figure 3.7, Table 3.1 below and Tables A.1 $\sim$ A.26 in Appendix A.2 respectively.




*(a)* $p = 50$ *(b)* $p = 100$

*Figure 3.7: Comparison Between ${}_{st}NCM_1$ and ${}_{s}DCM_1$ (Setting 1, n=100, $\rho = 0$)*

The results can be summarized as follows:

- On average, the Frobenius norm-based IRSE loss of each procedure is increasing in the dimension $p$ and in the degree of serial correlation $\rho$ while decreasing in sample size $n$.
- The degrees of sparsity and diagonal homogeneity in $\Sigma(u)$ have an effect on the performance of these four procedures. For example, when $(n, p, \rho) = (100, 300, 0)$, compared to the results in setting 1, the Frobenius norm-based IRSE loss of ${}_{\text{st}}\text{NCM}_0$ in setting 2 increases by 84%. This is not surprising as the degrees of sparsity and diagonal homogeneity in setting 2 lead to a higher dimensionality (i.e., the number of effective parameters in the model) than that in setting 1.
- Among the seven procedures, ${}_{\text{st}}\text{NCM}_1$ performs best in all three settings, followed by ${}_{\text{st}}\text{NCM}_0$, ${}_{\text{t}}\text{NCM}_1$, ${}_{\text{t}}\text{NCM}_0$, ${}_{\text{s}}\text{DCM}_1$, $\text{DCM}_1$ and $\text{DCM}_2$. In particular, the performance of $\text{DCM}_2$ is substantially worse than its competitors (see Section 3.2.1). For example, for $(n, p, \rho) = (100, 300, 0)$, in setting 1, compared

*Table 3.1: The Average (standard error in %) of Frobenius Norm-based IRSE for Setting 1*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $_{\text{t}}\text{NCM}_0$ | $_{\text{st}}\text{NCM}_0$ | $_{\text{t}}\text{NCM}_1$ | $_{\text{st}}\text{NCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0$ | | | | |
| 100 | 50 | 5.1307(25.04) | 0.5807(3.48) | 0.5634(3.49) | 0.4546(3.48) | 0.4482(3.61) | 0.4504(3.40) | 0.4431(3.50) |
| | 100 | 16.2142(47.02) | 0.6263(2.56) | 0.6116(2.54) | 0.4962(2.54) | 0.4878(2.64) | 0.4891(2.31) | 0.4798(2.40) |
| | 150 | 49.4335(75.62) | 0.6497(2.12) | 0.6363(2.12) | 0.5199(2.09) | 0.5112(2.19) | 0.5104(1.97) | 0.5006(2.05) |
| | 300 | 78.0434(48.33) | 0.7045(1.71) | 0.6935(1.71) | 0.5739(1.51) | 0.5654(1.56) | 0.5586(1.39) | 0.5488(1.43) |
| | 500 | 102.8816(39.13) | 0.7521(1.72) | 0.7417(1.72) | 0.6111(1.33) | 0.6021(1.38) | 0.5918(1.20) | 0.5812(1.23) |
| 200 | 50 | 2.8919(8.26) | 0.3650(2.69) | 0.3582(2.74) | 0.2821(2.42) | 0.2834(2.53) | 0.2808(2.40) | 0.2819(2.49) |
| | 100 | 8.8372(15.62) | 0.3868(1.79) | 0.3819(1.81) | 0.2976(1.56) | 0.2989(1.61) | 0.2967(1.52) | 0.2978(1.56) |
| | 150 | 18.5651(30.07) | 0.3915(1.64) | 0.3873(1.66) | 0.3031(1.50) | 0.3040(1.55) | 0.3021(1.47) | 0.3027(1.51) |
| | 300 | 70.8479(32.01) | 0.4194(1.26) | 0.4165(1.27) | 0.3201(1.00) | 0.3204(1.04) | 0.3198(0.97) | 0.3197(1.00) |
| | 500 | 84.8626(25.37) | 0.4508(1.08) | 0.4483(1.08) | 0.3315(0.92) | 0.3310(0.93) | 0.3310(0.87) | 0.3301(0.88) |
| 500 | 50 | 1.5780(3.56) | 0.2025(1.28) | 0.2018(1.37) | 0.1814(1.13) | 0.1831(1.19) | 0.1803(1.10) | 0.1820(1.15) |
| | 100 | 3.3680(4.40) | 0.2071(0.93) | 0.2070(0.94) | 0.1822(0.74) | 0.1840(0.78) | 0.1813(0.72) | 0.1831(0.76) |
| | 150 | 6.0579(6.45) | 0.2108(0.81) | 0.2109(0.82) | 0.1827(0.61) | 0.1845(0.64) | 0.1817(0.60) | 0.1835(0.63) |
| | 300 | 28.7062(24.38) | 0.2295(0.52) | 0.2304(0.52) | 0.1838(0.42) | 0.1856(0.44) | 0.1829(0.41) | 0.1846(0.43) |
| | 500 | 90.9963(20.61) | 0.2519(0.41) | 0.2535(0.43) | 0.1845(0.36) | 0.1863(0.38) | 0.1836(0.35) | 0.1853(0.37) |

to $\text{DCM}_2$, on average $\text{DCM}_1$ reduces the Frobenius norm-based IRSE loss by 99%. Compared to $\text{DCM}_1$, on average ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ reduce the Frobenius norm-based IRSE loss by 23% and 25% respectively. ${}_{\text{t}}\text{NCM}_1$ and ${}_{\text{st}}\text{NCM}_1$ perform slightly better than ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ in some settings. Compared to ${}_{\text{t}}\text{NCM}_0$, on average ${}_{\text{st}}\text{NCM}_0$ reduces the Frobenius norm-based IRSE loss by 3%. In setting 2, compared to $\text{DCM}_1$, on average ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ reduce the Frobenius norm-based IRSE loss by 12% and 16% respectively. ${}_{\text{t}}\text{NCM}_1$ and ${}_{\text{st}}\text{NCM}_1$ perform slightly better than ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$. Compared to $\text{DCM}_2$, on average $\text{DCM}_1$ reduces the Frobenius norm-based IRSE loss by 99%. Compared to ${}_{\text{t}}\text{NCM}_0$, on average ${}_{\text{st}}\text{NCM}_0$ reduces the Frobenius norm-based IRSE loss by 3%. In setting 3, compared to $\text{DCM}_1$, on average ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ reduce the Frobenius norm-based IRSE by 14% and 15% respectively. Compared to $\text{DCM}_2$, on average $\text{DCM}_1$ reduces the loss by 94%. Compared to ${}_{\text{t}}\text{NCM}_0$, on average ${}_{\text{st}}\text{NCM}_0$ reduces the Frobenius norm-based IRSE loss by 2%. ${}_{\text{t}}\text{NCM}_1$ and ${}_{\text{st}}\text{NCM}_1$ perform substantially better than their counterparts ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$. A similar conclusion can be made for dependent samples when $\rho = 0.3$ and 0.8. In particular, the optimal shrinkage can reduce the serial correlation effect on the proposed procedures ${}_{\text{st}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_1$. Furthermore, Spectral norm-based IRSE has the same performance as Frobenius norm-based IRSE, see Tables A.18 $\sim$ A.26 in Appendix A.2.

- Similar results are obtained in terms of ACC, see Tables A.9 $\sim$ A.17 in Appendix A.2.
- The CPU-time costs of ${}_{\text{t}}\text{NCM}_m$ and ${}_{\text{st}}\text{NCM}_m$, $m = 0, 1$, are less than those of $\text{DCM}_1$ and $\text{DCM}_2$. As example, for the 90 datasets simulated in setting 1 with $n = p = 100$, the CPU time required by $\text{DCM}_1$, ${}_{\text{s}}\text{DCM}_1$, $\text{DCM}_2$, ${}_{\text{t}}\text{NCM}_m$ and ${}_{\text{st}}\text{NCM}_m$, $m = 0, 1$ to estimate the covariance matrix function are reported in Figure 3.1.

### 3.5.3 Asset Return Data

Capital asset pricing model (CAPM) is a model that describes the relationship between systematic risk and expected return for assets, which is widely used throughout finance for the pricing of risky assets. However, the assumption that asset returns are linearly related to the market return is imposed on the model. The primary goal of this study is to extend the CAPM to the nonlinear setting. In particular, we are interested in how the volatility and co-volatility of a group of asset returns depend on the market return.

For this purpose, from the database of Yahoo Finance[1], we have collected monthly return data of 75 assets across 8 sectors over three time-periods, namely,

[1] https://finance.yahoo.com/?guccounter=1

before-financial-crisis period from 02/2001 to 01/2007, in-financial-crisis period from 02/2007 to 01/2010 and after-financial-crisis period from 02/2010 to 12/2017. The sector distribution of these assets is listed as follows. Technology: AAPL, AMD, HPQ, IBM, IIN, INTC, LNGY, LOGI, MSFT, NTAP, NVDA, SNE, TACT and WDC. Health care: AET, AMGN, AZN, BAX, CVS, GILD, GSK, HUM, IMMU, JNJ, LLY, MRK, NVS, PFE, TECH and UNH. Energy: BP, CVX, OXY, RDS-B, SU and XOM. Financial services: C, GS, HSBC, JPM, MS, PGR, RF and THG. Communication services: SHEN, T and TEO. Consumer defensive: BIG, DLTR, FRED, KO, TGT, TUES, UN and WMT. Consumer cyclical: AMZN, EMMS, KSS, SIRI and TM. Industrial: BA, CAJ, DY, EME, FIX, GE, GVA, IR, MMM, MTZ, PWR, SKYW, UPS, UTX and VMI. We have also collected the index return of S&P 500 which is treated as the market's return.

We apply the proposed ${}_{\text{st}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_1$ to the data for each time-period, obtaining almost the same result. Here, we report the corresponding estimates for mean $\boldsymbol{\mu}(u)$ and covariance matrix $\Sigma(u)$. Note that the diagonals of estimated $\Sigma(u)$ show the volatility of individual returns while estimated correlation coefficient matrix $C_0(u)$ captures cross-sectional relationships in these returns.

We plot the estimated individual mean functions and the estimated volatility functions in Figures A.1–A.3, revealing a number of assets which have nonlinear relationships to the market return. The degree of this non-linearity significantly decreases after financial crisis, indicating that the CAPM fitted to the market is better than that before the financial crisis, see Zhang & Li (2021) and the online Supplementary Material for more details. Figures A.1–A.3 also show that the individual volatility of the assets increases a lot during the financial crisis period but returns to normal after the financial crisis. The pattern of the dependence of the volatility on the market also changes a lot after financial crisis: Changes from non-constant volatility functions before the financial crisis to almost constant volatility functions after the financial crisis. We have also investigated effects of the financial crisis on the co-volatility of the selected assets by the estimated nonzero correlation coefficient functions. By use of the estimated covariance matrix functions, in each time-period, we have identified the associated pairs of assets that are of nonzero market-dependent conditional correlation coefficients (and nonzero conditional co-volatility).

We further conduct asymptotic tests for significance of co-volatility for these pairs as follows. For any pair of assets $(a, b)$, let $\text{Corr}_{(a,b)}(u)$ denote its correlation coefficient as a function of $u$ (the market's return) and with estimator $\hat{\text{Corr}}_{(a,b)}(u)$. Let $\hat{F}_{(a,b)}(u) = 0.5 \log((1 + \hat{\text{Corr}}_{(a,b)}(u))/(1 - \hat{\text{Corr}}_{(a,b)}(u)))$ be Fisher's Z transfor-

mation. To test $H_0 : \mathrm{Corr}_{(a,b)}(u) = 0$, we consider the test statistics

$$\mathrm{Avec}_{(a,b)} = \sum_{i=1}^{n} |\hat{F}_{(a,b)}(u_i)|/n \approx N\left(E[|F_{(a,b)}(U)|], \mathrm{Var}(|F_{(a,b)}(U)|)/n\right),$$

and calculate the approximate P-value

$$P\left(\sqrt{n}\mathrm{Avec}_{(a,b)}/\sqrt{\hat{\mathrm{Var}}(\mathrm{Corr}_{(a,b)}(U))}\Big| N(0,1)\right),$$

where the sample variance of $|\hat{F}_{(a,b)}(u_i)|, 1 \le i \le n$ is denoted by $\hat{\mathrm{Var}}(|F_{(a,b)}(U)|)$ and $P(\cdot|N(0,1))$ is the cumulative distribution function of the standard normal $N(0,1)$. Then, even after Bonferroni correction for multiple testing, these P-values are all significant ($< 10^{-2}$) for the above selected pairs of assets. The final list of significant pairs are as follows:

- *Before-financial-crisis.* There are 1, 14, 1 pairs existed within Technology, Energy and Consumer-Defensive respectively.
- *In-financial-crisis.* There are 4, 1, 8, 1, 4, 1, 1, 1, 1 pairs of correlated assets presented within Technology, Industrial, Energy, Consumer Defensive, Health Care and Financial Services respectively. Also, there is a pair of correlated assets belonging to different sectors: Industrial and Consumer cyclical, Consumer Cyclical and Consumer Defensive, and Financial Service and Industrial.
- *After-financial-crisis.* There are 3, 2, 10, 1, 11, and 12 pairs of assets within Technology, Industrial, Energy, Consumer defensive, Health care, and Financial services. There are 1, 1, 1, 1 and 2 pairs of assets between Financial service and Industrial, between consumer defensive and Financial services, between Consumer cyclical and Consumer defensive, between Technology and Industrial, and between Health Care and Consumer Defensive.

The results indicate that before financial crisis, there are only 16 significant within-sector co-volatility connections among these assets. In particular, there are no significant cross-sectional co-volatility connections among these assets. The number of co-volatility assets within and across sectors is significantly increasing during and after financial-crisis: The number of within-sector co-volatility connections increases from 16 to 22 during the financial crisis period and to 37 after the financial crisis. The number of between-sector co-volatility connections increases from 0 to 3 during the financial crisis period and to 7 after the financial crisis. This implies that in response to the financial crisis, the financial market has been more closely integrated than before the financial crisis.

## 3.6 Discussion and Conclusion

Estimating covariate-dependent covariance matrix $\Sigma(u)$ of a high-dimensional response vector poses a big challenge to contemporary statistical research. The existing kernel methods in Chen & Leng (2016) and Yin *et al.* (2010) might not be flexible enough to capture varying smoothness across key parts of the matrix as they used a single bandwidth for the entries of $\Sigma(u)$. Here, we have proposed a novel estimation procedure to overcome this obstacle, based on a variance-correlation factorization of $\Sigma(u)$, namely $\Sigma(u) = Q_0(u)C_0(u)Q_0^T(u)$, where $Q_0(u) = \text{Diag}(\Sigma(u))^{1/2}$ and the correlation matrix function $C_0(u)$ is further factorized into the product of multiple band matrices. The proposal has been implemented in two steps. In Step 1, we obtain robust estimators $Q_0(u)$ and $C_0(u)$ by use of separate bandwidths for band matrices, followed by thresholding entries of the estimated $C_0(u)$. In Step 2, we substitute these estimators in the above factorization formula to obtain a plug-in estimator, followed by an optimal shrinkage based on Frobenius norm.

We have conducted a set of simulations to demonstrate that the new proposal outperforms the existing DCM approach in terms of estimation loss and CPU-time cost. To illustrate our new proposal, we have applied it to a dataset of asset returns. We have developed a nonparametric capital asset pricing model to capture volatility and co-volatility among these risky assets. It shows that under some sparsity conditions, the proposed estimator is consistent with the underlying covariance matrix as both the sample size and the dimension tend to infinity. There are a few important topics which are remained to address but beyond the scope of this chapter, such as nonparametric nonlinear shrinkage.

# Chapter 4

# Divide-and-Combine Estimation of High-dimensional Nonparametric Covariance Models

## 4.1 Introduction

In Chapter 3, we have developed a factorized estimator of nonparametric covariance model. In this chapter, we will continue discussing the approaches that can further improve the estimation of nonparametric covariance model. Firstly, let us review the model in Chapter 3 again. Let $\mathbf{Y} = (Y_1, \ldots, Y_p)^T \in \mathbb{R}^p$ be a $p$-dimensional random vector and $U \in \mathbb{R}$ be the associated random variable. Denote $\boldsymbol{\mu}(u) = E[\mathbf{Y}|U = u]$ and $\Sigma(u) = \text{cov}(\mathbf{Y}|U = u)$ as the conditional mean and covariance matrix of $\mathbf{Y}$ given $U = u$. Each component of $\boldsymbol{\mu}(u)$ and $\Sigma(u)$ is assumed to be an unknown smooth function of $u$. Suppose that $(\mathbf{y}_i, u_i)_{i=1}^n$ are random observations from the model $\mathbf{y}_i = \boldsymbol{\mu}(u_i) + \Sigma(u_i)^{1/2}\varepsilon_i, i = 1, \ldots, n$ where $\varepsilon_i$ represents the noise.

As mentioned in Chapter 3, Chen & Leng (2016)'s method did not consider the effects of sparsity on the bandwidth selection, their covariance matrix estimation procedure includes two steps: the first step is bandwidth selection using the so-called *subset-y-variables* method; the second step is threshold using Bickel & Levina (2008b)'s method. The pilot study in Section 3.2.4 clearly illustrates how the zero entries in covariance matrix affect the bandwidth selection. In Chapter 3, we have proposed the factorized estimation of high-dimensional nonparametric covariance model which can solve the zero entries problem. Our factorized NCM includes five steps: standardization, factorization, bandwidth selection, threshold and shrinkage. Among these five steps, the factorization step plays a significant

role in solving zero entries problem by letting each factor $Q_k(u)$, $k = 1, \ldots, m$ own unique bandwidth, see the discussion in Section 3.3.

Generally, the estimator $\hat{\Sigma}(u)$ does not perform well given $p \gtrapprox n$ as it can generate the estimation error (Kan & Zhou, 2007). The main reason is that there exist many unknown parameters in $\Sigma(u)$ to be estimated but using only the finite observations. To eliminate this issue, Zou *et al.* (2017) pointed out that sparsity assumption is frequently imposed on either covariance matrix (Huang *et al.*, 2006; Bickel & Levina, 2008b) or its precision matrix (Dempster, 1972; Meinshausen & Bühlmann, 2006; Yuan & Lin, 2007; Friedman *et al.*, 2008). Bickel & Levina (2008b)'s threshold approach is used both in DCM and our factorized NCM methods. However, we notice that the threshold step is just following the bandwidth selection step in the DCM and our factorized NCM methods, this implies the zero entries still have effect on the bandwidth selection of nonzero entries.

Wang & Kolar (2014) proposed a method to implement bandwidth selection and zero entries identification simultaneously. They imposed the group graphical lasso penalty (Yuan & Lin, 2007) on $\Sigma(u)$ during the bandwidth selection step. To be concrete, the criterion of estimating the inverse matrix of $\Sigma(u_i)$ is

$$\min_{\{\Omega(u_i)\succ 0\}} \sum_{i=1}^{n} \left[ \operatorname{tr}\left(\hat{\Sigma}(u_i)\Omega(u_i)\right) - \log|\Omega(u_i)| \right] + \lambda \sum_{j_1 \neq j_2} \sqrt{\sum\nolimits_{i=1}^{n} \Omega^2_{j_1 j_2}(u_i)}, \quad (4.1)$$

where $\Omega(u_i)$ represents the matrix inversion of $\Sigma(u_i)$, $\Omega_{j_1 j_2}(u_i)$ is the entry located in $j_1$-th row and $j_2$-th column of $\Omega(u_i)$ and item $\sum_{j_1 \neq j_2} \sqrt{\sum_{i=1}^{n} \Omega^2_{j_1 j_2}(u_i)}$ is called the group graphical lasso penalty. Even though they proposed criterion (4.1), they did not select the optimal bandwidth via (4.1) in practice, because cross-validation and optimizing the penalty in high-dimensional setting will dramatically increase the computational complexity. They directly set $h = n^{-1/5}$ in (4.1) and concentrated on the penalty part throughout their research.

In this chapter, we propose the Divide-and-Combine estimation approach for both mean function and covariance matrix function. The key idea of Divide-and-Combine approach for the nonparametric covariance matrix is to identify the locations of zero entries before the bandwidth selection of the nonzero entries. To be concrete, our technical route has three steps:

1. The diagonal entries are estimated by using the local linear smoother, once we obtain the diagonal estimators, we can standardize the covariance matrix to get the correlation coefficient matrix.
2. We now only focus on the off-diagonal entries. There are $p(p-1)/2$ non-diagonal pairs $(j_1, j_2)$,$1 \leq j_1 < j_2 \leq p$ need to be identified. We assume that the positions of zero entries do not change with the condition $u_i, i =$

$1, \ldots, n$. If the underlying correlation coefficient $\rho_{j_1 j_2}(\cdot) = 0$ for $u_i, i = 1, \ldots, n$ at one specific pair $(j_1, j_2)$, then we can carry out the multiple null hypothesis: $\rho_{j_1 j_2}(\cdot) = 0$. If we can not reject the null hypothesis at the significant level $\alpha$, it is reasonable to treat this specific location $(j_1, j_2)$ as zero entry; otherwise, nonzero entry. This procedure is implemented just one time for one pair $(j_1, j_2)$. There are $p(p-1)/2$ non-diagonal pairs need to be tested using the same procedure. Typically, this is a multiple hypothesis testing, we use false discovery rate (FDR) here to keep the Type I error under an appropriate level. So far, we can identify the zero entries and nonzero entries.

3. Once obtaining the locations of nonzero entries, one can choose the bandwidth using nonparametric method based on the nonzero entries. Combining the main diagonal entries' estimators and off-diagonal nonzero entries' estimators, we can easily obtain the covariance matrix estimator. One can use shrinkage method (Ledoit & Wolf, 2004) or the method in Chen & Leng (2016) to make this covariance matrix estimator positive definite.

Next, we discuss the Divide-and-Combine method in the estimation of mean function.

### 4.1.1 Divide-and-Combine Estimation of Mean Function

In Chapter 3, we employ the local constant smoother (Nadaraya, 1964) to estimate both mean function and covariance matrix function. Fan & Gijbels (1996) suggested using the local linear smoother due to its minimax efficiency and the advantage of boundary effect auto-correction. However, if the components of underlying mean function consist of both the linear and nonlinear functions of $u$, then the common bandwidth of the component functions is sensitive to the proportion of linear functions. For example, if the mean function $\boldsymbol{\mu}(u)$ is composed of linear functions (including constant functions), then the common bandwidth of local linear smoother will go to infinity (e.g., Fan & Gijbels, 1996, p. 20) as the linear functions dominate the convergence of bandwidth. From now on, we call each component of $\boldsymbol{\mu}(u)$ as entry-wise function (EWF) with respect to $u$.

To solve the problem of infinite bandwidth, we first try to detect the linear and nonlinear EWFs of $\boldsymbol{\mu}(u)$, then split them into two groups: one is linear, and the other is nonlinear. For the linear EWFs of $\boldsymbol{\mu}(u)$, we directly use ordinary least square to evaluate them. For the nonlinear EWFs of $\boldsymbol{\mu}(u)$, we employ local linear smoother to estimate them. Hence, the key question is how to detect the linear EWFs of $\boldsymbol{\mu}(u)$ based on the observations.

We notice that Fan *et al.* (2001) proposed a *generalized likelihood ratio statistic* method, one application of this method is linearity test based on nonparamet-

ric maximum likelihood ratio statistic. The basic idea of generalized likelihood ratio statistic can be described as follows. Consider the null hypothesis:

$$\mathbf{H}_0 : m(x) = \alpha_0 + \alpha_1 x \quad \text{v.s.} \quad \mathbf{H}_1 : m(x) \neq \alpha_0 + \alpha_1 x,$$

where $x$ is a univariate variable, $\alpha_0$ and $\alpha_1$ are unknown parameters. Let $\hat{m}_h(x_i)$ be the local linear fits and $h$ is the nominal bandwidth, then generalized likelihood ratio test is given by

$$\lambda_n = \frac{n}{2} \log \frac{\text{RSS}_0}{\text{RSS}_1},$$

where $\text{RSS}_0 = \sum_{i=1}^n (y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i)^2$ and $\text{RSS}_1 = \sum_{i=1}^n (y_i - \hat{m}_h(x_i))^2$. Under the null hypothesis, one obtains

$$r_K \lambda_n \sim \chi^2_{a_K},$$

where $a_K$ is the degrees of freedom and $r_K$ is a positive constant. For more details, see Fan *et al.* (2001).

Suppose that we have obtained the linear and nonlinear EWFs, denote them as $\left(\mathbf{y}_i^{(1)}\right), \left(\mathbf{y}_i^{(2)}\right), i = 1, \ldots, n$ respectively, where $\mathbf{y}_i^{(1)}$ is $p_1$-dimensional column vector and $\mathbf{y}_i^{(2)}$ is $p_2$-dimensional column vector. $p_1$ and $p_2$ satisfy $p_1 + p_2 = p$. The linear EWFs are estimated directly by $\hat{\boldsymbol{\mu}}^{(1)}(u_i) = \hat{\alpha} + \hat{\beta} u_i$, where $\hat{\alpha}, \hat{\beta}$ are OLS estimators respectively. For the nonlinear EWFs, let $h_1$ be the bandwidth of kernel function. By the local linear smoother, the estimator can be expressed as

$$\hat{\boldsymbol{\mu}}^{(2)}(u) = \frac{\sum_{i=1}^n w_{h_1}(u_i - u)\mathbf{y}_i^{(2)}}{\sum_{i=1}^n w_{h_1}(u_i - u)}, \tag{4.2}$$

where $w_{h_1}(u_i - u) = K_{h_1}(u_i - u)[S_{n,2} - (u_i - u)S_{n,1}]$ represents the equivalent kernel, and $S_{n,j} = \sum_{i=1}^n K_{h_1}(u_i - u)(u_i - u)^j$, $j = 1, 2$. To demonstrate this idea, we set up a simple simulation in Section 4.2.1, Figure 4.5(a) shows that dividing the mean function estimation procedure into two steps performs better than estimating the mean function directly.

This idea is inspired by the computer algorithm divide-and-conquer (Cormen, 2009). We divide the mean function estimation into two sub-problems, then estimate each separately. Furthermore, we also apply this idea to the estimation of nonparametric covariance matrix, see the discussion in the previous subsection. Throughout this chapter, we call it Divide-and-Combine framework. Besides the Divide-and-Combine estimators of mean function and nonparametric covariance matrix, we also propose a nonparametric estimation of correlation coefficient by solving a cubic equation of correlation coefficient which will be discussed later.

### 4.1.2 Nonparametric Estimation of Correlation Coefficient

To demonstrate the motivation, we set up a simple example: Supposing $u \in [-1, 1]$ and the distribution of $(X_1, X_2)$ be a bivariate normal distribution,

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}1 & \rho(u)\\ \rho(u) & 1\end{bmatrix}\right),$$

where $\rho(u) = 1 - 2u^2$. Let $n = 200$, $u_i, i = 1, \ldots, n$ are randomly drawn from the uniform distribution $U(-1, 1)$. For each $u_i$, observation $(x_{i1}, x_{i2})$ is one random sample the above bivariate normal distribution. Given $u = u_0$, one can directly obtain the following nonparametric correlation coefficient estimator:

$$\hat{\rho}(u_0) = \frac{\sum_{i=1}^{n} x_{i1}x_{i2}K_h(u_i - u_0)}{\sum_{i=1}^{n} K_h(u_i - u_0)}. \tag{4.3}$$

However, estimator (4.3) may be larger than 1 or smaller than -1, for example, see Figure 4.1. Hence, we need to impose a constraint on (4.3) to guarantee $\rho(u_0) \in [-1, 1], \forall\ u_0 \in [-1, 1]$. We notice that the kernel weighted likelihood function of bivariate normal distribution at $u = u_0$ can be expressed as

$$L = \sum_{i=1}^{n} \left\{ \frac{x_{i1}^2 + x_{i2}^2 - 2\rho(u_0)x_{i1}x_{i2}}{1 - \rho^2(u_0)} + \log\left(1 - \rho^2(u_0)\right) \right\} K_h(u_i - u_0).$$

Let $\partial L/\partial\rho(u_0) = 0$, after some simple calculations, we get the so-called cubic equation of correlation coefficient as follows:

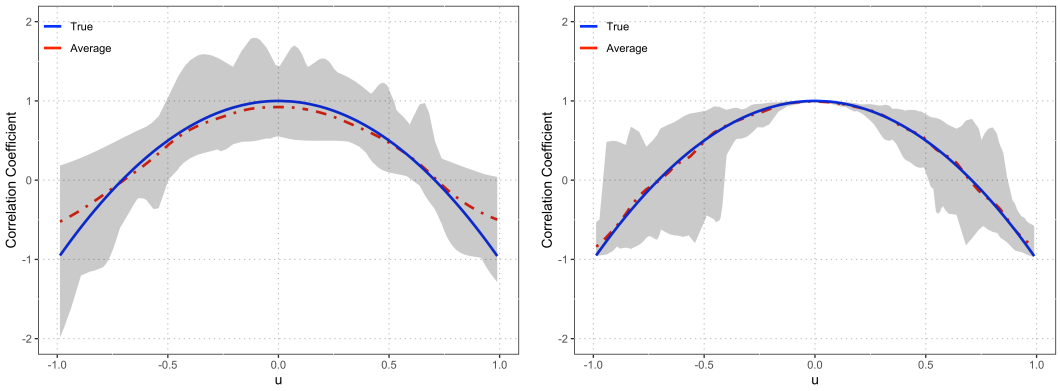
$$\rho^3(u_0) - B(u_0)\rho^2(u_0) + [A(u_0) - 1]\rho(u_0) - B(u_0) = 0, \tag{4.4}$$

where

$$A(u_0) = \frac{\sum_{i=1}^{n} \left[x_{i1}^2 + x_{i2}^2\right] K_h(u_i - u_0)}{\sum_{i=1}^{n} K_h(u_i - u_0)}, B(u_0) = \frac{\sum_{i=1}^{n} x_{i1}x_{i2}K_h(u_i - u_0)}{\sum_{i=1}^{n} K_h(u_i - u_0)}.$$

Especially, if $A(u_0) = 2$, one real root of (4.4) is equal to $B(u_0)$. Kendall *et al.* (1973) pointed out that there is at least one real root of equation (4.4) lying in the interval $[-1, 1]$. The details of bandwidth selection and roots of equation (4.4) are put off in Section 4.2.2. For each $u_i$, we can obtain two estimators: (4.3) and the real root of (4.4). We repeat the above simulation 100 times. Figure 4.1 summarizes this comparison study. The blue solid line represents the underlying correlation coefficient function. Each point of the red dash-dot line in Figures 4.1(a) and 4.1(b) represents the average of 100 estimators of (4.3) or (4.4). The grey

ribbons in Figures 4.1(a) and 4.1(b) are bounded by the maximum and minimum of 100 estimators of (4.3) and (4.4) respectively.




*(a) Estimator (4.3)* *(b) The real root of (4.4)*

*Figure 4.1: The Comparison of Two Correlation Estimators.*

Apparently, the correlation coefficient estimated by (4.3) could be greater than 1 or smaller than -1 while the real root of (4.4) are definitely lying in $[-1, 1]$ as shown in Figure 4.1(b). Even comparing the red dash-dot lines in Figures 4.1(a) and 4.1(b), the real root of (4.4) still preforms better than estimator (4.3).

Our Divide-and-Combine estimation for high-dimensional nonparametric covariance models needs to identify the zero entries based on the correlation matrix, see Section 4.2.2. To avoid the potential risk of estimator (4.3), we adopt the real root of cubic equation (4.4) as our nonparametric correlation coefficient estimator throughout this chapter.

The simulation result shows that our method is not only efficient for the sparse covariance matrix but also for the full covariance matrix. For example, the underlying full covariance matrix in scenario 4 decays at the exponential rate, however both the Frobenius norm-based IRSE and spectral norm-based IRSE are still less than those using the NCM method. In the previous paragraph, we assume the position of zero entries do not change with $u_i$. To evaluate the effect of varying-zero-position, we also design the scenario 6, the result illustrates that our method also performs well under this circumstance. Even the assumptions are not satisfied in scenario 4 and 6, our method still performs better. This is not surprising because through identifying zeros entries, the nonzero entries again play the main role in bandwidth selection. It also indicates that we do not need to identify the zeros entries exactly, we just need to control the zero entries effect below an acceptable tolerance. In essence, our method reduces the influence of zero entries so that the nonzero entries can take over the bandwidth selection.

The rest of this chapter is organized as follows: Section 4.2 shows the details of our three-step nonparametric covariance model. Simulation studies and a real

finance example are given in Section 4.3 and Section 4.4 while Section 4.5 concludes Chapter 4 with discussions. The detail of simulation results can be found in Appendix B. An R package is also developed on GitHub[1].

## 4.2 Methodology

We briefly review our model. Let $\mathbf{Y} = (Y_1, \ldots, Y_p)^T \in \mathbb{R}^p$ be a $p$-dimensional random vector and $U \in \mathbb{R}$ be the associated random variable. Denote $\boldsymbol{\mu}(u) = E[\mathbf{Y}|U = u]$ and $\Sigma(u) = \text{cov}(\mathbf{Y}|U = u)$ as the conditional mean and covariance matrix of $\mathbf{Y}$ given $U = u$. Each component of $\boldsymbol{\mu}(u)$ and $\Sigma(u)$ is assumed to be an unknown smooth function of $u$. Suppose that $(\mathbf{y}_i, u_i)_{i=1}^n$ are random observations from the following model

$$\mathbf{y}_i = \boldsymbol{\mu}(u_i) + \Sigma(u_i)^{1/2}\varepsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\mu}(u_i) = (\mu_1(u_i), \ldots, \mu_p(u_i))^T$ and $(u_i)_{i=1}^n$ is an independent random sample of $U$. Also, given $(u_i)_{i=1}^n$, $\varepsilon_i$'s are dependent on each other and with zero means and identity matrices (i.e., $E[\varepsilon_i|u_i] = 0_p$, $\text{cov}(\varepsilon_i|u_i) = I_p$ and $E[\varepsilon_i\varepsilon_j^T] \neq 0, i \neq j$). Let $K(u)$ be a kernel function, $K_h(u) = h^{-1}K(u/h)$ be the scaled kernel function with bandwidth $h > 0$ and $w_h(u_i - u) = K_h(u_i - u)/\sum_{k=1}^n K_h(u_k - u)$ be the weight function. Yin *et al.* (2010) considered the following kernel estimators of $\boldsymbol{\mu}(\cdot)$ and $\Sigma(\cdot)$:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}(u) &= \sum_{i=1}^n w_h(u_i - u)\mathbf{y}_i, \\
\hat{\Sigma}(u) &= \sum_{i=1}^n w_h(u_i - u)\left[\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i)\right]\left[\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i)\right]^T \triangleq (\hat{\sigma}_{j_1 j_2}(u))_{1 \leq j_1, j_2 \leq p},
\end{aligned}$$

under independent and $n > p$ conditions where $h$ is a nominated bandwidth for both mean and covariance matrix functions.

We aim to use Divide-and-Combine method to estimate both conditional mean function $\boldsymbol{\mu}(u)$ and conditional covariance matrix $\Sigma(u)$. In this section, we will introduce the estimation of mean function followed by a simple example to illustrate its performance. Thereafter, a new nonparametric covariance model to reduce the sparsity effects will be developed. Under the sparsity assumption, we adopt the FDR to keep Type I error under an appropriate level in identifying zero entries of covariance matrix estimator, then apply the local linear regression to those nonzero entries to obtain the estimator of covariance matrix.

[1]This repository (`https://github.com/Jieli12/llfdr`) is currently private, once this article is accepted online, this package will become open source under GPL-3 Licence.

### 4.2.1 Mean Function Estimation

We employ the generalized likelihood ratio statistic test to identify the linear and nonlinear mean functions. For more details of the generalized likelihood ratio statistic test, see Fan *et al.* (2001). We still use the notations introduced in Section 4.2.1. In terms of the bandwidth selection of nonlinear functions, we use the leave-one-out cross validation method to obtain the bandwidth. Define the cross validation function as

$$CV(h_1) = \frac{1}{np_2} \sum_{i=l_1}^{l_2} [\mathbf{y}_i^{(2)} - \hat{\boldsymbol{\mu}}_{-i}^{(2)}(u_i)]^T [\mathbf{y}_i^{(2)} - \hat{\boldsymbol{\mu}}_{-i}^{(2)}(u_i)],$$

where $l_1 = \lfloor 0.05n \rfloor$, $l_2 = \lfloor 0.95n \rfloor$ and $\hat{\boldsymbol{\mu}}_{-i}^{(2)}(\cdot)$ represents the duplication of equation (4.2) without the $i$-th observation. Finally, combining $\hat{\boldsymbol{\mu}}^{(1)}(u_i)$ and $\hat{\boldsymbol{\mu}}^{(2)}(u_i)$, we can obtain the mean function estimator $\hat{\boldsymbol{\mu}}(u_i)$.

To demonstrate this idea, we set up a simple simulation. Let $n = 200, p = 100$. Without loss of generality, the first $p_1 = 60$ EWFs are linear, the rest are nonlinear. The underlying mean function is $\mathbf{y}_i = \boldsymbol{\mu}(\mu_i) + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\mu}(u_i) = \left(\boldsymbol{\mu}^{(1)}(u_i)^T, \boldsymbol{\mu}^{(2)}(u_i)^T\right)^T$. The linear EWFs are $\boldsymbol{\mu}^{(1)}(u_i) = \alpha + \beta u_i$, where $\alpha = 0.01 \times \mathbf{1}$, $\mathbf{1}$ is a $p_1$ column vector with all entries 1, and $\beta = (\beta_1, \ldots, \beta_{p_1})^T$, the nonlinear EWFs $\mu_j^{(2)}(u_i)$ are $\sin(\pi(u_i + j/p))$, $j = 1, \ldots, p_2$ and $\boldsymbol{\mu}^{(2)}(u_i) = (\mu_1^{(2)}(u_i), \ldots, \mu_{p_2}^{(2)}(u_i))^T$.

We randomly draw $u_i$ from the uniform distribution $U(0, 1)$, $\beta$ is generated randomly from the uniform distribution $U(0.2, 0.5)$ and the noise $\boldsymbol{\varepsilon}_i$ is sampled from the multivariate normal distribution with zero mean and identity matrix. This procedure is repeated 90 times. We set up two different estimation methods, denoted as A and B respectively. Method A represents that we estimate the mean function without dividing the linear and nonlinear EWFs while Method B represents our Divide-and-Combine framework. For each method, we compute the average mean square error of mean function, Figure 4.5(a) displays the comparison results. Apparently, our new method performs better than Method A. Furthermore, we also calculate the sensitivity (or true positive rate) of linear functions identification, the average of 90 sensitivities is 0.9965 which means we can efficiently detect the linear EWFs by Fan *et al.* (2001)'s approach.

### 4.2.2 Covariance Matrix Estimation

#### 4.2.2.1 Identifying Off-diagonal Zero Entries

Without loss of generality, in the rest of this section, we always assume $\mathbf{y}_i, i = 1, \ldots, n$ is centralized by the Divide-and-Combine mean function esti-

mator in Section 4.2.1. First, we focus on the off-diagonal zero entries in covariance matrix $\Sigma(u)$, the diagonal and off-diagonal nonzero entries will be discussed afterwards. Let $(j_1, j_2)$, $j_1 \neq j_2, 1 \leq j_1, j_2 \leq p$ denote the row and column indices of one off-diagonal zero entries of $\Sigma(u)$, $u = u_1, \ldots, u_n$. For given $u_i$, let $\rho_{j_1 j_2}(u_i)$ be the conditional correlation of $y_{ij_1}$ and $y_{ij_2}$. Clearly, if the $(j_1, j_2)$-entry of $\Sigma(u_i), i = 1, \ldots, n$ is 0, i.e., $\rho_{j_1 j_2}(u_1) = \cdots = \rho_{j_1 j_2}(u_n) = 0$, then we can treat these correlations as 'unconditional' because $u_i, i = 1, \ldots, n$ does not affect the value of conditional correlation. This inspires us to employ the hypothesis test for correlation coefficient to identify the indices pair $(j_1, j_2)$ of off-diagonal zero entries in covariance matrix $\Sigma(u_i), i = 1, \ldots, n$. The null hypothesis is $H_0 : \rho_{j_1 j_2} = 0$ *v.s.* $H_1 : \rho_{j_1 j_2} \neq 0$. Under $H_0$, the test statistic is

$$t_{j_1 j_2} = \hat{\rho}_{j_1 j_2} \sqrt{\frac{n-2}{1 - \hat{\rho}^2_{j_1 j_2}}},$$

which follows $t$-distribution with $n-2$ degrees of freedom. The estimator of correlation coefficient is $\hat{\rho}_{j_1 j_2} = n^{-1} \sum_{i=1}^{n} y_{ij_1} y_{ij_2} / \sqrt{\sigma_{j_1 j_1}(u_i) \sigma_{j_2 j_2}(u_i)}$ where $\sigma_{j_1 j_1}(u_i)$ and $\sigma_{j_2 j_2}(u_i)$ are the variance of $y_{ij_1}$ and $y_{ij_2}$ respectively. If $H_0$ is rejected at the significant level $\alpha_0$, then we denote $e_{j_1 j_2} = 1$; otherwise $e_{j_1 j_2} = 0$.

Totally, there exist $m = p(p-1)/2$ combinations of $(j_1, j_2)$, $j_1 \neq j_2, 1 \leq j_1, j_2 \leq p$. This leads to a typical multiple comparison problem. False Discovery Rate (FDR) can control Type I Error of multiple hypothesis tests. After the FDR process, we obtain the indicator matrix $E = (e_{j_1 j_2})_{p \times p}$ (in graph theory, it is also called adjacency matrix), $E$ indicates the off-diagonal zero entries locations in covariance matrix $\Sigma$.

#### 4.2.2.2 Estimation of Diagonal Entries

During the FDR step, we notice that $\sigma_{j_1 j_1}(u)$ and $\sigma_{j_2 j_2}(u)$ are unknown parameters. In fact, we need to estimate the diagonal entries $\sigma_{jj}(u_i)$, $j = 1, \ldots, p$. In this section, we introduce two popular methods to estimate the diagonal entries: local linear and local maximum likelihood.

**Local Linear Method** For simplicity, we fix the index $j$, then the observations we need are $y_{ij}$, $i = 1, \ldots, n$. Given $u = u_0$, the objective function of local linear method can be expressed as

$$\sum\nolimits_{i=1}^{n} \left\{ y_{ij}^2 - \left[ \alpha_j^*(u_0) + \beta_j^*(u_0)(u_i - u_0) \right] \right\}^2 K_{h_2}(u_i - u_0),$$

where $\alpha_j^*(u_0) + \beta_j^*(u_0)(u_i - u_0)$ is the local linear approximation of $\sigma_{jj}(u_i)$ at $u_0$. One can easily obtain the local linear estimator, i.e.,

$$\hat{\sigma}_{jj}(u_0) = \hat{\alpha}_j^*(u_0) = \frac{\sum_{i=1}^n w_{h_2}(u_i - u_0) y_{ij}^2}{\sum_{i=1}^n w_{h_2}(u_i - u_0)}, \tag{4.5}$$

where $w_{h_2}(u_i - u_0) = K_{h_2}(u_i - u_0)[S_{n,2} - (u_i - u_0)S_{n,1}]$, and $S_{n,j} = \sum_{i=1}^n K_{h_2}(u_i - u_0)(u_i - u_0)^j$, $j = 1, 2$. We can get the estimators $\hat{\sigma}_{jj}(u_0)$ for $j = 1, \ldots, p$ when $u_0$ takes the values $u_1, \ldots, u_n$ respectively. The bandwidth $h_2$ can be selected by the leave-one-out cross validation method:

$$CV(h_2) = \sum_{j=1}^{p} \sum_{i=1}^{n} \left[ y_{ij}^2 - \hat{\sigma}_{jj(-i)}(u_i) \right]^2,$$

where $\hat{\sigma}_{jj(-i)}(u_i)$ is the version of equation (4.5) when $u_0 = u_i$ without the $i$-th observation. However, as we reviewed in Section 2.1.2, the equivalent kernel $w_{h_2}(u_i - u_0)$ may be negative which can lead to negative variance (see Figure 2.1). In practice, one can either use the interpolation method to adjust the negative variance or use the local maximum likelihood method to satisfy the positive variance.

**Local Maximum Likelihood Method** Given $u = u_0$, the likelihood objective function is

$$\sum\nolimits_{i=1}^{n} \left[ \frac{y_{ij}^2}{\sigma_{jj}(u_i)} + \log(\sigma_{jj}(u_i)) \right] K_{h_3}(u_i - u_0). \tag{4.6}$$

Fan & Yao (1998) and Yu & Jones (2004) treated the estimation of $\sigma_{jj}(\cdot)$ as a non-parametric regression problem. They applied the local linear regression smoother to $\sigma_{jj}(\cdot)$. Inspired by their idea, we let $\sigma_{jj}(u_i) = \exp(\alpha_j(u_0) + \beta_j(u_0)(u_i - u_0))$, which can guarantee the estimator of variance positive. The likelihood objective function $L_j$ can be expressed as

$$\sum_{i=1}^{n} \left[ \frac{y_{ij}^2}{\exp(\alpha_j(u_0) + \beta_j(u_0)(u_i - u_0))} + \alpha_j(u_0) + \beta_j(u_0)(u_i - u_0) \right] K_{h_3}(u_i - u_0),$$

where $h_3$ represents the bandwidth for diagonal entries. We take the partial differentiation of $L_j$ with respect to $\alpha_j(u_0)$ and $\beta_j(u_0)$ respectively, and let them equal to zero. After some simple calculations, we obtain

$$\frac{\sum_{i=1}^n \frac{y_{ij}^2 u_i}{\theta_i^{\beta_j(u_0)}} K_{h_3}(u_i - u_0)}{\sum_{i=1}^n \frac{y_{ij}^2}{\theta_i^{\beta_j(u_0)}} K_{h_3}(u_i - u_0)} = \frac{\sum_{i=1}^n u_i K_{h_3}(u_i - u_0)}{\sum_{i=1}^n K_{h_3}(u_i - u_0)}. \tag{4.7}$$

Equation (4.7) is a non-linear equation of $\beta_j(u_0)$ where $\theta_i = \exp(u_i)$.

Denote the real root of (4.7) as $\hat{\beta}_j(u_0)$, then

$$\hat{\alpha}_j(u_0) = \log\left[\frac{\sum_{i=1}^n \left[\frac{y_{ij}^2}{\exp(\hat{\beta}_j(u_0)(u_i-u_0))}\right] K_{h_3}(u_i-u_0)}{\sum_{i=1}^n K_{h_3}(u_i-u_0)}\right], \tag{4.8}$$

so

$$\hat{\sigma}_{jj}(u_i,u_0) = \exp[\hat{\alpha}_j(u_0) + \hat{\beta}_j(u_0)(u_i-u_0)], \tag{4.9}$$

where $\hat{\sigma}_{jj}(u_i,u_0)$ represents the exponential approximation estimator of $\sigma_{jj}(u_i)$ at $u=u_0$. In practice, we only need to compute the exponential approximation estimator of $\sigma_{jj}(u_i)$ at $u=u_i$, i.e., $\hat{\sigma}_{jj}(u_i,u_i)$. From now on, we denote $\hat{\sigma}_{jj}(u_i,u_i)$ as $\hat{\sigma}_{jj}(u_i)$. Replacing $u_0$ with $u_i$ in (4.8) and (4.9), after some simple calculations, we can obtain

$$\hat{\sigma}_{jj}(u_i) = \frac{\sum_{s=1}^n \left[\frac{y_{sj}^2}{\exp\left(\hat{\beta}_j(u_i)(u_s-u_i)\right)}\right] K_{h_3}(u_s-u_i)}{\sum_{s=1}^n K_{h_3}(u_s-u_i)}, \quad i=1,\ldots,n. \tag{4.10}$$

Hence, the key issue here is to find the real root of equation (4.7). Once this root is available, according to (4.10), we can finally get the estimators $\hat{\sigma}_{jj}(u_i)$, $i=1,\ldots,n$, $j=1,\ldots,p$. Furthermore, both $\hat{\beta}_j(u_i)$ and $\hat{\sigma}_{jj}(u_i)$ are related to the bandwidth $h_3$. In this section, we still use the leave-one-out cross validation method to construct the objective function

$$CV(h_3) = \sum_{j=1}^p \sum_{i=1}^n \left[\frac{y_{ij}^2}{\hat{\sigma}_{jj(-i)}(u_i)} + \log(\hat{\sigma}_{jj(-i)}(u_i))\right], \tag{4.11}$$

where $\hat{\sigma}_{jj(-i)}(u_i)$ is the estimator of $\sigma_{jj}(u_i)$ without the $i$-th observation. By equation (4.10), we have

$$\hat{\sigma}_{jj(-i)}(u_i) = \frac{\sum_{s=1,s\neq i}^n \left[\frac{y_{sj}^2}{\exp\left(\hat{\beta}_{j(-i)}(u_i)(u_s-u_i)\right)}\right] K_{h_3}(u_s-u_i)}{\sum_{s=1,s\neq i}^n K_{h_3}(u_s-u_i)},$$

where $\hat{\beta}_{j(-i)}(u_i)$ is the root of the function below:

$$\frac{\sum_{s=1,s\neq i}^n \frac{y_{sj}^2 u_s}{\theta_s^{\beta_{j(-i)}(u_i)}} K_{h_3}(u_s-u_i)}{\sum_{s=1,s\neq i}^n \frac{y_{sj}^2}{\theta_s^{\beta_{j(-i)}(u_i)}} K_{h_3}(u_s-u_i)} = \frac{\sum_{s=1,s\neq i}^n u_s K_{h_3}(u_s-u_i)}{\sum_{s=1,s\neq i}^n K_{h_3}(u_s-u_i)}. \tag{4.12}$$

Furthermore, we re-parameterize $\beta_{j(-i)}(u_i)$ as $\beta_{ji}^*$, then equation (4.12) can be

expressed as

$$f(\beta_{ji}^*) = \left[\sum_{s=1,s\neq i}^{n} \frac{y_{sj}^2 u_s}{\theta_s^{\beta_{ji}^*}} K_{h_3}(u_s - u_i)\right]\left[\sum_{s=1,s\neq i}^{n} K_{h_3}(u_s - u_i)\right] - \left[\sum_{s=1,s\neq i}^{n} \frac{y_{sj}^2}{\theta_s^{\beta_{ji}^*}} K_{h_3}(u_s - u_i)\right]\left[\sum_{s=1,s\neq i}^{n} u_s K_{h_3}(u_s - u_i)\right] = 0. \tag{4.13}$$

Minimizing (4.11) is time-consuming because for each candidate bandwidth $h_3$, we need to solve equation (4.13) $np$ times (the combinations of $i$ and $j$) to obtain one evaluation of $CV(h_3)$. We employ Newton-Raphson method to speed up the evaluation of $CV(h_3)$ and solve equation (4.13), for more details, see Appendix B.

#### 4.2.2.3 Estimation of Off-diagonal Nonzero Entries

Once we obtain the bandwidth $h_3$ using the leave-one-out cross validation method, we can estimate the diagonal entries of $\Sigma(u_i)$, $i = 1, \ldots, n$ using equation (4.10). The FDR step can be implemented to get the adjacency matrix $E$. Next, we estimate the off-diagonal nonzero entries in matrix $\Sigma(\cdot)$. Recall that the adjacency matrix $E$ is a symmetric matrix with zero diagonal entries. For simplicity, we only focus on the strictly lower triangular part of $E$, denote it as $tril(E)$, let $p^*$ denote the number of nonzero entries in $tril(E)$. Furthermore, we vectorize $tril(E)$ column-wise by omitting the zeros entries. At the same time, we also record the corresponding row and column indices of nonzero entries by the vectors $\mathbf{r} = (r_1, \ldots, r_{p^*})^T$ and $\mathbf{c} = (c_1, \ldots, c_{p^*})^T$, where the pair $(r_s, c_s)$ is the $s$-th element of the set $\mathcal{J} = \{(j_1, j_2) : e_{j_1 j_2} \neq 0, j_2 = 1, \ldots, p-1, j_2 < j_1 \leq p\}$.

We have completed the estimation of diagonal entries and the identification of zero entries. Next, we will develop a cubic equation-based method to achieve the estimation of off-diagonal nonzero entries.

Before introducing this method, let us concentrate on one pair $(j_1, j_2) \in \mathcal{J}$, i.e., the entry crossed at $j_1$-th row and $j_2$-th column of $\Sigma(\cdot)$ is nonzero. Given $u = u_0$, we notice that the kernel weighted likelihood function of bivariate normal distribution of $y_{ij_1}$ and $y_{ij_2}$ can be expressed as

$$L(u_0, h_4) = \sum_{i=1}^{n} \left\{ \frac{1}{1 - \rho_{j_1 j_2}^2(u_0)} \left[ \frac{y_{ij_1}^2}{\sigma_{j_1 j_1}(u_i)} + \frac{y_{ij_2}^2}{\sigma_{j_2 j_2}(u_i)} - \frac{2\rho_{j_1 j_2}(u_0) y_{ij_1} y_{ij_2}}{\sqrt{\sigma_{j_1 j_1}(u_i)\sigma_{j_2 j_2}(u_i)}} \right] + \log\left[1 - \rho_{j_1 j_2}^2(u_0)\right] \right\} K_{h_4}(u_i - u_0), \tag{4.14}$$

where $\rho_{j_1 j_2}(u_0)$ is the correlation coefficient of $y_{ij_1}$ and $y_{ij_2}$ at $u_0$, $h_4$ is a new bandwidth. Let $\partial L(u_0, h_4)/\partial \rho_{j_1 j_2}(u_0) = 0$, after some simple computations, we

get

$$\rho^3_{j_1j_2}(u_0) - B(u_0)\rho^2_{j_1j_2}(u_0) + [A(u_0) - 1]\rho_{j_1j_2}(u_0) - B(u_0) = 0, \tag{4.15}$$

where

$$A(u_0) = \frac{\sum_{i=1}^n \left[\frac{y^2_{ij_1}}{\sigma_{j_1j_1}(u_i)} + \frac{y^2_{ij_2}}{\sigma_{j_2j_2}(u_i)}\right] K_{h_4}(u_i - u_0)}{\sum_{i=1}^n K_{h_4}(u_i - u_0)},$$
$$B(u_0) = \frac{\sum_{i=1}^n \frac{y_{ij_1}y_{ij_2}}{\sqrt{\sigma_{j_1j_1}(u_i)\sigma_{j_2j_2}(u_i)}} K_{h_4}(u_i - u_0)}{\sum_{i=1}^n K_{h_4}(u_i - u_0)}.$$

Equation (4.15) is a cubic equation of $\rho_{j_1j_2}(u_0)$, Kendall *et al.* (1973) pointed out that there is at least one real root of (4.15) lying in the interval $[-1, 1]$. If there are two or more real roots in the interval $[-1, 1]$, we can use (4.14) to justify which is the maximum likelihood estimator of correlation coefficient. Especially, if $A(u_0) = 2$, then $\rho_{j_1j_2}(u_0) = B(u_0)$. According to the results illustrated in Figure 4.1, we adopt the real root of equation (4.15) as our correlation coefficient estimator at $u = u_0$. For the bandwidth selection, we use the leave-one-out cross validation criterion:

$$CV(h_4) = \sum_{(j_1,j_2)\in\mathcal{J}} \sum_{i=1}^n \left\{ \frac{1}{1 - \hat{\rho}^2_{j_1j_2(-i)}(u_i, h_4)} \left[ \frac{y^2_{ij_1}}{\hat{\sigma}_{j_1j_1}(u_i)} + \frac{y^2_{ij_2}}{\hat{\sigma}_{j_2j_2}(u_i)} - \frac{2\hat{\rho}_{j_1j_2(-i)}(u_i) y_{ij_1} y_{ij_2}}{\sqrt{\hat{\sigma}_{j_1j_1}(u_i)\hat{\sigma}_{j_2j_2}(u_i)}} \right] + \log\left[1 - \hat{\rho}^2_{j_1j_2(-i)}(u_i)\right] \right\}, \tag{4.16}$$

where $\hat{\rho}_{j_1j_2(-i)}(u_i, h_4)$ is the real root of equation (4.15) given $u_i$ and $h_4$ without the $i$-th observation. Once we obtain the bandwidth $\hat{h}_4$ by minimizing equation (4.16), we can solve equation (4.15) by substituting $h_4$ with $\hat{h}_4$ to obtain $\hat{\rho}_{j_1j_2}(u_i, \hat{h}_4)$, hence estimator $\hat{\sigma}_{j_1j_2}(u_i) = \hat{\rho}_{j_1j_2}(u_i, \hat{h}_4)\sqrt{\hat{\sigma}_{j_1j_1}(u_i)\hat{\sigma}_{j_2j_2}(u_i)}$. According to the pair of row and column indices $(r_s, c_s)$ in the set $\mathcal{J}$, we can obtain a strictly lower triangular matrix $\mathbf{L}_i$. Finally, given $u_i, i = 1, \ldots, n$, the estimator $\hat{\Sigma}(u_i) = \mathbf{L}_i + \text{diag}((\hat{\sigma}_{11}(u_i), \ldots, \hat{\sigma}_{pp}(u_i))) + \mathbf{L}_i^T$. If $\hat{\Sigma}(u_i)$ is negative definite, we modify it by subtracting $(\tau(u_i) - c_0)I_p$ where $\tau(u_i)$ is the smallest eigenvalue of $\hat{\Sigma}(u_i)$ and $c_0$ is a small positive constant, say $10^{-4}$.

In this section, we have developed Divide-and-Combine framework for both mean and covariance matrix functions estimation. For the mean function estimation, Fan *et al.* (2001)'s method can efficiently detect the linear EWFs. As for the covariance matrix estimation, we use three separate steps to estimate diagonal entries, identify zero entries and evaluate the off-diagonal nonzero entries respectively. Two estimators of diagonal entries are discussed, and we prefer the estimator (4.9) due to its positiveness. The FDR procedure is used to keep the Type I Error under an appropriate level in zero entries identification. Finally, we

obtain the estimator of off-diagonal nonzero entries by solving the cubic equations of correlation coefficient. To the best of our knowledge, there is no nonparametric theory related to our Divide-and-Combine estimation with solving the cubic equations. Even though the unavailability of theory, the seven scenarios in Section 4.3 show that our Divide-and-Combine method performs better than factorized NCM method in Chapter 3 in terms of Frobenius and spectral norm-based loss.

## 4.3 Numerical Study

In this section, we use seven scenarios to illustrate the performance of our method.

### Scenario 1

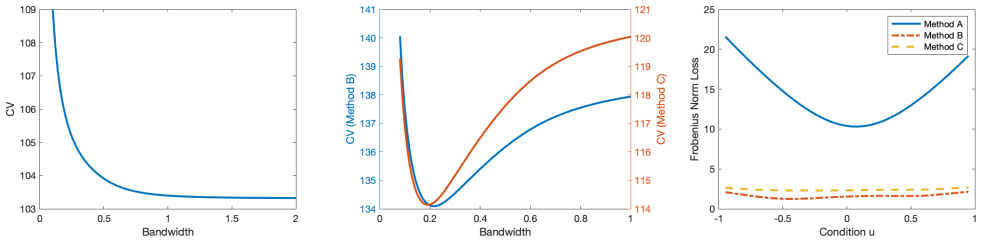
To illustrate the influence of sparsity on optimal bandwidth selection like the pilot study in Section 3.2.4, we design a simple example as well: let the sample size $n$ be 100, the number of variable $p$ be 100. Samples $u_i, i = 1, \ldots, n$ are evenly drawn from the uniform distribution over $[-0.95, 0.95]$. Then, given $u_i$, we define $\Sigma(u_i)$ through its square root matrix $R(u_i) = (r_{kj})_{p\times p}$. For a pre-selected $\theta \in [0, 1]$, we randomly select $p_\theta = \lfloor p_0\theta \rfloor$ entries from the strictly lower triangle part of $R(u_i)$ and assign zeros to them. To keep the symmetry of $R(u_i)$, we reflect these zero entries to the upper triangle part of $R(u_i)$. For the remaining entries, if $(k, j)$-th entry is nonzero, then we set $r_{kj}(u_i) = \exp(0.5 \times u_i \sin(kj)) \sin(\pi u_i)$. Therefore, $\Sigma(u_i) = R(u_i) \times R(u_i)$. Furthermore, we convert covariance matrix to correlation matrix by

$$\text{Corr}(u_i) = [\text{Diag}(\Sigma(u_i))]^{-\frac{1}{2}} \Sigma(u_i) [\text{Diag}(\Sigma(u_i))]^{-\frac{1}{2}}.$$

$\mathbf{y}_i$, $i = 1, \ldots, n$ are random samples from multivariate norm distribution with zero mean and covariance matrix $\text{Corr}(u_i)$. Let $\mathcal{S}_R$ be the sparsity of $R(u_i)$, then $\theta = \frac{p}{p-1}\mathcal{S}_R$. In this example, we take $\mathcal{S}_R = 0.98$, then the sparsity of $\Sigma(u)$ is 0.97.

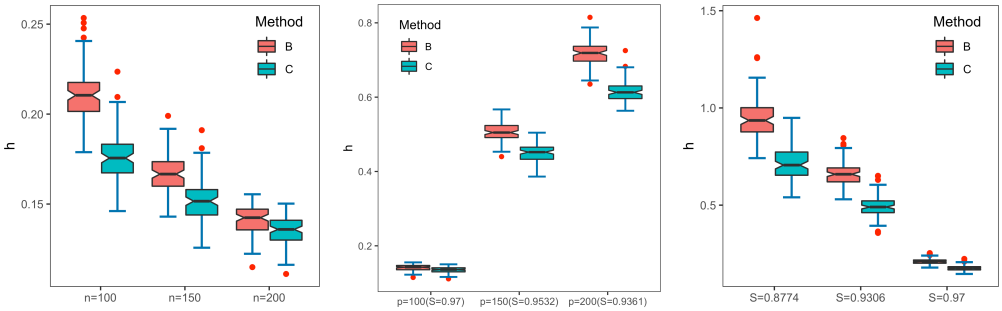
To discuss the effect of zero entries, we simply compare three methods here. Method A uses the local linear smoother to estimate the whole entries of $\Sigma(u)$ supposing the zero entries unknown. Method B uses the Divide-and-Combine approach without zero entries identification step supposing the zero entries are known. Method C implements the Divide-and-Combine approach steps supposing the zero entries unknown.

Figure 4.2(a) shows the value of cross validation objective function against the bandwidth if we use Method A. Figure 4.2(b) shows the same results of Method B and C, the left $y$-axis represents the CV values using Method B, the right




*(a) Method A* *(b) Method B and C* *(c) Loss Error*

*Figure 4.2: The Results of Method A, B and C with $S_R = 0.98 (S_\Sigma = 0.97)$*



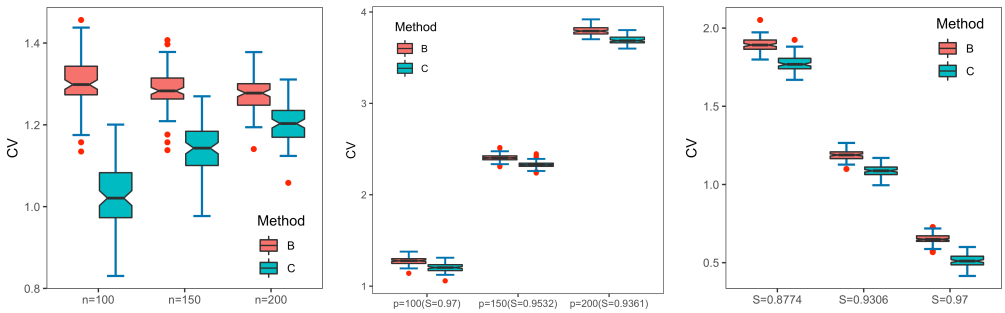

*(a) $p = 100, \mathcal{S}_R = 0.98$* *(b) $n = 200, \mathcal{S}_R = 0.98$* *(c) $n = 100, p = 100$*

*Figure 4.3: Bandwidth Comparison*

$y$-axis is for Method C. The optimal bandwidth in Figure 4.2(b) is finite while in Figure 4.2(a) the optimal bandwidth goes to infinity. Because massive zero entries dominate the convergence of bandwidth, i.e., $h \to \infty$ when one applies the local linear smoother to the entries of the covariance matrix. We can predict that Method B and C should significantly reduce the Frobenius norm-based loss, see the red dot dash line and yellow dash line in Figure 4.2(c).

Sample size $n$, variable dimension $p$ and the sparsity of covariance $\mathcal{S}_\Sigma$ also have effects on the bandwidth selection. To obtain insight into these effects, we design 3 cases: (1) Given $p = 100, \mathcal{S}_R = 0.98$, let the sample size $n$ be 100, 150 and 200 respectively, for each $n$ the procedure is repeated 90 times. The sparsity of covariance $\mathcal{S}_\Sigma$ is 0.97, see Figure 4.3(a); (2) Given $n = 200, \mathcal{S}_R = 0.98$, let $p$ be 100, 150 and 200 respectively, for each $p$ it is also repeated 90 times. The sparsity of covariance $\mathcal{S}_\Sigma$ are 0.97, 0.9532 and 0.9361, see Figure 4.3(b); (3) Given $n = 100, p = 100$, let $\mathcal{S}_R$ are 0.98, 0.96 and 0.94 respectively. The sparsity of covariance $\mathcal{S}_\Sigma$ are 0.97, 0.9306 and 0.8774, each case is repeated 90 times, see Figure 4.3(c).

Figure 4.3(c) clearly shows that the optimal bandwidth increases when the sparsity $\mathcal{S}_\Sigma$ decreases. This result is not surprising because the number of parameters (or the number of nonzero entries) increases if the sparsity decreases. In this circumstance, it needs more information from local neighbours which widens the optimal bandwidth. On the contrary, given the sparsity and $p$, the optimal




*(a)* $p = 100, \mathcal{S}_R = 0.98$ *(b)* $n = 200, \mathcal{S}_R = 0.98$ *(c)* $n = 100, p = 100$

*Figure 4.4: CV Values' Comparison*

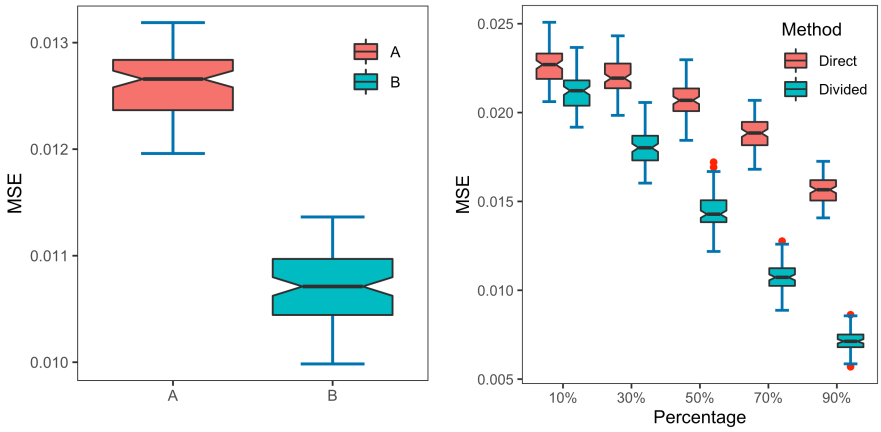
bandwidth becomes shorter if we enlarge the sample size $n$, see Figure 4.3(a). Because, large $n$ provides more local neighbours than small $n$, which indicates the bandwidth go to zero when $n \to \infty$. Lastly, given sparsity and sample size $n$, the optimal bandwidth increases when $p$ increases, see Figure 4.3(b). The number of parameters, i.e., $p^2 \times (1 - \mathcal{S}_\Sigma)$ and sample size $n$ have a great effect on the bandwidth selection. We also illustrate the corresponding CV values of previous examples, see Figure 4.4. We can conclude that the minimum CV value increases with the sample size $n$ and $p$ increasing, see Figures 4.4(a) and 4.4(b); The minimum CV value decreases with the sparsity $\mathcal{S}_\Sigma$ increasing.

## Scenario 2

Our basic idea of Divide-and-Combine framework is to divide the mean function and covariance function into different parts. For mean function, we use Fan *et al.* (2001)'s method to detect the linear and nonlinear EWFs, while for covariance function, we estimate the zero entries and nonzero entries separately. To illustrate the advantage of our method for mean function estimation, we introduce Scenario 2 here, the basic model is $\mathbf{y}_i = \boldsymbol{\mu}(u_i) + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n$, where $\boldsymbol{\mu}(u_i)$ is the mean function with $p$ components and $\boldsymbol{\varepsilon}_i$ represents Gaussian noise. We suppose the mean function components including linear and nonlinear EWFs. To satisfy our demands, without loss of generality, we always assume that the linear EWFs ranking before nonlinear EWFs. The portion of linear EWFs, denoted as $r$, is $10\%, 30\%, 50\%, 70\%, 90\%$ respectively. The linear EWFs are generated from the simple model $\mu_j(u_i) = \beta u_i$, where $\beta$ is the slope which is randomly generated from $U[2, 8]$. For the nonlinear EWFs, following Yuan & Cai (2010) and Chen & Leng (2016), each nonlinear component of $\boldsymbol{\mu}(u)$ is generated independently by

$$\mu_j(u) = \sum_{k=1}^{50} \frac{(-1)^{k+1}}{k^2} Z_j \cos(k\pi u),$$

where $\{Z_j : 1 \leq j \leq p\}$ are independently sampled from the uniform distribution over $[-5, 5]$. The noise $\boldsymbol{\varepsilon}$ is randomly generated from multivariate normal distribution with zero mean and covariance matrix $\Sigma(u)$. We assume that $\Sigma(u) = 0.5 \times \{\sigma_{ij}(u)\}_{1 \leq i,j \leq p}$, where $\sigma_{ij}(u) = \exp(u/2)[\{\phi(u) + 0.1\}I(|i-j| = 1) + \phi(u)I(|i-j| = 2) + I(i = j)]$ and $\phi(u)$ is the standard normal probability density function, the $u_i, i = 1, \ldots, n$ are randomly drawn from uniform distribution $U[-1, 1]$. Finally, we set the sample size $n = 200$, the dimension $p = 150$ and repeat the above procedure 90 times. Figure 4.5(b) displays the box-plots of mean square error for two different mean function estimations. Label Divided in Figure 4.5(b) stands for the method described in Section 4.2.1. Label Direct represents the direct local linear estimation of $\boldsymbol{\mu}(u_i)$ without linear EWFs detection. The MSE of our method consistently decreases deeper than the Direct method




*(a) Two Methods: A and B* *(b) Two Mean Function Estimators*

*Figure 4.5: Comparison*

with the linear portion increasing from 10% to 90%. Apparently, our two-step mean function estimation indeed performs better than Direct method. Throughout this chapter, we let $\text{DAC}_1$ and $\text{DAC}_2$ represent the Divide-and-Combine estimation method with the diagonal entries estimated by (4.5) and (4.9) respectively. Furthermore, we also obtain the Frobenius norm-based loss of covariance matrix estimated by ${}_{\text{st}}\text{NCM}_1$, $\text{DAC}_1$ and $\text{DAC}_2$ respectively, see Table B.1. We can see that the Divided method performs better (but not too much) than the Direct method for both $\text{DAC}_1$ and $\text{DAC}_2$ estimation procedure in terms of Frobenius norm-based loss. However, compared with ${}_{\text{st}}\text{NCM}_1$, the Frobenius norm-based loss of $\text{DAC}_1$ and $\text{DAC}_2$ are significantly smaller than ${}_{\text{st}}\text{NCM}_1$'s, and $\text{DAC}_2$ is a slightly better than $\text{DAC}_1$. The most contribution comes from our Divide-and-Combine estimation of nonparametric covariance matrix, see the loss comparison among ${}_{\text{st}}\text{NCM}_1$, $\text{DAC}_1$ and $\text{DAC}_2$ in Table B.1. These conclusions coincide our logical thinking of Divide-and-Combine both in mean and covariance matrix function estimation.

For simplicity, in the next five scenarios, we let the mean function compo-

nents be nonlinear functions, but we still adopt Divide-and-Combine method to estimate the mean function.

## Scenario 3

Following Yuan & Cai (2010) and Chen & Leng (2016), the component $\boldsymbol{\mu}(u) = (\mu_1(u), \ldots, \mu_p(u))^T$ is generated independently as follows:

$$\mu_j(u) = \sum_{k=1}^{50} \frac{(-1)^{k+1}}{k^2} Z_j \cos(k\pi u), \quad j = 1, \ldots, p, \tag{4.17}$$

where $\{Z_j : 1 \leq j \leq p\}$ are independently sampled from the uniform distribution over $[-5, 5]$. We assume that $\Sigma(u) = 0.5 \times \{\sigma_{ij}(u)\}_{1 \leq i,j \leq p}$, where $\sigma_{ij}(u) = \exp(u/2)[\{\phi(u) + 0.1\}I(|i-j| = 1) + \phi(u)I(|i-j| = 2) + I(i = j)]$ and $\phi(u)$ is the standard normal probability density function.

## Scenario 4

This scenario is originated from Zhang & Liu (2015) in simulation of the source signal in Beamforming method. In this circumstance, given $u$, we simulate $\boldsymbol{\mu}(u) = (\mu_1(u), \ldots, \mu_p(u))^T$ using the following model

$$\mu_j(u) = Z_j \exp\left(\frac{(u - \tau_j)^2}{4}\right) \sin(2\pi(u - \tau_j)), \qquad j = 1, \ldots, p,$$

where $Z_j, j = 1, \ldots, p$ are independently sampled from uniform distribution $U(-5, 5)$, $\tau = (\tau_1, \ldots, \tau_p)$ is a row vector of $p$ evenly spaced points between -1 and 1. Let $\Sigma(u) = \{\sigma_{ij}(u)\}_{1 \leq i,j \leq p}$, where $\sigma_{ij}(u) = 0.5 \times \exp(u/2)\phi(u)^{|i-j|}$.

## Scenario 5

This scenario is similar to Scenario 3 except the covariance $\Sigma(u) = 0.1 \times A^T(u)A(u)$, where the $(i, j)$-th entry of $A(u)$ is defined as:

$$\begin{aligned} a_{ij}(u) \quad = \quad & \exp\left(\frac{u \sin(ij)}{2}\right) \Big\{ [\sin(\pi u) + 0.1]\, I(|i-j| = 1) \\ & + \sin(\pi u) I(|i-j| = 2) + I(i = j) \Big\}, \end{aligned}$$

the mean function is the same as equation (4.17).

For each combination of $(n, p)$ with $n = 100, 200, 500$ and $p = 50, 100, 150, 300$, we repeat the experiment 90 times, generating 90 datasets of $(\mathbf{y}_i, u_i), 1 \leq i \leq n$. Each dataset is obtained in two steps. In step 1, we randomly draw $u_i, i = 1, \ldots, n$

from the uniform distribution $U(-1,1)$. In step 2, for each given $u_i$, we draw $\mathbf{y}_i$ from the covariance model $\mathbf{y}_i = \boldsymbol{\mu}(u_i) + \Sigma(u_i)^{1/2}\varepsilon_i$, where $\varepsilon_i, i = 1,\ldots,n$ are iteratively drawn from the vector VAR(1) model

$$\varepsilon_0 = \xi_0, \quad \varepsilon_i = \rho\varepsilon_{i-1} + \xi_i, \quad i = 1,\ldots,n,$$

with $0 \le \rho \le 1$ and $\xi_k, k = 0,1,\ldots$ are independently sampled from the standard $p$-dimensional normal $N(0, I_p)$. We consider $\rho = 0, 0.3, 0.8$ for the Scenarios 3–5.

## Scenario 6

The mean function is same as (4.17) in Scenario 3. The sparse covariance is generated by $\Sigma(u) = 0.1 \times R(u) \times R(u)$, where $R(u)$ is a sparse symmetry matrix and is generated by the method in Scenario 1. For specificity, we let $\mathcal{S}_R$ be 0.96, $n = 100, 200, 500$ and $p = 50, 100, 150, 300$. For each combination $(n, p)$, we repeat it 90 times in this simulation as well.

For simplicity, we compare three estimators ${}_{\text{st}}\text{NCM}_1$, $\text{DAC}_1$ and $\text{DAC}_2$ here. We use ${}_{\text{st}}\text{NCM}_1$ to represent our $Q_1$-based nonparametric covariance estimation using the local constant smoother in Chapter 3. We employ the spectral and Frobenius norm-based integrated root-squared error (IRSE) as the criteria. Specifically, we generate $u_i^*, i = 1,\ldots,25$ evenly from interval $[-0.9, 0.9]$ for Scenarios 3–6. Then for each $u_i^*$, the spectral and Frobenius norm-based IRSE are:

$$\text{IRSE}_F(u_i^*) = \frac{1}{K_0}\sum_{i=1}^{K_0}\left\|\hat{\Sigma}(u_i^*) - \Sigma(u_i^*)\right\|_F,$$
$$\text{IRSE}_S(u_i^*) = \frac{1}{K_0}\sum_{i=1}^{K_0}\left\|\hat{\Sigma}(u_i^*) - \Sigma(u_i^*)\right\|,$$

where $K_0 = 25$ and $\hat{\Sigma}(u_i^*)$ is the estimator of underlying covariance matrix $\Sigma(u_i^*)$.

In the step of estimating the off-diagonal zeros entries, the results of FDR procedure depend on the significant level $\alpha$. To test the effect of $\alpha$, we let $\alpha = 0.01, 0.02, \ldots, 0.1$ respectively. We apply ${}_{\text{st}}\text{NCM}_1$, $\text{DAC}_1$ and $\text{DAC}_2$ to the 90 datasets for each combination of $(n, p, \rho)$. Their Frobenius and spectral norm-based IRSE are also calculated at the same time for different $\alpha$. Furthermore, we also have a great interest in the accuracy of zero entries identification. Let $p_1$ ($p_2$) be the number of nonzero (zero) entries in $\Sigma(u)$. For any estimator $\hat{\Sigma}(u)$ of $\Sigma(u)$, let $n_{11}$ be the number of true discoveries of nonzero entries in $\Sigma(u)$ by $\hat{\Sigma}(u)$. Similarly, let $n_{22}$ denote the number of true discoveries of zero entries in $\Sigma(u)$ by $\hat{\Sigma}(u)$. Let

SEN, SPE and ACC denote sensitivity, specificity and accuracy in the above testing,

$$\text{SEN} = \frac{n_{11}}{p_1}, \quad \text{SPE} = \frac{n_{22}}{p_2}, \quad \text{ACC} = \frac{n_{11} + n_{22}}{p_1 + p_2}.$$

The significant level $\alpha$ is selected automatically by the minimum IRSE.

Furthermore, to evaluate the performance of our method when the locations of zero entries vary with $u$, we design the Scenario 7.

## Scenario 7

The idea of this case originates from the random graph model (Zhou *et al.*, 2010) with a slight modification. We assume the nonzero entries' locations will change at the points: -0.9, -0.7, -0.5, -0.3, -0.1, 0, 0.1, 0.3, 0.5, 0.7, 0.9. Firstly, we generate the full matrix $R(u)$ with

$$r_{ij}(u) = \exp(0.5u\sin(ij))(1 - u^2) + 0.1.$$

Let $u_k^* = -1.1 + 0.2k, k = 1, \ldots, 10$, $R_L(u)$ be strictly lower triangular matrix of $R(u)$. Denote the set of $R_L(u)$ indices as $S$, namely, $S = \{(i, j) : 1 \leq i < j \leq p\}$. When $k = 1$, we randomly choose $p$ elements from $S$, denote these $p$ elements as a new subset $S_1$, and let $S_0 = S \backslash S_1$. If $(i, j) \in S_0$, then let the $(i, j)$-th entry of $R_L(-0.9)$ be zero. Now $R_L(-0.9)$ is sparse and has $p$ nonzero entries. Next, we will discuss how to generate $R_L(u)$, $\forall\ u \in [-1, 1]$ through the following steps:

1. $\forall\ u \in [-1, -0.9)$, $R_L(u)$ shares the same nonzero locations as $R_L(-0.9)$;
2. When $k = 1$, for any $u \in (u_k^*, u_{k+1}^*]$, we randomly choose $p/10$ elements from current subset $S_1$, and let them decrease to zero by $(u_{k+1}^* - u) \times r_{ij}(u_k^*)/5$; on the other side, we randomly choose $p/10$ entries from the subset $S_0$ as well and let them increase to $r_{ij}(u_{k+1}^*)$ by $(u - u_k^*) \times r_{ij}(u_{k+1}^*)/5$. At $u_{k+1}^*$, update both $S_0$ and $S_1$;
3. Repeat the step 2 until $k = 9$;
4. For any $u \in (0.9, 1]$, $R_L(u)$ shares the same nonzero locations as $R_L(0.9)$.

After obtaining the sparse lower triangular matrix $R_L(u)$, we can easily get matrix $R(u)$. Note that, except the change-points $u_k^*$, there always are $3p + p/5$ nonzero entries in $R(u)$. Lastly, we let $\Sigma(u) = 0.1 \times R(u) \times R(u)$.

In this Scenario, we let $n = 100, p = 100, 150, 300$ and repeat each parameter setting 90 times. We compare the performance of the Frobenius and spectral norm-based IRSE, SEN, SPE and ACC in Tables 4.10 $\sim$ 4.18.

## Result

The average (standard error in %) of both Frobenius and spectral norm-based IRSE, SEN, SPE and ACC for Scenarios 1–7 are displayed in Tables 4.1 $\sim$ 4.18 and Tables B.1 $\sim$ B.22. Tables 4.1 $\sim$ 4.6 and Tables B.2 $\sim$ B.4 display the 12 combinations of pair $(n, p, \rho)$ in Scenario 3. Tables B.5 $\sim$ B.13 display the 12 combinations of pair $(n, p, \rho)$ in Scenario 4 and Tables B.14 $\sim$ B.22 display the 12 combinations of pair $(n, p, \rho)$ in Scenario 5. Tables 4.7 $\sim$ 4.9 display the 12 combinations of pair $(n, p, \rho)$ in Scenario 6 with $\mathcal{S}_R = 0.96$. The results can be summarized as follows:

- Both $\text{DAC}_1$ and $\text{DAC}_2$ perform consistently better than the method ${}_{\text{st}}\text{NCM}_1$ for each parameter sets $(n, p, \rho)$ of these seven simulations in terms of both Frobenius and spectral norm-based IRSE, see Tables 4.1 $\sim$ 4.18 and Tables B.1 $\sim$ B.22 in Appendix B.
- On average, the spectral and Frobenius norm-based IRSE of each parameter sets $(n, p, \rho)$ increase with the dimension $p$ and the degree of serial correlation $\rho$ but decrease with sample size $n$, see Tables 4.1 $\sim$ 4.3.
- The sparsity of $\Sigma(u)$ also has an effect on the performance of the spectral and Frobenius norm-based IRSE when one compares ${}_{\text{st}}\text{NCM}_1$ with $\text{DAC}_1$ and $\text{DAC}_2$. For instance, in Table 4.7, the sparsity of covariance matrix varies from 0.944 to 0.6413. We can see that both Frobenius and spectral norm-based IRSE increase when the sparsity decreases. Furthermore, Tables 4.7 and 4.9 show that the Frobenius and spectral norm-based IRSE of $\text{DAC}_1$ and $\text{DAC}_2$ are consistently better than those in ${}_{\text{st}}\text{NCM}_1$ method. On average, compared with ${}_{\text{st}}\text{NCM}_1$, the improvements of $\text{DAC}_1$ and $\text{DAC}_2$ are 20.06%, 33.18% respectively in Table 4.7.
- We also compare three criteria: SEN, SPE and ACC for each scenario. For example, in Table 4.4, the SPEs of $\text{DAC}_1$ and $\text{DAC}_2$ are almost equivalent to ${}_{\text{st}}\text{NCM}_1$, however, the SENs of $\text{DAC}_1$ and $\text{DAC}_2$ are significantly larger than the SENs of ${}_{\text{st}}\text{NCM}_1$ method. This is not surprising because we implement the zero entries detection before bandwidth selection step rather than after the bandwidth selection step as ${}_{\text{st}}\text{NCM}_1$ method. The similar conclusion can be made for dependent samples when $\rho = 0.3$ and 0.8, see Tables 4.5 and 4.6.
- In Scenario 6, we notice that the SEN of ${}_{\text{st}}\text{NCM}_1$ is quite small compared with $\text{DAC}_1$ and $\text{DAC}_2$, for example, when $n = 200, p = 100$ in Table 4.8, the value of SEN in ${}_{\text{st}}\text{NCM}_1$ is just 0.2097, while the value of $\text{DAC}_1$ is 0.8076. This means ${}_{\text{st}}\text{NCM}_1$ method can not identify the nonzero entries efficiently, this can also be confirmed by the ACC column in the same table.
- In Scenario 7, the global sparsity of underlying covariance matrix for $p =$

100, 150 and 300 are 0.8542, 0.8982 and 0.9471 respectively. The nonzero entries vary with the condition $u$ described in Scenario 7. Tables 4.10 $\sim$ 4.18 summarize the Frobenius and spectral norm-based IRSE, SEN, SPE and ACC at each changing point $u_k^* = -1.1 + 0.2k, k = 1, \ldots, 10$. We can conclude that $\text{DAC}_1$ and $\text{DAC}_2$ perform uniformly better than ${}_{\text{st}}\text{NCM}_1$ method even under location-varying nonzero entries circumstances.

*Table 4.1: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 3*

| $n$ | $p$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0$ | | | |
| 100 | 50 | 0.3412(4.67) | 0.2160(1.52) | 0.06 | 0.2132(1.39) | 0.05 |
| | 100 | 0.3280(1.08) | 0.2324(1.47) | 0.06 | 0.2290(1.18) | 0.05 |
| | 150 | 0.3471(1.02) | 0.2453(1.92) | 0.06 | 0.2395(0.99) | 0.05 |
| | 300 | 0.3672(0.55) | 0.2606(1.00) | 0.06 | 0.2564(0.59) | 0.06 |
| 200 | 50 | 0.2197(3.84) | 0.1325(1.11) | 0.01 | 0.1330(1.03) | 0.01 |
| | 100 | 0.2247(1.61) | 0.1399(0.75) | 0.02 | 0.1395(0.82) | 0.01 |
| | 150 | 0.2273(1.18) | 0.1418(0.70) | 0.02 | 0.1418(0.61) | 0.02 |
| | 300 | 0.2423(0.70) | 0.1484(1.01) | 0.02 | 0.1470(0.43) | 0.02 |
| 500 | 50 | 0.1070(0.80) | 0.0849(0.56) | 0.01 | 0.0863(0.52) | 0.01 |
| | 100 | 0.1092(0.69) | 0.0856(0.38) | 0.01 | 0.0873(0.37) | 0.01 |
| | 150 | 0.1109(0.55) | 0.0858(0.33) | 0.01 | 0.0873(0.33) | 0.01 |
| | 300 | 0.1198(0.49) | 0.0864(0.25) | 0.01 | 0.0877(0.24) | 0.01 |

## 4.4 Real Data Analysis

We apply our Divide-and-Combine estimation method to the stock prices. The dataset contains 421 stocks daily closed price and volume from 01/01/2005 to 31/12/2010. For each week, we calculate the volume weighted average price (VWAP) as the price of this week. We take logarithm of the ratio of VWAP at week $t$ and $t-1$. In this real data analysis, according to economic depression period, we divide the whole data into three periods, namely, before-financial-crisis period from 01/01/2005 to 31/12/2006, in-financial-crisis period from 01/01/2007 to 31/12/2008 and after-financial-crisis period from 01/01/2009 to 31/12/2010. We also collect the daily closed price and volume of index $S\&P$ 500.

We apply the proposed Divide-and-Combine estimation method to the data from each time-period, obtaining the corresponding estimators of mean $\boldsymbol{\mu}(u)$ and covariance matrix $\Sigma(u)$. Here, the diagonal estimators of $\Sigma(u)$ show the

*Table 4.2: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 3 (continued)*

| $n$ | $p$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.3$ | | | |
| 100 | 50 | 0.2908(2.02) | 0.2462(1.60) | 0.04 | 0.2464(1.55) | 0.04 |
| | 100 | 0.3049(1.36) | 0.2661(1.14) | 0.04 | 0.2633(1.13) | 0.05 |
| | 150 | 0.3143(1.00) | 0.2719(1.31) | 0.04 | 0.2699(0.87) | 0.05 |
| | 300 | 0.3346(0.63) | 0.2883(0.67) | 0.04 | 0.2864(0.59) | 0.04 |
| 200 | 50 | 0.2440(2.08) | 0.1610(0.95) | 0.01 | 0.1618(0.98) | 0.01 |
| | 100 | 0.2452(1.29) | 0.1669(0.96) | 0.01 | 0.1676(0.94) | 0.01 |
| | 150 | 0.2469(1.11) | 0.1719(0.69) | 0.01 | 0.1717(0.63) | 0.01 |
| | 300 | 0.2561(0.78) | 0.1820(0.53) | 0.01 | 0.1816(0.50) | 0.01 |
| 500 | 50 | 0.2284(1.28) | 0.1121(0.72) | 0.01 | 0.1163(0.76) | 0.01 |
| | 100 | 0.2352(0.84) | 0.1141(0.59) | 0.01 | 0.1181(0.61) | 0.01 |
| | 150 | 0.2332(0.74) | 0.1140(0.51) | 0.01 | 0.1181(0.53) | 0.01 |
| | 300 | 0.2340(0.49) | 0.1122(0.30) | 0.01 | 0.1157(0.31) | 0.01 |

*Table 4.3: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 3 (continued)*

| $n$ | $p$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.8$ | | | |
| 100 | 50 | 0.5888(3.20) | 0.4895(0.80) | 0.06 | 0.4884(0.83) | 0.05 |
| | 100 | 0.5928(2.68) | 0.5022(0.60) | 0.06 | 0.5014(0.65) | 0.04 |
| | 150 | 0.5901(2.39) | 0.5028(0.52) | 0.03 | 0.5026(0.39) | 0.03 |
| | 300 | 0.5980(2.35) | 0.5069(0.40) | 0.02 | 0.5066(0.32) | 0.02 |
| 200 | 50 | 0.5903(1.50) | 0.4637(0.77) | 0.01 | 0.4568(0.61) | 0.01 |
| | 100 | 0.5895(1.32) | 0.4737(0.58) | 0.01 | 0.4654(0.48) | 0.01 |
| | 150 | 0.5936(1.29) | 0.4746(0.46) | 0.01 | 0.4671(0.34) | 0.01 |
| | 300 | 0.5936(1.09) | 0.4812(0.34) | 0.01 | 0.4747(0.25) | 0.01 |
| 500 | 50 | 0.5887(0.93) | 0.4725(0.68) | 0.01 | 0.4679(0.53) | 0.01 |
| | 100 | 0.5920(0.34) | 0.4769(0.52) | 0.01 | 0.4696(0.26) | 0.01 |
| | 150 | 0.5923(0.49) | 0.4772(0.36) | 0.01 | 0.4704(0.23) | 0.01 |
| | 300 | 0.5931(0.19) | 0.4753(0.31) | 0.01 | 0.4676(0.14) | 0.01 |

*Table 4.4: The Average SEN, SPE and ACC for Scenario 3 ($\rho = 0$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.8377 | 0.9154 | 0.9168 | 0.8058 | 0.9871 | 0.9892 | 0.8089 | 0.9801 | 0.9822 | 0.06 | 0.05 |
| | 100 | 0.6487 | 0.9094 | 0.9010 | 0.9934 | 0.9896 | 0.9925 | 0.9764 | 0.9856 | 0.9880 | 0.06 | 0.05 |
| | 150 | 0.5845 | 0.8850 | 0.8705 | 0.9964 | 0.9929 | 0.9957 | 0.9828 | 0.9894 | 0.9915 | 0.06 | 0.05 |
| | 300 | 0.4159 | 0.8380 | 0.8279 | 0.9997 | 0.9969 | 0.9979 | 0.9900 | 0.9943 | 0.9950 | 0.06 | 0.06 |
| 200 | 50 | 0.9422 | 0.9856 | 0.9867 | 0.9307 | 0.9981 | 0.9982 | 0.9319 | 0.9969 | 0.9971 | 0.01 | 0.01 |
| | 100 | 0.8739 | 0.9839 | 0.9893 | 0.9924 | 0.9979 | 0.9968 | 0.9866 | 0.9972 | 0.9964 | 0.02 | 0.01 |
| | 150 | 0.8419 | 0.9797 | 0.9818 | 0.9973 | 0.9986 | 0.9987 | 0.9921 | 0.9980 | 0.9981 | 0.02 | 0.02 |
| | 300 | 0.8022 | 0.9806 | 0.9793 | 0.9990 | 0.9985 | 0.9989 | 0.9957 | 0.9982 | 0.9986 | 0.02 | 0.02 |
| 500 | 50 | 0.9949 | 1.0000 | 1.0000 | 0.9974 | 0.9982 | 0.9981 | 0.9971 | 0.9983 | 0.9983 | 0.01 | 0.01 |
| | 100 | 0.9918 | 1.0000 | 1.0000 | 0.9988 | 0.9989 | 0.9990 | 0.9984 | 0.9990 | 0.9990 | 0.01 | 0.01 |
| | 150 | 0.9858 | 1.0000 | 1.0000 | 0.9988 | 0.9993 | 0.9993 | 0.9984 | 0.9994 | 0.9994 | 0.01 | 0.01 |
| | 300 | 0.9570 | 1.0000 | 1.0000 | 0.9998 | 0.9997 | 0.9997 | 0.9990 | 0.9997 | 0.9997 | 0.01 | 0.01 |

*Table 4.5: The Average SEN, SPE and ACC for Scenario 3 ($\rho = 0.3$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.8462 | 0.9006 | 0.9081 | 0.9854 | 0.9841 | 0.9841 | 0.9718 | 0.9759 | 0.9767 | 0.04 | 0.04 |
| | 100 | 0.7700 | 0.8633 | 0.8755 | 0.9962 | 0.9916 | 0.9914 | 0.9850 | 0.9852 | 0.9857 | 0.04 | 0.05 |
| | 150 | 0.7309 | 0.8831 | 0.8726 | 0.9979 | 0.9898 | 0.9927 | 0.9890 | 0.9863 | 0.9887 | 0.04 | 0.05 |
| | 300 | 0.6229 | 0.8194 | 0.8299 | 0.9993 | 0.9960 | 0.9959 | 0.9931 | 0.9930 | 0.9932 | 0.04 | 0.04 |
| 200 | 50 | 0.9730 | 0.9802 | 0.9827 | 0.9959 | 0.9958 | 0.9957 | 0.9937 | 0.9943 | 0.9945 | 0.01 | 0.01 |
| | 100 | 0.9602 | 0.9827 | 0.9841 | 0.9982 | 0.9958 | 0.9957 | 0.9963 | 0.9951 | 0.9951 | 0.01 | 0.01 |
| | 150 | 0.9487 | 0.9762 | 0.9784 | 0.9988 | 0.9970 | 0.9970 | 0.9972 | 0.9963 | 0.9964 | 0.01 | 0.01 |
| | 300 | 0.9233 | 0.9717 | 0.9672 | 0.9995 | 0.9975 | 0.9983 | 0.9982 | 0.9971 | 0.9978 | 0.01 | 0.01 |
| 500 | 50 | 0.9998 | 1.0000 | 1.0000 | 0.9997 | 0.9956 | 0.9957 | 0.9997 | 0.9961 | 0.9961 | 0.01 | 0.01 |
| | 100 | 0.9995 | 1.0000 | 1.0000 | 0.9999 | 0.9978 | 0.9978 | 0.9999 | 0.9979 | 0.9979 | 0.01 | 0.01 |
| | 150 | 0.9991 | 1.0000 | 1.0000 | 0.9999 | 0.9984 | 0.9984 | 0.9999 | 0.9984 | 0.9984 | 0.01 | 0.01 |
| | 300 | 0.9988 | 1.0000 | 1.0000 | 1.0000 | 0.9991 | 0.9991 | 1.0000 | 0.9991 | 0.9991 | 0.01 | 0.01 |

*Table 4.6: The Average SEN, SPE and ACC for Scenario 3 ($\rho = 0.8$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.6549 | 0.8391 | 0.8424 | 0.9386 | 0.9652 | 0.9670 | 0.9109 | 0.9529 | 0.9549 | 0.06 | 0.05 |
| | 100 | 0.4841 | 0.8062 | 0.8095 | 0.9906 | 0.9784 | 0.9791 | 0.9656 | 0.9698 | 0.9707 | 0.06 | 0.04 |
| | 150 | 0.4208 | 0.7957 | 0.7974 | 0.9981 | 0.9816 | 0.9830 | 0.9790 | 0.9754 | 0.9768 | 0.03 | 0.03 |
| | 300 | 0.3087 | 0.6951 | 0.7255 | 0.9997 | 0.9934 | 0.9915 | 0.9882 | 0.9884 | 0.9871 | 0.02 | 0.02 |
| 200 | 50 | 0.8572 | 0.9511 | 0.9574 | 0.9666 | 0.9822 | 0.9794 | 0.9559 | 0.9792 | 0.9773 | 0.01 | 0.01 |
| | 100 | 0.7487 | 0.9365 | 0.9475 | 0.9933 | 0.9891 | 0.9861 | 0.9812 | 0.9865 | 0.9842 | 0.01 | 0.01 |
| | 150 | 0.6861 | 0.9279 | 0.9383 | 0.9957 | 0.9913 | 0.9887 | 0.9854 | 0.9892 | 0.9870 | 0.01 | 0.01 |
| | 300 | 0.5665 | 0.9087 | 0.9226 | 0.9988 | 0.9941 | 0.9921 | 0.9917 | 0.9927 | 0.9910 | 0.01 | 0.01 |
| 500 | 50 | 0.9940 | 0.9995 | 0.9999 | 0.9979 | 0.9833 | 0.9801 | 0.9975 | 0.9849 | 0.9820 | 0.01 | 0.01 |
| | 100 | 0.9905 | 0.9993 | 0.9996 | 0.9993 | 0.9895 | 0.9870 | 0.9988 | 0.9900 | 0.9876 | 0.01 | 0.01 |
| | 150 | 0.9890 | 0.9994 | 0.9996 | 0.9994 | 0.9915 | 0.9895 | 0.9991 | 0.9917 | 0.9898 | 0.01 | 0.01 |
| | 300 | 0.9839 | 0.9992 | 0.9995 | 0.9997 | 0.9938 | 0.9919 | 0.9995 | 0.9939 | 0.9920 | 0.01 | 0.01 |

*Table 4.7: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 6 with $\mathcal{S}_R = 0.96$*

| $n$ | $p$ | $\mathcal{S}_\Sigma$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0$ | | | |
| 100 | 50 | 0.9440 | 0.1295(0.58) | 0.0829(0.46) | 0.01 | 0.0814(0.47) | 0.01 |
| | 100 | 0.8828 | 0.3169(0.86) | 0.2261(0.69) | 0.05 | 0.2150(0.50) | 0.09 |
| | 150 | 0.8149 | 0.5118(1.24) | 0.4078(0.68) | 0.06 | 0.3828(0.54) | 0.10 |
| | 300 | 0.6431 | 1.0284(1.55) | 0.9075(0.72) | 0.01 | 0.9034(0.30) | 0.04 |
| 200 | 50 | 0.9424 | 0.0939(0.37) | 0.0663(0.94) | 0.01 | 0.0617(0.38) | 0.01 |
| | 100 | 0.8782 | 0.2628(0.48) | 0.1747(1.51) | 0.09 | 0.1639(0.49) | 0.09 |
| | 150 | 0.8152 | 0.4325(0.70) | 0.3157(2.52) | 0.10 | 0.2947(0.49) | 0.10 |
| | 300 | 0.6418 | 0.9734(1.33) | 0.8263(2.23) | 0.10 | 0.7842(0.76) | 0.10 |
| 500 | 50 | 0.9432 | 0.0745(0.22) | 0.0847(1.47) | 0.01 | 0.0477(0.24) | 0.01 |
| | 100 | 0.8804 | 0.2269(0.17) | 0.1944(3.19) | 0.01 | 0.1103(0.30) | 0.03 |
| | 150 | 0.8186 | 0.3854(0.11) | 0.3040(7.37) | 0.10 | 0.2021(0.39) | 0.07 |
| | 300 | 0.6413 | 0.8817(0.21) | 0.7289(8.39) | 0.10 | 0.5868(0.68) | 0.10 |

volatility of individual returns while correlation coefficient matrix estimator of $\Sigma(u)$ captures cross-sectional relationships in these returns. As the number of observations in this real dataset is 312, for each $S\&P$ 500 index in each period, we obtain the corresponding covariance matrix estimators of $\Sigma(u)$.

To analyse the structure and difference between these three periods, we employ four basic concepts of Graphic Model: Edge Density, Vertex Strength, Clustering Coefficient and Centrality, see the review in Section 2.4.3. Figure 4.6 shows these four terminologies for each period and makes comparison over three periods. For simplicity, Periods 1, 2 and 3 represent the before-financial-crisis period, in-financial-crisis period and after-financial-crisis period respectively. The number of observations in each period are not equal in this real data analysis. Hence, for each terminology, we use the Dwass-Steel-Crichtlow-Fligner pairwise ranking nonparametric method (Douglas & Michael, 1991) to compare the period difference. We employ the Bonferroni method (Dunn, 1961) to correct the $p$-value. Furthermore, we also implement the pairwise comparison for each graphical terminology. The comparison of Centrality Index, Clustering Coefficient and Vertex Strength are significantly different at the level 0.001. The Edge Density difference between Period 1 and 2 is not significant at level 0.001. However, they are both significantly different compared with Period 2. And the $p$-value for each comparison is also less than 0.001. That means the graphic structure or connection changed significantly before, during and after financial-crisis. Centrality, Edge Density and Vertex Strength have the same tendency of mean and median. We

*Table 4.8: The Average SEN, SPE and ACC for Scenario 6 with $\mathcal{S}_R = 0.96$*

| $n$ | $p$ | $\mathcal{S}_\Sigma$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.9440 | 0.7332 | 0.9043 | 0.9437 | 0.7819 | 0.9960 | 0.9874 | 0.7792 | 0.9908 | 0.9850 | 0.01 | 0.01 |
| | 100 | 0.8828 | 0.2637 | 0.6110 | 0.6177 | 0.8407 | 0.9618 | 0.9617 | 0.7731 | 0.9206 | 0.9214 | 0.05 | 0.09 |
| | 150 | 0.8149 | 0.1982 | 0.3544 | 0.3597 | 0.8421 | 0.9667 | 0.9669 | 0.7229 | 0.8534 | 0.8545 | 0.06 | 0.10 |
| | 300 | 0.6431 | 0.1278 | 0.0935 | 0.0904 | 0.8812 | 0.9822 | 0.9839 | 0.6123 | 0.6650 | 0.6650 | 0.01 | 0.04 |
| 200 | 50 | 0.9424 | 0.7229 | 0.9606 | 0.9878 | 0.8498 | 0.9747 | 0.9807 | 0.8425 | 0.9739 | 0.9812 | 0.01 | 0.01 |
| | 100 | 0.8782 | 0.2097 | 0.8076 | 0.8108 | 0.9178 | 0.9551 | 0.9617 | 0.8316 | 0.9372 | 0.9433 | 0.09 | 0.09 |
| | 150 | 0.8152 | 0.1378 | 0.5951 | 0.5979 | 0.9019 | 0.9556 | 0.9560 | 0.7607 | 0.8890 | 0.8899 | 0.10 | 0.10 |
| | 300 | 0.6418 | 0.0912 | 0.2311 | 0.2275 | 0.9178 | 0.9695 | 0.9722 | 0.6217 | 0.7050 | 0.7055 | 0.10 | 0.10 |
| 500 | 50 | 0.9432 | 0.7586 | 0.9726 | 0.9997 | 0.9080 | 0.6327 | 0.9947 | 0.8995 | 0.6520 | 0.9950 | 0.01 | 0.01 |
| | 100 | 0.8804 | 0.2878 | 0.8430 | 0.9664 | 0.8886 | 0.8206 | 0.9625 | 0.8167 | 0.8233 | 0.9630 | 0.01 | 0.03 |
| | 150 | 0.8186 | 0.1544 | 0.8503 | 0.8937 | 0.8884 | 0.6556 | 0.9359 | 0.7552 | 0.6909 | 0.9283 | 0.10 | 0.07 |
| | 300 | 0.6413 | 0.0968 | 0.5668 | 0.5453 | 0.9121 | 0.7421 | 0.9377 | 0.6196 | 0.6792 | 0.7969 | 0.10 | 0.10 |

*Table 4.9: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 6 with $\mathcal{S}_R = 0.96$*

| $n$ | $p$ | $\mathcal{S}_\Sigma$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|---|
| | | | $\rho = 0$ | | | | |
| 100 | 50 | 0.9440 | 0.1295(0.58) | 0.0829(0.46) | 0.01 | 0.0814(0.47) | 0.01 |
| | 100 | 0.8828 | 0.3169(0.86) | 0.2261(0.69) | 0.05 | 0.2150(0.50) | 0.09 |
| | 150 | 0.8149 | 0.5118(1.24) | 0.4078(0.68) | 0.06 | 0.3828(0.54) | 0.10 |
| | 300 | 0.6431 | 1.0284(1.55) | 0.9075(0.72) | 0.01 | 0.9034(0.30) | 0.04 |
| 200 | 50 | 0.9424 | 0.0939(0.37) | 0.0663(0.94) | 0.01 | 0.0617(0.38) | 0.01 |
| | 100 | 0.8782 | 0.2628(0.48) | 0.1747(1.51) | 0.09 | 0.1639(0.49) | 0.09 |
| | 150 | 0.8152 | 0.4325(0.70) | 0.3157(2.52) | 0.10 | 0.2947(0.49) | 0.10 |
| | 300 | 0.6418 | 0.9734(1.33) | 0.8263(2.23) | 0.10 | 0.7842(0.76) | 0.10 |
| 500 | 50 | 0.9432 | 0.0745(0.22) | 0.0847(1.47) | 0.01 | 0.0477(0.24) | 0.01 |
| | 100 | 0.8804 | 0.2269(0.17) | 0.1944(3.19) | 0.01 | 0.1103(0.30) | 0.03 |
| | 150 | 0.8186 | 0.3854(0.11) | 0.3040(7.37) | 0.10 | 0.2021(0.39) | 0.07 |
| | 300 | 0.6413 | 0.8817(0.21) | 0.7289(8.39) | 0.10 | 0.5868(0.68) | 0.10 |

*Table 4.10: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 7 with $p = 100$*

| $u_0$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|
| -0.9 | 0.4932(4.18) | 0.0358(0.32) | 0.01 | 0.0305(0.08) | 0.05 |
| -0.7 | 0.1807(1.32) | 0.1230(0.35) | 0.05 | 0.1269(0.42) | 0.10 |
| -0.5 | 0.3277(2.70) | 0.2194(0.58) | 0.09 | 0.2253(0.66) | 0.10 |
| -0.3 | 0.4059(2.90) | 0.2819(0.83) | 0.10 | 0.2837(0.80) | 0.10 |
| -0.1 | 0.3912(1.72) | 0.2967(1.09) | 0.10 | 0.2961(1.02) | 0.10 |
| 0.1 | 0.3695(1.10) | 0.2940(0.98) | 0.10 | 0.2920(1.00) | 0.10 |
| 0.3 | 0.3245(1.16) | 0.2583(0.77) | 0.10 | 0.2582(0.80) | 0.10 |
| 0.5 | 0.2458(0.99) | 0.1970(0.59) | 0.08 | 0.1981(0.60) | 0.10 |
| 0.7 | 0.1465(0.91) | 0.1170(0.37) | 0.07 | 0.1198(0.36) | 0.10 |
| 0.9 | 1.0921(7.05) | 0.0348(0.29) | 0.02 | 0.0308(0.11) | 0.08 |

*Table 4.11: The Average SEN, SPE and ACC for Scenario 7 with $p = 100$*

| $u_0$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | DAC$_1$ | DAC$_2$ |
| -0.9 | 0.9908 | 0.3498 | 0.4043 | 0.0085 | 0.9924 | 0.9837 | 0.0712 | 0.9514 | 0.9467 | 0.01 | 0.05 |
| -0.7 | 0.2296 | 0.4309 | 0.4839 | 0.9786 | 0.9867 | 0.9748 | 0.9308 | 0.9512 | 0.9434 | 0.05 | 0.10 |
| -0.5 | 0.2022 | 0.5072 | 0.5239 | 0.9835 | 0.9810 | 0.9771 | 0.9337 | 0.9510 | 0.9485 | 0.09 | 0.10 |
| -0.3 | 0.2237 | 0.5289 | 0.5356 | 0.9858 | 0.9812 | 0.9793 | 0.9372 | 0.9514 | 0.9501 | 0.10 | 0.10 |
| -0.1 | 0.2294 | 0.5818 | 0.5877 | 0.9933 | 0.9834 | 0.9814 | 0.9446 | 0.9580 | 0.9565 | 0.10 | 0.10 |
| 0.1 | 0.2128 | 0.5656 | 0.5710 | 0.9961 | 0.9838 | 0.9818 | 0.9461 | 0.9563 | 0.9548 | 0.10 | 0.10 |
| 0.3 | 0.1987 | 0.5476 | 0.5548 | 0.9953 | 0.9825 | 0.9807 | 0.9445 | 0.9539 | 0.9527 | 0.10 | 0.10 |
| 0.5 | 0.1884 | 0.4962 | 0.5267 | 0.9935 | 0.9841 | 0.9783 | 0.9421 | 0.9524 | 0.9489 | 0.08 | 0.10 |
| 0.7 | 0.1444 | 0.4403 | 0.4784 | 0.9837 | 0.9833 | 0.9755 | 0.9302 | 0.9474 | 0.9426 | 0.07 | 0.10 |
| 0.9 | 0.9893 | 0.3649 | 0.4327 | 0.0115 | 0.9904 | 0.9780 | 0.0739 | 0.9495 | 0.9424 | 0.02 | 0.08 |

*Table 4.12: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 7 with $p = 100$*

| $u_0$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|
| -0.9 | 4.7625(42.12) | 0.1240(2.43) | 0.01 | 0.1063(0.67) | 0.05 |
| -0.7 | 0.6000(11.82) | 0.4054(3.12) | 0.05 | 0.4505(2.94) | 0.10 |
| -0.5 | 1.1940(24.63) | 0.7872(6.28) | 0.09 | 0.8456(6.32) | 0.10 |
| -0.3 | 1.4632(23.33) | 1.0676(8.06) | 0.10 | 1.0876(7.81) | 0.10 |
| -0.1 | 1.4498(18.38) | 1.1199(12.62) | 0.10 | 1.1018(11.99) | 0.10 |
| 0.1 | 1.2996(9.91) | 1.0716(9.57) | 0.10 | 1.0474(8.62) | 0.10 |
| 0.3 | 1.1210(5.71) | 0.9068(6.28) | 0.10 | 0.9186(5.84) | 0.10 |
| 0.5 | 0.8850(7.53) | 0.7283(6.91) | 0.08 | 0.7486(6.83) | 0.10 |
| 0.7 | 0.5722(8.16) | 0.4567(4.85) | 0.07 | 0.5199(4.56) | 0.10 |
| 0.9 | 10.8086(71.34) | 0.1351(3.41) | 0.02 | 0.1271(1.10) | 0.08 |

*Table 4.13: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 7 with $p = 150$*

| $u_0$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|
| -0.9 | 0.5661(3.37) | 0.0347(0.21) | 0.01 | 0.0312(0.07) | 0.06 |
| -0.7 | 0.1761(1.27) | 0.1215(0.26) | 0.05 | 0.1235(0.25) | 0.10 |
| -0.5 | 0.3332(3.03) | 0.2142(0.46) | 0.09 | 0.2181(0.49) | 0.10 |
| -0.3 | 0.4135(2.63) | 0.2785(0.66) | 0.09 | 0.2809(0.68) | 0.10 |
| -0.1 | 0.4130(1.37) | 0.3210(0.76) | 0.09 | 0.3208(0.75) | 0.10 |
| 0.1 | 0.3882(0.85) | 0.3127(0.81) | 0.10 | 0.3113(0.82) | 0.10 |
| 0.3 | 0.3436(0.98) | 0.2750(0.65) | 0.09 | 0.2743(0.70) | 0.10 |
| 0.5 | 0.2714(1.11) | 0.2177(0.44) | 0.09 | 0.2205(0.50) | 0.10 |
| 0.7 | 0.1497(0.80) | 0.1227(0.34) | 0.05 | 0.1264(0.27) | 0.08 |
| 0.9 | 1.1283(5.18) | 0.0354(0.19) | 0.02 | 0.0325(0.07) | 0.07 |

*Table 4.14: The Average SEN, SPE and ACC for Scenario 7 with $p = 150$*

| $u_0$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | DAC$_1$ | DAC$_2$ |
| -0.9 | 0.9921 | 0.3138 | 0.3754 | 0.0076 | 0.9957 | 0.9899 | 0.0525 | 0.9646 | 0.9619 | 0.01 | 0.06 |
| -0.7 | 0.1881 | 0.3840 | 0.4290 | 0.9862 | 0.9921 | 0.9855 | 0.9498 | 0.9644 | 0.9602 | 0.05 | 0.10 |
| -0.5 | 0.1577 | 0.4572 | 0.4719 | 0.9877 | 0.9891 | 0.9872 | 0.9499 | 0.9653 | 0.9641 | 0.09 | 0.10 |
| -0.3 | 0.1829 | 0.4761 | 0.4906 | 0.9892 | 0.9905 | 0.9885 | 0.9525 | 0.9670 | 0.9658 | 0.09 | 0.10 |
| -0.1 | 0.1952 | 0.4819 | 0.4948 | 0.9954 | 0.9911 | 0.9890 | 0.9589 | 0.9675 | 0.9662 | 0.09 | 0.10 |
| 0.1 | 0.1789 | 0.5052 | 0.5098 | 0.9973 | 0.9901 | 0.9894 | 0.9600 | 0.9680 | 0.9675 | 0.10 | 0.10 |
| 0.3 | 0.1751 | 0.4960 | 0.5121 | 0.9970 | 0.9910 | 0.9891 | 0.9595 | 0.9688 | 0.9677 | 0.09 | 0.10 |
| 0.5 | 0.1581 | 0.4645 | 0.4819 | 0.9956 | 0.9897 | 0.9878 | 0.9574 | 0.9660 | 0.9650 | 0.09 | 0.10 |
| 0.7 | 0.1050 | 0.4119 | 0.4469 | 0.9969 | 0.9926 | 0.9890 | 0.9562 | 0.9673 | 0.9654 | 0.05 | 0.08 |
| 0.9 | 0.9925 | 0.3615 | 0.4169 | 0.0074 | 0.9950 | 0.9892 | 0.0523 | 0.9676 | 0.9645 | 0.02 | 0.07 |

*Table 4.15: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 7 with $p = 150$*

| $u_0$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|
| -0.9 | 6.7739(41.48) | 0.1348(1.54) | 0.01 | 0.1324(0.96) | 0.06 |
| -0.7 | 0.6844(11.88) | 0.4884(3.91) | 0.05 | 0.5423(3.51) | 0.10 |
| -0.5 | 1.4133(36.88) | 0.8983(7.51) | 0.09 | 0.9446(7.44) | 0.10 |
| -0.3 | 1.5481(24.57) | 1.0749(6.50) | 0.09 | 1.1218(6.94) | 0.10 |
| -0.1 | 1.6588(9.04) | 1.3373(8.57) | 0.09 | 1.3575(8.68) | 0.10 |
| 0.1 | 1.4995(9.55) | 1.2290(8.83) | 0.10 | 1.2485(9.08) | 0.10 |
| 0.3 | 1.3996(14.69) | 1.1495(12.59) | 0.09 | 1.1426(12.19) | 0.10 |
| 0.5 | 1.1909(9.53) | 0.9781(9.97) | 0.09 | 1.0351(9.27) | 0.10 |
| 0.7 | 0.5634(8.52) | 0.4783(4.14) | 0.05 | 0.5334(3.61) | 0.08 |
| 0.9 | 13.6906(63.99) | 0.1354(1.81) | 0.02 | 0.1365(0.83) | 0.07 |

*Table 4.16: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 7 with $p = 300$*

| $u_0$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|
| -0.9 | 0.7955(3.60) | 0.0369(0.21) | 0.01 | 0.0325(0.05) | 0.04 |
| -0.7 | 0.2348(1.55) | 0.1306(0.19) | 0.04 | 0.1337(0.20) | 0.07 |
| -0.5 | 0.4278(4.25) | 0.2293(0.32) | 0.06 | 0.2336(0.34) | 0.09 |
| -0.3 | 0.5150(3.78) | 0.3012(0.44) | 0.08 | 0.3033(0.46) | 0.09 |
| -0.1 | 0.4584(1.61) | 0.3334(0.50) | 0.08 | 0.3328(0.52) | 0.09 |
| 0.1 | 0.4103(0.99) | 0.3240(0.51) | 0.08 | 0.3228(0.53) | 0.09 |
| 0.3 | 0.3652(1.12) | 0.2898(0.44) | 0.08 | 0.2899(0.45) | 0.08 |
| 0.5 | 0.2800(1.22) | 0.2143(0.31) | 0.06 | 0.2170(0.34) | 0.08 |
| 0.7 | 0.1622(0.80) | 0.1212(0.18) | 0.04 | 0.1255(0.20) | 0.08 |
| 0.9 | 1.6857(5.89) | 0.0350(0.13) | 0.01 | 0.0323(0.06) | 0.06 |

*Table 4.17: The Average SEN, SPE and ACC for Scenario 7 with $p = 300$*

| $u_0$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | $_{st}$NCM$_1$ | DAC$_1$ | DAC$_2$ | DAC$_1$ | DAC$_2$ |
| -0.9 | 0.9952 | 0.2984 | 0.3389 | 0.0042 | 0.9979 | 0.9963 | 0.0267 | 0.9820 | 0.9813 | 0.01 | 0.04 |
| -0.7 | 0.1927 | 0.3590 | 0.3882 | 0.9850 | 0.9967 | 0.9950 | 0.9670 | 0.9823 | 0.9813 | 0.04 | 0.07 |
| -0.5 | 0.1530 | 0.4046 | 0.4329 | 0.9882 | 0.9963 | 0.9943 | 0.9692 | 0.9830 | 0.9818 | 0.06 | 0.09 |
| -0.3 | 0.1788 | 0.4392 | 0.4516 | 0.9896 | 0.9956 | 0.9948 | 0.9712 | 0.9831 | 0.9826 | 0.08 | 0.09 |
| -0.1 | 0.1883 | 0.4462 | 0.4584 | 0.9959 | 0.9960 | 0.9952 | 0.9775 | 0.9834 | 0.9829 | 0.08 | 0.09 |
| 0.1 | 0.1744 | 0.4524 | 0.4654 | 0.9981 | 0.9962 | 0.9954 | 0.9794 | 0.9837 | 0.9832 | 0.08 | 0.09 |
| 0.3 | 0.1625 | 0.4308 | 0.4364 | 0.9980 | 0.9959 | 0.9957 | 0.9790 | 0.9826 | 0.9826 | 0.08 | 0.08 |
| 0.5 | 0.1504 | 0.3989 | 0.4216 | 0.9968 | 0.9965 | 0.9953 | 0.9776 | 0.9826 | 0.9819 | 0.06 | 0.08 |
| 0.7 | 0.0971 | 0.3606 | 0.3993 | 0.9976 | 0.9971 | 0.9948 | 0.9771 | 0.9822 | 0.9809 | 0.04 | 0.08 |
| 0.9 | 0.9959 | 0.2977 | 0.3549 | 0.0038 | 0.9981 | 0.9954 | 0.0263 | 0.9816 | 0.9803 | 0.01 | 0.06 |

*Table 4.18: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 7 with $p = 300$*

| $u_0$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|
| -0.9 | 13.5284(62.80) | 0.1604(2.78) | 0.01 | 0.1500(1.15) | 0.04 |
| -0.7 | 1.2325(20.08) | 0.5689(4.58) | 0.04 | 0.6230(4.16) | 0.07 |
| -0.5 | 2.4460(61.52) | 1.0358(7.00) | 0.06 | 1.0676(7.11) | 0.09 |
| -0.3 | 2.6054(61.06) | 1.3276(8.01) | 0.08 | 1.3402(8.18) | 0.09 |
| -0.1 | 2.0575(13.55) | 1.7450(12.84) | 0.08 | 1.7585(12.83) | 0.09 |
| 0.1 | 1.7671(8.49) | 1.4855(8.09) | 0.08 | 1.5027(8.29) | 0.09 |
| 0.3 | 1.5712(7.03) | 1.3684(5.89) | 0.08 | 1.3804(6.11) | 0.08 |
| 0.5 | 1.1972(13.55) | 0.9737(4.44) | 0.06 | 1.0207(4.53) | 0.08 |
| 0.7 | 0.7810(18.00) | 0.4929(2.18) | 0.04 | 0.5547(1.92) | 0.08 |
| 0.9 | 29.0120(102.74) | 0.1468(2.22) | 0.01 | 0.1515(1.20) | 0.06 |

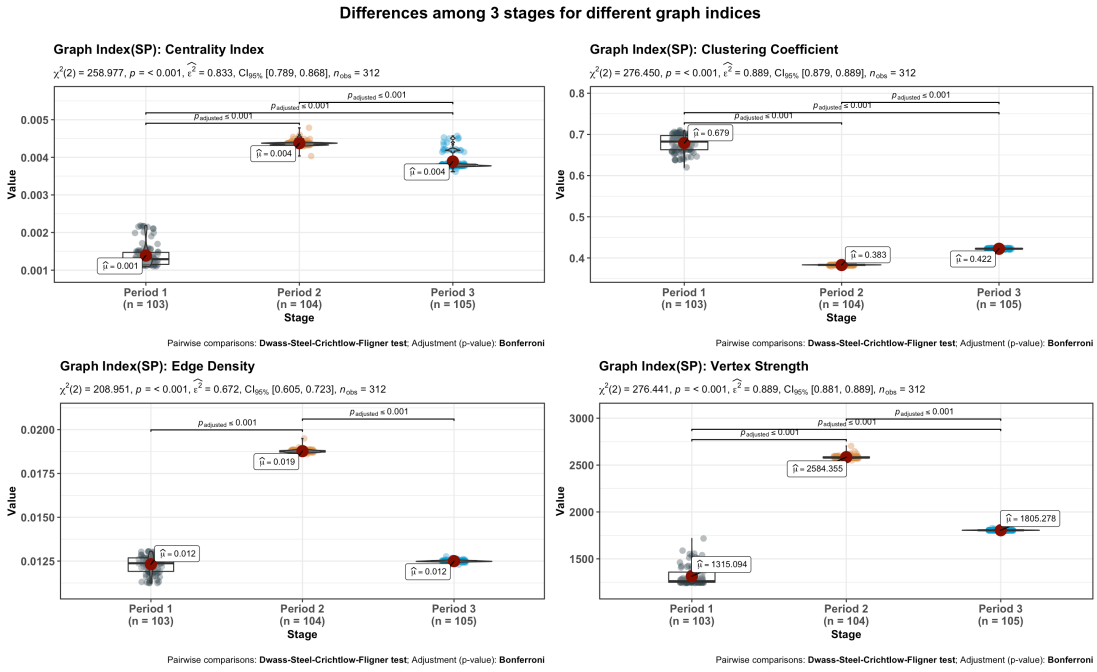
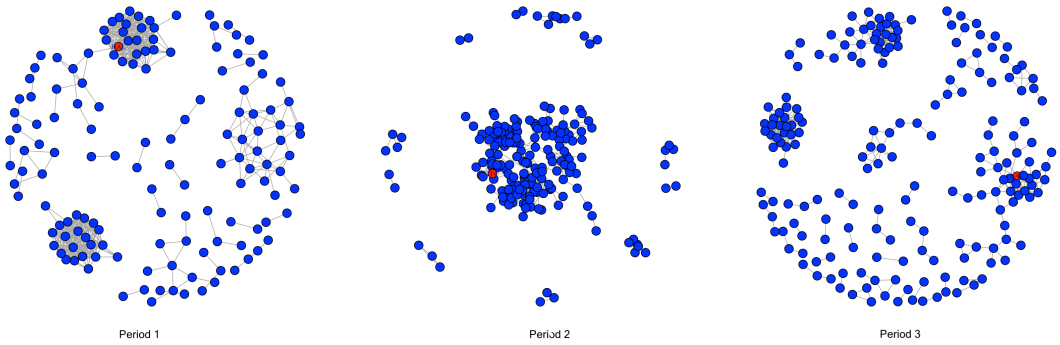



*Figure 4.6: Period Comparison*

can see that during financial-crisis, both the connection of stocks (Edges) and the correlation coefficient of stocks (Vertex Strength) increase significantly. This is not surprising because during financial-crisis, the economy recession can increase the connection and correlation among stocks. After financial-crisis, the market tries to recover from the economy depression.

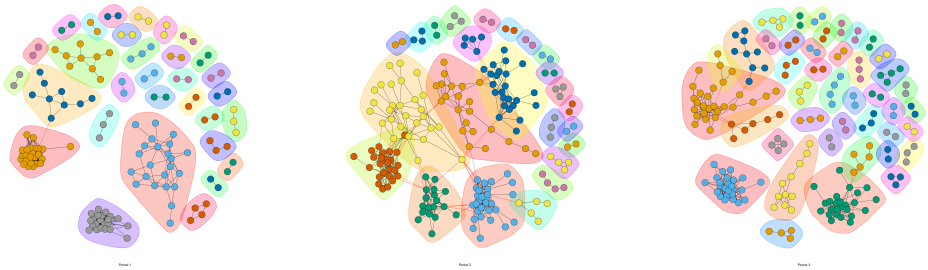
To verify this conjecture, we employ the network to represent the correlation coefficient matrix of the stocks prices. It is inconvenient to show all the networks obtained by the correlation coefficient matrices as one can obtain a network for

each day. Alternatively, we use the following way to show the changes of network in different periods. We use equation (2.29) to find the central stock of network given each $S\&P$ 500 index for every period. The central of Periods 1, 2, 3 are HAL, TFC and BAC respectively. Then we average the networks of which the centres are HAL, TFC and BAC. Figure 4.7 displays the average network for each period. The red nodes represent the central stocks. We can see that the edges in Periods 1 and 3 are more sparse than Period 2 which supports our conjecture.




Note: The initial graphical network is too dense. We trim the edge weight by 0.5 to display the networks clearly.

*Figure 4.7: Network Comparison*



*(a) Period 1* *(b) Period 2* *(c) Period 3*

*Figure 4.8: The Result of Clustering for Three Periods*

Furthermore, we apply the fast greedy clustering method (Clauset *et al.*, 2004) to the averaged networks (i.e., Figure 4.7) to find the communities in each period. The averaged networks in Period 1, 2 and 3 are divided into 31, 24 and 42 communities as shown in Figure 4.8. As the number of vertices in each period is different, the comparison among these three clustering results is unnecessary. However, we can see that the number of communities in before-financial-crisis and after-financial-crisis is bigger than the number of communities in-financial-crisis. This means that the market has recovered from economy depression and re-unions to more small communities, see Figure 4.8(c).

## 4.5 Discussion and Conclusion

In this chapter, inspired by the divide-and-conquer algorithm, we have developed the Divide-and-Combine framework of high-dimensional nonparametric covariance models combining False Discovery Rate procedure and local linear smoother. First, for mean function estimation, we identify the linear and nonlinear EWFs and split the mean function estimation into two parts. Second, we divide the procedure of covariance estimation into three steps: (1) estimation of the diagonal entries, (2) off-diagonal zero entries identification, (3) estimation of the off-diagonal nonzero entries.

Under the sparsity assumption, estimation of off-diagonal entries encounters the zero entries effect when we use the cross validation procedure to choose the bandwidth. Many zero entries will dominate the selection of bandwidth in local linear smoother, i.e., $h$ tends to infinity. It is necessary to eliminate the effect of zero entries before the application of local linear smoother. Hypothesis test can offer us a method to identify the zero entries. As there are $p_0$ null hypotheses, FDR is used in the second step to control the type I error.

We adopt the local maximum likelihood framework (Fan *et al.*, 1997; Yu & Jones, 2004) to estimate the diagonal entries to make sure the variance is positive. To solve the nonlinear equations, we have developed an algorithm based on Newton-Raphson iteration, see Appendix B.2.

Lastly, to satisfy the correlation coefficient constraint, we have developed a new nonparametric framework based on solving a nonparametric cubic equation. Solving a cubic equation to obtain the estimator of correlation coefficient is not our contribution, but applying it to the nonparametric setting is our contribution. The simulation in Figure 4.1 clearly shows that the correlation estimator with constraint is better than the empirical one.

Our method can also be extended to the non-sparse covariance matrix without any further requirement. In this circumstance, we can just concentrate on the estimator of diagonal and non-diagonal respectively. The kernel we used is the standard normal density function, one can also replace it with the other kernel functions, such as Epanechnikov, Biweight, Triweight, etc. This replacement is beyond the scope of our discussion and will be argued elsewhere.

# Chapter 5

# Change-point Detection in Time Series Segments

## 5.1 Introduction

As we reviewed in Chapter 1, intermittent isometric experiment consists of several repeated segmentations of time series in signal processing (Rhea *et al.*, 2011; Forrest *et al.*, 2014; Taylor *et al.*, 2016; Pethick *et al.*, 2016). For example, in sports science, intermittent isometric contractions are widely employed in the study of muscle fatigue (Agre & Rodriquez, 1991; Enoka & Duchateau, 2008; Katayama *et al.*, 2010; Pethick *et al.*, 2016). Figure 5.2(a) shows 659977 output torques of muscle contractions from only one participant. Data acquisition is the same as described in Pethick *et al.* (2016). For simplicity, the participant performed the designed exercise for six seconds. Between every two exercises, there is a period (four seconds) for short break (see the gaps among the spikes in Figure 5.2(a)). The output data is sampled at 1 kHz. The intermittent isometric contractions last until the task failure (Pethick *et al.*, 2016).

It is well known that the energy offered by Adenosine triphosphate (ATP) for muscle contraction will gradually reduce along the time. As a sequence, the output torques of muscle contraction could show different patterns, see Figure 5.4. The time series in Figure 5.4(c) are not as stable as those in Figures 5.4(a) and 5.4(b) because of the muscle fatigue. Therefore, sports scientists have a great interest in detecting the occurrence of muscle fatigue, i.e., when the muscle fatigue happens during a series of intermittent isometric contractions. It is easy to understand that the later muscle fatigue happens, the better athlete performs. Hence, it can evaluate the training effect by comparing the muscle fatigue change-points before and after training.

Generally, this problem can be described by the following mathematical notations. Let $\mathbb{x}_t = (x_{t1}, \ldots, x_{tN})$ be a univariate time series with length $N$ where $t$

represents the time. Notation $T$ are the number of intermittent isometric contractions. Let $\mathbb{x}_1, \mathbb{x}_2, \ldots, \mathbb{x}_T$ represent a series of intermittent isometric contractions. Mathematically, the change-point detection of muscle fatigue is equivalent to the change-points detection of $\{\mathbb{x}_1, \mathbb{x}_2, \ldots, \mathbb{x}_T\}$. In particular, if $N = 1$, it degenerates to the classical change-point detection problem (Page, 1954; Killick *et al.*, 2012; Fryzlewicz, 2014).

However, in this research, the $\mathbb{x}_t$ is no longer a scalar but a time series with length $N$. Therefore, the classical change-point detection approaches, i.e., CUSUM method (Page, 1954), multiple change-points detection approach (Killick *et al.*, 2012), multiple change-points detection for wild binary segmentation (Fryzlewicz, 2014) and frequentist change-points detection (Fryzlewicz, 2020) could not be applied to $\{\mathbb{x}_1, \mathbb{x}_2, \ldots, \mathbb{x}_T\}$ directly.

To coordinate with the classical change-point detection approaches, we need to find an appropriate statistic that can map the time series $\mathbb{x}_t$ to a scalar or score. Based on these scores, the classical change-point methods are applicable to address the above issue. Denote the statistic as the function $\mathcal{I} : \mathbb{R}^N \mapsto \mathbb{R}$. Next, we discuss the choice of $\mathcal{I}(\cdot)$. The minimum requirements we expected are the *transformation invariant* and *background-noise-free*. The property of *transformation invariant* was proved by Kullback & Leibler (1951) as follows: Suppose the transformation $\mathbb{y}_t = h(\mathbb{x}_t)$ makes $\mathbb{y}_t$ is a sufficient statistics of $\mathbb{x}_t$, then the relative entropies (5.6) of $\mathbb{y}_t$ and $\mathbb{x}_t$ are same. To the best of our knowledge, we have not found the relevant articles about background noise in terms of relative entropy. To be specific, *background-noise-free* refers to $\mathcal{I}(\cdot)$ *is independent of the variance of noise* throughout this thesis. The former property could eliminate the impact of unit while the latter property guarantees that the $\mathcal{I}(\cdot)$ does not include the background noise.

The reasons that we use *transformation invariant* and *background-noise-free* to choose function $\mathcal{I}(\cdot)$ are summarized in Table 5.1.

*Table 5.1: Potential Choices of $\mathcal{I}(\cdot)$*

| | Mean | Variance | En | CoEn | RlEn |
|---|---|---|---|---|---|
| Transformation invariant | ✘ | ✘ | ✘ | ✘ | ✔ |
| Background-noise-free | ✔ | ✘ | ✘ | ✘ | ✔ |

En, CoEn and RlEn represent entropy, conditional entropy and relative entropy respectively. We also suppose the noise has zero mean and variance $\sigma^2$.

*1. Mean and Variance.* Suppose $\mathcal{I}(\cdot)$ represents mean function, $\mathbb{y}_t = h(\mathbb{x}_t) = \alpha\mathbb{x}_t$ is a linear transformation, $\alpha \neq 0$. Then $\bar{\mathbb{y}}_t = \alpha\bar{\mathbb{x}}_t \neq \bar{\mathbb{x}}_t$. If $\mathbb{x}_t$ is independent of $\sigma^2$, then for any transformation $h(\cdot)$, $h(\mathbb{x}_t)$ is also independent of $\sigma^2$. Similarly, one can easily verify that variance does not have these two properties.

*2. Entropy (En) and Conditional Entropy (CoEn).* Entropy and conditional entropy are inappropriate choice of $\mathcal{I}(\cdot)$.

- Entropy is scale variant, for example, let $\mathrm{En}(x)$ represent the entropy of variable $x$, for any scale transformation $y = \alpha x$, $\alpha \in \mathbb{R}$ and $\alpha \neq 0$ then the entropy of variable $y$ is $\mathrm{En}(x) + \log|\alpha|$. More generally, entropy is not transformation invariant under change of variable as well, see Ihara (1993, p. 18).
- Conditional entropy is neither transformation invariant. In nonparametric settings, the four entropies: *Approximate Entropy* (ApEn), *Sample Entropy* (SpEn), *Multi-scale Entropy* (MsEn) and *Fuzzy Entropy* (FzEn) are the special cases of conditional entropy, see their reviews in Section 2.3. For instance, when one uses the multivariate uniform kernel to estimate the nonparametric CoEn, the difference between ApEn and CoEn is $\log(2h) =$ CoEn $-$ApEn, see more details in Appendix C.5. In fact, the term $\log(2h)$ comes from the scale transformation in kernel function.
- Besides, entropy and conditional entropy are not background-noise-free. A counterfactual example can be found in Section 5.2.1. Equations (5.3) and (5.4) are entropy and conditional entropy respectively, however both are related to the $\sigma^2$.

*3. Relative Entropy (RlEn).* Kullback & Leibler (1951) have proved that RlEn has the property of *transformation invariant*. The discussion of *background-noise-free* property is put off in Propositions 5.1, 5.4 and 5.5.

Besides, in some specific circumstances, the mean and variance are not suitable choices of $\mathcal{I}(\cdot)$. Suppose there are two stationary AR(*2*) processes:

$$\begin{aligned} &\text{Process 1:} \quad x_i = \phi_{11} x_{i-1} + \phi_{12} x_{i-2} + \varepsilon_{1i}, \\ &\text{Process 2:} \quad y_i = \phi_{21} y_{i-1} + \phi_{22} y_{i-2} + \varepsilon_{2i}, \end{aligned}$$

where $\varepsilon_{1i}$ and $\varepsilon_{2i}$ are white noises with zero means and variances $\sigma_1^2$ and $\sigma_2^2$ respectively. Let $N = 500, T = 100$, we randomly generate 60 time series $\mathbb{x}_1, \ldots, \mathbb{x}_{60}$ from Process 1. The last 40 time series $\mathbb{x}_{61}, \ldots, \mathbb{x}_{100}$ are from Process 2.

If $E(x_i) = E(y_i) = 0$, then we cannot use the means of $\mathbb{x}_t$,$1 \leq t \leq 100$ to detect the change-point 61. Furthermore, if equation (5.1) holds,

$$\sigma_2^2 = \sigma_1^2 \frac{(\phi_{12} - 1)(\phi_{22} + 1)({\phi_{21}}^2 - {\phi_{22}}^2 + 2\,\phi_{22} - 1)}{(\phi_{22} - 1)(\phi_{12} + 1)({\phi_{11}}^2 - {\phi_{12}}^2 + 2\,\phi_{12} - 1)}, \tag{5.1}$$

then $\mathrm{Var}(x_i) = \mathrm{Var}(y_i)$. In this case, it is also difficult to detect the change-point based on the variances of time series $\mathbb{x}_t$, $1 \leq t \leq 100$ as they are identical in theory. Note that (5.1) stands for numerous combinations of Process 1 and

Process 2 as long as $\phi_{11}, \phi_{12}, \phi_{21}$ and $\phi_{22}$ satisfy (5.1).

In this chapter, we will use the relative entropy (RlEn) as the statistic for $\mathbb{x}_t$ in ARMA processes and nonparametric settings. Relative entropy is also called Kullback-Leibler divergence (Kullback & Leibler, 1951). It is a measure to describe the distance between two probability distributions. In Section 5.2.1, we reveal the nature and superiority of relative entropy in the context of autoregressive-moving-average processes. The relative entropy is not only transformation invariant but also background-noise-free. For instance, the relative entropy (5.5) of the AR(*2*) in Section 5.2.1 is only determined by the autoregression coefficient $\phi_1$ and $\phi_2$. More generally, we extend the relative entropy to the nonparametric case. It employs the kernel density estimation (KDE) method to complete the estimation of relative entropy. We have not only clarified the detailed steps of the nonparametric RlEn estimation, but also developed a consistency theory of nonparametric RlEn. Under certain assumptions, the limiting distribution of nonparametric RlEn is Gaussian with convergence rate $\sqrt{n}h^{(m+1)/2}$ where $m$ has an upper bound. Furthermore, we recommand using the BIC criterion to select the pre-determined parameter $m$. The consistency theory of BIC is developed to ensure that the estimator of lag order converges to the true order with probability 1. The theories are summarized in Section 5.3, and the detailed proofs are put off into Appendix C. In Section 5.4, we list the simulation studies and the results. The results show that our algorithms for lag order selection and change-point detection using the RlEn are efficient in nonparametric settings. Lastly, we apply our method to two real datasets: muscle contraction and Covid-19 dataset respectively to verify the performance of our approach in practice.

## 5.2 Methodology

As aforementioned, the change-point detection in time series segments has two steps: the determinant of function $\mathcal{I}(\cdot)$ and change-point detection. In this section, we first discuss the properties of relative entropy as function $\mathcal{I}(\cdot)$ for stationary ARMA process and nonparametric settings. Second, we propose a BIC criterion for the selection of lag order and develop a consistency theory for relative entropy. Finally, the optimal change-point detection method (Killick *et al.*, 2012) is applied to the scores to find the change-points.

### 5.2.1 Relative Entropy

#### 5.2.1.1 Relative Entropy for Stationary ARMA Process

First, we introduce the relative entropy for stationary AR(2) process, then we extend it to the general stationary AR($p$), MA($q$) and ARMA($p, q$) processes.

**AR(2) Process** Without loss of generality, let

$$x_i = \phi_1 x_{i-1} + \phi_2 x_{i-2} + \varepsilon_i, \tag{5.2}$$

represent the AR(2) process without intercept, where $\varepsilon_i$ is Gaussian white noise with zero mean and variance $\sigma^2$. Suppose $-1 < \phi_2 < 1 - |\phi_1|$, then process (5.2) is stationary. Let $\gamma_0 = E(X_i^2)$, $\gamma_1 = E(X_i X_{i-1})$ and $\gamma_2 = E(X_i X_{i-2})$. By (5.2), we have the following Yule-Walker equations: $\gamma_0 = \gamma_0 \phi_1{}^2 + 2\gamma_1 \phi_1 \phi_2 + \gamma_0 \phi_2{}^2 + \sigma^2$, $\gamma_1 = \gamma_0 \phi_1 + \gamma_1 \phi_2$ and $\gamma_2 = \gamma_1 \phi_1 + \gamma_0 \phi_2$. Solving the above linear equations, we can get $\gamma_0 = \sigma^2 (\phi_2 - 1)/\phi_c$, $\gamma_1 = -\phi_1 \sigma^2/\phi_c$ and $\gamma_2 = -\sigma^2 (\phi_1{}^2 - \phi_2{}^2 + \phi_2)/\phi_c$ where $\phi_c = (\phi_2 + 1)\left(\phi_1{}^2 - \phi_2{}^2 + 2\phi_2 - 1\right)$. As $\varepsilon_i$ is the Gaussian white noise, $X_i, X_{i-1}, X_{i-2}$ have the following density function, namely,

$$f(x_i, x_{i-1}, x_{i-2}) = (2\pi)^{-\frac{3}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-(x_i, x_{i-1}, x_{i-2})\Sigma^{-1}(x_i, x_{i-1}, x_{i-2})^T/2\right),$$

where

$$\Sigma = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}, \qquad \Sigma^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} 1 & -\phi_1 & -\phi_2 \\ -\phi_1 & \phi_1{}^2 - \phi_2{}^2 + 1 & -\phi_1 \\ -\phi_2 & -\phi_1 & 1 \end{pmatrix},$$

and $|\Sigma| = -\sigma^6/(\phi_c(\phi_2 + 1))$. Therefore, the entropy of $f(x_i, x_{i-1}, x_{i-2})$ is

$$\begin{aligned} \mathrm{En}(f) &= -\iiint f(x_i, x_{i-1}, x_{i-2}) \log\left(f(x_i, x_{i-1}, x_{i-2})\right) \mathrm{d}x_i \, \mathrm{d}x_{i-1} \, \mathrm{d}x_{i-2}, \\ &= 2^{-1} \log\left((2\pi e)^3 |\Sigma|\right). \end{aligned} \tag{5.3}$$

By (5.2) and given $x_{i-1}, x_{i-2}$, we can obtain the conditional density:

$$g(x_i | x_{i-1}, x_{i-2}) = (2\pi)^{-1/2} |\sigma|^{-1} \exp\left(-(x_i - \phi_1 x_{i-1} - \phi_2 x_{i-2})^2/(2\sigma^2)\right).$$

The conditional entropy is

$$\begin{aligned}\mathrm{CoEn}(f,g) &= -\iiint f(x_i,x_{i-1},x_{i-2})\log\left(g(x_i|x_{i-1},x_{i-2})\right)\,\mathrm{d}x_i\,\mathrm{d}x_{i-1}\,\mathrm{d}x_{i-2},\\ &= \frac{1}{2}\log(2\pi)+\log(\sigma)+\frac{1}{2\sigma^2}E\left[(x_i-\phi_1x_{i-1}-\phi_2x_{i-2})^2\right],\\ &= \frac{1}{2}\log(2\pi)+\log(\sigma)+\frac{1}{2\sigma^2}E\left[y^2\right],\end{aligned}$$

where $y = c^T\mathbf{x}$, $c^T = (1,-\phi_1,-\phi_2)$, $\mathbf{x} = (x_i,x_{i-1},x_{i-2})^T$. Apparently, $y \sim N(0,c^T\Sigma c)$. It is easy to verify that $c^T\Sigma c = \sigma^2$, so

$$\mathrm{CoEn}(f,g) = 2^{-1}\log(2\pi e)+\log(\sigma). \tag{5.4}$$

We also notice that the density of $x_i$ is $g(x_i) = (2\pi)^{-1/2}|\gamma_0|^{-1/2}\exp\left(-x_i^2/2\gamma_0\right)$. Finally, one can obtain the relative entropy

$$\mathrm{RlEn}(f,g) = 2^{-1}\log\left((\phi_2-1)/\phi_c\right). \tag{5.5}$$

Comparing Entropy (5.3), Conditional Entropy (5.4) with Relative Entropy (5.5), we conclude that RlEn is determined by the coefficients of autoregression coefficients and does not include the variance of noise in the AR(*2*) process. Next, we give more general RlEn results for AR($p$), MA($q$) and ARMA($p,q$) processes.

We first generalize the relative entropy in the context of AR($p$) process, then extend the theory to MA($q$) and ARMA($p,q$) processes. From now on, we simply use $\mathcal{I}(\cdot)$ to represent the RlEn. For the consecutive variable vector $\mathbf{x}^{(m+1)} = (x_i,x_{i+1},\ldots,x_{i+m})^T$, $m \geq 1$, we try to find a statistic $\mathcal{I}(\cdot)$ such that $\mathcal{I}(\mathbf{x}^{(m+1)})$ is transformation invariant and background-noise-free (i.e., independent of $\sigma^2$). From AR(2) process, we know the relative entropy is transformation invariant and background-noise-free, see equation (5.5). Based on this fact, we divide variable vector $\mathbf{x}^{(m+1)}$ into two consecutive parts, i.e., $\mathbf{x}^{(m+1)} = ((\mathbf{x}^{(m+1-s)})^T,(\mathbf{x}^{(s)})^T)^T$, where $\mathbf{x}^{(m+1-s)} = (x_i,x_{i+1},\ldots,x_{i+m-s})^T$ and $\mathbf{x}^{(s)} = (x_{i+m+1-s},x_{i+m+2-s},\ldots,x_{i+m})^T$, $1 \leq s \leq m$. Under the stationary assumption, we define the relative entropy as

$$\begin{aligned}\mathcal{I}_s\left(\mathbf{x}^{(m+1)}\right) &= \mathrm{RlEn}_s(f,g)\\ &= \int_{\mathbb{R}^{(m+1)}} f\left(\mathbf{x}^{(m+1)}\right)\log\left(\frac{f\left(\mathbf{x}^{(m+1)}\right)}{g\left(\mathbf{x}^{(m+1-s)}\right)g\left(\mathbf{x}^{(s)}\right)}\right)\,\mathrm{d}\mathbf{x}^{(m+1)},\end{aligned} \tag{5.6}$$

where $f(\cdot)$ and $g(\cdot)$ are the corresponding probability density functions. The RlEn (5.6) defines the divergence between $f\left(\mathbf{x}^{(m+1)}\right)$ and $g\left(\mathbf{x}^{(m+1-s)}\right)g\left(\mathbf{x}^{(s)}\right)$. Furthermore, let $R_m, R_{11;ms}, R_{22;ms}$ be the autocorrelation matrices of vectors $\mathbf{x}^{(m+1)}$, $\mathbf{x}^{(m+1-s)}$ and $\mathbf{x}^{(s)}$ respectively, see the explicit expression in Appendix C.1.

**AR($p$) Process** Without loss of generality, let the AR($p$) process with zero mean be

$$x_i = \phi_1 x_{i-1} + \phi_2 x_{i-2} + \cdots + \phi_p x_{i-p} + \varepsilon_i, \tag{5.7}$$

where $\phi_1, \ldots, \phi_p$ are autoregression coefficients, $\varepsilon_i$ is the Gaussian white noise with zero mean and variance $\sigma^2$, $\{x_i\}_{1\le i\le n}$ and $\{\varepsilon_i\}_{1\le i\le n}$ are dependent. Let $\gamma_k = E(x_i x_{i-k}), k = 0, \pm 1, \pm 2, \ldots$ represent the auto-covariance functions, apparently $\gamma_k = \gamma_{-k}$. Next, define $\rho_k = \gamma_k/\gamma_0, k = 0, \pm 1, \pm 2, \ldots$ as the auto-correlation functions. Hence, we have the following proposition.

**Proposition 5.1.** *Supposed $\{x_i\}$ is a time series from the stationary AR($p$) process defined in (5.7), $\varepsilon_i$ is the Gaussian white noise with zero mean and variance $\sigma^2$, then we have*

$$\mathcal{I}_s\left(\mathbf{x}^{(m+1)}\right) = \frac{1}{2}\log\left(\frac{|R_{11;ms}|\,|R_{22;ms}|}{|R_m|}\right), \quad 1 \le s \le m, \tag{5.8}$$

*which is independent of $\sigma^2$, where $|\cdot|$ is a matrix determinant operator.*

The proof of Proposition 5.1 can be found in Appendix C.1. Proposition 5.1 demonstrates the *background-noise-free* property. Clearly, $\mathcal{I}_s(\mathbf{x}^{(m+1)})$ depends on two parameters: $m$ and $s$, $m$ is the lag order and $s$ represents the partition way of two consecutive variable vectors. It seems that for any $m \ge 1$, $1 \le s \le m$, the relative entropy (5.8) can characterize the information of AR($p$) process. However, Proposition 5.2 indicates that when $m < p$, the relative entropy only contains part information of AR($p$) process.

**Proposition 5.2.** *For $\mathbf{x}^{(m+1)}$, $\mathbf{x}^{(m+1-s)}$ and $\mathbf{x}^{(s)}$ defined in equation (5.6), $p$ is the order of AR($p$) process (5.7). If $m < p$, then for any $1 \le s \le m$, $\mathcal{I}_s(\mathbf{x}^{(m+1)})$ is a function of $\rho_1, \ldots, \rho_m$ only.*

Proposition 5.2 can be directly proved by Proposition 5.1 and Yule-Walker equations. Proposition 5.2 implies one cannot distinguish the change-point using the relative entropy in practice, if the choice of $m$ is inappropriate. For instance, the two AR(2) processes in (5.33) and (5.34), if we let $\phi_{11} = 3/4$, $\phi_{21} = 1/4$ and $\phi_{12} = -1/4$, $\phi_{22} = 7/12$, then Process 1 and Process 2 represent two different processes. However, if we let $m = 1$, then $s$ must be 1. Following Proposition 5.1, the relative entropy $\mathcal{I}_1(\mathbf{x}^{(2)})$ of Processes 1 and 2 are both equal to $\log(5/4)$, hence one fails to separate Process 1 from Process 2. However, if we let $m = 2$, $s = 1$, $\mathcal{I}_1(\mathbf{x}^{(3)})$ of Process 1 is 2.3684 while $\mathcal{I}_1(\mathbf{x}^{(3)})$ of Process 2 is 1.6667, then we can distinguish Process 1 and Process 2. Proposition 5.2 also means that $\mathcal{I}_s(\mathbf{x}^{(m+1)})$ is a function of $\rho_1, \ldots, \rho_p$ when $m \ge p$ as the partial auto-correlation function (PACF) of AR($p$) is cut-off at $p$. In practice, we can let $m = p$ and $p$ is determined by AIC or BIC in the context of stationary AR processes.

$\mathcal{I}_s(\mathbf{x}^{(m+1)})$ is not only dependent on $m$ but also on $s$. Since $R$ is symmetric,

for $s = 1$ and $s = m$, we have $\mathcal{I}_1(\mathbf{x}^{(m+1)}) = \mathcal{I}_m(\mathbf{x}^{(m+1)})$. More generally, if $m$ is odd, then we have $(m+1)/2$ ways to divide $\mathbf{x}^{(m+1)}$ into two consecutive parts, if $m$ is even, then the number of division ways is $m/2$. Proposition 5.3 gives an explicit relative entropy when $s = 1$.

**Proposition 5.3.** *For AR($p$) process and variable vector $\mathbf{x}^{(m+1)}$, if $s = 1$ and $m = p$, then the relative entropy defined in (5.6) is*

$$\mathcal{I}_1\left(\mathbf{x}^{(p+1)}\right) = -2^{-1}\log\left(1 - \sum\nolimits_{k=1}^{p}\phi_k\rho_k\right),$$

*where $\phi_k$ and $\rho_k$, $k = 1, \ldots, p$ are the autoregression and autocorrelation coefficient respectively. Furthermore, for any $m > p$, $\mathcal{I}_1(\mathbf{x}^{(m+1)}) = \mathcal{I}_1(\mathbf{x}^{(p+1)})$.*

The proof of Proposition 5.3 can be found in Appendix C.1. Proposition 5.3 gives an explicit form of relative entropy for the stationary AR($p$) process. When $m \geq p$ and $s = 1$, $\mathcal{I}_1(\mathbf{x}^{(m+1)})$ is no longer relevant to $m$, this result is not surprising because the partial correlation of AR($p$) process is cut-off at order $p$. This property can also be extended to $s = 2, 3, \ldots, \lceil p/2 \rceil$. For example,

**Corollary 5.1.** *For AR($p$) process and variable vector $\mathbf{x}^{(m+1)}$, if $s = 2$ and $m = p + 1$, then the relative entropy defined in (5.6) is*

$$\mathcal{I}_2\left(\mathbf{x}^{(p+2)}\right) = \frac{1}{2}\log\left(\frac{1-\rho_1^2}{\left(1-\sum_{k=1}^{p}\phi_k\rho_k\right)^2 - \left(\rho_1 - \sum_{k=1}^{p}\phi_k\rho_{k+1}\right)^2}\right),$$

*where $\phi_k$ and $\rho_k$ are the autoregression and autocorrelation coefficient respectively. Furthermore, for any $m > p + 1$, $\mathcal{I}_2(\mathbf{x}^{(m+1)}) = \mathcal{I}_2(\mathbf{x}^{(p+2)})$.*

The proof of Corollary 5.1 is similar to that of Proposition 5.3, we will omit the proof. $\mathcal{I}_2(\mathbf{x}^{(m+1)})$ is more complex than $\mathcal{I}_1(\mathbf{x}^{(m+1)})$ when $m > p + 1$. In practice, we suggest using $m = p$ and $s = 1$.

**MA($q$) Process** Without loss of generality, let the MA($q$) process with zero mean be

$$x_i = \varepsilon_i + \theta_1\varepsilon_{i-1} + \theta_2\varepsilon_{i-2} + \cdots + \theta_q\varepsilon_{i-q}, \tag{5.9}$$

where $\theta_1, \ldots, \theta_q$ are parameters. $\{\varepsilon_i\}_{1\leq i\leq n}$ are the i.i.d. Gaussian process with zero mean and variance $\sigma^2$. Let $\gamma_k$ and $\rho_k$, $k = 0, \pm1, \pm2, \ldots$ still represent the auto-covariance and auto-correlation functions of MA($q$), then we have

**Proposition 5.4.** *If $x_i$ is a stationary moving average process of order $q$ defined as (5.9), $R_m^{(1)}, R_{11;ms}^{(1)}, R_{22;ms}^{(1)}$ are the autocorrelation matrices of vectors $\mathbf{x}^{(m+1)}$,*

$\mathbf{x}^{(m+1-s)}$ *and* $\mathbf{x}^{(s)}$ *in MA(q) process, then we have*

$$\mathcal{I}_s\left(\mathbf{x}^{(m+1)}\right) = \frac{1}{2}\log\left(\frac{\left|R^{(1)}_{11;ms}\right|\left|R^{(1)}_{22;ms}\right|}{\left|R^{(1)}_{m}\right|}\right), \quad 1 \le s \le m, \tag{5.10}$$

*which is independent of* $\sigma^2$*, where* $|\cdot|$ *is a matrix determinant operator. Furthermore, let* $q_1 = q + 1$*, if* $s = 1$ *and* $m \ge q_1$*, then* $\mathcal{I}_s\left(\mathbf{x}^{(m+1)}\right) = -2^{-1}\log(1 - R^{(1)}_{12;q_1 1}(R^{(1)}_{11;q_1 1})^{-1}R^{(1)}_{21;q_1 1})$ *where* $R^{(1)}_{12;q_1 1} = (\rho_1, \ldots, \rho_{q_1})$ *and* $R^{(1)}_{21;q_1 1} = (R^{(1)}_{12;q_1 1})^T$.

The proof of Proposition 5.4 can be found in Appendix C.1. Similar to AR($p$), when $s = 1$ and $m \ge q_1$, the RlEn is irrelevant to $m$ and no longer changes. Next, we show that ARMA($p, q$) process also has the *background-noise-free* property.

**ARMA**$(p, q)$ **Process** Suppose the stationary ARMA($p$,$q$) be

$$x_i = \phi_1 x_{i-1} + \phi_2 x_{i-2} + \cdots + \phi_p x_{i-p} + \varepsilon_i + \theta_1\varepsilon_{i-1} + \theta_2\varepsilon_{i-2} + \cdots + \theta_q\varepsilon_{i-q},$$

where $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are parameters. $\{\varepsilon_i\}_{1\le i\le n}$ are the i.i.d. Gaussian process with zero mean and variance $\sigma^2$. Let $\gamma_k$ and $\rho_k$ still represent the auto-covariance and auto-correlation functions of ARMA($p$,$q$), then we have

**Proposition 5.5.** *For stationary ARMA(p, q) process, let* $R^{(2)}_m, R^{(2)}_{11;ms}, R^{(2)}_{22;ms}$ *be the autocorrelation matrices of vectors* $\mathbf{x}^{(m+1)}$*,* $\mathbf{x}^{(m+1-s)}$ *and* $\mathbf{x}^{(s)}$ *in ARMA(p,q) process, then we have*

$$\mathcal{I}_s\left(\mathbf{x}^{(m+1)}\right) = \frac{1}{2}\log\left(\frac{\left|R^{(2)}_{11;ms}\right|\left|R^{(2)}_{22;ms}\right|}{\left|R^{(2)}_{m}\right|}\right), \quad 1 \le s \le m, \tag{5.11}$$

*which is independent of* $\sigma^2$*, where* $|\cdot|$ *is a matrix determinant operator. Furthermore, if* $s = 1$*, then* $\mathcal{I}_s\left(\mathbf{x}^{(m+1)}\right) = -2^{-1}\log(1 - R^{(2)}_{12;m1}(R^{(2)}_{11;m1})^{-1}R^{(2)}_{21;m1})$.

The proof of Proposition 5.5 can be found in Appendix C.1. Note that when $s = 1$, the RlEn of ARMA($p, q$) process depends on $m$ even $m \ge \max(p, q + 1)$.

**Remark 5.1.** The orders $p$ and $q$ are finite in Propositions 5.5. In fact, by Wold representation (Wold, 1948) and stationary assumption, the RlEns in equations (5.8), (5.10) and (5.11) still hold when $p$ or/and $q$ are infinite. So the *background-noise-free* property is true. Furthermore, the form of RlEns in equations (5.8), (5.10) and (5.11) can be regarded as the statistics for testing the independency between $\mathbf{x}^{(m+1-s)}$ and $\mathbf{x}^{(s)}$. The RlEn (or Kullback–Leibler divergence) is 0 if $\mathbf{x}^{(m+1-s)}$ and $\mathbf{x}^{(s)}$ are independent.

So far, we have discussed the relative entropy for stationary AR($p$), MA($q$) and ARMA($p, q$) processes. In literature, there are plenty of nonlinear time series topics, for example, Fan & Yao (2003) listed various models and estimation

approaches for nonlinear time series in their book. Next, we will introduce the method of relative entropy estimation under the nonparametric circumstance. As discussed above, the tuning parameters $m$ and $s$ are also related to the relative entropy. For simplicity, we let $s = 1$ in the nonparametric settings.

#### 5.2.1.2 Nonparametric Relative Entropy

Let $X_1, \ldots, X_N$ represent the time-varying scalar measurements, which form a strictly stationary process. Denote $\mathbf{X}^{(m)} = (X_i, \ldots, X_{i+m-1})^T$ as the $m$ consecutive variables vector where $m$ could be sufficiently large but be bounded by $M$. The density function of $\mathbf{X}^{(m)}$ is defined as $g(\mathbf{X}^{(m)})$. Furthermore, let $\mathbf{X}^{(m+1)} = (X_i, \ldots, X_{i+m})^T$ and $f(\mathbf{X}^{(m+1)})$ be the $m+1$ consecutive variables vector and its probability density function. Note that $\mathbf{X}^{(m+1)} = \left(\mathbf{X}^{(m)^T}, X_{i+m}\right)^T$, given the first vector $\mathbf{X}^{(m)}$, the conditional probability density function can be expressed as $f(X_{i+m} \mid \mathbf{X}^{(m)}) = f(\mathbf{X}^{(m+1)})/g(\mathbf{X}^{(m)})$. And let $g_1(X_{i+m})$ be the density function of $X_{i+m}$, the relative entropy of system can be expressed as

$$\mathrm{RlEn} = E\left[\log\left(\frac{f\left(\mathbf{X}^{(m+1)}\right)}{g\left(\mathbf{X}^{(m)}\right) g_1(X_{i+m})}\right)\right].$$

Estimation of RlEn can be divided into two parts: the density estimation and expectation estimation. For density estimation, we use nonparametric kernel method to estimate $f(\mathbf{X}^{(m+1)})$, $g(\mathbf{X}^{(m)})$ and $g_1(X_{i+m})$. Let $x_1, \ldots, x_N$ be the observations of $X_1, \ldots, X_N$, $\mathbf{x}_i^{(m)} = (x_i, \ldots, x_{i+m-1})^T$, $\mathbf{x}_i^{(m+1)} = (\mathbf{x}_i^{(m)^T}, x_{i+m})^T$, $i = 1, \ldots, n$ where $n = N - m$. Next we employ the Jackknife kernel to estimate the densities. Jackknife kernel has been proposed to eliminate the boundaries effect for the kernel with bounded support (e.g., John, 1984; Härdle, 1990; Jones, 1993, and the references therein). Without loss of generality, we suppose Assumption 1 holds throughout this chapter.

**Assumption 1.** *The domain of kernel function $K(\cdot)$ is $[-1, 1]$ and $K(\cdot)$ satisfies $K(-x) = K(x)$ for any $x \in [-1, 1]$, $\int_{-1}^{+1} K(x)\,\mathrm{d}x = 1$ and $\int_{-1}^{+1} x^2 K(x)\,\mathrm{d}x < +\infty$.*

In fact, the Jackknife kernel is a linear combination of two different self-normalized kernel, namely,

$$k_\rho(u) = (1 + \beta(\rho))\frac{K(u)}{\omega_0(\rho)} - \frac{\beta(\rho)}{\alpha}\frac{K(u/\alpha)}{\omega_0(\rho/\alpha)},$$

where $\omega_l(\rho) = \int_{-1}^{\rho} u^l K(u)\,\mathrm{d}u$, $l = 0, 1, 2$, $0 \le \rho \le 1$ and $\beta(\rho) = \frac{R_1(\rho)}{\alpha R_1(\rho/\alpha) - R_1(\rho)}$, where $R_l(\rho) = \omega_l(\rho)/\omega_0(\rho)$, $l = 1, 2$. In this chapter, we follow the choice of $\alpha$ in John (1984) and let $\alpha = 2 - \rho$. Finally, for univariate $x$ and $y$, the Jackknife

kernel is

$$K_h^J(x-y) = \begin{cases} h^{-1}k_{(x/h)}\left(\frac{x-y}{h}\right), & \text{if } x \in [0,h). \\ h^{-1}K\left(\frac{x-y}{h}\right), & \text{if } x \in [h,1-h]. \\ h^{-1}k_{[(1-x)/h]}\left(\frac{x-y}{h}\right), & \text{if } x \in (1-h,1]. \end{cases}$$

For more details of Jackknife kernel, see Section 2.4.4 and Hong & White (2005). Next, for vectors $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_m)$, we define the scaled multivariate kernel as

$$\mathcal{K}_h^{(m)}(\mathbf{x}-\mathbf{y}) = K_h^J(x_1-y_1) \times K_h^J(x_2-y_2) \times \cdots \times K_h^J(x_m-y_m). \tag{5.12}$$

The bandwidths of $x_1, \ldots, x_m$ in equation (5.12) are same following the assumption in Hong & White (2005).

Define the "leave-one-out" kernel density estimators:

$$\begin{aligned} \hat{f}\left(\mathbf{x}_i^{(m+1)}\right) &= \frac{1}{n-1}\sum\nolimits_{j=1}^{n} \mathcal{K}_h^{(m+1)}\left(\mathbf{x}_i^{(m+1)} - \mathbf{x}_j^{(m+1)}\right)\mathbb{1}(j \neq i), \\ \hat{g}\left(\mathbf{x}_i^{(m)}\right) &= \frac{1}{n-1}\sum\nolimits_{j=1}^{n} \mathcal{K}_h^{(m)}\left(\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)}\right)\mathbb{1}(j \neq i), \\ \hat{g}_1(x_{i+m}) &= \frac{1}{n-1}\sum\nolimits_{j=1}^{n} K_h^J(x_{i+m} - x_{j+m})\mathbb{1}(j \neq i), \end{aligned}$$

then the nonparametric estimator of RlEn can be expressed as

$$\hat{\mathcal{I}}_n(m,h) = \frac{1}{n}\sum_{i \in S_n(m)} \log \frac{\hat{f}\left(\mathbf{x}_i^{(m+1)}\right)}{\hat{g}\left(\mathbf{x}_i^{(m)}\right)\hat{g}_1(x_{i+m})}, \tag{5.13}$$

where $S_n(m) = \{i \in \mathbb{N} : 1 \leq i \leq n, \hat{f}(\mathbf{x}_i^{(m+1)}) > 0, \hat{g}(\mathbf{x}_i^{(m)}) > 0, \hat{g}_1(x_{i+m}) > 0\}$. We select the bandwidth by maximizing estimator (5.13) given $m$, namely, $\hat{h} = \arg\max_h \hat{\mathcal{I}}_n(m,h)$.

However, given $h$, maximization of estimator (5.13) with respect to $m$ is an inappropriate criterion to select $m$, because the curve of $\hat{\mathcal{I}}_n(m,h)$ against $m$ changes dramatically for different bandwidths, see Figure C.1. In practice, the lag order $m$ should be determined before relative entropy computation. In next section, we use BIC criterion to select the optimal lag order based on the general nonlinear autoregression model.

#### 5.2.1.3 Lag Order Selection

For simplicity, the general nonlinear autoregression model studied in this circumstance is

$$x_{i+m} = \mathfrak{F}\left(\mathbf{x}_i^{(m)}\right) + \varepsilon_i, \tag{5.14}$$

where $1 \leq m \leq M$, $\varepsilon_i$ is Gaussian white noise and $\mathfrak{F}(\cdot)$ is an unknown function. The Nadaraya-Watson estimator of $\mathfrak{F}(\mathbf{x}_i^{(m)})$ can be expressed as:

$$\hat{\mathfrak{F}}\left(\mathbf{x}_i^{(m)}, h_*\right) = \sum\nolimits_{j=1}^{n} l_j\left(\mathbf{x}_i^{(m)}, h_*\right) x_{j+m}, \tag{5.15}$$

where

$$l_j\left(\mathbf{x}_i^{(m)}, h_*\right) = \frac{\mathcal{K}_{h_*}^{(m)}\left(\mathbf{x}_j^{(m)} - \mathbf{x}_i^{(m)}\right)}{\sum_{s=1}^{n} \mathcal{K}_{h_*}^{(m)}\left(\mathbf{x}_s^{(m)} - \mathbf{x}_i^{(m)}\right)}, \quad j = 1, \ldots, n.$$

Denote $L(h_*)$ as

$$L(h_*) = \begin{bmatrix} l_1\left(\mathbf{x}_1^{(m)}, h_*\right) & l_2\left(\mathbf{x}_1^{(m)}, h_*\right) & \cdots & l_n\left(\mathbf{x}_1^{(m)}, h_*\right) \\ l_1\left(\mathbf{x}_2^{(m)}, h_*\right) & l_2\left(\mathbf{x}_2^{(m)}, h_*\right) & \cdots & l_n\left(\mathbf{x}_2^{(m)}, h_*\right) \\ \vdots & \vdots & \cdots & \vdots \\ l_1\left(\mathbf{x}_n^{(m)}, h_*\right) & l_2\left(\mathbf{x}_n^{(m)}, h_*\right) & \cdots & l_n\left(\mathbf{x}_n^{(m)}, h_*\right) \end{bmatrix}. \tag{5.16}$$

Then we have the following Lemma.

**Lemma 5.1.** *For the multivariate kernel $\mathcal{K}_{h_*}^{(m)}(\cdot)$ and Nadaraya-Watson estimator (5.15), $L(h_*)$ is defined as equation (5.16), the effective degrees of freedom $v$ can be explicitly expressed as*

$$v(m, h_*) = \boldsymbol{tr}(L(h_*)) = \mathcal{K}_{h_*}^{(m)}(\mathbf{0}) \sum\nolimits_{i=1}^{n} \left(\sum\nolimits_{s=1}^{n} \mathcal{K}_{h_*}^{(m)}\left(\mathbf{x}_s^{(m)} - \mathbf{x}_i^{(m)}\right)\right)^{-1} = O(h_*^{-m}).$$

The proof is straightforward which will be omitted here. The bandwidth is selected by so-called leave-one-out cross validation method, i.e., minimizing

$$CV(m, h_*) = \sum\nolimits_{i=1}^{n} \left(x_{i+m} - \hat{\mathfrak{F}}_{-i}\left(\mathbf{x}_i^{(m)}, h_*\right)\right)^2,$$

where

$$\hat{\mathfrak{F}}_{-i}\left(\mathbf{x}_i^{(m)}, h_*\right) = \sum_{j=1}^{n} \frac{\mathcal{K}_{h_*}^{(m)}\left(\mathbf{x}_j^{(m)} - \mathbf{x}_i^{(m)}\right)}{\sum_{s=1, s\neq i}^{n} \mathcal{K}_{h_*}^{(m)}\left(\mathbf{x}_s^{(m)} - \mathbf{x}_i^{(m)}\right)} x_{j+m}.$$

Given $m = 1, 2, \ldots, M$, let $\hat{h}_m = \arg\min_{h_*} CV(m, h_*)$ be the optimal bandwidth, define the average square predict error as

$$\hat{\sigma}_e^2(m) = \frac{1}{n} \sum\nolimits_{i=1}^{n} \left(x_{i+m} - \hat{\mathfrak{F}}_{-i}\left(\mathbf{x}_i^{(m)}, \hat{h}_m\right)\right)^2 = n^{-1} CV(m, \hat{h}_m),$$

and the BIC is

$$BIC(m) = n \log(\hat{\sigma}_e^2(m)) + v(m, \hat{h}_m) \log(n), \quad m = 1, 2, \ldots, M. \tag{5.17}$$

Supposing $m_0$ be the underlying lag order and $m_0 \in \{1, 2, \ldots, M\}$. Let $\hat{m} = \arg\min_m BIC(m)$, then we have Theorem 5.1.

**Theorem 5.1.** *Under conditions (C11)–(C17), $\hat{m}$ converges to $m_0$ in probability, i.e.,*

$$P(\hat{m} = m_0) \to 1.$$

The proof details of Theorem 5.1 can be found in Appendix C.2. There exist various criteria proposed to address the lag order selection problem (e.g., Shibata, 1981; Vieu, 1995; Shao, 1997). This proof follows the framework of Vieu (1995) combining the discussion in Shao (1997). We can use criterion (5.17) to choose $m$ in advance, then implement the computation of relative entropy.

### 5.2.2 Change-points Detection

In the previous subsection, we have discussed the relative entropy of a time series segment for ARMA processes and nonparametric settings. Similarly, we can apply the same procedure to the other time series segments. Once we obtain the relative entropies of time series segments, denoted as $\mathrm{RlEn}_1, \mathrm{RlEn}_2, \ldots, \mathrm{RlEn}_J$ where $J$ represents the number of time series segments. Then, we can apply the existing detection methods such as CUSUM (Page, 1954) and its variants (Inclán & Tiao, 1994; Picard *et al.*, 2011), quasi-likelihood (Braun *et al.*, 2000). In this chapter, we employ the proposed detection method (Killick *et al.*, 2012) to search the change-points as they pointed out that the optimal change-points can be detected with a linear computational cost. Furthermore, their method is officially adopted in the function *findchangepts* by MATLAB since 2018, which is convenient in the context of our algorithms below.

### 5.2.3 Algorithms

In practice, let $\mathcal{X} = (x_{ij})_{N \times J}$. Each column of $\mathcal{X}$ represents a time series with length $N$. Suppose $\mathcal{X}$ has been transformed by the following logistic function,

$$\mathcal{X} = \frac{1}{1 + \exp(-\mathcal{Y})}, \tag{5.18}$$

where $\mathcal{Y}$ represents the original time series observations without bounded support. Similar to Hong & White (2005), we use the logistic function (5.18) to ensure the compact support in Assumption 2 throughout this chapter.

Finally, we summarize our approach using the following two algorithms:

In Algorithm 1, one can specify the initial value of $M$. We take 10 as the default value following the suggestion from Section 4.5 in Wasserman (2006). The bandwidth selection consumes most computation time. To reduce the time

**Algorithm 1:** $m$-selection step

**Data:** Matrix observation $\mathcal{X}$ with size $N \times J$.
**Result:** $m^*$, the optimal lag order of $\mathcal{X}$.
**Init:** $M \leftarrow 10$

**for** $j \leftarrow 1$ **to** $J$ **do**
  $m_j \leftarrow 0$
  $\mathcal{X}_j \leftarrow \mathcal{X}(:, j)$
  **for** $m \leftarrow 1$ **to** $M$ **do**
    $n \leftarrow N - m$
    **for** $i \leftarrow 1$ **to** $n$ **do**
      $\mathbf{x}_i^{(m)} \leftarrow \mathcal{X}_j(i : (i + m - 1))$
      $x_{i+m} \leftarrow \mathcal{X}_j(i + m)$
    **end**
    $h_* \leftarrow \arg\min_{h_*} 1/n \sum_{i=1}^{n} \left( x_{i+m} - \hat{g}_{-i} \left( \mathbf{x}_i^{(m)}, h_* \right) \right)^2$
    $v(m, h_*) \leftarrow \mathbf{tr}(L(h_*))$
    $\hat{\sigma}_e^2 \leftarrow 1/n \sum_{i=1}^{n} \left( x_{i+m} - \hat{g} \left( \mathbf{x}_i^{(m)}, h_* \right) \right)^2$
    $\mathrm{BIC}_j(m) \leftarrow n \log(\hat{\sigma}_e^2) + v(m, h_*) \log(n)$
  **end**
**end**
$\overline{\mathrm{BIC}}(m) = \frac{1}{J} \sum_{j=1}^{J} \mathrm{BIC}_j(m)$
$m^* \leftarrow \arg\min_m \{\overline{\mathrm{BIC}}(m), m = 1, \ldots, M\}$

**Algorithm 2:** RlEn Step

**Data:** Matrix observation $\mathcal{X}$ with size $N \times J$.
**Result:** $j^*$, the change-points.
**Init:** $m \leftarrow m^*$ from **Algorithm 1**

**for** $j \leftarrow 1$ **to** $J$ **do**
  $\mathrm{rlen}_j \leftarrow 0$
  $n \leftarrow N - m$
  $\mathcal{X}_j \leftarrow \mathcal{X}(:, j)$
  $h \leftarrow \arg\min_h \left\{ \hat{\mathcal{I}}_n(m, h) \right\}$ from equation (5.13)
  $\mathrm{rlen}_j \leftarrow \hat{\mathcal{I}}_n(m, h)$
**end**
$j^* \leftarrow$ change point detected from $\mathrm{rlen}_j, j = 1, \ldots, J$ using the proposed method (Killick *et al.*, 2012)

of selection $h_*$, one can choose $h_*$ moderately at the order $O(n^{-1/(4+m)})$ as an initial value, see Section 8.2 in Fan & Yao (2003) for more details.

In Algorithm 2, we implement the bandwidth selection as well. The main difference is that the Algorithm 1 includes the multivariate nonparametric regression but in Algorithm 2, RlEn includes the multivariate nonparametric kernel density estimation.

## 5.3 Theory

Hong & White (2005) have proved that the relative entropy of pairwise variable $(X_t, X_{t-j})$ has a normal limiting distribution. The basic idea of Hong & White (2005)'s proof is to decompose the relative entropy into some different items, then expand each item to different parts by neglecting the smaller ones. Heuristically, the main parts can be expressed by the $U$-statistics. By discussing the limiting distribution of these $U$-statistics, they finally established consistency theory of relative entropy for pairwise variables.

In this section, we develop a consistency theory of the relative entropy for $m$ consecutive variables. By using their proving skills and ideas, we show that the limiting distribution of consecutive variable is Gaussian as well if $m$ has an upper bound, say $M$. The framework of our proof is very similar to that of Hong & White (2005)'s proof. Hence, the notations and most Lemmas and Theorems below originate from the theory and Appendix in Hong & White (2005). However, our theory is not a straightforward extension from pairwise variables to $m$ consecutive variable. There are some key points that need to be emphasized in our theory because they are different from those in Hong & White (2005)'s proof. In the following Lemmas, Theorems and the proofs in Appendix C.3, we will highlight these key points where they need to be emphasized.

From now on, we abbreviate $\hat{\mathcal{I}}_n(m,h)$ as $\hat{\mathcal{I}}_n(m)$. Next we rewrite the estimator (5.13), namely

$$\begin{aligned}\hat{\mathcal{I}}_n(m) = \frac{1}{n}\sum_{i\in S_n(m)} \Bigg\{ &\log\left[\frac{f\left(\mathbf{x}_i^{(m+1)}\right)}{g\left(\mathbf{x}_i^{(m)}\right) g_1(x_{i+m})}\right] + \log\left[\frac{\hat{f}\left(\mathbf{x}_i^{(m+1)}\right)}{f\left(\mathbf{x}_i^{(m+1)}\right)}\right] \\ &- \log\left[\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right] - \log\left[\frac{\hat{g}_1(x_{i+m})}{g_1(x_{i+m})}\right]\Bigg\}, \\ = \hat{I}_{nm}(f, g\cdot g_1) &+ \hat{I}_{nm}(\hat{f}, f) - \hat{I}_{nm}(\hat{g}, g) - \hat{I}_{n1}(\hat{g}_1, g_1).\end{aligned} \tag{5.19}$$

Under the following null hypothesis:

$$\mathbb{H}_0 : f\left(\mathbf{X}_i^{(m+1)}\right) = g\left(\mathbf{X}_i^{(m)}\right) g_1(X_{i+m}),$$

the first term in equation (5.19), $\hat{I}_{nm}(f, g \cdot g_1) = 0$ almost surely. Note that for $|x| < 1$, we have the inequality $\left|\log(1+x) - x + \frac{1}{2}x^2\right| \leq |x|^3$, so the third term in equation (5.19) can be expressed as

$$\begin{aligned}
\hat{I}_{nm}(\hat{g}, g) &= \frac{1}{n} \sum_{i \in S_n(m)} \log\left[1 + \frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right], \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right] \\
&\quad - \frac{1}{2} \sum_{i=1}^{n} \left[\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right]^2 + \text{remainder}, \\
&= \hat{W}_1(m) - \frac{1}{2}\hat{W}_2(m) + \text{remainder}.
\end{aligned} \tag{5.20}$$

To obtain the order of the remainder term in equation (5.20), we need the following assumption.

**Assumption 2.** *Suppose $\{X_t\}$ is strictly stationary time series with the support $\mathbb{I} = [0, 1]$. Let $\mathcal{G} : \mathbb{I} \to \mathbb{R}^+$ be the marginal density of $X_t$. On support $\mathbb{I}$, $\mathcal{G}$ is away from 0 and has twice continuously differentiation $\mathcal{G}^{(2)}(\cdot)$. Furthermore, $\mathcal{G}^{(2)}(\cdot)$ satisfies the Lipschitz condition, i.e., for any $x_1, x_2 \in \mathbb{I}$, $\left|\mathcal{G}^{(2)}(x_1) - \mathcal{G}^{(2)}(x_2)\right| \leq \mathcal{L}\,|x_1 - x_2|$, where $\mathcal{L}$ is the Lipschitz constant.*

Moreover, at the bounds 0 and 1, the first and second derivatives of $\mathcal{G}(\cdot)$ are defined by their right-hand derivative and left-hand derivative respectively. Assumption 2 is quite general and can avoid the slower convergence rate at the bounds of $\mathbb{I}$ (e.g., Hall, 1988; Robinson, 1991; Hong & White, 2005).

For the remainder term in equation (5.20), we have

**Lemma 5.2.** *Given $\mathbb{H}_0$, under Assumptions 1 and 2, if $nh^m / \log n \to \infty$, $h \to 0$ and $m < M$. The order of the remainder term in equation (5.20) is*

$$O_p\left(n^{-3/2} h^{-3m/2} (\log n)^{1/2} + m^2 h^6\right).$$

**Remark 5.2.** The powers of $h$ and $\log(n)$ in Lemma 5.2 are different to the powers in Lemma A.5 (Appendix A, Hong & White, 2005, p. 871), because Equation (B11) (Hong & White, 2005, p. 897) quoted the results of Theorem 5.3 (Fan &

Yao, 2003, p. 208)[1]. However, Hong & White (2005) claimed the uniform convergence rate (for univariate) as $O_p(n_j^{-1/2}h^{-1}\log(n_j)+h^2)[n_j = n-j]$. According to Li & Racine (2007, pp. 30–32), the uniform convergence rate for univariate should be $O_p(n^{-1/2}h^{-1/2}\log(n)^{1/2}+h^2)$. We also notice that Hong & White (2005) put $\max_{1\leq t\leq n}$ in Equation (B11) where index $t$ indicates a density function depending on $t$. By reading the detailed proof of uniform rate of convergence for kernel density estimation (KDE), the uniform rate does not depend on $f(\cdot)$, see Section 1.12 in Li & Racine (2007). Considering that Hong & White (2005) did not clarify how to obtain the Equation (B11) and our theory did not include the parameter $t$, so we adopt the uniform rate of convergence result proposed for univariate (Li & Racine, 2007, p. 32), see also equation (C.9). Hence, from now on, even our framework of theory is as same as Hong & White (2005)'s, but the convergence rate is different (not just bringing in $m$) for each Lemma and Theorem below.

To expand the term $\hat{W}_1(m)$, we need to introduce some notations: for any vector $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{I}^m$, define $\bar{g}(\mathbf{z}_1) = \int_{\mathbb{I}^m} \mathcal{K}_h^{(m)}(\mathbf{z}_1,\mathbf{z}_2)g(\mathbf{z}_2)d\mathbf{z}_2$, where $\mathcal{K}_h^{(m)}(\mathbf{z}_1,\mathbf{z}_2) = \mathcal{K}_h^{(m)}(\mathbf{z}_1-\mathbf{z}_2)$. Let

$$\tilde{\mathcal{K}}_h^{(m)}(\mathbf{z}_1,\mathbf{z}_2) = \mathcal{K}_h^{(m)}(\mathbf{z}_1,\mathbf{z}_2) - \int_{\mathbb{I}^m} \mathcal{K}_h^{(m)}(\mathbf{z},\mathbf{z}_2)d\mathbf{z},$$

$$\tilde{A}_{nm}(\mathbf{z}_1,\mathbf{z}_2) = \left[\tilde{\mathcal{K}}_h^{(m)}(\mathbf{z}_1,\mathbf{z}_2) - \int_{\mathbb{I}^m} \tilde{\mathcal{K}}_h^{(m)}(\mathbf{z}_1,\mathbf{z})g(\mathbf{z})d\mathbf{z}\right]/g(\mathbf{z}_1),$$

$$A_{nm}(\mathbf{z}_1,\mathbf{z}_2) = \left[\mathcal{K}_h^{(m)}(\mathbf{z}_1,\mathbf{z}_2) - \int_{\mathbb{I}^m} \mathcal{K}_h^{(m)}(\mathbf{z}_1,\mathbf{z})g(\mathbf{z})d\mathbf{z}\right]/g(\mathbf{z}_1), \tag{5.21}$$

$$\gamma_{nm}(\mathbf{z}_1,\mathbf{z}_2) = \int_{\mathbb{I}^m}\left[\mathcal{K}_h^{(m)}(\mathbf{z},\mathbf{z}_2) - \int_{\mathbb{I}^m} \mathcal{K}_h^{(m)}(\mathbf{z},\mathbf{z}^*)g(\mathbf{z}^*)d\mathbf{z}^*\right]d\mathbf{z}/g(\mathbf{z}_1),$$

$$B_{nm}(\mathbf{z}_1) = \left[\int_{\mathbb{I}^m} \mathcal{K}_h^{(m)}(\mathbf{z}_1,\mathbf{z})g(\mathbf{z})d\mathbf{z} - g(\mathbf{z}_1)\right]/g(\mathbf{z}_1),$$

$$H_{1nm}(\mathbf{z}_1,\mathbf{z}_2) = \tilde{A}_{nm}(\mathbf{z}_1,\mathbf{z}_2) + \tilde{A}_{nm}(\mathbf{z}_2,\mathbf{z}_1), \tag{5.22}$$

$$H_{2nm}(\mathbf{z}_1,\mathbf{z}_2) = \int_{\mathbb{I}^m} A_{nm}(\mathbf{z},\mathbf{z}_1)A_{nm}(\mathbf{z},\mathbf{z}_2)g(\mathbf{z})\,\mathrm{d}\mathbf{z},$$

$$\hat{C}_n(m) = \frac{1}{n}\sum\nolimits_{i=1}^{n}\int_{\mathbf{z}\in\mathbb{I}^m} A_{nm}\left(\mathbf{z},\mathbf{x}_i^{(m)}\right)B_{nm}(\mathbf{z})g(\mathbf{z})\,\mathrm{d}\mathbf{z}, \tag{5.23}$$

[1]In fact, the order in Theorem 5.3 should be $O_p(\log^{1/2}(T)/(Th)^{1/2})$, the power should be 1/2 which is mended in Section 5.7 in Fan & Yao (2003, p. 212).

then, we have

$$\begin{aligned}\hat{W}_1(m) &= \frac{1}{2}\binom{n}{2}^{-1}\sum_{j=2}^{n}\sum_{i=1}^{j-1}\left[\tilde{A}_{nm}\left(\mathbf{x}_i^{(m)},\mathbf{x}_j^{(m)}\right)+\tilde{A}_{nm}\left(\mathbf{x}_j^{(m)},\mathbf{x}_i^{(m)}\right)\right]\\ &\quad+\frac{1}{2}\binom{n}{2}^{-1}\sum_{j=2}^{n}\sum_{i=1}^{j-1}\left[\gamma_{nm}\left(\mathbf{x}_i^{(m)},\mathbf{x}_j^{(m)}\right)+\gamma_{nm}\left(\mathbf{x}_j^{(m)},\mathbf{x}_i^{(m)}\right)\right]\\ &\quad+\frac{1}{n}\sum_{i=1}^{n}B_{nm}\left(\mathbf{x}_i^{(m)}\right),\\ &=\frac{1}{2}\hat{H}_{1n}(m)+\frac{1}{2}\hat{\Gamma}_n(m)+\hat{B}_n(m).\end{aligned} \tag{5.24}$$

Next, we discuss the expansion of second term in equation (5.20). We write

$$\hat{W}_2(m)=\hat{W}_{21}(m)+\hat{W}_{22}(m)+\hat{W}_{23}(m), \tag{5.25}$$

where

$$\begin{aligned}\hat{W}_{21}(m)&=n^{-1}\sum\nolimits_{i=1}^{n}\left[\left(\hat{g}\left(\mathbf{x}_i^{(m)}\right)-\bar{g}\left(\mathbf{x}_i^{(m)}\right)\right)/g\left(\mathbf{x}_i^{(m)}\right)\right]^2,\\ \hat{W}_{22}(m)&=n^{-1}\sum\nolimits_{i=1}^{n}\left[\left(\bar{g}\left(\mathbf{x}_i^{(m)}\right)-g\left(\mathbf{x}_i^{(m)}\right)\right)/g\left(\mathbf{x}_i^{(m)}\right)\right]^2,\end{aligned}$$

and

$$\hat{W}_{23}(m)=\frac{2}{n}\sum_{i=1}^{n}\left[\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right)-\bar{g}\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right]\left[\frac{\bar{g}\left(\mathbf{x}_i^{(m)}\right)-g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right].$$

Let $D_{nm}(\mathbf{z}_1,\mathbf{z}_2)=A_{nm}^2(\mathbf{z}_1,\mathbf{z}_2)+A_{nm}^2(\mathbf{z}_2,\mathbf{z}_1)$ and

$$\begin{aligned}\tilde{H}_{2nm}(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3)&=A_{nm}(\mathbf{z}_1,\mathbf{z}_2)A_{nm}(\mathbf{z}_1,\mathbf{z}_3)+A_{nm}(\mathbf{z}_2,\mathbf{z}_3)A_{nm}(\mathbf{z}_2,\mathbf{z}_1)\\ &\quad+A_{nm}(\mathbf{z}_3,\mathbf{z}_1)A_{nm}(\mathbf{z}_3,\mathbf{z}_2),\end{aligned}$$

then after some simple calculations, we can obtain

$$\hat{W}_{21}(m)=\frac{1}{2(n-1)}\hat{D}_n(m)+\frac{1}{3}\frac{n-2}{n-1}\tilde{H}_{2n}(m), \tag{5.26}$$

where $\hat{D}_n(m)=\binom{n}{2}^{-1}\sum_{j=2}^{n}\sum_{i=1}^{j-1}D_{nm}(\mathbf{x}_i^{(m)},\mathbf{x}_j^{(m)})$ and $\tilde{H}_{2n}(m)=\binom{n}{3}^{-1}\sum_{k=3}^{n}\sum_{j=2}^{k-1}\sum_{i=1}^{j-1}\tilde{H}_{2nm}(\mathbf{x}_k^{(m)},\mathbf{x}_i^{(m)},\mathbf{x}_j^{(m)})$. Combining Equations

(5.20), (5.24), (5.25) and (5.26), we finally have

$$\begin{aligned}\hat{I}_{nm}(\hat{g}, g) = \frac{1}{2}\hat{H}_{1n}(m) + \frac{1}{2}\hat{\Gamma}_n(m) + \hat{B}_n(m) - \frac{1}{4(n-1)}\hat{D}_n(m) \\ - \frac{n-2}{6(n-1)}\tilde{H}_{2n}(m) - \frac{1}{2}\hat{W}_{22}(m) - \frac{1}{2}\hat{W}_{23}(m) \\ + O_p\left(\frac{(\log n)^{1/2}}{n^{3/2}h^{3m/2}} + m^2h^6\right).\end{aligned} \tag{5.27}$$

Next, we will consider each term in equation (5.27) one by one.

**Lemma 5.3.** *Given Assumptions 1 and 2, if $h \to 0$ and $m < M$. Then under $\mathbb{H}_0$, we have $P(\lim_{n\to\infty} \hat{\Gamma}_n(m) = 0) = 1$.*

**Remark 5.3.** Lemma 5.3 is the version of Lemma A.6 in Hong & White (2005).

**Lemma 5.4.** *Given Assumptions 1 and 2, if $nh^m \to \infty$, $h \to 0$ and $m < M$, then under $\mathbb{H}_0$,*

$$\hat{D}_n(m) = 2EA^2_{nm}(\mathbf{z}_1, \mathbf{z}_2) + O_p(n^{-1/2}m^{1/2}h^{-m}), \tag{5.28}$$

*where $\mathbf{z}_1, \mathbf{z}_2$ have no overlap variable.*

**Remark 5.4.** In Lemma A.7 (Hong & White, 2005, p. 872), the remainder of $\hat{D}_n(j)$ has of order $O_p(n_j^{-1}h^{-3})$. However, the proof of Lemma A.7 (Hong & White, 2005, p. 898) shows that the order is $O_p(n_j^{-1/2}h^{-2})$. We have checked and confirmed that the proof of Lemma A.7 (Hong & White, 2005, p. 898) is correct. Therefore, the remainder in Lemma 5.4 is of order $O_p(n^{-1/2}m^{1/2}h^{-m})$.

**Lemma 5.5.** *Given Assumptions 1 and 2, if $nh^m \to \infty$, $h \to 0$ and $m < M$, then under $\mathbb{H}_0$,*

$$\tilde{H}_{2n}(m) = 3\hat{H}_{2n}(m) + O_p(n^{-3/2}m^{3/2}h^{-m}), \tag{5.29}$$

*where $\hat{H}_{2n}(m)$ is defined in equation (5.30).*

$$\hat{H}_{2n}(m) = \binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1}^{j-1} H_{2nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right). \tag{5.30}$$

**Remark 5.5.** Lemma 5.5 is the version of Lemma A.8 in Hong & White (2005).

**Lemma 5.6.** *Given Assumptions 1 and 2, if $nh^m \to \infty$, $h \to 0$ and $m < M$, then under $\mathbb{H}_0$,*

$$\hat{W}_{22}(m) = EB^2_{nm}\left(\mathbf{x}_1^{(m)}\right) + O_p(n^{-1/2}h^4).$$

**Remark 5.6.** Lemma 5.6 is the version of Lemma A.9 in Hong & White (2005).

**Lemma 5.7.** *Given Assumptions 1 and 2, if $nh^m \to \infty$, $h \to 0$ and $m < M$,*

*then under* $\mathbb{H}_0$,

$$\hat{W}_{23}(m) = 2\hat{C}_n(m) + O_p(n^{-1}mh^{2-m/2}),$$

*where* $\hat{C}_n(m)$ *is defined in equation* (5.23).

**Remark 5.7.** Lemma 5.7 is the version of Lemma A.10 in Hong & White (2005).

Based on Lemma 5.3-Lemma 5.7, we immediately have Theorem 5.2.

**Theorem 5.2.** *Given Assumptions 1 and 2, if* $2 \leq m < M$, $nh^m \to \infty$, $nh^{m+12} \to 0$ *and* $(\log n)^{1/2}/(nh^m) \to 0$, *then under* $\mathbb{H}_0$,

$$\hat{I}_{nm}(\hat{g}, g) = \frac{1}{2}\hat{H}_n(m) - \frac{1}{2}L_n(m) + \left[\hat{B}_n(m) - \hat{C}_n(m)\right] + o_p(n^{-1/2}h^{-m/2}),$$

*where* $\hat{H}_n(m) = \hat{H}_{1n}(m) - (n-2)/(n-1)\hat{H}_{2n}(m)$, $L_n(m) = (n-1)^{-1}EA^2_{nm}(\mathbf{z}_1, \mathbf{z}_2) + EB^2_{nm}(\mathbf{z}_1)$, $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{I}^m$ *and* $\mathbf{z}_1, \mathbf{z}_2$ *have no overlap variable,*

$$\hat{H}_{1n}(m) = \binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1}^{j-1} H_{1nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right),$$
$$\hat{H}_{2n}(m) = \binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1}^{j-1} H_{2nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right).$$

The proof of Theorem 5.2 is straightforward which we omit here. Note that, due to Remarks 5.1 and 5.2, the remainder in Theorem 5.2 is of order $o_p(n^{-1/2}h^{-1})$ when $m = 2$ rather than $o_p(n_j^{-1}h^{-1})$ (Theorem A.1 Hong & White, 2005, p. 868). It needs to point out that if the uniform rate of convergence claimed in Equation (B11) (Hong & White, 2005, p. 897) is correct, then the condition $nh^4/\log(n) \to \infty$ in Theorem A.1 (Hong & White, 2005, p. 868) should be $nh^4/(\log(n))^2 \to \infty$.

Similarly, we can obtain the corresponding results with $g(\cdot)$ being replaced by $g_1(\cdot)$ and $f(\cdot)$ respectively. Specifically, we have the following two theorems.

**Theorem 5.3.** *Given Assumptions 1 and 2, if* $nh^{m+1} \to \infty$, $nh^{m+13} \to 0$, $(\log n)^{1/2}/(nh^{m+1}) \to 0$ *and* $1 \leq m < M$, *for any vector* $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{I}^{m+1}$, *define* $\bar{f}(\mathbf{z}_1) = \int_{\mathbb{I}^{m+1}} \mathcal{K}_h^{(m+1)}(\mathbf{z}_1, \mathbf{z}_2) f(\mathbf{z}_2) d\mathbf{z}_2$, *where* $\mathcal{K}_h^{(m+1)}(\mathbf{z}_1, \mathbf{z}_2) = \mathcal{K}_h^{(m+1)}(\mathbf{z}_1 - \mathbf{z}_2)$. *Let*

$$\tilde{K}_h^J(\mathbf{z}_1, \mathbf{z}_2) = \mathcal{K}_h^{(m+1)}(\mathbf{z}_1, \mathbf{z}_2) - \int_{\mathbb{I}^{m+1}} \mathcal{K}_h^{(m+1)}(\mathbf{z}, \mathbf{z}_2) d\mathbf{z},$$

$$\tilde{A}_{n(m+1)}(\mathbf{z}_1, \mathbf{z}_2) = \left[\tilde{K}_h^J(\mathbf{z}_1, \mathbf{z}_2) - \int_{\mathbb{I}^{m+1}} \tilde{K}_h^J(\mathbf{z}_1, \mathbf{z}) f(\mathbf{z}) d\mathbf{z}\right] / f(\mathbf{z}_1),$$

$$A_{n(m+1)}(\mathbf{z}_1, \mathbf{z}_2) = \left[\mathcal{K}_h^{(m+1)}(\mathbf{z}_1, \mathbf{z}_2) - \int_{\mathbb{I}^{m+1}} \mathcal{K}_h^{(m+1)}(\mathbf{z}_1, \mathbf{z}) f(\mathbf{z}) d\mathbf{z}\right] / f(\mathbf{z}_1),$$

$$\gamma_{n(m+1)}(\mathbf{z}_1, \mathbf{z}_2) = \int_{\mathbb{I}^{m+1}} \left[\mathcal{K}_h^{(m+1)}(\mathbf{z}, \mathbf{z}_2) - \int_{\mathbb{I}^{m+1}} \mathcal{K}_h^{(m+1)}(\mathbf{z}, \mathbf{z}^*) f(\mathbf{z}^*) d\mathbf{z}^*\right] d\mathbf{z} / f(\mathbf{z}_1),$$

$$B_{n(m+1)}(\mathbf{z}_1) = \left[\int_{\mathbb{I}^{m+1}} \mathcal{K}_h^{(m+1)}(\mathbf{z}_1, \mathbf{z}) f(\mathbf{z}) d\mathbf{z} - f(\mathbf{z}_1)\right] / f(\mathbf{z}_1),$$

$$H_{1n(m+1)}(\mathbf{z}_1,\mathbf{z}_2)=\tilde{A}_{n(m+1)}(\mathbf{z}_1,\mathbf{z}_2)+\tilde{A}_{n(m+1)}(\mathbf{z}_2,\mathbf{z}_1),$$

$$H_{2n(m+1)}(\mathbf{z}_1,\mathbf{z}_2)=\int_{\mathbb{I}^{m+1}}A_{n(m+1)}(\mathbf{z},\mathbf{z}_1)A_{n(m+1)}(\mathbf{z},\mathbf{z}_2)f(\mathbf{z})\,\mathrm{d}\mathbf{z},$$

*then under* $\mathbb{H}_0$*, we have*

$$\hat{I}_{nm}(\hat{f},f)=\frac{1}{2}\hat{H}_n(m+1)-\frac{1}{2}L_n(m+1)+\left[\hat{B}_n(m+1)-\hat{C}_n(m+1)\right]+o_p(n^{-1/2}h^{-(m+1)/2}),$$

*where* $\hat{H}_n(m+1)=\hat{H}_{1n}(m+1)-(n-2)/(n-1)\hat{H}_{2n}(m+1)$*,* $L_n(m+1)=(n-1)^{-1}EA^2_{n(m+1)}(\mathbf{z}_1,\mathbf{z}_2)+EB^2_{n(m+1)}(\mathbf{z}_1)$*,* $\mathbf{z}_1,\mathbf{z}_2\in\mathbb{I}^{m+1}$ *and* $\mathbf{z}_1,\mathbf{z}_2$ *have no overlap variable,*

$$\hat{H}_{1n}(m+1)=\binom{n}{2}^{-1}\sum\nolimits_{j=2}^{n}\sum\nolimits_{i=1}^{j-1}H_{1n(m+1)}\left(\mathbf{x}_i^{(m+1)},\mathbf{x}_j^{(m+1)}\right),$$

$$\hat{H}_{2n}(m+1)=\binom{n}{2}^{-1}\sum\nolimits_{j=2}^{n}\sum\nolimits_{i=1}^{j-1}H_{2n(m+1)}\left(\mathbf{x}_i^{(m+1)},\mathbf{x}_j^{(m+1)}\right),$$

$$\hat{B}_{n(m+1)}=n^{-1}\sum\nolimits_{i=1}^{n}B_{n(m+1)}\left(\mathbf{x}_i^{(m+1)}\right),$$

*and* $\hat{C}_n(m+1)=n^{-1}\sum_{i=1}^{n}\int_{\mathbf{z}\in\mathbb{I}^{m+1}}A_{n(m+1)}\left(\mathbf{z},\mathbf{x}_i^{(m+1)}\right)B_{n(m+1)}(\mathbf{z})f(\mathbf{z})\,\mathrm{d}\mathbf{z}$.

**Theorem 5.4.** *Given Assumptions 1 and 2, if* $nh\to\infty$*,* $(\log n)^{1/2}/(nh)\to 0$*,* $nh^{13}\to 0$*, for any vector* $z_1,z_2\in\mathbb{I}$*, define* $\bar{g}_1(z_1)=\int_0^1 K_h^J(z_1,z_2)g_1(z_2)dz_2$*, where* $K_h^J(z_1,z_2)=K_h^J(z_1-z_2)$*. Let*

$$\tilde{K}_h^J(z_1,z_2)=K_h^J(z_1,z_2)-\int_0^1 K_h^J(z,z_2)dz,$$

$$\tilde{a}_n(z_1,z_2)=\left[\tilde{K}_h^J(z_1,z_2)-\int_0^1\tilde{K}_h^J(z_1,z)g_1(z)dz\right]/g_1(z_1),$$

$$a_n(z_1,z_2)=\left[K_h^J(z_1,z_2)-\int_0^1 K_h^J(z_1,z)g_1(z)dz\right]/g_1(z_1),$$

$$\gamma_n(z_1,z_2)=\int_0^1\left[K_h^J(z,z_2)-\int_0^1 K_h^J(z,z^*)g_1(z^*)dz^*\right]dz/g_1(z_1),$$

$$b_n(z_1)=\left[\int_0^1 K_h^J(z_1,z)g_1(z)dz-g_1(z_1)\right]/g_1(z_1),$$

$$v_{1n}(z_1,z_2)=\tilde{a}_n(z_1,z_2)+\tilde{a}_n(z_2,z_1),$$

$$v_{2n}(z_1,z_2)=\int_0^1 a_n(z,z_1)a_n(z,z_2)g_1(z)\,\mathrm{d}z,$$

*then under* $\mathbb{H}_0$*, we have*

$$\hat{I}_{n1}(\hat{g}_1, g_1) = 2^{-1}(\hat{V}_n - l_n) + [\hat{b}_n - \hat{c}_n] + o_p(n^{-1/2}h^{-1/2}),$$

*where* $\hat{V}_n = \hat{V}_{1n} - (n-2)/(n-1)\hat{V}_{2n}$, $l_n = (n-1)^{-1}Ea_n^2(z_1, z_2) + Eb_n^2(z_1)$, $z_1, z_2 \in \mathbb{I}$,

$$\begin{aligned}
\hat{V}_{1n} &= \binom{n}{2}^{-1} \sum\nolimits_{j=2}^{n} \sum\nolimits_{i=1}^{j-1} v_{1n}\left(x_{i+m}, x_{j+m}\right), \\
\hat{V}_{2n} &= \binom{n}{2}^{-1} \sum\nolimits_{j=2}^{n} \sum\nolimits_{i=1}^{j-1} v_{2n}\left(x_{i+m}, x_{j+m}\right), \\
\hat{b}_n &= n^{-1} \sum\nolimits_{i=1}^{n} b_n\left(x_{i+m}\right),
\end{aligned}$$

*and* $\hat{c}_n = n^{-1}\sum_{i=1}^{n}\int_{z\in\mathbb{I}} a_n(z, x_{i+m})b_n(z)g_1(z)\,\mathrm{d}z$.

By Theorems 5.2, 5.3 and 5.4, we have

$$\begin{aligned}
2\hat{\mathcal{I}}_n(m) &= \hat{I}_{nm}(f, g\cdot g_1) + \hat{I}_{nm}(\hat{f}, f) - \hat{I}_{nm}(\hat{g}, g) - \hat{I}_{n1}(\hat{g}_1, g_1), \\
&= \left[\hat{H}_n(m+1) - \hat{H}_n(m) - \hat{V}_n\right] \\
&\quad - (n-1)^{-1}\left[\mathrm{E}A^2_{n(m+1)}(\mathbf{z}_1, \mathbf{z}_2) - \mathrm{E}A^2_{n(m)}(\mathbf{y}_1, \mathbf{y}_2) - \mathrm{E}a_n^2(z_{1m}, z_{2m})\right] \\
&\quad - \left[\mathrm{E}B^2_{n(m+1)}(\mathbf{z}_1) - \mathrm{E}B^2_{n(m)}(\mathbf{y}_1) - \mathrm{E}b_n^2(z_{1m})\right] \\
&\quad + 2\left[\hat{B}_n(m+1) - \hat{B}_n(m) - \hat{b}_n\right] \\
&\quad - 2\left[\hat{C}_n(m+1) - \hat{C}_n(m) - \hat{c}_n\right] \\
&\quad + o_p(n^{-1/2}h^{-(m+1)/2}),
\end{aligned} \tag{5.31}$$

where $\mathbf{z}_1 = (z_{10}, \ldots, z_{1(m-1)}, z_{1m})^T = (\mathbf{y}_1^T, z_{1m})^T$, $\mathbf{z}_2 = (z_{20}, \ldots, z_{2(m-1)}, z_{2m})^T = (\mathbf{y}_2^T, z_{2m})^T$. Next, we summarize the expansion of the items in equation (5.31) in Lemmas 5.8, 5.9 and 5.10 respectively. The proofs can be found in Appendix C.3.

**Lemma 5.8.** *Given Assumptions 1 and 2, under* $\mathbb{H}_0$*, we have*

$$(n-1)^{-1}\left[EA^2_{n(m+1)}(\mathbf{z}_1, \mathbf{z}_2) - EA^2_{n(m)}(\mathbf{y}_1, \mathbf{y}_2) - Ea_n^2(z_{1m}, z_{2m})\right] = d_0 + O(n^{-1}h^{-m}),$$

*where* $\kappa = \int_{-1}^{1} K^2(u)\,\mathrm{d}u$ *and* $d_0 = (n-1)^{-1}\kappa^{m+1}h^{-(m+1)}$.

**Remark 5.8.** Lemma 5.8 is the version of Lemma A.1 in Hong & White (2005).

**Lemma 5.9.** *Given* $\mathbb{H}_0$ *and* $1 \le m < M$,

$$\begin{aligned}
&2\left[\hat{B}_n(m+1) - \hat{B}_n(m) - \hat{b}_n\right] \\
&\quad - \left[EB^2_{n(m+1)}(\mathbf{z}_1) - EB^2_{n(m)}(\mathbf{y}_1) - Eb_n^2(z_{1m})\right] \\
&= O(h^6) + O_p(n^{-1/2}h^4).
\end{aligned}$$

**Remark 5.9.** Lemma 5.9 is the combination of Lemma A.2 and Lemma A.3

in Hong & White (2005).

**Lemma 5.10.** *Given* $\mathbb{H}_0$ *and* $1 \leq m < M$,

$$\hat{C}_n(m+1) - \hat{C}_n(m) - \hat{c}_n = O_p(n^{-1/2}h^4),$$

*where* $\hat{C}_n(m+1) = n^{-1}\sum_{i=1}^n \breve{C}_{m+1}(\mathbf{x}_i^{(m+1)})$, $\hat{C}_n(m) = n^{-1}\sum_{i=1}^n \breve{C}_m(\mathbf{x}_i^{(m)})$, $\hat{c}_n = n^{-1}\sum_{i=1}^n \breve{c}(x_{i+m})$, $\mathbf{z}_1 = (z_{10}, \ldots, z_{1(m-1)}, z_{1m})^T = (\mathbf{y}_1^T, z_{1m})^T$,

$$\begin{aligned}
\breve{C}_{m+1}\left(\mathbf{x}_i^{(m+1)}\right) &= \int_{\mathbf{z}_1 \in \mathbb{I}^{m+1}} A_{n(m+1)}\left(\mathbf{z}_1, \mathbf{x}_i^{(m+1)}\right) B_{n(m+1)}(\mathbf{z}_1) f(\mathbf{z}_1)\, \mathrm{d}\mathbf{z}_1, \\
\breve{C}_m\left(\mathbf{x}_i^{(m)}\right) &= \int_{\mathbf{y}_1 \in \mathbb{I}^m} A_{nm}\left(\mathbf{y}_1, \mathbf{x}_i^{(m)}\right) B_{nm}(\mathbf{y}_1) g(\mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1, \\
\breve{c}(x_{i+m}) &= \int_0^1 a_n\left(z_{1m}, x_{i+m}\right) b_n(z_{1m}) g_1(z_{1m})\, \mathrm{d}z_{1m}.
\end{aligned}$$

**Remark 5.10.** Lemma 5.10 is the version of Lemma A.4 in Hong & White (2005).

By Lemmas (5.8)–(5.10), we have

$$2\hat{\mathcal{I}}_n(m) + d_0 = \hat{H}_n(m+1) - \hat{H}_n(m) - \hat{V}_n + o_p(n^{-1/2}h^{-(m+1)/2}).$$

Recalling that $\mathbf{z_1}$ and $\mathbf{z}_2$ may have common variables in multivariate $U$-statistics, it is impossible to apply the central limit theorem of $U$-statistics to our case directly. So we need to divide $\hat{H}_n(m+1) - \hat{H}_n(m) - \hat{V}_n$ into two parts: one part includes independent components of $\mathbf{z_1}$ and $\mathbf{z}_2$, the other part includes dependent components of $\mathbf{z_1}$ and $\mathbf{z}_2$. We rewrite

$$\begin{aligned}
\hat{H}_{1n}(m) &= \binom{n}{2}^{-1} \sum_{j=1+m}^{n} \sum_{i=1}^{j-m} H_{1nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right) \\
&\quad + \binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1 \vee (j-m+1)}^{j-1} H_{1nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right), \\
&= T_{1n}(m) + T_{1n0}(m),
\end{aligned}$$

$$\begin{aligned}
\hat{H}_{2n}(m) &= \binom{n}{2}^{-1} \sum_{j=1+m}^{n} \sum_{i=1}^{j-m} H_{2nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right) \\
&\quad + \binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1 \vee (j-m+1)}^{j-1} H_{2nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right), \\
&= T_{2n}(m) + T_{2n0}(m).
\end{aligned}$$

Similarly, $\hat{H}_{1n}(m+1) = T_{1n}(m+1) + T_{1n0}(m+1)$, $\hat{H}_{2n}(m+1) = T_{2n}(m+1) +$

$T_{2n0}(m+1)$. We have the following Lemma:

**Lemma 5.11.** *Given* $\mathbb{H}_0$, *Assumptions 1 and 2, if* $nh^{m+1} \to \infty$, $nh^{m+13} \to 0$, $(\log n)^{1/2}/(nh^{m+1}) \to 0$ *and* $1 \le m < M$, *we have* $E\breve{H}_n = -d_1 + o_p(n^{-1/2}h^{-(m+1)/2})$ *where* $\tau = \int_{-1}^{1}\int_{-1}^{1} K(u)K(u+v)\,\mathrm{d}u\,\mathrm{d}v$, $d_1 = (n-2)/(n-1)[c_1(\tau^{m+1}-1) - c_2(\tau^m - 1)]$ *and* $\breve{H}_n = \hat{H}_n(m+1) - \hat{H}_n(m) - \hat{V}_n$

Finally, we prove the limiting distribution of RlEn is Gaussian with the rate $\sqrt{n}h^{(m+1)/2}$ in Theorem 5.5.

**Theorem 5.5.** *Given Assumptions 1 and 2, if* $nh^{m+1} \to \infty$, $nh^{m+13} \to 0$, $(\log n)^{1/2}/(nh^{m+1}) \to 0$ *and* $1 \le m < M$, *under* $\mathbb{H}_0$, *we have*

$$\sqrt{n}h^{(m+1)/2}\left[2\hat{\mathcal{I}}_n(m) + d_0 + d_1\right] \xrightarrow{d} N(0, \sigma_*^2), \tag{5.32}$$

*where* $\sigma_*^2 = 2\beta\kappa^m + \beta_1\tau_1^m + 2\beta_2\tau_2^m$, $\tau_1 = \int_{-1}^{1}\left[\int_{-1}^{1} K(u)K(u+v)\,\mathrm{d}u\right]^2 \mathrm{d}v$, $\tau_2 = \int_{-1}^{1}\int_{-1}^{1} K(u)K(v)K(u+v)\,\mathrm{d}u\,\mathrm{d}v$ *and* $\beta = 2n(n-m)(n-m+1)/[n^2(n-1)^2]$, $\beta_1 = \beta(n-2)^2/(n-1)^2$ *and* $\beta_2 = \beta(n-2)/(n-1)$.

The proofs of above Lemmas and Theorems are in Appendix C.3. $U$-statistics plays a significant role in the proof of RlEn consistency, Appendix C.4 includes two lemmas about the second and third-order $U$-statistics. In this section, we assume $m$ can be arbitrarily large but be bounded by $M$. It is desirable to relieve this limitation and let $M$ go to infinity at a suitable rate, say $M = O(\log(\log(n)))$. This type extension of our theory is trivial and the effect of $M \to \infty$ needs to be carefully scrutinized, which will not be discussed here. Next, we carry out several numerical examples and real dataset analysis to evaluate our theory.

## 5.4 Numerical Study

### 5.4.1 Case 1

This case is related to the nonlinear time series. Model 1 comes from Section 8.4 in Fan & Yao (2003), we change Model 2 according to Mode 1 to make them different.

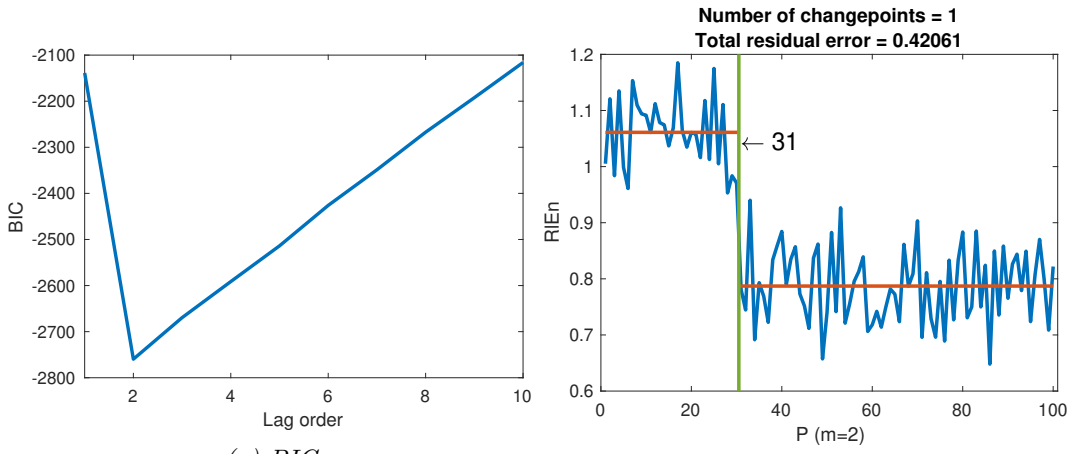
$$\begin{aligned}
&\text{Model 1:} \quad x_i = -x_{i-2}\exp\left(-x_{i-2}^2/2\right) + (1+x_{i-2}^2)^{-1}\cos(\alpha x_{i-2})x_{i-1} + \varepsilon_{1i},\\
&\text{Model 2:} \quad y_i = -y_{i-2}\exp\left(-y_{i-2}^2/2\right) + (1+y_{i-2}^2)^{-1}\sin(\alpha y_{i-2})y_{i-1} + \varepsilon_{2i},
\end{aligned}$$

where $\varepsilon_{1i}$ and $\varepsilon_{2i}$ are Gaussian white noise with zero mean and variance $0.1^2$ and $\alpha = 1.5$. Let $N = 400$, we generate $P_1 = 30$ time series from Model 1 and another $P_2 = 70$ time series from Model 2. Let $P = P_1 + P_2$, in this case, the change-point is 31. The initial values of $x_1, x_2, y_1, y_2$ are all 1. In the first step, Algorithm 1 is

implemented, and we found this algorithm can choose the correct lag order, i.e., $\hat{m} = 2$. Next, we apply our relative entropy to the simulated dataset, compute the RlEn values for each time series. Figure 5.1 shows our method can exactly identify the change point. Furthermore, we randomly draw $\alpha$ from interval $[0, \pi]$ for 150 times, and repeat the previous procedures for each $\alpha$ using ApEn and RlEn methods. The change-points detected by RlEn are 28, 29, 30, 31, 32, 33 and 34, see Table 5.2. This result shows that the accuracy of change-point detection based on RlEn and ApEn are 89.33% and 16% respectively. In this case, RlEn performs better than ApEn.

*Table 5.2: The Change-point Detection Based on ApEn and RlEn*

| Change-point | 28 | 29 | 30 | 31 | 32 | 33 | 34 |
|---|---|---|---|---|---|---|---|
| ApEn | 5 | 6 | 6 | 24 | 13 | 5 | 33 |
| RlEn | 1 | 0 | 5 | 134 | 7 | 2 | 1 |




*(a) BIC* *(b) Change-point*

*Figure 5.1: Result of Case 1*

### 5.4.2 Case 2

Suppose there are two AR(3) processes:

$$\text{Process 1:} \quad x_i = \phi_1 x_{i-1} + \phi_2 x_{i-2} + \phi_3 x_{i-3} + \varepsilon_{1i}, \tag{5.33}$$

$$\text{Process 2:} \quad y_i = \phi_1^* y_{i-1} + \phi_2^* y_{i-2} + \phi_3^* y_{i-3} + \varepsilon_{2i}, \tag{5.34}$$

where $\varepsilon_{1i}$, $\varepsilon_{2i}$ are white noise with zero mean and variance $\sigma_1^2$, $\sigma_2^2$ respectively. It is easy to verify that the variance of $x_i$ is $\sigma_1^2/(1-\phi_1\rho_1-\phi_2\rho_2-\phi_3\rho_3)$ where

$$\begin{aligned}\rho_1 &= -(\phi_1+\phi_2\,\phi_3)/\phi_d,\\ \rho_2 &= -(\phi_1^2+\phi_3\,\phi_1-{\phi_2}^2+\phi_2)/\phi_d,\\ \rho_3 &= -(\phi_1^3+{\phi_1}^2\,\phi_3+c_1\phi_1+c_2)/\phi_d,\end{aligned} \tag{5.35}$$

and $\phi_d = \phi_3^2+\phi_1\,\phi_3+\phi_2-1$, $c_1 = -{\phi_2}^2+2\,\phi_2-{\phi_3}^2$, $c_2 = {\phi_2}^2\,\phi_3-\phi_2\,\phi_3-{\phi_3}^3+\phi_3$. Suppose $x_i$ and $y_i$ have the same variance, then

$$\sigma_2^2 = \sigma_1^2\frac{1-\phi_1^*\rho_1^*-\phi_2^*\rho_2^*-\phi_3^*\rho_3^*}{1-\phi_1\rho_1-\phi_2\rho_2-\phi_3\rho_3}, \tag{5.36}$$

where $\rho_1^*, \rho_2^*, \rho_3^*$ are the expressions of equation (5.35) with $\phi_1, \phi_2, \phi_3$ replaced by $\phi_1^*, \phi_2^*, \phi_3^*$. We let $\phi_1 = 0.8$, $\phi_1^* = 0.7$, $\phi_2 = \phi_2^* = -0.3$, $\phi_3 = \phi_3^* = 0.1$ and $\sigma_1^2 = 0.1$, $\sigma_2^2$ is obtained according to equation (5.36), i.e., 0.1168. We let $N = 500$, then generate $P_1 = 60$ and $P_2 = 40$ time series from Process 1 and Process 2 respectively. Denote $P = P_1+P_2$, the change point is 61. To investigate the robustness of RlEn with respect to the selection of $m$, we appropriately allow $m$ to change from 1 to 6. For each $m$, both RlEn and ApEn are calculated using the same time series. ApEn uses the algorithm in Pincus (1991) except the pre-specified $m$. Last, repeat the above estimation procedure $J = 150$ times. Let $\tau$ represent the change-point, we define the mean absolute distance (MAD) as $\bar{\tau} = J^{-1}\sum_{j=1}^{J}|\tau_j - 61|$.

Table 5.3 shows the comparison results between RlEn and ApEn. The RlEn's MAD is consistently smaller than that of ApEn for $m = 1, \ldots, 6$. The 'failure' columns in Table 5.3 represent the number of no change-point detected. The rest columns list the number of exactly detecting $\tau = 61$. RlEn method can identify the change-point for the 150 repetitions, out of which there are at least 105 exact detections. However, ApEn is not as robust as RlEn when $m$ is large, for instance, when $m = 6$, the number of exactly detecting $\tau = 61$ is 0 and the failure number of finding change-point is 116 for ApEn method. Especially, as Pincus (1991)'s suggestion, $m = 2$ is not a suitable choice in this simulation. Furthermore, this study also verifies that our RlEn is robust with respect to the lag order. Even $m$ is misspecified, the MAD is still less than 0.45. This conclusion coincides with our theorems in the ARMA processes, see Section 5.2.1.

*Table 5.3: The Comparison Between RlEn and ApEn for Different m in Case 2*

| $m$ | RlEn | | | ApEn | | |
|---|---|---|---|---|---|---|
| | MAD | Failure | $\tau = 61$ | MAD | Failure | $\tau = 61$ |
| 1 | 0.3667 | 0 | 112/150 | 0.5733 | 0 | 101/150 |
| 2 | 0.4200 | 0 | 111/150 | 34.1591 | 106 | 0/150 |
| 3 | 0.3600 | 0 | 112/150 | 1.5200 | 0 | 62/150 |
| 4 | 0.4267 | 0 | 110/150 | 2.4698 | 1 | 51/150 |
| 5 | 0.4067 | 0 | 110/150 | 11.3868 | 44 | 10/150 |
| 6 | 0.4400 | 0 | 105/150 | 32.6471 | 116 | 0/150 |

### 5.4.3 Case 3

This case is designed to evaluate the performance of change-point detection in nonlinear time series models:

$$\text{Model 1:} \quad x_i = 0.138 + (0.316 + 0.982x_{i-1})e^{-3.89x_{i-1}^2} + \varepsilon_{1i},$$
$$\text{Model 2:} \quad y_i = -0.437 - (0.659 + 1.260y_{i-1})e^{-3.89y_{i-1}^2} + \varepsilon_{2i}.$$

We let $N = 500$, then generate $P_1 = 160$ and $P_2 = 80$ time series from Model 1 and Model 2 respectively. Denote $P = P_1 + P_2$, the change point is 161. To investigate the robustness of RlEn with respect to the selection of $m$, we appropriately allow $m$ to change from 1 to 8. The other settings are as same as Case 2. Table 5.4 summaries the comparison between RlEn and ApEn. We can obtain the same conclusion as Case 2.

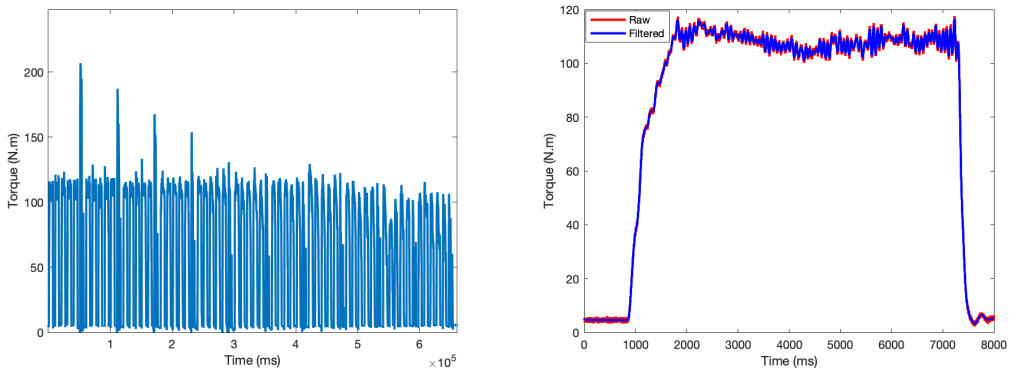
*Table 5.4: The Comparison Between RlEn and ApEn for Different m in Case 3*

| $m$ | RlEn | | | ApEn | | |
|---|---|---|---|---|---|---|
| | MAD | Failure | $\tau = 161$ | MAD | Failure | $\tau = 161$ |
| 1 | 0.2333 | 0 | 123/150 | 0.3067 | 0 | 114/150 |
| 2 | 0.2200 | 0 | 123/150 | 85.80 | 110 | 0/150 |
| 3 | 0.3000 | 0 | 116/150 | 1.120 | 0 | 79/150 |
| 4 | 0.4333 | 0 | 106/150 | 2.4467 | 0 | 51/150 |
| 5 | 0.4467 | 0 | 104/150 | 12.1727 | 11 | 19/150 |
| 6 | 0.3533 | 0 | 111/150 | 38.4800 | 100 | 2/150 |
| 7 | 0.4133 | 0 | 107/150 | 88.5385 | 137 | 0/150 |
| 8 | 0.4800 | 0 | 104/150 | 105.667 | 144 | 0/150 |

## 5.5 Real Data Analysis

### 5.5.1 Muscle Contraction Data from Single Subject

The real data contains 659977 observations which are recorded at each millisecond. Figure 5.2(a) shows the dataset. Each contraction can be identified by a rise in torque output. From Figure 5.2(a), we can also obtain the fact that there is a short sharp rise after every five tests. There are 10 short sharp rises which divide the data into 11 small periods. Figure 5.2(a) shows that there are lots of noise



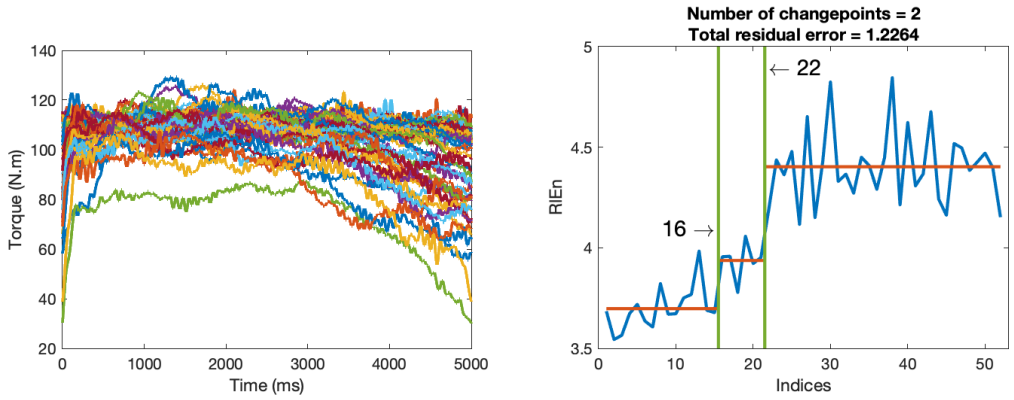

*(a) All contractions* *(b) One contraction*

*Figure 5.2: Muscle Contraction Data*

data in the observations. We need to extract the useful observations. We first cut the small periods into five pieces, each piece contains about 10000 (depends on the situation) observations. For each piece, see Figure 5.2(b), we extract 5000 consecutive observations which the moving variance is minimum. For details, let $a_i, i = 1, \ldots, 10000$ represent the 10000 observations, $A_i = (a_i, \ldots, a_{i+4999})$ denotes the consecutive 5000 observations, then the minimum moving variance of $A_s$ can be found at $s = \arg\min_i(\mathrm{Var}(A_i), i = 1, \ldots, 5001)$. Furthermore, we also use Butterworth method (Butterworth, 1930) to filter the time series before extraction. Figure 5.3(a) shows 52 extractions after using Butterworth Filter. Figure 5.3(b) shows the result of change-point detection, the two change points are 16 and 22.

To verify the performance of RlEn, we further divide the time series into three groups based on Figure 5.3(b), namely, Group 1, 2 and 3. For each group, we obtain a seasonal ARIMA$(p, d, q)$ process. Again, 52 new time series are generated from the new seasonal ARIMA processes. Then we regard them as observations and apply our RlEn to these new observations to check whether our approach can detect the change-points correctly.

First, we need to estimate three seasonal ARIMA processes, for simplicity, let $L$ be the lag operator notation, i.e., $L^i x_t = x_{t-i}$. We found that this sport dataset




*(a) All Extractions* *(b) RlEn*

*Figure 5.3: The Results of Extractions and Change-point Detection*

is more complex than we expected, the degree of integration for three groups are 2, 2 and 2 respectively according to the Augmented Dickey-Fuller test. The real sport dataset contains seasonal effect and seasonal difference for the three groups as well, so it is a better choice to build the seasonal ARIMA processes[2]:

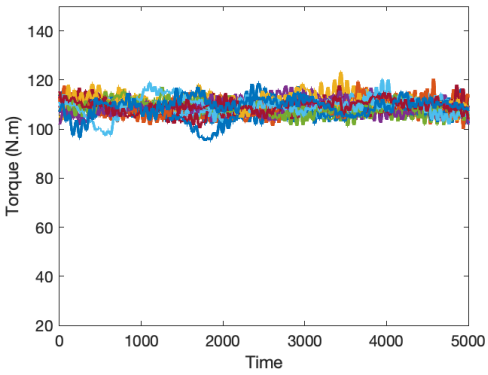
$$\phi(L)\Phi(L)(1-L)^{D}(1-L^{s})^{D_s}x_t = c + \theta(L)\varepsilon_t,$$

where $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$ and $\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$ represent the AR and MA operator polynomials. $\Phi(L) = 1 - \Phi_{p_1} L^{p_1} - \Phi_{p_2} L^{p_2} - \cdots - \Phi_{p_s} L^{p_s}$ is seasonal auto-regressive operator polynomials. $(1-L^s)^{D_s}$ is the so-called Seasonal Difference factor, for more details of seasonal ARIMA, see Section 9.9 in Hyndman & Athanasopoulos (2013). The order of $\Phi(L)$ is determined by the spectrum analysis of time series. We use Bayesian Information Criterion (BIC) to choose the order $p$ and $q$ in $\phi(L)$ and $\theta(L)$.

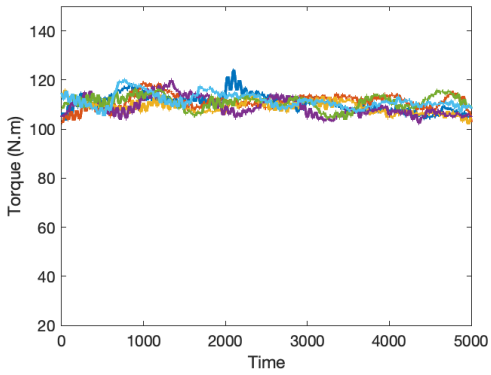
Based on the average time series of each group, we have got three processes:

$$\begin{aligned}
\textbf{Process 1}: \; & D = 2, \quad D_s = 1, \quad \hat{p} = 2, \\
& \hat{q} = 2, \quad s = 75, \quad c = 2.9993 \times 10^{-6}, \\
& \hat{\phi}(L) = 1 - 1.9414L + 0.693L^2, \\
& \hat{\theta}(L) = 1 + 1.82984L + 0.9931L^2, \\
& \hat{\Phi}(L) = 1 - 0.02037L^{75}, \; \hat{\sigma}^2 = 2 \times 10^{-7}.
\end{aligned}$$

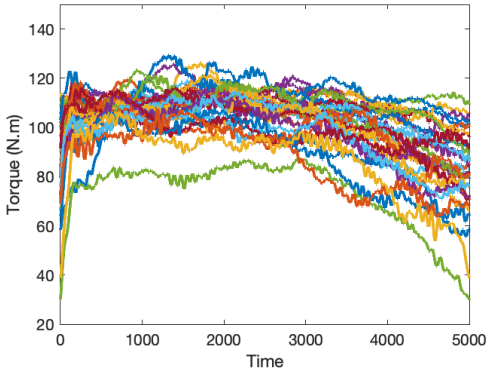
[2] https://uk.mathworks.com/help/econ/seasonal-arima-sarima-model.html




*(a) Group 1*




*(b) Group 2*




*(c) Group 3*

*Figure 5.4: The Divided Groups*

$$\textbf{Process 2}: D = 2, \quad D_s = 1, \quad \hat{p} = 2,$$
$$\hat{q} = 2, \quad s = 67, \quad c = 2.1477 \times 10^{-6},$$
$$\hat{\phi}(L) = 1 - 1.9631L + 0.9851L^2,$$
$$\hat{\theta}(L) = 1 + 1.9619L + 0.9910L^2,$$
$$\hat{\Phi}(L) = 1 + 0.2818L^{67}, \ \hat{\sigma}^2 = 2 \times 10^{-7}.$$

$$\textbf{Process 3}: D = 2, \quad D_s = 1, \quad \hat{p} = 2,$$
$$\hat{q} = 1, \quad s = 81, \quad c = 3.9159 \times 10^{-7},$$
$$\hat{\phi}(L) = 1 - 1.9768L + 0.98801L^2,$$
$$\hat{\Phi}(L) = 1 - 0.1474L^{81},$$
$$\hat{\theta}(L) = 1 + 0.3421L, \ \hat{\sigma}^2 = 2 \times 10^{-7}.$$

The details of Processes 1, 2 and 3 can be found in Appendix C.6. The number of time series generated from Processes 1, 2 and 3 are 15, 6 and 31 respectively. Figure 5.5 shows our method can detect the change-points exactly at 16 and 22.

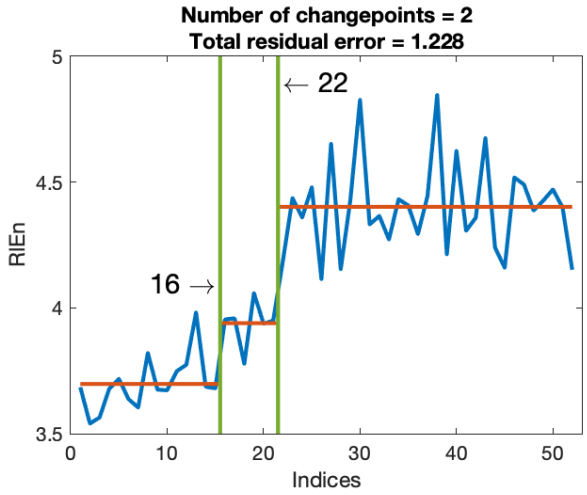



*Figure 5.5: Change-point Detection for the New Simulation Dataset*

### 5.5.2 Multi-subjects Muscle Contraction Dataset

This dataset consists of 11 subjects' muscle contraction observations. Each subject needs to perform a series of intermittent isometric contractions (six seconds for contraction and four seconds for rest) until to task failure (Pethick *et al.*, 2016). Therefore, the number of each subject contractions is not consistent, see Table 5.5. The sampling frequency is 1 kHz. We found that the Figures 5.4(a) and 5.4(b) share the similar patterns, and both are significantly different to Figure 5.4(c). Hence, in this study, we only find one change-point. Furthermore, based on the analysis of selection of $m$ in Cases 2 and 3, the selection of $m$ is not sensitive to the change-point detection. In many research fields, ApEn is frequently employed to evaluate the complexity of signals (e.g., Richman & Moorman, 2000; Burioka *et al.*, 2005; Pethick *et al.*, 2016, and among others). Considering the computational complexity, we set $m = 2$ to coordinate with ApEn. The change-point detections based on RlEn for each subject are summarized in Table 5.5. In contrast, we also obtain the change-point results based on ApEn, see Table 5.6. The parameter settings for ApEn follow the suggestions in Pincus (1991).

In Tables 5.5 and 5.6, $N_c$ represents the number of contractions in the series of experiments. $CP$ stands for the change-point detected by ApEn or RlEn. $CP/N_c$ is the relative location of change-point (in percentage) compared with $N_c$. $\overline{RlEn}_1(std.)$, $\overline{RlEn}_2(std.)$, $\overline{ApEn}_1(std.)$ and $\overline{ApEn}_2(std.)$ stand for the two groups entropy averages (standard deviation) for RlEn and ApEn respectively. The last column shows the $p$-values of $t$-test for mean comparison of two groups.

In Table 5.5, we can conclude that the intermittent isometric contractions of each subject can be divided into two groups which are supported by the $p$-values in the last column. The averages of RlEn in the first group $\overline{RlEn}_1$ are

*Table 5.5: Result of Change-point Detection Based on RlEn*

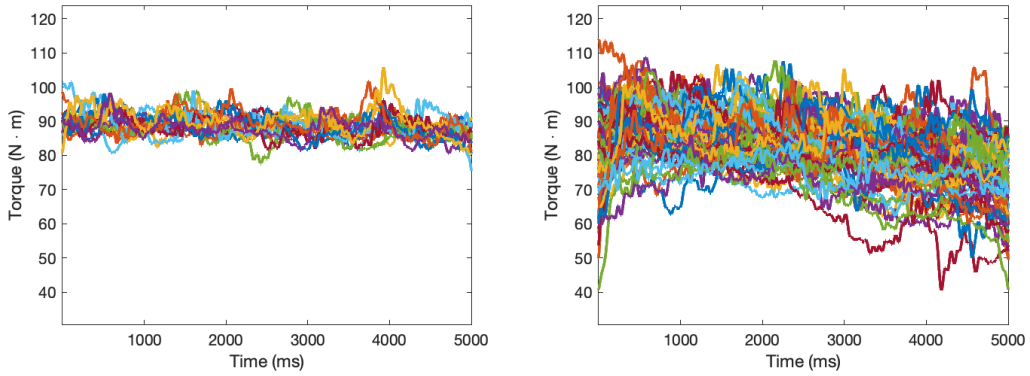
| Subject | $N_c$ | $CP$ | $CP/N_c$ | $\overline{RlEn}_1(std.)$ | $\overline{RlEn}_2(std.)$ | $p$-value |
|---|---|---|---|---|---|---|
| 1 | 70 | 22 | 31.43% | 4.0315(0.2065) | 4.2406(0.2129) | 4.30e-04 |
| 2 | 38 | 11 | 28.95% | 3.6582(0.1889) | 4.3606(0.2218) | 1.22e-08 |
| 3 | 54 | 23 | 42.59% | 3.8257(0.2571) | 4.5578(0.2871) | 4.52e-13 |
| 4 | 79 | 54 | 68.35% | 4.1137(0.2065) | 4.3692(0.2502) | 5.12e-05 |
| 5 | 289 | 236 | **81.66**% | 3.4292(0.2286) | 3.8779(0.2563) | 1.07e-18 |
| 6 | 54 | 40 | 74.07% | 3.8749(0.2395) | 4.6409(0.1624) | 5.78e-16 |
| 7 | 80 | 49 | 61.25% | 3.6241(0.2112) | 4.1233(0.2889) | 2.81e-11 |
| 8 | 177 | 78 | 44.07% | 4.1200(0.2251) | 4.4312(0.1759) | 4.03e-18 |
| 9 | 52 | 23 | 44.23% | 3.7092(0.1310) | 4.3786(0.2141) | 1.31e-18 |
| 10 | 87 | 19 | **21.84**% | 3.9454(0.1881) | 4.3561(0.1665) | 1.05e-08 |
| 11 | 89 | 38 | 42.70% | 3.9879(0.1999) | 4.4219(0.2508) | 3.62e-14 |

*Table 5.6: Result of Change-point Detection Based on ApEn*

| Subject | $N_c$ | $CP$ | $CP/N_c$ | $\overline{ApEn}_1(std.)$ | $\overline{ApEn}_2(std.)$ | $p$-value |
|---|---|---|---|---|---|---|
| 1 | 70 | 68 | 97.14% | 0.0062(0.0023) | 0.0114(0.0050) | 0.215 |
| 2 | 38 | 10 | 26.32% | 0.0134(0.0043) | 0.0037(0.0018) | 1.15e-04 |
| 3 | 54 | 18 | 33.33% | 0.0181(0.0067) | 0.0040(0.0025) | 1.33e-07 |
| 4 | 79 | – | – | – | – | – |
| 5 | 289 | 239 | 82.70% | 0.0139(0.0053) | 0.0073(0.0031) | 1.64e-22 |
| 6 | 54 | 26 | 48.15% | 0.0144(0.0046) | 0.0040(0.0031) | 5.22e-12 |
| 7 | 80 | 48 | 60.00% | 0.0129(0.0042) | 0.0059(0.0030) | 4.54e-13 |
| 8 | 177 | 78 | 44.07% | 0.0068(0.0019) | 0.0047(0.0013) | 8.36e-13 |
| 9 | 52 | 19 | 36.54% | 0.0231(0.0059) | 0.0046(0.0031) | 1.94e-11 |
| 10 | 87 | 33 | 37.93% | 0.0098(0.0023) | 0.0063(0.0026) | 5.80e-09 |
| 11 | 89 | 39 | 43.82% | 0.0123(0.0031) | 0.0047(0.0021) | 4.41e-19 |

consistently smaller than $\overline{RlEn}_2$. It is not surprising because muscle fatigue will increase the entropy of contraction signals (Pethick *et al.*, 2016). According to $CP/N_c$, Subject 5 has the largest relative location of change-point while Subject 2's is just 21.84%. Compared to other subjects, it means that the contraction torques are stable and Subject 5 can keep the stable contraction for a long time.

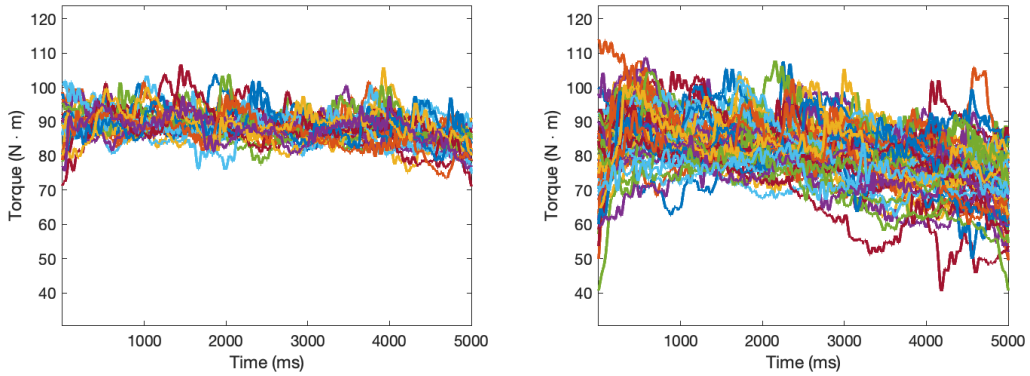
In Table 5.6, the "–" represents the failure of change-point detection based on ApEn. Besides, the $p$-value of $t$-test for Subject 1 is even larger than 0.1, which means the change-point, 68, is not statistically reliable. It is also worth pointing out that for Subject 10, the change-point based on ApEn is 33 but is 19 based on RlEn. Figure 5.6 shows the two divided groups using ApEn and RlEn respectively. It is clear that the group in Figure 5.6(a) is more stable than the group illustrated by Figure 5.6(c). Moreover, there is no need to compare the averages of ApEn and RlEn because ApEn has two free parameters and is not transformation invariant. The change-points of other subjects are almost the same for both ApEn and RlEn.




*(a) RlEn, First 18 Contractions*




*(b) RlEn, Last 69 Contractions*




*(c) ApEn, First 32 Contractions*




*(d) ApEn, Last 55 Contractions*

*Figure 5.6: Divided Groups for Subject 10*

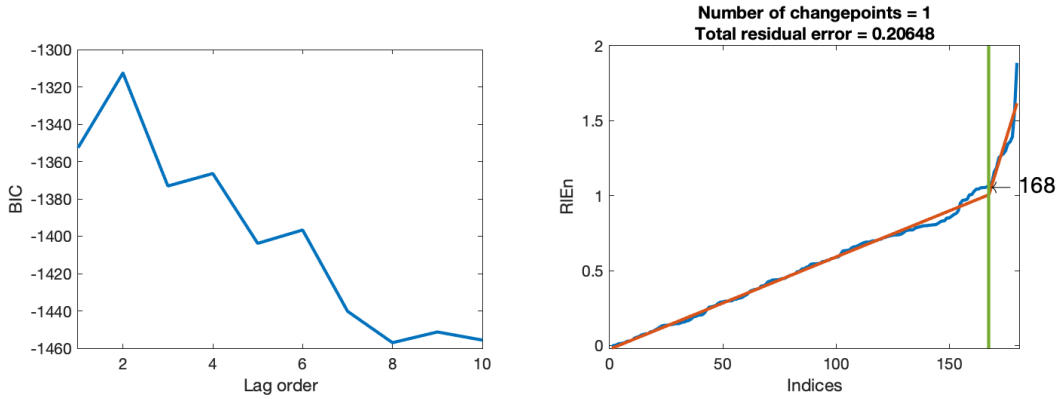
Cases 1-3 show that the RlEn is less sensitive to the lag order $m$ and better than ApEn. Combining the results of muscle contraction data, i.e., Figure 5.6, Tables 5.5 and 5.6, our RlEn performs better than the ApEn.

### 5.5.3 Covid-19 Dataset Analysis

We collect the daily confirmed cases data of each Country (Region) all over the world from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[3]. As of this thesis writing, the Covid-19 virus is in the midst of a global pandemic. This dataset includes daily confirmed report data from January 20th 2020 to February 1st 2021 only. This dataset excludes the confirmed cases in Diamond Princess, Grand Princess and MS Zaandam cruise ships. We also delete the Country (Region) whose total confirmed cases are less than 100 until February 1st 2021. There are 180 Countries (Regions) left in this dataset.

Since the time of the first confirmed case in each country is different and the ability to spread Covid-19 virus varies from country to country, each Country's (Region's) time series starts from the date on which this Country's (Region's) total confirmed cases are larger than 100. Therefore, this dataset is unbalanced. Each time series is self-normalized by its maximum daily confirmed-cases number. We apply nonparametric relative entropy method to each Country's (Region's) time series, choose 8 as the lag order for the Countries (Regions), see Figure 5.7(a). Because the Country's (Region's) RlEns in this case are not time-related, we cannot find the change-point directly. Our goal is to divide the Countries (Regions) into two groups according to their RlEns. Why we divide it into two groups rather than three groups or other number of groups? we divided them into two groups based on the result of Figure 5.7(b). It seems that dividing two groups is more appropriate than dividing three groups. So, we sort the RlEns in ascending order, then find the change-point by detecting the changes both in mean and slope of ranked RlEns, see Lavielle (2005) and Killick *et al.* (2012) for more details.

The change point is 168, see Figure 5.7(b). To illustrate the two groups,




*(a) the selection of m* *(b) change-point detection*

*Figure 5.7: Covid-19 Dataset Analysis*

we use bubbles to represent the RlEn in Figure 5.8. The red colour shows the

[3] https://github.com/CSSEGISandData/COVID-19

Countries (Regions) which have a high RlEn value. The blue colour represents the rest. The diameter of bubble is proportional to the RlEn value. Relative

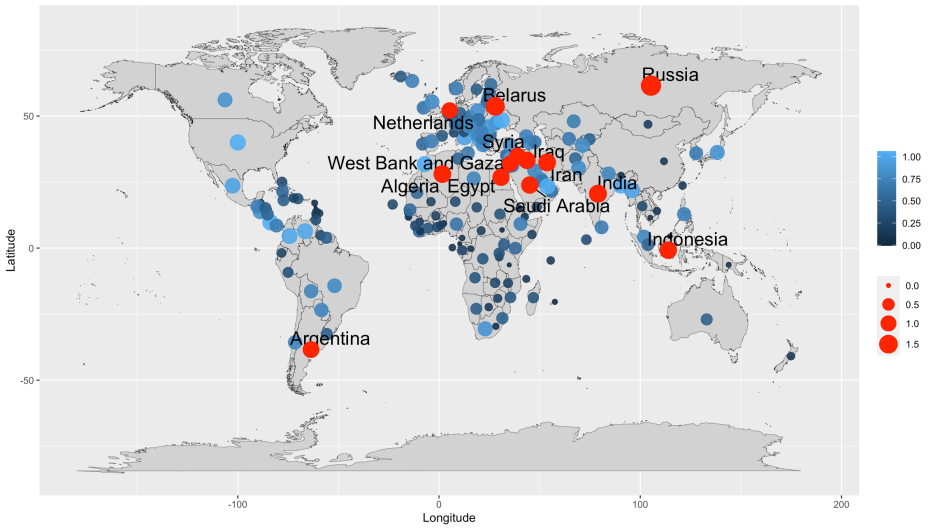



*Figure 5.8: The Global RlEn*

entropy describes the divergence between two different distributions. A large value of RlEn implies a big divergence. The red bubbles in Figure 5.8 represent Russia, Belarus, India, Iran, Algeria, Indonesia, Saudi Arabia, Iraq, Egypt, Syria, Argentina, Netherlands, West Bank and Gaza. In Appendix C.6, Figures C.4(a)–C.4(d) show the normalized daily new cases for countries: Russia, Belarus, India and Iran respectively. These countries are amid the second wave of pandemic. Moreover, we also compare US, UK, Singapore, China with the previous four countries, see Figures C.4(e)–C.4(h). The RlEn of United States is 1.055, the UK's is 0.831 because UK kept the curve under control (flatter than US's during the first wave) before September 2020. It is not surprising because the flat part of time series is less complex than the rest of time series, hence UK has a lower RlEn comparing with US which implies that the curve of new daily confirmed cases in UK has higher predicability than that in US, because entropy can be treated as a measure of chaos. Here, *predictability* is a quantitative degree to describe how a time series can be correctly predicted. This conclusion can also be verified by the time series of Singapore and China in Figures C.4(g) and C.4(h). It needs to be clarified that the application of Covid-19 dataset is just to compare the relative entropy of time series. It should not be used to determine which country is better in battling with Covid-19 because many variables such as ability of government management, medical capacity, technology level, etc are excluded from our simple application. The evaluation of government performances all over the world is a complicated task which we will not discuss here.

## 5.6 Conclusion

In this chapter, we have proposed a nonparametric relative entropy as a testing statistic to detect the change-points of time series segments. For the ARMA processes under strictly stationary assumption, the relative entropy is free of background noise and only be determined by the auto-regressive coefficients. Especially, when the lag order $m$ is larger than the underlying lag order $p$ of auto-regression, the relative entropy no longer changes. This merit in essence originates from the cut-off tail property of partial auto-correlation function in AR($p$) process. In nonparametric setting, we have developed a type of leave-one-out relative entropy. Given Assumptions 1 and 2, we have proved that the relative entropy has a limiting normal distribution with of order $\sqrt{n}h^{(m+1)/2}$. Similarly, we have also discussed the selection of lag order $m$. We suggest using BIC to select $m$ and if $m$ has an upper bound, a theory of the selection of $m$ can ensure the consistency based on BIC from the point view of nonparametric regression. Three simulations have shown that the relative entropy is appropriate to summaries the information of a time series. Based on RlEn, one can find the change-points of time series segments with high accurateness. Two real examples have shown that our approach is effective in terms of change-point detection in practice as well.

There are some interesting issues such as: In nonparametric setting, could the relative entropy be a constant like the ARMA processes when $m$ is enough large? How to speed up the computation of relative entropy? These questions are beyond the scope of this chapter, we will not discuss them further.

# Chapter 6

# Conclusions and Future Works

In this chapter, we will draw conclusions on the two nonparametric models and relative entropy we established in previous chapters and briefly introduce the related future works.

## 6.1 Conclusions

In this thesis, we have carried out two novel nonparametric covariance models for high-dimensional settings in Chapter 3 and Chapter 4 respectively. In Chapter 5, we have proposed the RlEn as the statistic to detect the time of muscle fatigue during a series of intermittent isometric contractions in sports science.

In particular, nonparametric covariance estimation is a big challenge in contemporary high-dimensional statistics. One of the critical issues for nonparametric covariance estimation is the effect of sparsity on the bandwidth selection. A pilot study in Section 3.2.4 clearly shows that the bandwidth will go to infinity if the sparsity becomes larger and larger. The wrong-selected bandwidth can bring in extra errors in covariance matrix estimator if one uses a common bandwidth to guarantee the positive definiteness. To address this issue, we have developed a novel framework that includes multiple bandwidths in factorized band matrices of the correlation coefficient matrix. Compared to the existing kernel methods proposed by Yin *et al.* (2010) and Chen & Leng (2016), the straightforward benefit is the improvement in reducing the Frobenius norm-based error, see the results in Appendix A.

Moreover, we have employed the Frobenius norm-based criterion to avoid the computation of precision matrix in high-dimensional settings. Our algorithm is more efficient than DCM, see Figure 3.1. Furthermore, we have also developed a consistency theory for factorized NCM. Under some sparsity conditions, our proposal is consistent with the underlying covariance matrix as both the sample size and the dimension tend to infinity. It is worthy noting that our theory holds

not only for the i.i.d. case but also for the non i.i.d. case which means factorized NCM could be applied to more flexible and complex scenarios. Numerical simulations (including non i.i.d. case) shows that factorized NCM and its variants are consistently better than DCM in terms of the Frobenius norm-based loss. We have also applied factorized NCM to an asset returns example to illustrate its application in Finance.

The Factorized NCM method solves the sparsity effect problem from importing multiple bandwidths in the band matrix factors. In contrast, the Divide-and-Combine NCM addresses this problem from another point of view. Literally, we divide the covariance matrix estimation into three steps: diagonal entry estimation, zero-entry detection and off-diagonal nonzero entry estimation. Once completing these three steps, we put them together to form a new covariance matrix estimator.

There are two advantages of Divide-and-Combine NCM. (1) The zero-entry detection happens before the bandwidth selection. This means less zero entries will affect the bandwidth selection of the off-diagonal nonzero entries. The essence is to let the nonzero entries take over the bandwidth selection again via identifying most zero entries. (2) We develop a framework for the nonparametric correlation coefficient estimation with constraints. The core idea of this framework is to solve a nonparametric cubic equation of correlation. Figure 4.1 clearly shows that the constrained correlation estimator performs better than empirical correlation estimator. The controversial issue is the positive definiteness of covariance matrix using Divide-and-Combine framework. In Chapter 4, we modify the negative definite covariance matrix by adding a suitable identity matrix to itself.

Furthermore, we have also applied the Divide-and-Combine framework to the mean function estimation. The choice of local polynomial order will affect the bandwidth selection. For instance, when the mean function partially consists of constant functions, if one chooses the order of local polynomial be 0 (local constant smoother), then the bandwidths of constant functions will tend to infinity. Similarly, local linear smoother cannot be applied to bandwidth selection for the linear functions. In Chapter 4, we have used local linear smoothers for the nonlinear parts in mean function but detected the linear function by *generalized likelihood ratio statistics* (Fan *et al.*, 2001).

Considering the contributions from the variant methods, the framework of Divide-and-Combine NCM are consistently better than Factorized NCM, see the results in Appendix B.4.

In Chapter 5, we have suggested using the RlEn as the statistic to detect the change-point for the segments of time series. Because of the *transformation invariant* and *background-noise-free* properties, the RlEn is a more appropriate

statistic to summary the information of a time series comparing ApEn, SpEn, FzEn, etc. For ARMA$(p, q)$ process, the RlEn is only determined by the auto-regression and moving-averaging coefficients. Moreover, if the lag order of RlEn, $m$, is larger than $p$ in AR$(p)$ process or larger than $q+1$ in MA$(q)$ process, then the RlEn is no longer change, see the Lemmas and Theorems in §5.2.1.

In nonparametric settings, the selection of $m$ is also critical. For completeness, we have proposed a BIC criterion and developed a consistency theory of $\hat{m}$ which tends to the true underlying lag order $m$ as $n \to \infty$. We have also developed the consistency theory for the RlEn and obtained its limiting distribution. The convergence rate is $\sqrt{n}h^{(m+1)/2}$. Several numerical studies show that the RlEn performs better than ApEn. Tables 5.3 and 5.4 imply that the RlEn is not sensitive to $m$. In practice, if one can relax the selection of $m$, then we suggest $m = 2$ following Pincus (1991). To show the application of RlEn, we have implemented the whole procedure in two real world datasets.

## 6.2 Future Works

For the Divide-and-Combine NCM approach, we did not develop a corresponding theory for the constrained nonparametric correlation estimator. Even the results in Appendix B.4 are better than the Factorized NCM, the consistency property and the convergence rate remain unsolved. Figure 4.1(b) clearly shows that the convergence rate is much faster than that in the middle when the correlation is near 1 or -1. To the best of our knowledge, there is no theory about the nonparametric correlation with constraint in literature. In the future, we will aim at the development of the consistency theory.

Another research direction is the positive definiteness of covariance matrix estimation. As the Remark 2 in Yin *et al.* (2010) pointed out, to satisfy the positive definite property, the entries of nonparametric covariance matrix share the same bandwidth. In this thesis, Factorized NCM employs band matrix factors with different bandwidths to guarantee the positive definite property. However, there is no general criterion to choose the number of factors. In contrast, the Divide-and-Combine NCM approach divides the covariance matrix into three parts. Even we keep the off-diagonal nonzero entries sharing the same bandwidth, the combination of these three steps cannot always make the estimator positive definite. Thus, the conflict between one single bandwidth and positive definiteness in nonparametric covariance model is not perfectly solved in high-dimensional settings. We will seek the potential framework that can avoid this conflict.

As for the RlEn, the first task we concerned is to extend the theory to more general case. In Theorem 5.5, the lag order $m$ has an upper bound $M$ where

$M$ could be sufficient large but not tend to infinity as $n$. Theorem 5.1 inherits this condition. However, without considering the Theorem 5.5, one can relax the $M$ to be of order $O(\sqrt{\log(n)})$ or $O(\log(\log(n)))$ in Theorem 5.1. Therefore, this relaxation of $M$ is determined by Theorem 5.5. From the proof of Theorem 5.5, it is not straightforward because each convergence rate in Lemmas and previous Theorems needs to be carefully scrutinized and modified to keep the final convergence rate.

The second task is the application of RlEn in Time Series Classification (TSC). We have proved that the RlEn owns two desirable properties: *transformation invariant* and *background-noise-free* in theory. Therefore, RlEn summarizes the information of time series. It could be regarded as a feature of time series itself. So far, the feature-based TSC approaches did not consider the RlEn. It is unclear whether the application to TSC could improve the accuracy of classification. In machine learning, we will try to embed RlEn into the problem of Time Series Classification in our future research.

# Appendix A

# Results of Chapter 3

## A.1 Deriving the Plug-in Optimal Shrinkage Estimator and Factorization

The derivation of the optimal shrinkage can be divided into two steps.

Step 1: we find a population version, namely a linear combination of $I_p$ and $\hat{\Sigma}^{(t)}(u)$, denoted as $\Sigma^*(u) = \rho a I_p + (1-\rho)\hat{\Sigma}^{(t)}(u)$, whose expected Frobenius loss $\mathrm{E}\|\Sigma^*(u) - \Sigma(u)\|_F^2$ attains the minimum with respect to $0 \leq \rho \leq 1$ and $a \in \mathbb{R}$. For this purpose, we decompose the above expected quadratic loss as follows:

$$\begin{aligned}\mathrm{E}\|\Sigma^*(u) - \Sigma(u)\|_F^2 &= \mathrm{E}\|\Sigma^*(u) - \mathrm{E}[\Sigma^*(u)] + \mathrm{E}[\Sigma^*(u)] - \Sigma(u)\|^2, \\ &= (1-\rho)^2 \mathrm{E}\left\|\hat{\Sigma}^{(t)}(u) - \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]\right\|_F^2 \\ &\quad + \left\|\rho\left(aI_p - \mathrm{E}\left[\Sigma^{(t)}(u)\right]\right) + \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right] - \Sigma(u)\right\|_F^2.\end{aligned} \tag{A.1}$$

Differentiating the above loss with respect to $a$ and setting it to zero, we have

$$\begin{aligned}\mathrm{d}\mathrm{E}\|\Sigma^*(u) - \Sigma(u)\|_F^2/\mathrm{d}a = 2\rho\Big\langle I_p, \rho\left(aI_p - \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]\right) \\ + \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right] - \Sigma(u)\Big\rangle = 0,\end{aligned}$$

which yields

$$a(u) = \left\langle I_p, \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]\right\rangle - \rho^{-1}\left\langle I_p, \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right] - \Sigma(u)\right\rangle.$$

Substituting it back to (A.1), we have

$$\begin{aligned}\mathrm{E}\|\Sigma^*(u)-\Sigma(u)\|_F^2 &= (1-\rho)^2\mathrm{E}\left\|\hat{\Sigma}^{(t)}(u)-\mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]\right\|^2 \\ &\quad + \|(1-\rho)A_h - A\|_F^2, \\ &= (1-\rho)^2\mathrm{E}\|\Sigma^*(u)-\Sigma(u)\|^2 + \rho^2\|A_h\|_F^2 \\ &\quad + \|A_h - A\|_F^2 - 2\rho\langle A_h, A_h - A\rangle,\end{aligned} \tag{A.2}$$

where

$$\begin{aligned}A_h(u) &= \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right] - \left\langle I_p, \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]\right\rangle I_p, \\ A(u) &= \Sigma(u) - \langle I_p, \Sigma(u)\rangle I_p.\end{aligned}$$

Differentiating (A.2) with respect to $\rho$ and setting it to zero, we have

$$\begin{aligned}&-2(1-\rho)\mathrm{E}\left\|\hat{\Sigma}^{(t)}(u)-\Sigma(u)\right\|^2 + 2\rho\|A_h(u)\|_F^2 \\ &-2\langle A_h(u), A_h(u)-A(u)\rangle = 0.\end{aligned}$$

Solving the above equation, we have the solution

$$\rho_h(u) = \left(0 \vee \frac{\beta_h^2(u)+Q_h(u)}{\beta_h^2(u)+\alpha_h^2(u)}\right) \wedge 1, \tag{A.3}$$

where

$$\begin{aligned}\alpha_h^2(u) &= \|A_h\|_F^2, \quad \beta_h^2(u) = \mathrm{E}\left\|\hat{\Sigma}^{(t)}(u)-\mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]\right\|_F^2, \\ Q_h(u) &= \langle A_h(u), A_h(u)-A(u)\rangle.\end{aligned}$$

It is easy to see that $\alpha_h(u)$ is a Frobenius norm of the residual of $\mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]$ after its projection to the space spanned by the identity matrix $I_p$ while $\beta_h^2(u)$ is a Frobenius-type variance of $\hat{\Sigma}^{(t)}(u)$. And $Q_h(u)$ is a bias effect of the kernel smoothing. If replacing $\rho$ in $a(u)$ by $\rho_h(u)$, then we have the solution

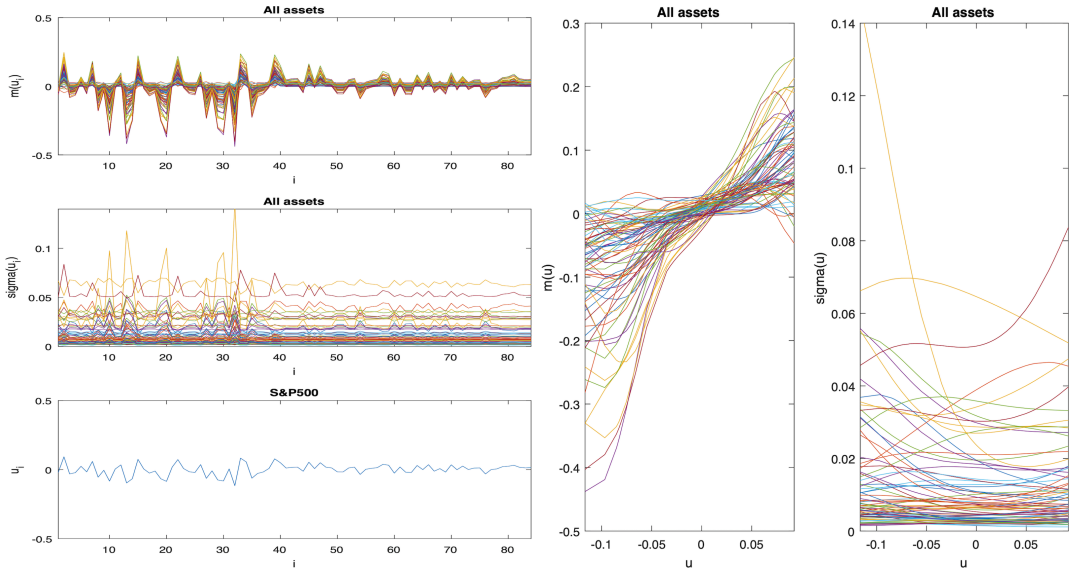
$$a_h(u) = \left\langle I_p, \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]\right\rangle - \rho_h^{-1}(u)\left\langle I_p, \mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right] - \Sigma(u)\right\rangle.$$

Therefore, the optimal solution $\hat{\Sigma}^*(u)$ to the above covariance optimization has the form:

$$\hat{\Sigma}^*(u) = \rho_h(u)a_h(u)I_p + (1-\rho_h(u))\hat{\Sigma}^{(t)}(u).$$

Note that $\alpha_h^2(u)$, $\beta_h^2(u)$ and $Q_h(u)$ in (A.3) depend on the unknown matrices $\mathrm{E}\left[\hat{\Sigma}^{(t)}(u)\right]$ and $\Sigma(u)$. So, in Step 2, we estimate them by the plug-in estimators $\hat{\alpha}_p^2(u)$ and $\hat{\beta}_p^2(u)$. It is easy to see that $\hat{\beta}_p^2(u)$ is the squared Frobenius-norm of

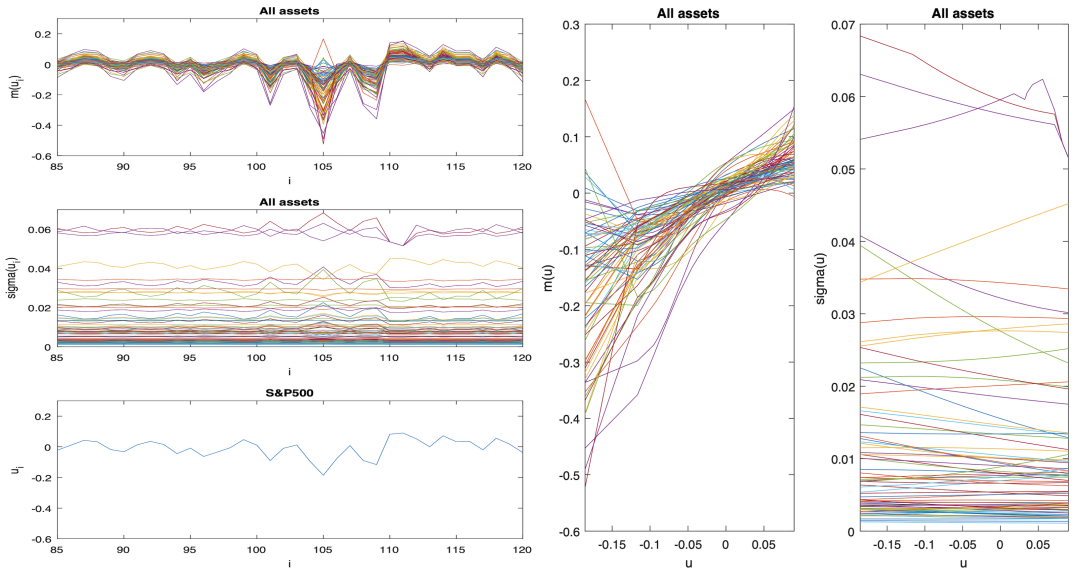
the variance estimators of $\hat{\sigma}_{jk}(u)$'s. For simplicity, we shrink $Q_h(u)$ to zero, since $Q_h(u) = o(\alpha_h^2(u))$. Combining the above two steps gives the desired estimator of $\Sigma(u)$. The derivation is completed.




*(a)* *(b)*

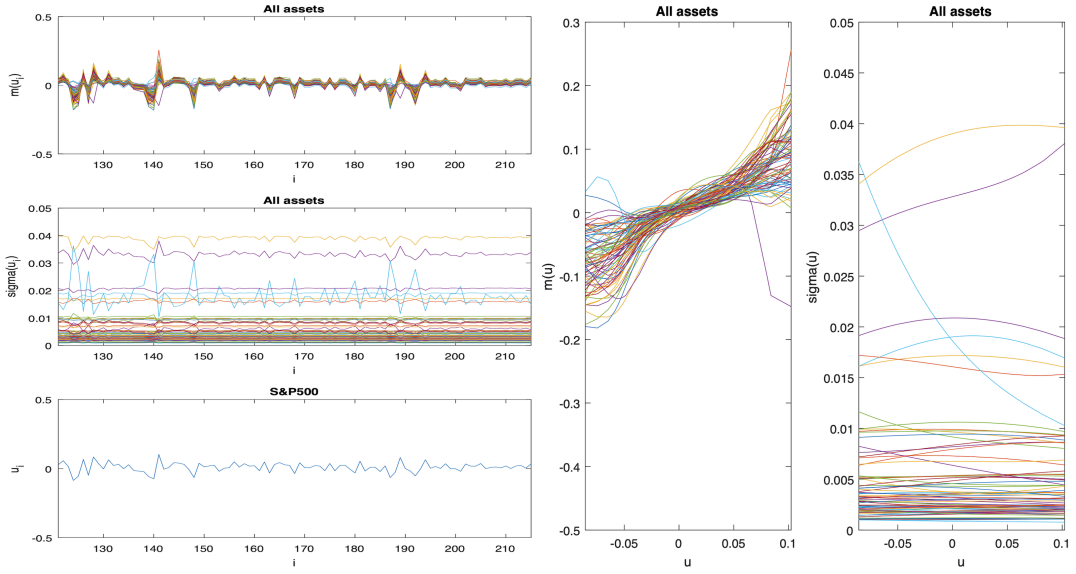
Note: Before-financial-crisis: (a) Plots of estimated means $\hat{\mu}_k(u_i)$ against $i$ (top), estimated individual volatility $\hat{\sigma}_{kk}(u_i)$ against $i$ (middle) and $u_i$ against $i$ (bottom). (b) Plots of estimated $\hat{\mu}_k(u)$ against $u$ (left) and estimated individual volatility $\hat{\sigma}_{kk}(u)$ against $u$ right. Similarly, (a) and (b) in Figure A.2 for the in-financial-crisis period while (a) and (b) in Figure A.3 for the after-financial-crisis.

*Figure A.1: Before-financial-crisis*




*(a)* *(b)*

*Figure A.2: In-financial-crisis*




*(a)* *(b)*

*Figure A.3: After-financial-crisis*

## A.2 Tables

*Table A.1: The Average (standard error in %) of Frobenius Norm-based IRSE for Setting 1 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | ${}_{\text{t}}\text{NCM}_0$ | ${}_{\text{st}}\text{NCM}_0$ | ${}_{\text{t}}\text{NCM}_1$ | ${}_{\text{st}}\text{NCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | $\rho = 0.3$ | | | |
| 100 | 50 | 5.7885(28.45) | 0.6171(3.45) | 0.5952(3.16) | 0.4934(2.92) | 0.4816(2.98) | 0.4870(2.87) | 0.4745(2.87) |
| | 100 | 18.6343(54.94) | 0.6686(3.01) | 0.6520(2.86) | 0.5375(2.45) | 0.5263(2.48) | 0.5312(2.32) | 0.5189(2.32) |
| | 150 | 55.3409(74.86) | 0.7056(2.84) | 0.6905(2.79) | 0.5627(2.12) | 0.5512(2.23) | 0.5552(1.94) | 0.5424(2.03) |
| | 300 | 80.5517(48.35) | 0.7648(2.06) | 0.7522(2.03) | 0.6157(1.62) | 0.6047(1.65) | 0.6025(1.54) | 0.5898(1.54) |
| | 500 | 102.5949(50.96) | 0.8198(2.95) | 0.8084(2.95) | 0.6532(1.08) | 0.6422(1.10) | 0.6369(0.99) | 0.6239(1.00) |
| 200 | 50 | 3.0460(9.37) | 0.3980(2.51) | 0.3887(2.52) | 0.3079(2.12) | 0.3069(2.18) | 0.3066(2.10) | 0.3054(2.15) |
| | 100 | 8.1745(15.68) | 0.4188(1.84) | 0.4114(1.82) | 0.3239(1.51) | 0.3229(1.54) | 0.3228(1.51) | 0.3214(1.53) |
| | 150 | 17.6292(28.08) | 0.4274(1.80) | 0.4210(1.73) | 0.3310(1.47) | 0.3294(1.47) | 0.3305(1.47) | 0.3286(1.46) |
| | 300 | 72.9782(32.37) | 0.4608(1.52) | 0.4561(1.51) | 0.3533(1.01) | 0.3510(1.03) | 0.3530(0.96) | 0.3503(0.98) |
| | 500 | 93.0913(29.22) | 0.4972(1.21) | 0.4934(1.20) | 0.3697(0.83) | 0.3671(0.86) | 0.3695(0.82) | 0.3665(0.84) |
| 500 | 50 | 1.5613(4.10) | 0.2161(1.23) | 0.2147(1.25) | 0.1902(0.97) | 0.1915(1.04) | 0.1893(0.94) | 0.1906(1.00) |
| | 100 | 3.3541(4.84) | 0.2191(1.02) | 0.2182(1.01) | 0.1904(0.75) | 0.1917(0.77) | 0.1894(0.74) | 0.1907(0.75) |
| | 150 | 6.4276(6.60) | 0.2250(0.95) | 0.2243(0.95) | 0.1918(0.72) | 0.1930(0.73) | 0.1910(0.71) | 0.1922(0.72) |
| | 300 | 27.2019(24.76) | 0.2416(0.60) | 0.2421(0.61) | 0.1932(0.48) | 0.1946(0.49) | 0.1923(0.48) | 0.1938(0.48) |
| | 500 | 92.5289(25.49) | 0.2630(0.43) | 0.2641(0.43) | 0.1932(0.38) | 0.1946(0.40) | 0.1924(0.39) | 0.1937(0.40) |

*Table A.2: The Average (standard error in %) of Frobenius Norm-based IRSE for Setting 1 (continued)*

| $n$ | $p$ | $\mathrm{DCM}_2$ | $\mathrm{DCM}_1$ | $\mathrm{sDCM}_1$ | ${}_{\mathrm{t}}\mathrm{NCM}_0$ | ${}_{\mathrm{st}}\mathrm{NCM}_0$ | ${}_{\mathrm{t}}\mathrm{NCM}_1$ | ${}_{\mathrm{st}}\mathrm{NCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0.8$ | | | | |
| 100 | 50 | 5.6548(36.49) | 1.3968(12.61) | 1.1879(10.99) | 1.3139(11.78) | 1.1319(10.38) | 1.3122(11.89) | 1.1298(10.49) |
| | 100 | 18.0835(51.43) | 1.8898(12.64) | 1.5739(11.18) | 1.7843(11.77) | 1.4853(10.51) | 1.7828(11.83) | 1.4834(10.57) |
| | 150 | 55.7814(61.03) | 2.3268(12.93) | 1.9282(11.72) | 2.1991(12.25) | 1.8063(10.94) | 2.1979(12.29) | 1.8048(10.99) |
| | 300 | 80.8167(41.57) | 3.2234(12.61) | 2.6555(11.55) | 3.0622(11.99) | 2.4755(10.86) | 3.0612(12.01) | 2.4742(10.89) |
| | 500 | 102.7016(42.63) | 4.1306(10.81) | 3.3822(10.04) | 3.9262(10.21) | 3.1492(9.36) | 3.9253(10.22) | 3.1482(9.37) |
| 200 | 50 | 3.0830(14.28) | 1.1094(7.83) | 0.9556(6.63) | 1.0305(7.09) | 0.9078(6.20) | 1.0298(7.16) | 0.9069(6.28) |
| | 100 | 8.1824(19.64) | 1.5004(7.31) | 1.2544(6.30) | 1.4002(6.77) | 1.1875(5.95) | 1.3995(6.82) | 1.1866(6.00) |
| | 150 | 17.3394(30.18) | 1.8256(6.69) | 1.5113(5.94) | 1.7130(6.20) | 1.4254(5.56) | 1.7124(6.22) | 1.4248(5.60) |
| | 300 | 73.0011(32.85) | 2.5373(5.88) | 2.0875(5.42) | 2.4022(5.54) | 1.9570(5.04) | 2.4017(5.56) | 1.9564(5.06) |
| | 500 | 93.1479(26.67) | 3.2475(6.85) | 2.6655(6.26) | 3.0812(6.45) | 2.4830(5.83) | 3.0807(6.45) | 2.4825(5.83) |
| 500 | 50 | 1.6169(6.21) | 0.7379(4.05) | 0.6636(3.19) | 0.6859(3.19) | 0.6324(2.70) | 0.6856(3.23) | 0.6320(2.74) |
| | 100 | 3.3779(6.24) | 1.0280(3.19) | 0.8847(2.51) | 0.9495(2.79) | 0.8410(2.40) | 0.9493(2.81) | 0.8407(2.43) |
| | 150 | 6.4755(8.86) | 1.2453(2.95) | 1.0496(2.46) | 1.1566(2.80) | 1.0007(2.46) | 1.1564(2.81) | 1.0004(2.47) |
| | 300 | 26.5926(28.80) | 1.7210(3.00) | 1.4213(2.67) | 1.6184(2.84) | 1.3545(2.53) | 1.6182(2.85) | 1.3544(2.54) |
| | 500 | 92.7346(25.02) | 2.1956(2.68) | 1.8023(2.34) | 2.0834(2.54) | 1.7137(2.22) | 2.0832(2.55) | 1.7135(2.22) |

*Table A.3: The Average (standard error in %) of Frobenius norm-based IRSE for Setting 2*

| $n$ | $p$ | $\mathrm{DCM}_2$ | $\mathrm{DCM}_1$ | $\mathrm{sDCM}_1$ | ${}_{\mathrm{t}}\mathrm{NCM}_0$ | ${}_{\mathrm{st}}\mathrm{NCM}_0$ | ${}_{\mathrm{t}}\mathrm{NCM}_1$ | ${}_{\mathrm{st}}\mathrm{NCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0$ | | | | |
| 100 | 50 | 11.8865(25.91) | 0.5261(1.94) | 0.5019(1.99) | 0.4534(2.29) | 0.4338(2.32) | 0.4322(2.47) | 0.4117(2.41) |
| | 100 | 37.3349(126.67) | 0.5494(1.44) | 0.5281(1.50) | 0.4834(1.48) | 0.4642(1.61) | 0.4563(1.92) | 0.4357(1.87) |
| | 150 | 62.9012(59.70) | 0.5609(1.49) | 0.5415(1.54) | 0.4980(1.61) | 0.4797(1.68) | 0.4702(1.53) | 0.4501(1.52) |
| | 300 | 87.7651(59.14) | 0.5902(1.83) | 0.5738(1.88) | 0.5248(1.07) | 0.5077(1.13) | 0.4977(1.30) | 0.4780(1.31) |
| | 500 | 114.8459(50.70) | 0.6096(1.71) | 0.5942(1.74) | 0.5449(0.71) | 0.5286(0.76) | 0.5193(0.80) | 0.5000(0.80) |
| 200 | 50 | 6.1169(21.48) | 0.3867(1.86) | 0.3719(1.81) | 0.3177(1.61) | 0.3095(1.64) | 0.3078(1.59) | 0.2996(1.57) |
| | 100 | 16.9859(23.04) | 0.3995(1.14) | 0.3869(1.15) | 0.3277(1.10) | 0.3195(1.11) | 0.3146(1.15) | 0.3063(1.10) |
| | 150 | 35.6152(50.72) | 0.4049(1.16) | 0.3934(1.15) | 0.3336(0.96) | 0.3253(0.96) | 0.3170(0.94) | 0.3087(0.90) |
| | 300 | 90.4787(44.91) | 0.4235(1.05) | 0.4142(1.05) | 0.3447(0.80) | 0.3364(0.82) | 0.3238(0.84) | 0.3153(0.82) |
| | 500 | 116.3211(39.69) | 0.4442(0.89) | 0.4358(0.89) | 0.3529(0.71) | 0.3446(0.71) | 0.3293(0.72) | 0.3207(0.71) |
| 500 | 50 | 2.8192(7.88) | 0.2626(0.89) | 0.2576(0.93) | 0.2367(0.94) | 0.2341(1.02) | 0.2350(0.94) | 0.2324(1.01) |
| | 100 | 7.2953(8.71) | 0.2672(0.49) | 0.2632(0.52) | 0.2420(0.51) | 0.2395(0.56) | 0.2403(0.50) | 0.2378(0.55) |
| | 150 | 13.8324(9.62) | 0.2708(0.44) | 0.2675(0.45) | 0.2453(0.35) | 0.2430(0.38) | 0.2437(0.35) | 0.2414(0.37) |
| | 300 | 84.4015(80.99) | 0.2825(0.30) | 0.2803(0.30) | 0.2478(0.24) | 0.2457(0.25) | 0.2462(0.23) | 0.2441(0.24) |
| | 500 | 117.4588(26.18) | 0.2974(0.23) | 0.2960(0.24) | 0.2488(0.17) | 0.2469(0.18) | 0.2471(0.17) | 0.2452(0.18) |

*Table A.4: The Average (standard error in %) of Frobenius Norm-based IRSE for Setting 2 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0.3$ | | | | |
| 100 | 50 | 11.9070(26.32) | 0.5488(2.29) | 0.5224(2.30) | 0.4846(3.00) | 0.4628(2.80) | 0.4660(3.27) | 0.4431(2.97) |
| | 100 | 37.5512(133.29) | 0.5643(1.64) | 0.5426(1.63) | 0.5049(1.65) | 0.4850(1.71) | 0.4871(1.92) | 0.4653(1.88) |
| | 150 | 62.9160(54.14) | 0.5831(1.59) | 0.5631(1.60) | 0.5190(1.36) | 0.4999(1.45) | 0.5029(1.47) | 0.4815(1.45) |
| | 300 | 87.6334(53.64) | 0.6114(1.87) | 0.5943(1.91) | 0.5456(0.83) | 0.5280(0.93) | 0.5333(1.03) | 0.5128(1.06) |
| | 500 | 114.8919(59.94) | 0.6266(2.22) | 0.6105(2.23) | 0.5603(0.71) | 0.5433(0.74) | 0.5542(0.79) | 0.5339(0.79) |
| 200 | 50 | 6.2194(21.03) | 0.4067(1.68) | 0.3894(1.61) | 0.3343(1.71) | 0.3243(1.63) | 0.3256(1.83) | 0.3154(1.67) |
| | 100 | 16.9228(21.61) | 0.4186(1.61) | 0.4044(1.58) | 0.3465(1.61) | 0.3364(1.55) | 0.3342(1.67) | 0.3238(1.58) |
| | 150 | 35.6480(52.16) | 0.4275(1.14) | 0.4144(1.12) | 0.3534(0.93) | 0.3434(0.89) | 0.3374(1.05) | 0.3271(0.98) |
| | 300 | 90.4008(44.07) | 0.4491(1.17) | 0.4381(1.19) | 0.3677(0.80) | 0.3575(0.79) | 0.3522(0.94) | 0.3414(0.89) |
| | 500 | 116.3462(37.37) | 0.4703(1.21) | 0.4605(1.22) | 0.3787(0.78) | 0.3688(0.78) | 0.3608(0.88) | 0.3501(0.85) |
| 500 | 50 | 2.8119(8.47) | 0.2714(0.83) | 0.2658(0.86) | 0.2456(0.77) | 0.2423(0.81) | 0.2440(0.76) | 0.2407(0.79) |
| | 100 | 7.3496(7.72) | 0.2751(0.65) | 0.2706(0.65) | 0.2495(0.53) | 0.2467(0.54) | 0.2479(0.52) | 0.2450(0.53) |
| | 150 | 13.7107(9.75) | 0.2798(0.53) | 0.2759(0.52) | 0.2519(0.40) | 0.2492(0.40) | 0.2502(0.40) | 0.2475(0.39) |
| | 300 | 84.6477(54.83) | 0.2909(0.37) | 0.2883(0.37) | 0.2536(0.29) | 0.2514(0.29) | 0.2517(0.27) | 0.2495(0.27) |
| | 500 | 117.4813(26.74) | 0.3046(0.29) | 0.3028(0.28) | 0.2539(0.22) | 0.2517(0.22) | 0.2519(0.22) | 0.2498(0.22) |

*Table A.5: The Average (standard error in %) of Frobenius Norm-based IRSE for Setting 2 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0.8$ | | | | |
| 100 | 50 | 11.8065(29.64) | 1.3110(10.46) | 1.0854(8.86) | 1.2273(9.32) | 1.0305(8.07) | 1.2246(9.41) | 1.0271(8.17) |
| | 100 | 37.5451(152.25) | 1.7824(11.25) | 1.4565(10.03) | 1.6758(10.54) | 1.3707(9.42) | 1.6737(10.59) | 1.3680(9.48) |
| | 150 | 62.8919(56.96) | 2.1604(11.16) | 1.7605(10.14) | 2.0378(10.50) | 1.6487(9.43) | 2.0361(10.54) | 1.6465(9.47) |
| | 300 | 87.5892(52.03) | 3.0189(9.12) | 2.4601(8.38) | 2.8635(8.70) | 2.2906(7.94) | 2.8622(8.71) | 2.2889(7.95) |
| | 500 | 114.9517(56.58) | 3.8808(9.67) | 3.1538(9.00) | 3.6842(9.39) | 2.9324(8.65) | 3.6832(9.40) | 2.9308(8.66) |
| 200 | 50 | 6.0032(26.41) | 1.0720(5.62) | 0.8902(4.94) | 0.9904(5.28) | 0.8423(4.79) | 0.9891(5.33) | 0.8407(4.84) |
| | 100 | 16.6911(29.23) | 1.4519(5.47) | 1.1827(4.82) | 1.3521(5.27) | 1.1169(4.70) | 1.3510(5.30) | 1.1155(4.73) |
| | 150 | 36.6918(64.70) | 1.7727(5.91) | 1.4389(5.35) | 1.6604(5.57) | 1.3564(5.07) | 1.6595(5.58) | 1.3553(5.09) |
| | 300 | 90.4006(45.64) | 2.4649(4.91) | 2.0040(4.49) | 2.3305(4.62) | 1.8768(4.21) | 2.3298(4.63) | 1.8759(4.23) |
| | 500 | 116.2670(38.76) | 3.1504(5.99) | 2.5646(5.54) | 2.9914(5.71) | 2.3935(5.23) | 2.9907(5.71) | 2.3927(5.25) |
| 500 | 50 | 2.7977(10.20) | 0.7284(3.24) | 0.6240(2.50) | 0.6723(2.67) | 0.5950(2.35) | 0.6717(2.71) | 0.5943(2.39) |
| | 100 | 7.1270(11.76) | 1.0120(2.68) | 0.8334(2.24) | 0.9328(2.50) | 0.7952(2.27) | 0.9323(2.52) | 0.7946(2.29) |
| | 150 | 13.3994(13.79) | 1.2240(2.46) | 0.9963(2.17) | 1.1367(2.35) | 0.9520(2.15) | 1.1363(2.36) | 0.9515(2.16) |
| | 300 | 82.8582(86.72) | 1.6915(2.56) | 1.3677(2.31) | 1.5929(2.42) | 1.3054(2.20) | 1.5926(2.42) | 1.3050(2.21) |
| | 500 | 117.4724(26.92) | 2.1614(2.18) | 1.7491(1.93) | 2.0525(2.09) | 1.6643(1.85) | 2.0522(2.09) | 1.6638(1.85) |

*Table A.6: The Average (standard error in %) of Frobenius Norm-based IRSE for Setting 3*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | $\rho = 0$ | | | |
| 100 | 50 | 7.2002(40.09) | 3.5156(16.37) | 3.4092(13.85) | 3.0263(13.51) | 2.9973(14.18) | 2.9914(22.51) | 2.9435(22.49) |
| | 100 | 16.4443(45.65) | 3.8041(7.26) | 3.7854(6.61) | 3.2530(5.51) | 3.2200(5.89) | 3.1908(5.74) | 3.1422(6.01) |
| | 150 | 50.7710(57.89) | 3.9366(6.49) | 3.9182(6.12) | 3.3912(4.14) | 3.3586(4.32) | 3.3442(4.00) | 3.2950(4.11) |
| | 300 | 78.2423(68.10) | 4.1044(6.24) | 4.0766(5.86) | 3.5757(2.97) | 3.5416(2.92) | 3.5215(2.89) | 3.4716(2.90) |
| | 500 | 102.9735(69.49) | 4.2842(7.93) | 4.2468(7.81) | 3.7076(2.20) | 3.6691(2.18) | 3.6139(2.20) | 3.5627(2.25) |
| 200 | 50 | 4.7672(22.41) | 2.7869(12.92) | 2.7160(10.91) | 2.3683(9.90) | 2.3456(10.21) | 2.2827(8.97) | 2.2546(9.12) |
| | 100 | 10.3670(27.26) | 2.9617(9.63) | 2.9178(8.62) | 2.5836(5.53) | 2.5661(5.74) | 2.4739(4.69) | 2.4487(4.81) |
| | 150 | 16.7314(25.15) | 3.5027(5.79) | 3.4986(5.75) | 2.7180(4.50) | 2.6976(4.64) | 2.5932(3.50) | 2.5661(3.65) |
| | 300 | 71.1626(46.56) | 3.7042(3.68) | 3.6938(3.57) | 2.9883(2.97) | 2.9567(3.05) | 2.8087(2.45) | 2.7782(2.52) |
| | 500 | 85.0637(43.23) | 3.9768(4.48) | 3.9578(4.40) | 3.1697(2.12) | 3.1307(2.17) | 2.9480(1.79) | 2.9156(1.82) |
| 500 | 50 | 3.0828(9.90) | 2.0581(6.36) | 2.0208(5.87) | 1.6115(6.19) | 1.6079(6.34) | 1.5504(5.23) | 1.5457(5.35) |
| | 100 | 5.4027(12.83) | 2.1808(5.77) | 2.1587(5.34) | 1.8712(4.35) | 1.8677(4.43) | 1.7403(3.81) | 1.7349(3.85) |
| | 150 | 7.9711(9.29) | 2.2271(4.73) | 2.2119(4.47) | 2.0339(3.33) | 2.0311(3.41) | 1.8751(2.92) | 1.8697(3.00) |
| | 300 | 20.0680(14.89) | 2.3369(3.75) | 2.3315(3.60) | 2.3138(2.21) | 2.3083(2.26) | 2.1054(1.95) | 2.0984(1.99) |
| | 500 | 90.3515(26.78) | 3.2536(1.29) | 3.2554(1.31) | 2.5211(1.75) | 2.5125(1.79) | 2.2781(1.59) | 2.2692(1.63) |

*Table A.7: The Average (standard error in %) of Frobenius Norm-based IRSE for Setting 3 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0.3$ | | | | |
| 100 | 50 | 7.9703(43.26) | 3.5611(18.76) | 3.4437(15.56) | 3.0472(11.33) | 3.0116(11.88) | 2.9809(7.19) | 2.9300(7.52) |
| | 100 | 18.1734(47.03) | 3.8359(8.43) | 3.8136(7.77) | 3.2654(5.16) | 3.2310(5.51) | 3.2208(5.52) | 3.1713(5.72) |
| | 150 | 56.4342(68.37) | 3.9699(7.78) | 3.9492(7.23) | 3.4091(4.33) | 3.3762(4.61) | 3.3814(4.58) | 3.3317(4.71) |
| | 300 | 80.4879(72.18) | 4.1319(7.47) | 4.1037(7.09) | 3.5784(2.85) | 3.5443(2.84) | 3.5553(2.67) | 3.5046(2.65) |
| | 500 | 102.5341(77.28) | 4.2849(9.87) | 4.2484(9.76) | 3.6996(2.33) | 3.6641(2.28) | 3.6535(2.15) | 3.6020(2.15) |
| 200 | 50 | 4.7921(20.95) | 2.8340(11.46) | 2.7577(9.68) | 2.3977(8.45) | 2.3708(8.79) | 2.3188(7.62) | 2.2867(7.87) |
| | 100 | 9.8312(23.52) | 3.0478(9.96) | 2.9959(9.15) | 2.6016(5.75) | 2.5786(5.94) | 2.5043(4.64) | 2.4745(4.75) |
| | 150 | 16.2014(23.87) | 3.5476(6.40) | 3.5398(6.53) | 2.7332(4.92) | 2.7097(5.05) | 2.6156(4.42) | 2.5863(4.50) |
| | 300 | 73.1892(46.65) | 3.7509(4.18) | 3.7385(4.09) | 2.9873(3.07) | 2.9555(3.13) | 2.8271(2.67) | 2.7953(2.69) |
| | 500 | 93.0226(41.03) | 3.9992(4.87) | 3.9801(4.80) | 3.1546(2.72) | 3.1171(2.80) | 2.9657(1.91) | 2.9327(1.99) |
| 500 | 50 | 2.9875(12.13) | 2.0968(8.05) | 2.0573(7.09) | 1.6410(5.65) | 1.6351(5.76) | 1.5813(4.96) | 1.5746(4.95) |
| | 100 | 5.3832(15.91) | 2.2157(6.10) | 2.1910(5.51) | 1.8861(4.43) | 1.8810(4.58) | 1.7637(3.79) | 1.7569(3.88) |
| | 150 | 7.9702(7.86) | 2.2675(4.99) | 2.2508(4.73) | 2.0482(3.52) | 2.0437(3.60) | 1.8954(3.01) | 1.8884(3.10) |
| | 300 | 18.9870(14.29) | 2.3773(4.36) | 2.3707(4.23) | 2.3161(2.38) | 2.3101(2.43) | 2.1223(2.05) | 2.1148(2.10) |
| | 500 | 91.6005(30.58) | 3.2760(1.79) | 3.2772(1.81) | 2.5101(1.64) | 2.5011(1.67) | 2.2881(1.49) | 2.2790(1.53) |

*Table A.8: The Average (standard error in %) of Frobenius Norm-based IRSE for Setting 3 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | ${}_{\text{t}}\text{NCM}_0$ | ${}_{\text{st}}\text{NCM}_0$ | ${}_{\text{t}}\text{NCM}_1$ | ${}_{\text{st}}\text{NCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | $\rho = 0.8$ | | | |
| 100 | 50 | 8.2587(57.99) | 4.6369(50.98) | 4.1911(35.72) | 4.1809(30.17) | 3.7168(17.94) | 4.1655(30.88) | 3.6989(18.77) |
| | 100 | 18.1379(49.44) | 5.9809(40.96) | 5.1925(28.01) | 5.2965(31.18) | 4.4277(18.98) | 5.2827(31.04) | 4.4106(19.12) |
| | 150 | 50.7099(628.67) | 7.1100(50.62) | 5.9898(37.28) | 6.2984(34.98) | 5.0698(22.49) | 6.2891(35.05) | 5.0575(22.87) |
| | 300 | 78.9593(66.06) | 9.3633(36.83) | 7.4882(28.51) | 8.4308(34.41) | 6.4292(23.22) | 8.4141(34.33) | 6.4097(23.30) |
| | 500 | 101.1399(67.26) | 11.5664(33.17) | 9.0710(28.03) | 10.5839(29.16) | 7.8342(22.32) | 10.5646(29.39) | 7.8125(22.67) |
| 200 | 50 | 5.5375(37.75) | 3.8373(39.96) | 3.5554(28.96) | 3.5263(20.55) | 3.1336(13.01) | 3.4737(22.94) | 3.0878(15.32) |
| | 100 | 10.3672(25.50) | 4.8094(44.44) | 4.4571(31.96) | 4.5007(17.61) | 3.7962(10.82) | 4.4291(19.16) | 3.7320(12.44) |
| | 150 | 16.4920(28.07) | 6.2207(44.57) | 5.3995(32.21) | 5.3075(15.38) | 4.3213(9.67) | 5.2388(16.68) | 4.2620(11.03) |
| | 300 | 70.7900(46.41) | 8.2310(20.65) | 6.6645(16.62) | 7.0400(15.83) | 5.4331(11.33) | 6.9690(16.48) | 5.3750(12.09) |
| | 500 | 91.2056(37.82) | 10.1027(20.81) | 7.9993(16.90) | 8.8042(17.84) | 6.5836(13.20) | 8.7358(18.29) | 6.5305(13.80) |
| 500 | 50 | 3.4523(21.60) | 2.6353(20.05) | 2.5270(16.89) | 2.4554(14.28) | 2.3046(10.97) | 2.4273(15.32) | 2.2800(11.91) |
| | 100 | 6.0072(16.10) | 2.9998(20.64) | 2.9297(18.78) | 3.2643(9.77) | 2.9233(5.64) | 3.2120(10.94) | 2.8790(6.81) |
| | 150 | 8.2472(11.43) | 3.3002(18.75) | 3.2351(17.25) | 3.8406(7.02) | 3.3285(4.37) | 3.7832(7.80) | 3.2831(5.12) |
| | 300 | 19.1704(15.33) | 4.2317(36.24) | 4.1544(32.15) | 5.0320(7.66) | 4.1276(5.25) | 4.9842(8.50) | 4.0989(6.17) |
| | 500 | 40.0499(33.96) | 7.2277(8.40) | 5.9138(6.42) | 6.2086(7.10) | 4.9125(4.87) | 6.1813(7.80) | 4.9114(5.65) |

*Table A.9: The Average SEN, SPE and ACC for Setting 1*

| $n$ | $p$ | $\rho = 0$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.7093 | 0.9931 | 0.9654 | 0.8674 | 0.9938 | 0.9815 |
| | 100 | 0.6660 | 0.9976 | 0.9812 | 0.8093 | 0.9976 | 0.9883 |
| | 150 | 0.6410 | 0.9987 | 0.9869 | 0.7723 | 0.9987 | 0.9913 |
| | 300 | 0.5782 | 0.9996 | 0.9926 | 0.6839 | 0.9996 | 0.9944 |
| | 500 | 0.5152 | 0.9998 | 0.9950 | 0.6240 | 0.9998 | 0.9961 |
| 200 | 50 | 0.9074 | 0.9949 | 0.9863 | 0.9790 | 0.9984 | 0.9965 |
| | 100 | 0.8924 | 0.9980 | 0.9928 | 0.9684 | 0.9991 | 0.9976 |
| | 150 | 0.8936 | 0.9989 | 0.9954 | 0.9620 | 0.9994 | 0.9982 |
| | 300 | 0.8914 | 0.9996 | 0.9978 | 0.9447 | 0.9997 | 0.9988 |
| | 500 | 0.8818 | 0.9998 | 0.9986 | 0.9316 | 0.9998 | 0.9992 |
| 500 | 50 | 0.9863 | 0.9988 | 0.9976 | 0.9998 | 1.0000 | 0.9999 |
| | 100 | 0.9875 | 0.9996 | 0.9990 | 0.9998 | 1.0000 | 1.0000 |
| | 150 | 0.9894 | 0.9998 | 0.9995 | 0.9998 | 1.0000 | 1.0000 |
| | 300 | 0.9927 | 0.9999 | 0.9998 | 0.9996 | 1.0000 | 1.0000 |
| | 500 | 0.9946 | 1.0000 | 0.9999 | 0.9996 | 1.0000 | 1.0000 |

*Table A.10: The Average SEN, SPE and ACC for Setting 1 (continued)*

| $n$ | $p$ | $\rho = 0.3$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.7088 | 0.9880 | 0.9607 | 0.8597 | 0.9885 | 0.9759 |
| | 100 | 0.6442 | 0.9963 | 0.9789 | 0.7844 | 0.9962 | 0.9858 |
| | 150 | 0.6057 | 0.9980 | 0.9850 | 0.7375 | 0.9980 | 0.9894 |
| | 300 | 0.5293 | 0.9993 | 0.9915 | 0.6436 | 0.9994 | 0.9935 |
| | 500 | 0.4623 | 0.9997 | 0.9944 | 0.5773 | 0.9997 | 0.9955 |
| 200 | 50 | 0.8940 | 0.9923 | 0.9827 | 0.9720 | 0.9960 | 0.9937 |
| | 100 | 0.8782 | 0.9969 | 0.9911 | 0.9570 | 0.9984 | 0.9963 |
| | 150 | 0.8764 | 0.9983 | 0.9943 | 0.9469 | 0.9990 | 0.9973 |
| | 300 | 0.8660 | 0.9992 | 0.9970 | 0.9248 | 0.9995 | 0.9983 |
| | 500 | 0.8486 | 0.9996 | 0.9981 | 0.9053 | 0.9997 | 0.9988 |
| 500 | 50 | 0.9838 | 0.9981 | 0.9967 | 0.9998 | 0.9998 | 0.9998 |
| | 100 | 0.9847 | 0.9993 | 0.9986 | 0.9996 | 0.9999 | 0.9999 |
| | 150 | 0.9858 | 0.9997 | 0.9992 | 0.9994 | 1.0000 | 0.9999 |
| | 300 | 0.9895 | 0.9999 | 0.9997 | 0.9989 | 1.0000 | 1.0000 |
| | 500 | 0.9914 | 1.0000 | 0.9999 | 0.9988 | 1.0000 | 1.0000 |

*Table A.11: The Average SEN, SPE and ACC for Setting 1 (continued)*

| $n$ | $p$ | $\rho = 0.8$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.9835 | 0.1074 | 0.1929 | 0.9895 | 0.0802 | 0.1690 |
| | 100 | 0.9930 | 0.0517 | 0.0982 | 0.9943 | 0.0441 | 0.0910 |
| | 150 | 0.9957 | 0.0341 | 0.0659 | 0.9960 | 0.0314 | 0.0633 |
| | 300 | 0.9972 | 0.0213 | 0.0375 | 0.9977 | 0.0197 | 0.0359 |
| | 500 | 0.9982 | 0.0147 | 0.0245 | 0.9985 | 0.0136 | 0.0234 |
| 200 | 50 | 0.9935 | 0.1367 | 0.2203 | 0.9979 | 0.0810 | 0.1705 |
| | 100 | 0.9976 | 0.0651 | 0.1112 | 0.9988 | 0.0490 | 0.0959 |
| | 150 | 0.9986 | 0.0435 | 0.0751 | 0.9992 | 0.0318 | 0.0638 |
| | 300 | 0.9993 | 0.0224 | 0.0386 | 0.9995 | 0.0200 | 0.0363 |
| | 500 | 0.9996 | 0.0153 | 0.0251 | 0.9997 | 0.0137 | 0.0236 |
| 500 | 50 | 0.9971 | 0.2226 | 0.2982 | 1.0000 | 0.0840 | 0.1734 |
| | 100 | 0.9997 | 0.1050 | 0.1492 | 1.0000 | 0.0522 | 0.0990 |
| | 150 | 0.9999 | 0.0626 | 0.0936 | 1.0000 | 0.0356 | 0.0675 |
| | 300 | 1.0000 | 0.0280 | 0.0442 | 1.0000 | 0.0204 | 0.0367 |
| | 500 | 1.0000 | 0.0174 | 0.0272 | 1.0000 | 0.0150 | 0.0248 |

*Table A.12: The Average SEN, SPE and ACC for Setting 2*

| $n$ | $p$ | $\rho = 0$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.0425 | 0/0 | 0.0425 | 0.0583 | 0/0 | 0.0583 |
| | 100 | 0.0189 | 0/0 | 0.0189 | 0.0251 | 0/0 | 0.0251 |
| | 150 | 0.0120 | 0/0 | 0.0120 | 0.0154 | 0/0 | 0.0154 |
| | 300 | 0.0054 | 0/0 | 0.0054 | 0.0065 | 0/0 | 0.0065 |
| | 500 | 0.0029 | 0/0 | 0.0029 | 0.0034 | 0/0 | 0.0034 |
| 200 | 50 | 0.0580 | 0/0 | 0.0580 | 0.0647 | 0/0 | 0.0647 |
| | 100 | 0.0280 | 0/0 | 0.0280 | 0.0314 | 0/0 | 0.0314 |
| | 150 | 0.0186 | 0/0 | 0.0186 | 0.0206 | 0/0 | 0.0206 |
| | 300 | 0.0090 | 0/0 | 0.0090 | 0.0099 | 0/0 | 0.0099 |
| | 500 | 0.0053 | 0/0 | 0.0053 | 0.0058 | 0/0 | 0.0058 |
| 500 | 50 | 0.0659 | 0/0 | 0.0659 | 0.0747 | 0/0 | 0.0747 |
| | 100 | 0.0320 | 0/0 | 0.0320 | 0.0350 | 0/0 | 0.0350 |
| | 150 | 0.0212 | 0/0 | 0.0212 | 0.0225 | 0/0 | 0.0225 |
| | 300 | 0.0104 | 0/0 | 0.0104 | 0.0107 | 0/0 | 0.0107 |
| | 500 | 0.0062 | 0/0 | 0.0062 | 0.0063 | 0/0 | 0.0063 |

*Table A.13: The Average SEN, SPE and ACC for Setting 2 (continued)*

| $n$ | $p$ | $\rho = 0.3$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.0419 | 0/0 | 0.0419 | 0.0591 | 0/0 | 0.0591 |
| | 100 | 0.0189 | 0/0 | 0.0189 | 0.0251 | 0/0 | 0.0251 |
| | 150 | 0.0117 | 0/0 | 0.0117 | 0.0150 | 0/0 | 0.0150 |
| | 300 | 0.0051 | 0/0 | 0.0051 | 0.0061 | 0/0 | 0.0061 |
| | 500 | 0.0028 | 0/0 | 0.0028 | 0.0033 | 0/0 | 0.0033 |
| 200 | 50 | 0.0604 | 0/0 | 0.0604 | 0.0668 | 0/0 | 0.0668 |
| | 100 | 0.0285 | 0/0 | 0.0285 | 0.0322 | 0/0 | 0.0322 |
| | 150 | 0.0186 | 0/0 | 0.0186 | 0.0209 | 0/0 | 0.0209 |
| | 300 | 0.0091 | 0/0 | 0.0091 | 0.0100 | 0/0 | 0.0100 |
| | 500 | 0.0052 | 0/0 | 0.0052 | 0.0057 | 0/0 | 0.0057 |
| 500 | 50 | 0.0665 | 0/0 | 0.0665 | 0.0776 | 0/0 | 0.0776 |
| | 100 | 0.0324 | 0/0 | 0.0324 | 0.0353 | 0/0 | 0.0353 |
| | 150 | 0.0214 | 0/0 | 0.0214 | 0.0226 | 0/0 | 0.0226 |
| | 300 | 0.0105 | 0/0 | 0.0105 | 0.0107 | 0/0 | 0.0107 |
| | 500 | 0.0062 | 0/0 | 0.0062 | 0.0063 | 0/0 | 0.0063 |

*Table A.14: The Average SEN, SPE and ACC for Setting 2 (continued)*

| $n$ | $p$ | $\rho = 0.8$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.8940 | 0/0 | 0.8940 | 0.9316 | 0/0 | 0.9316 |
| | 100 | 0.9485 | 0/0 | 0.9485 | 0.9610 | 0/0 | 0.9610 |
| | 150 | 0.9641 | 0/0 | 0.9641 | 0.9693 | 0/0 | 0.9693 |
| | 300 | 0.9792 | 0/0 | 0.9792 | 0.9813 | 0/0 | 0.9813 |
| | 500 | 0.9853 | 0/0 | 0.9853 | 0.9864 | 0/0 | 0.9864 |
| 200 | 50 | 0.8771 | 0/0 | 0.8771 | 0.9281 | 0/0 | 0.9281 |
| | 100 | 0.9383 | 0/0 | 0.9383 | 0.9583 | 0/0 | 0.9583 |
| | 150 | 0.9597 | 0/0 | 0.9597 | 0.9683 | 0/0 | 0.9683 |
| | 300 | 0.9778 | 0/0 | 0.9778 | 0.9811 | 0/0 | 0.9811 |
| | 500 | 0.9852 | 0/0 | 0.9852 | 0.9859 | 0/0 | 0.9859 |
| 500 | 50 | 0.8152 | 0/0 | 0.8152 | 0.9213 | 0/0 | 0.9213 |
| | 100 | 0.9157 | 0/0 | 0.9157 | 0.9503 | 0/0 | 0.9503 |
| | 150 | 0.9435 | 0/0 | 0.9435 | 0.9647 | 0/0 | 0.9647 |
| | 300 | 0.9723 | 0/0 | 0.9723 | 0.9789 | 0/0 | 0.9789 |
| | 500 | 0.9834 | 0/0 | 0.9834 | 0.9863 | 0/0 | 0.9863 |

*Table A.15: The Average SEN, SPE and ACC for Setting 3*

| $n$ | $p$ | $\rho = 0$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.2947 | 0.9914 | 0.8716 | 0.3408 | 0.9939 | 0.8816 |
| | 100 | 0.3245 | 0.9979 | 0.9387 | 0.2876 | 0.9981 | 0.9356 |
| | 150 | 0.2858 | 0.9992 | 0.9571 | 0.2531 | 0.9990 | 0.9549 |
| | 300 | 0.2360 | 0.9998 | 0.9771 | 0.2058 | 0.9996 | 0.9760 |
| | 500 | 0.2160 | 0.9999 | 0.9859 | 0.1856 | 0.9998 | 0.9852 |
| 200 | 50 | 0.3642 | 0.9939 | 0.8856 | 0.4960 | 0.9940 | 0.9083 |
| | 100 | 0.3403 | 0.9983 | 0.9404 | 0.4453 | 0.9977 | 0.9491 |
| | 150 | 0.4732 | 0.9985 | 0.9675 | 0.4300 | 0.9983 | 0.9647 |
| | 300 | 0.4757 | 0.9994 | 0.9838 | 0.3883 | 0.9991 | 0.9809 |
| | 500 | 0.4760 | 0.9997 | 0.9903 | 0.3510 | 0.9995 | 0.9878 |
| 500 | 50 | 0.4915 | 0.9931 | 0.9068 | 0.6694 | 0.9940 | 0.9381 |
| | 100 | 0.4512 | 0.9979 | 0.9497 | 0.6402 | 0.9981 | 0.9666 |
| | 150 | 0.4422 | 0.9989 | 0.9660 | 0.6155 | 0.9990 | 0.9763 |
| | 300 | 0.4776 | 0.9995 | 0.9840 | 0.5770 | 0.9994 | 0.9869 |
| | 500 | 0.6483 | 0.9998 | 0.9935 | 0.5489 | 0.9995 | 0.9914 |

*Table A.16: The Average SEN, SPE and ACC for Setting 3 (continued)*

| $n$ | $p$ | $\rho = 0.3$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.2909 | 0.9912 | 0.8707 | 0.3440 | 0.9900 | 0.8789 |
| | 100 | 0.3110 | 0.9978 | 0.9373 | 0.2866 | 0.9976 | 0.9351 |
| | 150 | 0.2727 | 0.9990 | 0.9561 | 0.2474 | 0.9988 | 0.9544 |
| | 300 | 0.2233 | 0.9998 | 0.9767 | 0.2040 | 0.9996 | 0.9759 |
| | 500 | 0.2020 | 0.9999 | 0.9856 | 0.1822 | 0.9998 | 0.9852 |
| 200 | 50 | 0.3622 | 0.9932 | 0.8847 | 0.4872 | 0.9935 | 0.9064 |
| | 100 | 0.3410 | 0.9978 | 0.9400 | 0.4471 | 0.9969 | 0.9485 |
| | 150 | 0.4648 | 0.9982 | 0.9666 | 0.4264 | 0.9978 | 0.9640 |
| | 300 | 0.4588 | 0.9992 | 0.9831 | 0.3806 | 0.9989 | 0.9805 |
| | 500 | 0.4598 | 0.9996 | 0.9899 | 0.3465 | 0.9993 | 0.9876 |
| 500 | 50 | 0.4872 | 0.9923 | 0.9054 | 0.6647 | 0.9929 | 0.9364 |
| | 100 | 0.4477 | 0.9977 | 0.9493 | 0.6311 | 0.9977 | 0.9654 |
| | 150 | 0.4418 | 0.9986 | 0.9657 | 0.6101 | 0.9986 | 0.9756 |
| | 300 | 0.4731 | 0.9994 | 0.9838 | 0.5680 | 0.9992 | 0.9864 |
| | 500 | 0.6402 | 0.9996 | 0.9932 | 0.5391 | 0.9994 | 0.9911 |

*Table A.17: The Average SEN, SPE and ACC for Setting 3 (continued)*

| $n$ | $p$ | $\rho = 0.8$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\text{DCM}_1$ and $_{\text{s}}\text{DCM}_1$ | | | $_{\text{t}}\text{NCM}_1$ and $_{\text{st}}\text{NCM}_1$ | | |
| | | SEN | SPE | ACC | SEN | SPE | ACC |
| 100 | 50 | 0.4976 | 0.7072 | 0.6712 | 0.8213 | 0.2994 | 0.3892 |
| | 100 | 0.8604 | 0.2424 | 0.2967 | 0.9117 | 0.1550 | 0.2216 |
| | 150 | 0.8883 | 0.1771 | 0.2191 | 0.9344 | 0.1094 | 0.1582 |
| | 300 | 0.9804 | 0.0520 | 0.0796 | 0.9720 | 0.0541 | 0.0814 |
| | 500 | 0.9903 | 0.0286 | 0.0458 | 0.9864 | 0.0308 | 0.0479 |
| 200 | 50 | 0.4877 | 0.8299 | 0.7710 | 0.9114 | 0.2552 | 0.3681 |
| | 100 | 0.4795 | 0.8178 | 0.7880 | 0.9563 | 0.1256 | 0.1987 |
| | 150 | 0.7759 | 0.3762 | 0.3998 | 0.9716 | 0.0809 | 0.1335 |
| | 300 | 0.9857 | 0.0503 | 0.0781 | 0.9856 | 0.0417 | 0.0698 |
| | 500 | 0.9919 | 0.0306 | 0.0478 | 0.9903 | 0.0281 | 0.0453 |
| 500 | 50 | 0.5392 | 0.9389 | 0.8702 | 0.8696 | 0.4875 | 0.5532 |
| | 100 | 0.4640 | 0.9823 | 0.9367 | 0.9512 | 0.2192 | 0.2836 |
| | 150 | 0.4564 | 0.9862 | 0.9548 | 0.9750 | 0.1162 | 0.1669 |
| | 300 | 0.4930 | 0.9796 | 0.9652 | 0.9904 | 0.0474 | 0.0755 |
| | 500 | 0.9917 | 0.0470 | 0.0639 | 0.9938 | 0.0308 | 0.0480 |

*Table A.18: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 1*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0$ | | | | |
| 100 | 50 | 33.3518(182.27) | 1.4542(11.85) | 1.3952(11.36) | 1.1688(12.54) | 1.1610(12.47) | 1.1400(11.85) | 1.1259(11.90) |
| | 100 | 159.8834(472.42) | 1.6117(9.73) | 1.5489(9.18) | 1.3504(10.60) | 1.3283(10.45) | 1.3028(9.14) | 1.2761(9.08) |
| | 150 | 603.8583(928.01) | 1.6948(8.61) | 1.6302(7.95) | 1.4346(10.17) | 1.4117(9.78) | 1.3705(8.63) | 1.3428(8.06) |
| | 300 | 1348.6526(837.11) | 1.8412(7.45) | 1.7733(6.43) | 1.5710(7.62) | 1.5451(7.15) | 1.5011(6.76) | 1.4690(6.35) |
| | 500 | 2296.6220(876.96) | 1.9485(7.21) | 1.8729(6.16) | 1.6820(5.85) | 1.6485(5.45) | 1.5972(5.78) | 1.5595(5.46) |
| 200 | 50 | 17.3622(65.65) | 1.0140(11.41) | 0.9936(11.43) | 0.7835(9.76) | 0.7828(10.04) | 0.7765(9.37) | 0.7729(9.62) |
| | 100 | 85.5035(160.45) | 1.1463(10.27) | 1.1271(9.83) | 0.8784(7.53) | 0.8775(7.46) | 0.8697(7.20) | 0.8659(7.17) |
| | 150 | 225.5423(371.28) | 1.1957(9.38) | 1.1746(9.06) | 0.9221(6.45) | 0.9187(6.45) | 0.9125(6.35) | 0.9067(6.38) |
| | 300 | 1223.9888(555.04) | 1.3408(8.24) | 1.3171(8.05) | 1.0183(6.92) | 1.0124(6.69) | 1.0036(6.45) | 0.9953(6.22) |
| | 500 | 1892.3089(569.99) | 1.4421(7.56) | 1.4142(7.30) | 1.0740(6.53) | 1.0657(6.11) | 1.0507(6.20) | 1.0397(5.76) |
| 500 | 50 | 7.9786(30.98) | 0.5769(5.85) | 0.5769(6.02) | 0.4983(5.15) | 0.5023(5.21) | 0.4945(5.13) | 0.4974(5.19) |
| | 100 | 29.9248(45.26) | 0.6442(5.36) | 0.6417(5.20) | 0.5368(3.84) | 0.5405(3.79) | 0.5328(3.84) | 0.5353(3.78) |
| | 150 | 69.4453(78.39) | 0.6969(6.58) | 0.6933(6.49) | 0.5715(4.53) | 0.5742(4.51) | 0.5679(4.58) | 0.5695(4.54) |
| | 300 | 495.0249(422.64) | 0.7759(5.39) | 0.7719(5.33) | 0.6133(4.02) | 0.6143(3.97) | 0.6098(4.03) | 0.6098(3.97) |
| | 500 | 2030.2082(460.90) | 0.8503(5.37) | 0.8457(5.30) | 0.6422(4.14) | 0.6428(3.99) | 0.6386(4.13) | 0.6379(3.97) |

*Table A.19: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 1 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | ${}_{\text{t}}\text{NCM}_0$ | ${}_{\text{st}}\text{NCM}_0$ | ${}_{\text{t}}\text{NCM}_1$ | ${}_{\text{st}}\text{NCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0.3$ | | | | |
| 100 | 50 | 38.0291(203.66) | 1.5765(17.83) | 1.4878(13.82) | 1.2672(11.81) | 1.2477(11.01) | 1.2351(11.63) | 1.2097(10.65) |
| | 100 | 184.4051(553.37) | 1.7098(12.54) | 1.6271(10.96) | 1.4365(10.37) | 1.4106(9.57) | 1.3880(9.62) | 1.3566(8.70) |
| | 150 | 676.5004(919.42) | 1.8208(13.10) | 1.7284(11.17) | 1.4961(8.72) | 1.4677(8.47) | 1.4560(7.87) | 1.4201(7.60) |
| | 300 | 1392.4414(841.55) | 1.9503(8.59) | 1.8612(6.98) | 1.6559(6.88) | 1.6243(6.48) | 1.5914(6.50) | 1.5535(6.29) |
| | 500 | 2289.5151(1140.37) | 2.0855(14.35) | 1.9836(11.62) | 1.7537(6.04) | 1.7162(5.37) | 1.6820(6.02) | 1.6392(5.22) |
| 200 | 50 | 18.6342(72.43) | 1.0596(9.12) | 1.0311(8.85) | 0.8240(7.76) | 0.8228(7.72) | 0.8176(7.64) | 0.8131(7.64) |
| | 100 | 78.7297(159.87) | 1.2160(10.22) | 1.1822(9.56) | 0.9339(8.24) | 0.9286(8.21) | 0.9260(7.89) | 0.9180(7.83) |
| | 150 | 213.8932(347.45) | 1.2594(9.26) | 1.2283(8.87) | 0.9813(8.52) | 0.9747(8.31) | 0.9680(7.50) | 0.9586(7.33) |
| | 300 | 1261.2552(562.67) | 1.4158(8.84) | 1.3843(8.54) | 1.0986(8.49) | 1.0871(8.32) | 1.0800(7.46) | 1.0658(7.30) |
| | 500 | 2077.0729(656.70) | 1.5408(8.18) | 1.5053(7.62) | 1.1658(6.10) | 1.1525(5.85) | 1.1394(5.93) | 1.1234(5.65) |
| 500 | 50 | 7.8815(35.08) | 0.6138(5.57) | 0.6123(5.42) | 0.5194(4.64) | 0.5229(4.72) | 0.5158(4.62) | 0.5181(4.68) |
| | 100 | 29.7787(48.96) | 0.6829(6.34) | 0.6798(6.08) | 0.5639(4.43) | 0.5666(4.28) | 0.5600(4.50) | 0.5614(4.36) |
| | 150 | 74.8449(75.58) | 0.7414(6.60) | 0.7364(6.49) | 0.5951(4.87) | 0.5960(4.78) | 0.5919(4.85) | 0.5915(4.76) |
| | 300 | 468.8686(429.63) | 0.8171(5.03) | 0.8124(4.95) | 0.6422(4.56) | 0.6422(4.42) | 0.6389(4.55) | 0.6376(4.40) |
| | 500 | 2064.7836(571.18) | 0.8858(5.03) | 0.8805(4.96) | 0.6722(4.12) | 0.6709(4.13) | 0.6696(4.08) | 0.6670(4.08) |

*Table A.20: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 1 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | $\rho = 0.8$ | | | |
| 100 | 50 | 36.8596(257.84) | 5.6624(112.58) | 4.5693(100.80) | 5.1430(113.75) | 4.1817(102.46) | 5.1731(113.96) | 4.2121(102.69) |
| | 100 | 178.1774(517.33) | 9.9206(135.77) | 7.9774(119.47) | 9.1877(131.77) | 7.3518(116.51) | 9.2162(131.86) | 7.3802(116.63) |
| | 150 | 681.4811(748.91) | 14.7183(175.65) | 11.9087(154.47) | 13.7770(173.05) | 11.0124(150.02) | 13.8053(173.18) | 11.0410(150.17) |
| | 300 | 1396.1786(724.92) | 27.1927(213.67) | 22.0948(186.41) | 25.7680(206.12) | 20.5154(177.49) | 25.7959(206.19) | 20.5434(177.57) |
| | 500 | 2290.4654(954.95) | 43.7580(262.87) | 35.5141(227.48) | 41.6585(255.14) | 33.0900(216.43) | 41.6859(255.19) | 33.1172(216.39) |
| 200 | 50 | 18.4530(97.24) | 3.9942(64.39) | 3.2515(57.09) | 3.5650(63.07) | 2.9618(56.77) | 3.5858(63.20) | 2.9824(56.94) |
| | 100 | 78.2292(194.31) | 6.6562(70.78) | 5.3341(61.50) | 6.0134(70.45) | 4.8720(61.90) | 6.0331(70.56) | 4.8916(62.02) |
| | 150 | 209.4273(370.72) | 9.4771(84.66) | 7.6062(73.36) | 8.6694(81.76) | 6.9711(71.29) | 8.6890(81.84) | 6.9910(71.39) |
| | 300 | 1261.0717(571.30) | 17.0229(111.23) | 13.7540(95.55) | 15.7842(98.78) | 12.6036(84.32) | 15.8035(98.83) | 12.6232(84.39) |
| | 500 | 2077.3738(599.25) | 26.6635(149.13) | 21.6283(126.73) | 24.8934(137.62) | 19.8062(115.61) | 24.9122(137.61) | 19.8251(115.56) |
| 500 | 50 | 7.9958(42.44) | 2.3173(29.94) | 1.9504(25.75) | 2.0615(28.14) | 1.7884(25.18) | 2.0749(28.21) | 1.8005(25.38) |
| | 100 | 29.4594(61.20) | 3.8539(31.63) | 3.1322(26.97) | 3.4190(31.06) | 2.8655(27.73) | 3.4315(31.08) | 2.8781(27.76) |
| | 150 | 74.6545(101.75) | 5.2741(38.48) | 4.2411(32.60) | 4.7632(40.31) | 3.9332(35.26) | 4.7760(40.33) | 3.9459(35.28) |
| | 300 | 457.4873(499.85) | 8.8302(39.47) | 7.0687(33.93) | 8.0394(40.25) | 6.5147(34.42) | 8.0519(40.26) | 6.5277(34.42) |
| | 500 | 2068.9668(559.93) | 13.2767(50.26) | 10.6681(41.85) | 12.2830(50.93) | 9.8812(42.26) | 12.2957(50.93) | 9.8939(42.30) |

*Table A.21: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 2*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0$ | | | | |
| 100 | 50 | 79.6652(152.01) | 1.3308(8.09) | 1.2711(6.96) | 1.1706(8.90) | 1.1385(7.82) | 1.1081(9.05) | 1.0729(7.87) |
| | 100 | 338.8887(901.25) | 1.4396(8.40) | 1.3662(6.88) | 1.2944(6.16) | 1.2505(5.41) | 1.2288(7.97) | 1.1819(6.89) |
| | 150 | 637.0929(687.27) | 1.5044(8.71) | 1.4253(7.19) | 1.3573(6.25) | 1.3065(5.08) | 1.2946(6.30) | 1.2422(5.36) |
| | 300 | 1206.0978(936.25) | 1.6026(12.14) | 1.5033(9.49) | 1.4373(6.01) | 1.3745(4.15) | 1.3871(7.00) | 1.3283(5.50) |
| | 500 | 1980.8181(1113.52) | 1.6644(12.33) | 1.5434(8.91) | 1.4801(4.36) | 1.4078(2.80) | 1.4630(5.00) | 1.4022(3.97) |
| 200 | 50 | 39.6853(142.34) | 1.0217(7.57) | 0.9987(7.04) | 0.8421(7.80) | 0.8399(7.41) | 0.7985(6.81) | 0.7964(6.35) |
| | 100 | 166.0280(227.60) | 1.1167(5.71) | 1.0880(5.13) | 0.9103(5.54) | 0.9023(5.36) | 0.8643(5.54) | 0.8539(5.18) |
| | 150 | 415.7393(527.09) | 1.1782(6.62) | 1.1462(6.00) | 0.9671(5.41) | 0.9526(5.04) | 0.9053(5.52) | 0.8890(4.93) |
| | 300 | 1306.9273(727.94) | 1.2630(6.63) | 1.2294(6.10) | 1.0572(5.30) | 1.0357(4.81) | 0.9823(5.66) | 0.9588(5.14) |
| | 500 | 2087.4556(813.07) | 1.3434(5.90) | 1.3029(5.22) | 1.0995(5.49) | 1.0751(4.90) | 1.0333(5.32) | 1.0060(4.61) |
| 500 | 50 | 16.5562(60.13) | 0.6816(4.56) | 0.6854(4.41) | 0.6080(3.93) | 0.6129(4.01) | 0.6013(3.91) | 0.6056(3.99) |
| | 100 | 69.6720(91.13) | 0.7485(3.82) | 0.7471(3.59) | 0.6524(2.94) | 0.6554(2.83) | 0.6446(2.98) | 0.6471(2.86) |
| | 150 | 166.4749(114.24) | 0.7960(5.09) | 0.7917(4.76) | 0.6843(3.02) | 0.6856(2.85) | 0.6770(2.98) | 0.6778(2.79) |
| | 300 | 1258.2511(1100.21) | 0.8627(4.65) | 0.8561(4.40) | 0.7132(2.69) | 0.7128(2.46) | 0.7042(2.49) | 0.7036(2.27) |
| | 500 | 2209.7829(568.41) | 0.9027(4.72) | 0.8955(4.46) | 0.7321(2.68) | 0.7306(2.43) | 0.7222(2.47) | 0.7204(2.20) |

*Table A.22: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 2 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | ${}_{\text{t}}\text{NCM}_0$ | ${}_{\text{st}}\text{NCM}_0$ | ${}_{\text{t}}\text{NCM}_1$ | ${}_{\text{st}}\text{NCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0.3$ | | | | |
| 100 | 50 | 80.2647(163.26) | 1.3662(9.05) | 1.2981(7.78) | 1.2168(7.93) | 1.1817(7.48) | 1.1572(8.41) | 1.1186(7.56) |
| | 100 | 340.3248(936.91) | 1.4672(10.06) | 1.3882(7.89) | 1.3223(6.84) | 1.2783(6.03) | 1.2715(7.31) | 1.2246(6.31) |
| | 150 | 637.2886(659.40) | 1.5591(18.84) | 1.4644(14.81) | 1.3833(8.18) | 1.3324(6.36) | 1.3460(8.89) | 1.2937(7.11) |
| | 300 | 1204.8271(860.83) | 1.6715(21.20) | 1.5598(16.31) | 1.4483(5.00) | 1.3873(3.34) | 1.4355(5.95) | 1.3788(4.57) |
| | 500 | 1981.5834(1288.48) | 1.7120(13.70) | 1.5813(10.80) | 1.4886(4.89) | 1.4195(3.37) | 1.4983(5.60) | 1.4405(4.66) |
| 200 | 50 | 40.6091(139.67) | 1.0507(7.59) | 1.0220(6.73) | 0.8664(7.63) | 0.8604(7.13) | 0.8344(7.14) | 0.8266(6.51) |
| | 100 | 165.1892(215.65) | 1.1607(7.32) | 1.1269(6.49) | 0.9559(6.40) | 0.9424(6.01) | 0.9035(6.47) | 0.8880(5.85) |
| | 150 | 416.0394(544.43) | 1.2262(7.67) | 1.1889(6.78) | 1.0126(5.87) | 0.9938(5.10) | 0.9582(6.06) | 0.9368(5.25) |
| | 300 | 1305.6412(699.49) | 1.3396(8.10) | 1.2966(7.17) | 1.1049(6.57) | 1.0773(5.93) | 1.0556(6.17) | 1.0244(5.47) |
| | 500 | 2089.6536(848.93) | 1.4024(8.49) | 1.3553(7.17) | 1.1566(5.03) | 1.1281(4.60) | 1.1033(5.05) | 1.0721(4.49) |
| 500 | 50 | 16.5653(63.14) | 0.6969(4.90) | 0.6990(4.84) | 0.6217(3.96) | 0.6256(4.00) | 0.6152(3.91) | 0.6183(3.97) |
| | 100 | 70.1519(80.52) | 0.7777(5.05) | 0.7742(4.73) | 0.6701(3.34) | 0.6720(3.18) | 0.6630(3.28) | 0.6642(3.12) |
| | 150 | 164.9559(119.44) | 0.8146(4.79) | 0.8089(4.47) | 0.6911(3.34) | 0.6917(3.09) | 0.6834(3.36) | 0.6833(3.11) |
| | 300 | 1261.2489(752.81) | 0.8980(5.63) | 0.8897(5.31) | 0.7312(3.52) | 0.7302(3.24) | 0.7199(3.30) | 0.7186(3.01) |
| | 500 | 2209.9376(552.32) | 0.9348(4.51) | 0.9266(4.27) | 0.7498(3.04) | 0.7471(2.82) | 0.7357(2.67) | 0.7328(2.41) |

*Table A.23: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 2 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | $\rho = 0.8$ | | | |
| 100 | 50 | 79.6800(170.18) | 5.1153(85.30) | 4.0723(74.11) | 4.5971(84.20) | 3.6862(74.39) | 4.6237(84.27) | 3.7130(74.47) |
| | 100 | 339.9926(1042.58) | 9.2936(122.37) | 7.4132(107.17) | 8.5878(121.62) | 6.8284(106.29) | 8.6125(121.75) | 6.8536(106.43) |
| | 150 | 638.1674(644.79) | 13.3220(137.84) | 10.6708(120.53) | 12.4193(134.61) | 9.8478(116.57) | 12.4443(134.74) | 9.8734(116.66) |
| | 300 | 1207.0931(819.44) | 25.0121(163.91) | 20.1925(141.96) | 23.6962(171.27) | 18.7551(145.33) | 23.7219(171.28) | 18.7810(145.31) |
| | 500 | 1981.0598(1250.53) | 40.7397(262.32) | 32.9162(225.01) | 38.7855(257.31) | 30.6721(217.61) | 38.8130(257.34) | 30.6973(217.64) |
| 200 | 50 | 39.1294(165.27) | 3.7494(45.86) | 2.9940(39.97) | 3.3292(44.54) | 2.7151(39.76) | 3.3478(44.54) | 2.7337(39.77) |
| | 100 | 163.1501(292.05) | 6.2660(52.48) | 4.9691(45.17) | 5.6186(52.90) | 4.5064(45.76) | 5.6368(52.89) | 4.5246(45.74) |
| | 150 | 426.4054(663.79) | 8.9803(63.45) | 7.1548(54.29) | 8.1288(65.36) | 6.5088(56.15) | 8.1462(65.36) | 6.5262(56.15) |
| | 300 | 1308.2968(685.17) | 16.2039(92.04) | 13.0352(77.94) | 14.9151(89.47) | 11.8796(75.01) | 14.9329(89.46) | 11.8973(75.02) |
| | 500 | 2093.3160(790.69) | 25.6097(144.19) | 20.7070(121.97) | 23.7782(125.85) | 18.9027(105.66) | 23.7952(125.87) | 18.9199(105.70) |
| 500 | 50 | 16.2307(73.27) | 2.1358(22.06) | 1.7304(18.49) | 1.8927(22.59) | 1.5920(20.03) | 1.9044(22.66) | 1.6032(20.15) |
| | 100 | 68.1315(117.79) | 3.5846(23.19) | 2.8344(20.12) | 3.1624(23.28) | 2.5922(20.37) | 3.1733(23.28) | 2.6032(20.37) |
| | 150 | 161.2885(163.21) | 4.9223(30.38) | 3.8850(25.77) | 4.4173(31.75) | 3.5878(26.89) | 4.4279(31.76) | 3.5983(26.89) |
| | 300 | 1236.9190(1174.24) | 8.3181(35.67) | 6.5999(29.77) | 7.5837(36.28) | 6.0998(30.77) | 7.5940(36.28) | 6.1098(30.76) |
| | 500 | 2210.8352(568.24) | 12.6637(42.36) | 10.1224(34.81) | 11.6621(40.44) | 9.3409(33.24) | 11.6725(40.43) | 9.3508(33.20) |

*Table A.24: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 3*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{tNCM}_0$ | $\text{stNCM}_0$ | $\text{tNCM}_1$ | $\text{stNCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0$ | | | | |
| 100 | 50 | 33.7549(372.74) | 9.4937(80.84) | 8.7683(58.47) | 8.2683(48.40) | 8.1245(49.48) | 8.1492(131.70) | 7.9617(128.64) |
| | 100 | 146.4512(498.19) | 11.0308(58.66) | 10.6847(50.63) | 9.2852(32.72) | 9.1165(32.11) | 8.9970(38.04) | 8.8084(37.79) |
| | 150 | 616.3399(715.99) | 11.6164(52.99) | 11.2265(44.97) | 9.7749(22.08) | 9.5835(20.78) | 9.5105(27.25) | 9.3152(25.61) |
| | 300 | 1346.4954(1189.73) | 12.0237(60.80) | 11.5701(49.16) | 10.2837(20.80) | 10.0549(19.06) | 10.0631(25.43) | 9.8417(24.22) |
| | 500 | 2290.5995(1563.26) | 12.9481(73.20) | 12.3455(57.36) | 10.6572(16.21) | 10.3883(14.11) | 10.4322(22.31) | 10.1897(20.49) |
| 200 | 50 | 18.1520(196.61) | 7.2289(62.43) | 6.7238(47.00) | 6.6584(38.99) | 6.5366(39.38) | 6.2671(39.57) | 6.1453(39.93) |
| | 100 | 84.8202(264.55) | 8.0585(61.71) | 7.6458(50.76) | 7.6315(30.85) | 7.5017(29.84) | 7.1147(32.99) | 6.9808(31.94) |
| | 150 | 189.1150(343.15) | 10.3223(40.87) | 10.1205(38.46) | 8.1637(24.45) | 8.0100(23.71) | 7.5854(24.99) | 7.4427(24.69) |
| | 300 | 1225.2410(813.14) | 10.9175(40.16) | 10.6879(36.44) | 8.8911(21.91) | 8.6853(21.21) | 8.1727(23.91) | 8.0078(23.24) |
| | 500 | 1890.4326(971.83) | 11.7357(45.64) | 11.4608(41.13) | 9.3754(17.24) | 9.1440(16.00) | 8.5833(20.50) | 8.4092(19.41) |
| 500 | 50 | 9.0113(70.83) | 5.1414(33.55) | 4.8219(28.28) | 4.6374(29.08) | 4.6031(29.13) | 4.3953(25.94) | 4.3627(26.51) |
| | 100 | 29.8554(177.06) | 5.6780(37.61) | 5.4100(31.68) | 5.6586(24.32) | 5.6014(24.30) | 5.1702(25.09) | 5.1191(24.99) |
| | 150 | 66.5288(194.72) | 5.9399(30.57) | 5.7154(26.98) | 6.2597(21.81) | 6.1980(21.68) | 5.6807(23.05) | 5.6231(23.13) |
| | 300 | 331.3584(274.54) | 6.6691(26.33) | 6.5251(24.95) | 6.9688(15.80) | 6.8907(15.70) | 6.2318(18.83) | 6.1662(18.50) |
| | 500 | 2012.3872(602.47) | 9.6349(21.67) | 9.5702(20.74) | 7.5105(16.22) | 7.4184(15.92) | 6.6846(17.92) | 6.6148(17.63) |

*Table A.25: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 3 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | $\text{t}^{\text{NCM}_0}$ | $\text{st}^{\text{NCM}_0}$ | $\text{t}^{\text{NCM}_1}$ | $\text{st}^{\text{NCM}_1}$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0.3$ | | | | |
| 100 | 50 | 42.5725(387.81) | 9.7541(101.56) | 8.9544(71.31) | 8.3634(45.30) | 8.1987(44.55) | 8.0760(38.76) | 7.8829(38.85) |
| | 100 | 167.2741(454.10) | 11.1980(60.81) | 10.8311(51.70) | 9.3025(31.84) | 9.1333(30.67) | 9.0816(34.83) | 8.8876(34.53) |
| | 150 | 686.2410(848.33) | 11.7635(59.88) | 11.3575(50.66) | 9.8619(26.88) | 9.6663(25.62) | 9.6472(30.31) | 9.4441(29.29) |
| | 300 | 1385.5918(1263.88) | 12.3104(60.14) | 11.8060(48.40) | 10.3519(21.29) | 10.1187(19.31) | 10.2207(26.98) | 9.9935(25.60) |
| | 500 | 2279.5026(1736.48) | 13.0251(84.49) | 12.3973(65.09) | 10.6818(16.74) | 10.4249(14.30) | 10.5816(25.08) | 10.3391(22.34) |
| 200 | 50 | 18.7703(183.63) | 7.3868(60.40) | 6.8384(45.49) | 6.7798(37.68) | 6.6431(37.62) | 6.3987(37.34) | 6.2613(37.37) |
| | 100 | 77.5118(225.10) | 8.4184(71.81) | 7.9150(59.20) | 7.7259(31.68) | 7.5770(31.41) | 7.2374(32.84) | 7.0884(32.18) |
| | 150 | 179.2545(327.00) | 10.5025(51.21) | 10.2798(48.31) | 8.2325(26.10) | 8.0698(24.95) | 7.6522(27.19) | 7.5008(25.58) |
| | 300 | 1260.5702(814.43) | 11.1745(45.03) | 10.9270(41.05) | 8.9400(21.90) | 8.7339(20.89) | 8.2469(24.05) | 8.0746(22.61) |
| | 500 | 2069.3326(927.66) | 11.9235(44.32) | 11.6278(39.61) | 9.4110(16.62) | 9.1780(15.47) | 8.6526(18.66) | 8.4744(17.55) |
| 500 | 50 | 8.5698(79.81) | 5.3028(45.38) | 4.9564(37.19) | 4.7093(28.30) | 4.6740(28.03) | 4.4597(26.91) | 4.4273(26.62) |
| | 100 | 29.7056(206.78) | 5.8531(35.14) | 5.5519(28.43) | 5.7263(25.38) | 5.6675(25.74) | 5.2754(27.51) | 5.2208(27.41) |
| | 150 | 75.4650(133.92) | 6.1151(31.66) | 5.8554(27.10) | 6.3191(18.85) | 6.2524(18.87) | 5.7469(21.38) | 5.6862(21.42) |
| | 300 | 311.7880(256.68) | 6.7666(25.20) | 6.6109(23.92) | 7.0164(15.00) | 6.9382(15.01) | 6.2996(17.19) | 6.2328(17.14) |
| | 500 | 2040.5197(687.59) | 9.7485(19.86) | 9.6772(18.92) | 7.5230(14.42) | 7.4291(14.14) | 6.7439(17.17) | 6.6690(16.89) |

*Table A.26: The Average (standard error in %) of Spectral Norm-based IRSE for Setting 3 (continued)*

| $n$ | $p$ | $\text{DCM}_2$ | $\text{DCM}_1$ | $\text{sDCM}_1$ | ${}_{\text{t}}\text{NCM}_0$ | ${}_{\text{st}}\text{NCM}_0$ | ${}_{\text{t}}\text{NCM}_1$ | ${}_{\text{st}}\text{NCM}_1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | $\rho = 0.8$ | | | |
| 100 | 50 | 44.0482(390.97) | 17.5425(418.21) | 14.3574(301.39) | 13.8504(254.07) | 11.2706(155.57) | 14.0664(269.78) | 11.4083(173.55) |
| | 100 | 164.2430(484.69) | 27.4742(414.80) | 21.6247(285.97) | 23.4012(351.73) | 16.9324(244.78) | 23.8219(353.61) | 17.3178(252.25) |
| | 150 | 611.7285(7924.24) | 40.7790(701.45) | 30.7643(538.90) | 35.5976(495.29) | 25.2827(376.93) | 36.0318(496.42) | 25.7477(380.36) |
| | 300 | 1353.7699(1170.09) | 74.7893(677.76) | 55.4119(541.34) | 67.1574(572.51) | 47.2723(427.15) | 67.5872(572.75) | 47.6786(428.03) |
| | 500 | 2239.9087(1516.11) | 117.1434(977.98) | 87.0191(789.34) | 108.2073(740.20) | 75.8882(569.97) | 108.6379(739.53) | 76.2645(571.68) |
| 200 | 50 | 23.9602(217.58) | 13.4203(285.43) | 11.3627(220.58) | 11.5626(160.36) | 9.5405(100.70) | 11.5987(184.91) | 9.5039(126.44) |
| | 100 | 81.1553(245.01) | 19.7198(381.61) | 16.5282(278.95) | 17.8691(208.85) | 13.2696(134.04) | 18.2654(216.68) | 13.5147(157.18) |
| | 150 | 179.3693(363.09) | 31.0124(430.86) | 24.3665(309.45) | 25.2955(236.04) | 17.8988(178.11) | 25.8623(232.63) | 18.5699(183.29) |
| | 300 | 1214.1012(821.13) | 54.7451(429.61) | 40.8074(352.07) | 45.1306(324.45) | 31.5505(249.85) | 45.6746(317.88) | 32.2594(248.93) |
| | 500 | 2021.1184(858.12) | 82.6793(467.03) | 61.5751(375.20) | 70.1296(384.06) | 48.9165(292.79) | 70.6236(381.60) | 49.6008(296.57) |
| 500 | 50 | 11.2020(142.56) | 8.1112(129.25) | 7.2656(106.94) | 7.5152(79.90) | 6.7626(57.16) | 7.5319(92.58) | 6.7323(68.24) |
| | 100 | 36.8988(182.87) | 9.7052(138.68) | 8.9240(121.68) | 11.3344(88.61) | 9.3589(50.04) | 11.5436(112.46) | 9.4029(73.40) |
| | 150 | 76.6572(123.08) | 11.0363(128.19) | 10.1980(112.81) | 14.8076(100.24) | 11.3120(61.45) | 15.3864(106.86) | 11.7633(78.14) |
| | 300 | 308.1191(292.62) | 15.2102(283.28) | 14.1278(234.32) | 24.2464(129.32) | 17.3188(100.55) | 25.1207(129.61) | 18.4088(104.79) |
| | 500 | 872.4070(784.97) | 43.4568(192.84) | 32.1728(151.01) | 35.8148(149.90) | 25.4052(114.35) | 36.9197(150.00) | 26.8189(118.76) |

# Appendix B

# Proofs and Results of Chapter 4

## B.1 The Derivative of CV Function

The objective function is

$$\begin{aligned} cv(h_3) &= \sum_{j=1}^{p}\sum_{i=1}^{n}\left[\frac{y_{ij}^2}{\hat{\sigma}_{jj(-i)}(u_i)} + \log\left(\hat{\sigma}_{jj(-i)}(u_i)\right)\right], \\ &= \sum_{j=1}^{p}\sum_{i=1}^{n}\left[\frac{y_{ij}^2}{\exp\left[\hat{\alpha}_{j(-i)}(u_i)\right]} + \hat{\alpha}_{j(-i)}(u_i)\right], \end{aligned}$$

where

$$\begin{aligned} \hat{\alpha}_{j(-i)}(u_i) &= \log\left\{\frac{\sum_{s=1,s\neq i}^{n}\left[\frac{y_{sj}^2}{\exp\left(\hat{\beta}_{j(-i)}(u_i)(u_s-u_i)\right)}\right]K_{h_3}(u_s-u_i)}{\sum_{s=1,s\neq i}^{n}K_{h_3}(u_s-u_i)}\right\}, \\ &= \log(A_{ji}). \end{aligned}$$

Suppose, given index $j$ and for each $u_i$, we have already obtained the $\hat{\beta}_{j(-i)}(u_i)$. Furthermore, let $B_{ji}(u_s) = y_{sj}^2/\exp(\hat{\beta}_{j(-i)}(u_i)(u_s-u_i))$, then

$$A_{ji} = \frac{\sum_{s=1,s\neq i}^{n}B_{ji}(u_s)K_{h_3}(u_s-u_i)}{\sum_{s=1,s\neq i}^{n}K_{h_3}(u_s-u_i)}.$$

The first and second derivative of function $cv(h_3)$ are:

$$\frac{\partial cv(h_3)}{\partial h_3} = \sum_{j=1}^{p}\sum_{i=1}^{n}\left[-\frac{y_{ij}^2}{A_{ji}^2} + \frac{1}{A_{ji}}\right]\frac{\partial A_{ji}}{\partial h_3},$$

$$\frac{\partial^2 cv(h_3)}{\partial h_3^2} = \sum_{j=1}^{p}\sum_{i=1}^{n}\left\{\left[\frac{2y_{ij}^2}{A_{ji}^3} - \frac{1}{A_{ji}^2}\right]\left[\frac{\partial A_{ji}}{\partial h_3}\right]^2 + \left[-\frac{y_{ij}^2}{A_{ji}^2} + \frac{1}{A_{ji}}\right]\frac{\partial^2 A_{ji}}{\partial h_3^2}\right\}.$$

Before we calculate these two parts: $\partial A_{ji}/\partial h_3$ and $\partial^2 A_{ji}/\partial h_3^2$, we need to introduce some notations. The kernel function we adopted here is standard normal density function, i.e.,

$$K_{h_3}(u_s - u_i) = \frac{1}{h_3\sqrt{2\pi}}\exp\left[-\frac{(u_s - u_i)^2}{2h_3^2}\right],$$

then the first and second derivative of kernel function with respect to $h_3$ can be expressed as:

$$\begin{aligned} K'_{h_3}(u_s - u_i) &= \frac{1}{h_3}\left[-1 + \left(\frac{u_s - u_i}{h_3}\right)^2\right]K_{h_3}(u_s - u_i), \\ &= C_i(u_s)K_{h_3}(u_s - u_i), \end{aligned}$$

and

$$\begin{aligned} K''_{h_3}(u_s - u_i) &= \frac{1}{h_3^2}\left[2 - 5\left(\frac{u_s - u_i}{h_3}\right)^2 + \left(\frac{u_s - u_i}{h_3}\right)^4\right]K_{h_3}(u_s - u_i), \\ &= D_i(u_s)K_{h_3}(u_s - u_i). \end{aligned}$$

Furthermore, denote

$$\begin{aligned} SBK_{ji} &= \sum_{s=1,s\neq i}^{n} B_{ji}(u_s)K_{h_3}(u_s - u_i), \\ SK_i &= \sum_{s=1,s\neq i}^{n} K_{h_3}(u_s - u_i). \end{aligned}$$

For simplicity, let $\vartheta_{ji} = \hat{\beta}_{j(-i)}(u_i)$, then we have

$$B_{ji}(u_s) = \frac{y_{sj}^2}{\exp(\vartheta_{ji}(u_s - u_i))},$$

$$\begin{aligned} \frac{\partial B_{ji}(u_s)}{\partial h_3} &= -\frac{y_{sj}^2(u_s - u_i)}{\exp(\vartheta_{ji}(u_s - u_i))} * \frac{\partial \vartheta_{ji}}{\partial h_3}, \\ &= -B_{ji}(u_s) * (u_s - u_i) * \frac{\partial \vartheta_{ji}}{\partial h_3}, \end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 B_{ji}(u_s)}{\partial h_3^2} &= -\frac{\partial B_{ji}(u_s)}{\partial h_3} * (u_s - u_i) * \frac{\partial \vartheta_{ji}}{\partial h_3} - B_{ji}(u_s) * (u_s - u_i) * \frac{\partial^2 \vartheta_{ji}}{\partial h_3^2}, \\ &= B_{ji}(u_s) * (u_s - u_i)^2 * \left(\frac{\partial \vartheta_{ji}}{\partial h_3}\right)^2 - B_{ji}(u_s) * (u_s - u_i) * \frac{\partial^2 \vartheta_{ji}}{\partial h_3^2}.\end{aligned}$$

Based on these notations, the first derivative of $SBK_{ji}$ and $SK_{ji}$ with respect to $h_3$ are

$$\begin{aligned}\frac{\partial SBK_{ji}}{\partial h_3} &= -\frac{\partial \vartheta_{ji}}{\partial h_3} \sum_{s=1, s\neq i}^{n} B_{ji}(u_s)(u_s - u_i) K_{h_3}(u_s - u_i) \\ &\quad + \sum_{s=1, s\neq i}^{n} B_{ji}(u_s) C_i(u_s) K_{h_3}(u_s - u_i), \\ &= -\frac{\partial \vartheta_{ji}}{\partial h_3} SBUK_{ji} + SBCK_{ji},\end{aligned}$$

and

$$\frac{\partial SK_i}{\partial h_3} = \sum_{s=1, s\neq i}^{n} C_i(u_s) K_{h_3}(u_s - u_i) = SCK_i.$$

So

$$\frac{\partial A_{ji}}{\partial h_3} = -\frac{\partial \vartheta_{ji}}{\partial h_3} \frac{SBUK_{ji}}{SK_i} + \frac{SBCK_{ji}}{SK_i} - A_{ji} \frac{SCK_i}{SK_i}.$$

To compute $\partial^2 A_{ji}/\partial h_3^2$, we need $\partial^2 \vartheta_{ji}/\partial h_3^2$, $\partial SBUK_{ji}/\partial h_3$, $\partial SBCK_{ji}/\partial h_3$ and $\partial SCK_i/\partial h_3$ respectively. The $\partial^2 \vartheta_{ji}/\partial h_3^2$ is remained to discuss later.

$$SBUK_{ji} = \sum_{j=1, j\neq i}^{n} B_{ji}(u_s)(u_s - u_i) K_{h_3}(u_s - u_i),$$

$$\begin{aligned}\frac{\partial SBUK_{ji}}{\partial h_3} &= -\frac{\partial \vartheta_{ji}}{\partial h_3} \sum_{j=1, j\neq i}^{n} B_{ji}(u_s)(u_s - u_i)^2 K_{h_3}(u_s - u_i), \\ &\quad + \sum_{j=1, j\neq i}^{n} B_{ji}(u_s)(u_s - u_i) C_i(u_s) K_{h_3}(u_s - u_i) \\ &= -\frac{\partial \vartheta_{ji}}{\partial h_3} SBU2K_{ji} + SBUCK_{ji},\end{aligned}$$

$$\begin{aligned}\frac{\partial SBCK_{ji}}{\partial h_3} &= -\frac{\partial \vartheta_{ji}}{\partial h_3}\sum_{j=1,j\neq i}^{n} B_{ji}(u_s)(u_s-u_i)C_i(u_s)K_{h_3}(u_s-u_i)+\\ &\quad \sum_{j=1,j\neq i}^{n} B_{ji}(u_s)D_i(u_s)K_{h_3}(u_s-u_i),\\ &= -\frac{\partial \vartheta_{ji}}{\partial h_3}SBUCK_{ji}+SBDK_{ji},\end{aligned}$$

$$\begin{aligned}SCK_i &= \sum_{j=1,j\neq i}^{n} C_i(u_s)K_{h_3}(u_s-u_i),\\ \frac{\partial SCK_i}{\partial h_3} &= \sum_{j=1,j\neq i}^{n} D_i(u_s)K_{h_3}(u_s-u_i)=SDK_i,\end{aligned}$$

then

$$\begin{aligned}\frac{\partial \frac{SBUK_{ji}}{SK_i}}{\partial h_3} &= -\frac{\partial \vartheta_{ji}}{\partial h_3}\frac{SBU2K_{ji}}{SK_i}+\frac{SBUCK_{ji}}{SK_i}-\frac{SBUK_{ji}}{SK_i}\frac{SCK_i}{SK_i},\\ \frac{\partial \frac{SBCK_{ji}}{SK_i}}{\partial h_3} &= -\frac{\partial \vartheta_{ji}}{\partial h_3}\frac{SBUCK_{ji}}{SK_i}+\frac{SBDK_{ji}}{SK_i}-\frac{SBCK_{ji}}{SK_i}\frac{SCK_i}{SK_i},\\ \frac{\partial \frac{SCK_i}{SK_i}}{\partial h_3} &= \frac{SDK_i}{SK_i}-\frac{SCK_i}{SK_i}\frac{SCK_i}{SK_i},\end{aligned}$$

so

$$\begin{aligned}\frac{\partial^2 A_{ji}}{\partial h_3^2} = &-\frac{\partial^2 \vartheta_{ji}}{\partial h_3^2}\frac{SBUK_{ji}}{SK_i}+\left(\frac{\partial \vartheta_{ji}}{\partial h_3}\right)^2\frac{SBU2K_{ji}}{SK_i}-2\frac{\partial \vartheta_{ji}}{\partial h_3}\frac{SBUCK_{ji}}{SK_i}\\ &+\frac{SBDK_{ji}}{SK_i}-2\frac{\partial A_{ji}}{\partial h_3}\frac{SCK_i}{SK_i}-A_{ji}\frac{SDK_i}{SK_i}.\end{aligned}$$

Basically, the unknown parts here are just $\partial\vartheta_{ji}/\partial h_3$ and $\partial^2\vartheta_{ji}/\partial h_3^2$. We notice that $SBUK_{ji}/SBK_{ji} = SUK_i/SK_i$, $f(\vartheta_{ji}) = SBUK_{ji}\cdot SK_i - SBK_{ji}\cdot SUK_i = 0$ and

$$\begin{aligned}\frac{\partial f(\vartheta_{ji})}{\partial h_3} = &-\frac{\partial \vartheta_{ji}}{\partial h_3}SBU2K_{ji}\cdot SK_i+SBUCK_{ji}\cdot SK_i+SBUK_{ji}\cdot SCK_i\\ &+\frac{\partial \vartheta_{ji}}{\partial h_3}SBUK_{ji}\cdot SUK_i-SCK_{ji}\cdot SUK_i-SBK_{ji}\cdot SUCK_i,\\ &=0.\end{aligned}$$

Let $P_1 = SBUK_{ji}\cdot SUK_i - SBU2K_{ji}\cdot SK_i$ and $P_2 = SCK_{ji}\cdot SUK_i + SBK_{ji}\cdot SUCK_i - SBUCK_{ji}\cdot SK_i - SBUK_{ji}\cdot SCK_i$, then $\partial\vartheta_{ji}/\partial h_3 = P_2/P_1$. Taking

the further derivation of both sides of $P_2/P_1$ with respect to $h_3$, we obtain

$$\begin{aligned}\frac{\partial^2\vartheta_{ji}}{\partial h_3^2}P_1+\frac{\partial\vartheta_{ji}}{\partial h_3}\frac{\partial P_1}{\partial h_3} &= \frac{\partial P_2}{\partial h_3},\\ \frac{\partial^2\vartheta_{ji}}{\partial h_3^2} &= \frac{\frac{\partial P_2}{\partial h_3}-\frac{\partial\vartheta_{ji}}{\partial h_3}\frac{\partial P_1}{\partial h_3}}{P_1}.\end{aligned}$$

Next, we calculate $\partial P_1/\partial h_3$ and $\partial P_2/\partial h_3$, denote

$$\begin{aligned}\frac{\partial SUK_i}{\partial h_3} &= SUCK_i,\\ \frac{\partial SBU2K_{ji}}{\partial h_3} &= -\frac{\partial\vartheta_{ji}}{\partial h_3}SBU3K_{ji}+SBU2CK_{ji},\end{aligned}$$

$$\begin{aligned}\frac{\partial P_1}{\partial h_3} &= \frac{\partial\vartheta_{ji}}{\partial h_3}\left(SBU3K_{ji}\cdot SK_i-SBU2K_{ji}\cdot SUK_i\right)\\ &\quad+\left(SBUCK_{ji}\cdot SUK_i+SBUK_{ji}\cdot SUCK_i\right.\\ &\quad\left.-SBU2CK_{ji}\cdot SK_i-SBU2K_{ji}\cdot SCK_i\right),\\ &= \frac{\partial\vartheta_{ji}}{\partial h_3}G_1+G_2.\end{aligned}$$

We also denote

$$\begin{aligned}\frac{\partial SUCK_i}{\partial h_3} &= SUDK_i,\\ \frac{\partial SBUCK_{ji}}{\partial h_3} &= -\frac{\partial\vartheta_{ji}}{\partial h_3}SBU2CK_{ji}+SBUDK_{ji},\end{aligned}$$

then

$$\begin{aligned}\frac{\partial P_2}{\partial h_3} &= -\frac{\partial\vartheta_{ji}}{\partial h_3}G_2\\ &\quad+\left(SBDK_{ji}\cdot SUK_i+2SCK_i\cdot SUCK_i+SBK_{ji}\cdot SUDK_i\right.\\ &\quad\left.+-SBUDK_{ji}\cdot SK_i-2SBUCK_{ji}\cdot SCK_i-SBUK_{ji}\cdot SDK\right),\\ &= -\frac{\partial\vartheta_{ji}}{\partial h_3}G_2+G_3,\end{aligned}$$

so

$$\frac{\partial^2\vartheta_{ji}}{\partial h_3^2}=\frac{G_3-2G_2\frac{\partial\vartheta_{ji}}{\partial h_3}-G_1\left(\frac{\partial\vartheta_{ji}}{\partial h_3}\right)^2}{P_1}.$$

Finally, we obtain the parts except the $\hat{\beta}_{j(-i)}(u_i)$ to compute the first and second derivative of $cv(h_3)$. We will show the details of solving equation (4.13) using Newton-Raphson algorithm in the next section.

## B.2 The Details of Solving Nonlinear Equation

Recall that in the previous section, we briefly note $\hat{\beta}_{j(-i)}(u_i)$ using $\vartheta_{ji}$. We construct new objective function $g(\vartheta_{ji})$ based on equation (4.13) and the previous notations

$$g(\vartheta_{ji}) = \left(\frac{SBUK_{ji}}{SBK_{ji}} - \frac{SUK_i}{SK_i}\right)^2.$$

Notice that

$$\frac{\partial B_{ji}(u_s)}{\partial \vartheta_{ji}} = -B_{ji}(u_s) * (u_s - u_i),$$

then the first derivation of $g(\vartheta_{ji})$ is

$$\frac{\partial g}{\partial \vartheta_{ji}} = 2\left(\frac{SBUK_{ji}}{SBK_{ji}} - \frac{SUK_i}{SK_i}\right)\frac{\partial \frac{SBUK_{ji}}{SBK_{ji}}}{\partial \vartheta_{ji}},$$

where

$$\frac{\partial \frac{SBUK_{ji}}{SBK_{ji}}}{\partial \vartheta_{ji}} = \left(\frac{SBUK_{ji}}{SBK_{ji}}\right)^2 - \frac{SBU2K_{ji}}{SBK_{ji}}.$$

So

$$\frac{\partial g}{\partial \vartheta_{ji}} = 2\left(\frac{SBUK_{ji}}{SBK_{ji}} - \frac{SUK_i}{SK_i}\right)\left[\left(\frac{SBUK_{ji}}{SBK_{ji}}\right)^2 - \frac{SBU2K_{ji}}{SBK_{ji}}\right].$$

Then the second derivation is

$$\begin{aligned}\frac{\partial^2 g}{\partial \vartheta_{ji}^2} = {} & 2\left[\left(\frac{SBUK_{ji}}{SBK_{ji}}\right)^2 - \frac{SBU2K_{ji}}{SBK_{ji}}\right]^2 + 4\frac{\partial g}{\partial \vartheta_{ji}}\frac{SBUK_{ji}}{SBK_{ji}} \\ & + 2\left(\frac{SBUK_{ji}}{SBK_{ji}} - \frac{SUK_i}{SK_i}\right)\left(\frac{SBU3K_{ji}}{SBK_{ji}} - \frac{SBU2K_{ji}}{SBK_{ji}}\frac{SBUK_{ji}}{SBK_{ji}}\right).\end{aligned}$$

Hence, at the $t$-th iteration, we can use the following equation to update $\vartheta_{ji}^{(t+1)}$:

$$\vartheta_{ji}^{(t+1)} = \vartheta_{ji}^{(t)} - \left.\frac{\frac{\partial g}{\partial \vartheta_{ji}}}{\frac{\partial^2 g}{\partial \vartheta_{ji}^2}}\right|_{\vartheta_{ji}=\vartheta_{ji}^{(t)}}.$$

## B.3 The Details of Bandwidth $h_3$ Selection

According to objective function, we can easily get

$$\begin{aligned}
\frac{\partial cv(h_3)}{\partial h_3} &= -\frac{2}{n}\sum_{i=1}^{n}\left[\varsigma(u_i)-\hat{\theta}_{-i}(u_i)\right]^T\frac{\partial\hat{\theta}_{-i}(u_i)}{\partial h_3},\\
\frac{\partial^2 cv(h_3)}{\partial h_3^2} &= \frac{2}{n}\sum_{i=1}^{n}\left\{\left[\frac{\partial\hat{\theta}_{-i}(u_i)}{\partial h_3}\right]^T\frac{\partial\hat{\theta}_{-i}(u_i)}{\partial h_3}-\left[\varsigma(u_i)-\hat{\theta}_{-i}(u_i)\right]^T\frac{\partial^2\hat{\theta}_{-i}(u_i)}{\partial h_3^2}\right\},
\end{aligned}$$

so, we only need to compute two parts $\partial\hat{\theta}_{-i}(u_i)/\partial h_3$ and $\partial^2\hat{\theta}_{-i}(u_i)/\partial h_3^2$. For simplicity, we denote $\hat{\theta}_{-i}(u_i)$ as

$$\hat{\theta}_{-i}(u_i)=\frac{SKV_{-i}\cdot SKU2-SKUV\cdot SKU}{SK_{-i}\cdot SKU2-SKU^2}=\frac{Q_1}{Q_2},$$

then $\partial\hat{\theta}_{-i}(u_i)/\partial h_3$ and $\partial^2\hat{\theta}_{-i}(u_i)/\partial h_3^2$ can be obtained as follows

$$\begin{aligned}
\frac{\partial\hat{\theta}_{-i}(u_i)}{\partial h_3} &= \frac{Q_1'}{Q_2}-\hat{\theta}_{-i}(u_i)\frac{Q_2'}{Q_2},\\
\frac{\partial^2\hat{\theta}_{-i}(u_i)}{\partial h_3^2} &= \frac{Q_1''}{Q_2}-2\frac{\partial\hat{\theta}_{-i}(u_i)}{\partial h_3}\frac{Q_2'}{Q_2}-\hat{\theta}_{-i}(u_i)\frac{Q_2''}{Q_2}.
\end{aligned}$$

Hence, at last we only need to compute $Q_1'$, $Q_2'$, $Q_1''$ and $Q_2''$:

$$\begin{aligned}
Q_1 &= SKV_{-i}\cdot SKU2-SKUV\cdot SKU,\\
Q_1' &= SCKV_{-i}\cdot SKU2+SKV_{-i}\cdot SCKU2\\
&\quad -SCKUV\cdot SKU-SKUV\cdot SCKU,\\
Q_1'' &= SDKV_{-i}\cdot SKU2+2SCKV_{-i}\cdot SCKU2+SKV_{-i}\cdot SDKU2\\
&\quad -SDKUV\cdot SKU-2SCKUV\cdot SCKU-SKUV\cdot SDKU,\\
Q_2 &= SK_{-i}\cdot SKU2-SKU^2,\\
Q_2' &= SCK_{-i}\cdot SKU2+SK_{-i}\cdot SCKU2-2SKU\cdot SCKU,\\
Q_2'' &= SDK_{-i}\cdot SKU2+2SCK_{-i}\cdot SCKU2+SK_{-i}\cdot SDKU2\\
&\quad -2SCKU^2-2SKU\cdot SDKU,
\end{aligned}$$

where $SK_{-i}=SK-k_{h_3}(0)$, $SCK_{-i}=SCK+k_{h_3}(0)/h_3$, $SDK_{-i}=SDK-2k_{h_3}(0)/h_3^2$, $SKV_{-i}=SKV-k_{h_3}(0)\varsigma(u_i)$, $SCKV_{-i}=SCKV+k_{h_3}(0)\varsigma(u_i)/h_3$ and $SDKV_{-i}=SDKV-2k_{h_3}(0)\varsigma(u_i)/h_3^2$.

## B.4 Tables

*Table B.1: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 2 with $n = 200, p = 150$*

| *pct.* | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | | $\text{DAC}_2$ | |
|---|---|---|---|---|---|
| | | Divided | Direct | Divided | Direct |
| 10% | 0.2234(1.35) | 0.1421(0.68) | 0.1430(0.84) | 0.1414(0.63) | 0.1419(0.63) |
| 30% | 0.2154(1.07) | 0.1409(0.68) | 0.1431(0.67) | 0.1403(0.64) | 0.1417(0.63) |
| 50% | 0.2038(0.81) | 0.1397(0.65) | 0.1414(0.68) | 0.1391(0.62) | 0.1411(0.63) |
| 70% | 0.1888(0.67) | 0.1386(0.67) | 0.1414(0.66) | 0.1382(0.64) | 0.1408(0.64) |
| 90% | 0.1576(0.60) | 0.1381(0.67) | 0.1407(0.70) | 0.1375(0.65) | 0.1398(0.64) |

*Table B.2: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 3*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0$ | | | |
| 100 | 50 | 1.1428(45.63) | 0.5597(5.52) | 0.06 | 0.5586(4.81) | 0.05 |
| | 100 | 0.8839(9.15) | 0.6271(4.82) | 0.06 | 0.6312(4.62) | 0.05 |
| | 150 | 0.8971(7.53) | 0.6666(4.92) | 0.06 | 0.6721(4.80) | 0.05 |
| | 300 | 0.8831(2.52) | 0.7213(3.53) | 0.06 | 0.7270(3.60) | 0.06 |
| 200 | 50 | 0.6796(26.54) | 0.3664(4.70) | 0.01 | 0.3761(4.30) | 0.01 |
| | 100 | 0.6297(8.63) | 0.4124(3.92) | 0.02 | 0.4227(3.64) | 0.01 |
| | 150 | 0.6386(5.25) | 0.4290(3.63) | 0.02 | 0.4387(3.32) | 0.02 |
| | 300 | 0.6897(3.24) | 0.4644(2.99) | 0.02 | 0.4720(2.70) | 0.02 |
| 500 | 50 | 0.2933(3.17) | 0.2300(2.59) | 0.01 | 0.2410(2.41) | 0.01 |
| | 100 | 0.3285(3.43) | 0.2485(1.86) | 0.01 | 0.2606(1.87) | 0.01 |
| | 150 | 0.3504(3.73) | 0.2619(2.34) | 0.01 | 0.2755(2.22) | 0.01 |
| | 300 | 0.3894(2.21) | 0.2794(1.88) | 0.01 | 0.2920(1.76) | 0.01 |

*Table B.3: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 3 (continued)*

| $n$ | $p$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.3$ | | | |
| 100 | 50 | 0.7345(6.03) | 0.6365(5.86) | 0.04 | 0.6367(5.78) | 0.04 |
| | 100 | 0.7955(4.40) | 0.7155(5.17) | 0.04 | 0.7130(4.66) | 0.05 |
| | 150 | 0.8302(3.50) | 0.7412(3.79) | 0.04 | 0.7394(4.00) | 0.05 |
| | 300 | 0.8791(2.73) | 0.7980(3.74) | 0.04 | 0.7957(3.78) | 0.04 |
| 200 | 50 | 0.6045(4.96) | 0.4361(4.33) | 0.01 | 0.4371(4.26) | 0.01 |
| | 100 | 0.6369(3.67) | 0.4796(4.26) | 0.01 | 0.4819(4.11) | 0.01 |
| | 150 | 0.6612(3.58) | 0.5087(3.67) | 0.01 | 0.5117(3.81) | 0.01 |
| | 300 | 0.7017(2.95) | 0.5511(3.27) | 0.01 | 0.5525(3.07) | 0.01 |
| 500 | 50 | 0.5283(2.78) | 0.3030(2.77) | 0.01 | 0.3116(2.80) | 0.01 |
| | 100 | 0.5599(2.09) | 0.3330(2.25) | 0.01 | 0.3406(2.30) | 0.01 |
| | 150 | 0.5653(2.17) | 0.3371(1.89) | 0.01 | 0.3451(1.83) | 0.01 |
| | 300 | 0.5817(1.70) | 0.3472(1.67) | 0.01 | 0.3543(1.62) | 0.01 |

*Table B.4: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 3 (continued)*

| $n$ | $p$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.8$ | | | |
| 100 | 50 | 1.2338(5.37) | 1.0927(2.74) | 0.06 | 1.0764(2.18) | 0.05 |
| | 100 | 1.2501(3.86) | 1.1304(2.21) | 0.06 | 1.1156(1.93) | 0.04 |
| | 150 | 1.2492(3.38) | 1.1436(2.59) | 0.03 | 1.1295(1.58) | 0.03 |
| | 300 | 1.2654(3.05) | 1.1687(4.64) | 0.02 | 1.1502(1.42) | 0.02 |
| 200 | 50 | 1.2210(3.88) | 1.0536(3.27) | 0.01 | 1.0220(1.75) | 0.01 |
| | 100 | 1.2230(3.37) | 1.0950(3.23) | 0.01 | 1.0553(1.70) | 0.01 |
| | 150 | 1.2362(3.22) | 1.1071(2.92) | 0.01 | 1.0691(0.93) | 0.01 |
| | 300 | 1.2390(2.66) | 1.1334(3.60) | 0.01 | 1.0938(1.77) | 0.01 |
| 500 | 50 | 1.2261(2.62) | 1.0501(2.92) | 0.01 | 1.0281(1.97) | 0.01 |
| | 100 | 1.2422(0.62) | 1.0902(3.49) | 0.01 | 1.0543(1.01) | 0.01 |
| | 150 | 1.2463(1.36) | 1.0980(2.63) | 0.01 | 1.0630(0.70) | 0.01 |
| | 300 | 1.2533(0.33) | 1.1263(3.95) | 0.01 | 1.0780(1.31) | 0.01 |

*Table B.5: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 4*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0$ | | | |
| 100 | 50 | 0.2214(1.69) | 0.2142(1.15) | 0.03 | 0.2141(1.14) | 0.03 |
| | 100 | 0.2360(1.21) | 0.2289(0.86) | 0.02 | 0.2290(0.81) | 0.02 |
| | 150 | 0.2432(1.25) | 0.2372(0.83) | 0.02 | 0.2369(0.82) | 0.02 |
| | 300 | 0.2514(0.74) | 0.2466(0.71) | 0.02 | 0.2478(0.57) | 0.02 |
| 200 | 50 | 0.1554(0.84) | 0.1501(0.68) | 0.01 | 0.1500(0.68) | 0.01 |
| | 100 | 0.1599(0.73) | 0.1543(0.58) | 0.01 | 0.1538(0.58) | 0.01 |
| | 150 | 0.1619(0.57) | 0.1582(0.48) | 0.01 | 0.1572(0.48) | 0.01 |
| | 300 | 0.1660(0.53) | 0.1636(0.70) | 0.01 | 0.1618(0.36) | 0.01 |
| 500 | 50 | 0.1169(0.45) | 0.1100(0.44) | 0.03 | 0.1102(0.45) | 0.03 |
| | 100 | 0.1195(0.27) | 0.1134(0.31) | 0.03 | 0.1137(0.27) | 0.02 |
| | 150 | 0.1221(0.23) | 0.1152(0.22) | 0.01 | 0.1155(0.22) | 0.01 |
| | 300 | 0.1237(0.15) | 0.1170(0.14) | 0.01 | 0.1172(0.14) | 0.01 |

*Table B.6: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 4 (continued)*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.3$ | | | |
| 100 | 50 | 0.2672(1.40) | 0.2319(1.08) | 0.03 | 0.2327(1.08) | 0.03 |
| | 100 | 0.2707(1.11) | 0.2442(0.87) | 0.02 | 0.2443(0.84) | 0.02 |
| | 150 | 0.2784(0.85) | 0.2517(0.86) | 0.02 | 0.2515(0.71) | 0.02 |
| | 300 | 0.2873(0.57) | 0.2621(0.50) | 0.02 | 0.2623(0.50) | 0.02 |
| 200 | 50 | 0.2330(1.04) | 0.1726(0.85) | 0.02 | 0.1732(0.84) | 0.02 |
| | 100 | 0.2375(0.95) | 0.1768(0.62) | 0.01 | 0.1774(0.62) | 0.01 |
| | 150 | 0.2372(0.66) | 0.1802(0.46) | 0.01 | 0.1803(0.44) | 0.01 |
| | 300 | 0.2428(0.56) | 0.1865(0.48) | 0.01 | 0.1862(0.38) | 0.01 |
| 500 | 50 | 0.2185(0.91) | 0.1260(0.46) | 0.01 | 0.1273(0.46) | 0.02 |
| | 100 | 0.2232(0.62) | 0.1307(0.33) | 0.01 | 0.1321(0.34) | 0.01 |
| | 150 | 0.2232(0.55) | 0.1322(0.27) | 0.01 | 0.1337(0.28) | 0.01 |
| | 300 | 0.2240(0.33) | 0.1337(0.19) | 0.01 | 0.1350(0.20) | 0.01 |

*Table B.7: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 4 (continued)*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.8$ | | | |
| 100 | 50 | 0.4906(2.45) | 0.4161(0.74) | 0.02 | 0.4137(0.59) | 0.02 |
| | 100 | 0.4969(2.50) | 0.4229(0.55) | 0.01 | 0.4212(0.50) | 0.01 |
| | 150 | 0.4961(2.28) | 0.4310(0.40) | 0.01 | 0.4293(0.43) | 0.01 |
| | 300 | 0.4932(1.66) | 0.4455(0.47) | 0.01 | 0.4462(0.38) | 0.01 |
| 200 | 50 | 0.5154(1.13) | 0.3986(0.64) | 0.01 | 0.3934(0.53) | 0.01 |
| | 100 | 0.5129(1.01) | 0.4024(0.50) | 0.01 | 0.3978(0.39) | 0.01 |
| | 150 | 0.5157(1.07) | 0.4038(0.28) | 0.01 | 0.3999(0.25) | 0.01 |
| | 300 | 0.5154(0.94) | 0.4085(0.21) | 0.01 | 0.4053(0.17) | 0.01 |
| 500 | 50 | 0.5177(0.41) | 0.3912(0.68) | 0.01 | 0.3850(0.27) | 0.01 |
| | 100 | 0.5184(0.34) | 0.3939(0.42) | 0.01 | 0.3892(0.20) | 0.01 |
| | 150 | 0.5192(0.20) | 0.3957(0.35) | 0.01 | 0.3914(0.18) | 0.01 |
| | 300 | 0.5193(0.16) | 0.3992(0.23) | 0.01 | 0.3956(0.14) | 0.01 |

*Table B.8: The Average SEN, SPE and ACC for Scenario 4 ($\rho = 0$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.0674 | 0.0635 | 0.0594 | 0/0 | 0/0 | 0/0 | 0.0674 | 0.0635 | 0.0594 | 0.03 | 0.03 |
| | 100 | 0.0304 | 0.0254 | 0.0258 | 0/0 | 0/0 | 0/0 | 0.0304 | 0.0254 | 0.0258 | 0.02 | 0.02 |
| | 150 | 0.0183 | 0.0164 | 0.0167 | 0/0 | 0/0 | 0/0 | 0.0183 | 0.0164 | 0.0167 | 0.02 | 0.02 |
| | 300 | 0.0075 | 0.0077 | 0.0078 | 0/0 | 0/0 | 0/0 | 0.0075 | 0.0077 | 0.0078 | 0.02 | 0.02 |
| 200 | 50 | 0.0690 | 0.0746 | 0.0749 | 0/0 | 0/0 | 0/0 | 0.0690 | 0.0746 | 0.0749 | 0.01 | 0.01 |
| | 100 | 0.0336 | 0.0353 | 0.0338 | 0/0 | 0/0 | 0/0 | 0.0336 | 0.0353 | 0.0338 | 0.01 | 0.01 |
| | 150 | 0.0220 | 0.0223 | 0.0224 | 0/0 | 0/0 | 0/0 | 0.0220 | 0.0223 | 0.0224 | 0.01 | 0.01 |
| | 300 | 0.0106 | 0.0109 | 0.0109 | 0/0 | 0/0 | 0/0 | 0.0106 | 0.0109 | 0.0109 | 0.01 | 0.01 |
| 500 | 50 | 0.0795 | 0.0903 | 0.0905 | 0/0 | 0/0 | 0/0 | 0.0795 | 0.0903 | 0.0905 | 0.03 | 0.03 |
| | 100 | 0.0382 | 0.0385 | 0.0363 | 0/0 | 0/0 | 0/0 | 0.0382 | 0.0385 | 0.0363 | 0.03 | 0.02 |
| | 150 | 0.0240 | 0.0236 | 0.0237 | 0/0 | 0/0 | 0/0 | 0.0240 | 0.0236 | 0.0237 | 0.01 | 0.01 |
| | 300 | 0.0114 | 0.0115 | 0.0116 | 0/0 | 0/0 | 0/0 | 0.0114 | 0.0115 | 0.0116 | 0.01 | 0.01 |

*Table B.9: The Average SEN, SPE and ACC for Scenario 4 ($\rho = 0.3$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.0632 | 0.0791 | 0.0757 | 0/0 | 0/0 | 0/0 | 0.0632 | 0.0791 | 0.0757 | 0.03 | 0.03 |
| | 100 | 0.0263 | 0.0323 | 0.0327 | 0/0 | 0/0 | 0/0 | 0.0263 | 0.0323 | 0.0327 | 0.02 | 0.02 |
| | 150 | 0.0160 | 0.0168 | 0.0171 | 0/0 | 0/0 | 0/0 | 0.0160 | 0.0168 | 0.0171 | 0.02 | 0.02 |
| | 300 | 0.0065 | 0.0080 | 0.0081 | 0/0 | 0/0 | 0/0 | 0.0065 | 0.0080 | 0.0081 | 0.02 | 0.02 |
| 200 | 50 | 0.0650 | 0.0731 | 0.0734 | 0/0 | 0/0 | 0/0 | 0.0650 | 0.0731 | 0.0734 | 0.02 | 0.02 |
| | 100 | 0.0313 | 0.0368 | 0.0369 | 0/0 | 0/0 | 0/0 | 0.0313 | 0.0368 | 0.0369 | 0.01 | 0.01 |
| | 150 | 0.0202 | 0.0257 | 0.0259 | 0/0 | 0/0 | 0/0 | 0.0202 | 0.0257 | 0.0259 | 0.01 | 0.01 |
| | 300 | 0.0096 | 0.0119 | 0.0120 | 0/0 | 0/0 | 0/0 | 0.0096 | 0.0119 | 0.0120 | 0.01 | 0.01 |
| 500 | 50 | 0.0724 | 0.0862 | 0.0869 | 0/0 | 0/0 | 0/0 | 0.0724 | 0.0862 | 0.0869 | 0.01 | 0.02 |
| | 100 | 0.0326 | 0.0393 | 0.0395 | 0/0 | 0/0 | 0/0 | 0.0326 | 0.0393 | 0.0395 | 0.01 | 0.01 |
| | 150 | 0.0211 | 0.0239 | 0.0240 | 0/0 | 0/0 | 0/0 | 0.0211 | 0.0239 | 0.0240 | 0.01 | 0.01 |
| | 300 | 0.0103 | 0.0118 | 0.0118 | 0/0 | 0/0 | 0/0 | 0.0103 | 0.0118 | 0.0118 | 0.01 | 0.01 |

*Table B.10: The Average SEN, SPE and ACC for Scenario 4 ($\rho = 0.8$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $_{st}NCM_1$ | $DAC_1$ | $DAC_2$ | $_{st}NCM_1$ | $DAC_1$ | $DAC_2$ | $_{st}NCM_1$ | $DAC_1$ | $DAC_2$ | $DAC_1$ | $DAC_2$ |
| 100 | 50 | 0.1797 | 0.1227 | 0.1394 | 0/0 | 0/0 | 0/0 | 0.1797 | 0.1227 | 0.1394 | 0.02 | 0.02 |
| | 100 | 0.0341 | 0.0473 | 0.0548 | 0/0 | 0/0 | 0/0 | 0.0341 | 0.0473 | 0.0548 | 0.01 | 0.01 |
| | 150 | 0.0231 | 0.0346 | 0.0434 | 0/0 | 0/0 | 0/0 | 0.0231 | 0.0346 | 0.0434 | 0.01 | 0.01 |
| | 300 | 0.0067 | 0.0222 | 0.0304 | 0/0 | 0/0 | 0/0 | 0.0067 | 0.0222 | 0.0304 | 0.01 | 0.01 |
| 200 | 50 | 0.0683 | 0.0873 | 0.0797 | 0/0 | 0/0 | 0/0 | 0.0683 | 0.0873 | 0.0797 | 0.01 | 0.01 |
| | 100 | 0.0266 | 0.0397 | 0.0437 | 0/0 | 0/0 | 0/0 | 0.0266 | 0.0397 | 0.0437 | 0.01 | 0.01 |
| | 150 | 0.0162 | 0.0276 | 0.0310 | 0/0 | 0/0 | 0/0 | 0.0162 | 0.0276 | 0.0310 | 0.01 | 0.01 |
| | 300 | 0.0064 | 0.0152 | 0.0180 | 0/0 | 0/0 | 0/0 | 0.0064 | 0.0152 | 0.0180 | 0.01 | 0.01 |
| 500 | 50 | 0.0708 | 0.0914 | 0.0967 | 0/0 | 0/0 | 0/0 | 0.0708 | 0.0914 | 0.0967 | 0.01 | 0.01 |
| | 100 | 0.0323 | 0.0488 | 0.0524 | 0/0 | 0/0 | 0/0 | 0.0323 | 0.0488 | 0.0524 | 0.01 | 0.01 |
| | 150 | 0.0210 | 0.0344 | 0.0378 | 0/0 | 0/0 | 0/0 | 0.0210 | 0.0344 | 0.0378 | 0.01 | 0.01 |
| | 300 | 0.0102 | 0.0199 | 0.0226 | 0/0 | 0/0 | 0/0 | 0.0102 | 0.0199 | 0.0226 | 0.01 | 0.01 |

*Table B.11: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 4*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0$ | | | |
| 100 | 50 | 0.6392(7.05) | 0.5687(4.80) | 0.03 | 0.5848(4.86) | 0.03 |
| | 100 | 0.8031(11.48) | 0.6515(5.39) | 0.02 | 0.6743(5.65) | 0.02 |
| | 150 | 0.8870(15.00) | 0.6967(6.06) | 0.02 | 0.7235(6.26) | 0.02 |
| | 300 | 0.9513(13.70) | 0.7237(5.58) | 0.02 | 0.7709(4.88) | 0.02 |
| 200 | 50 | 0.4235(3.16) | 0.3893(3.08) | 0.01 | 0.4024(2.92) | 0.01 |
| | 100 | 0.4767(5.06) | 0.4249(3.17) | 0.01 | 0.4391(3.19) | 0.01 |
| | 150 | 0.5142(4.82) | 0.4551(2.59) | 0.01 | 0.4683(2.40) | 0.01 |
| | 300 | 0.5770(7.68) | 0.4952(3.25) | 0.01 | 0.5070(3.14) | 0.01 |
| 500 | 50 | 0.3217(2.15) | 0.2708(2.16) | 0.03 | 0.2847(2.09) | 0.03 |
| | 100 | 0.3501(1.98) | 0.2878(1.31) | 0.03 | 0.3045(1.25) | 0.02 |
| | 150 | 0.3773(2.10) | 0.3033(1.53) | 0.01 | 0.3196(1.57) | 0.01 |
| | 300 | 0.4037(2.35) | 0.3157(1.29) | 0.01 | 0.3329(1.15) | 0.01 |

*Table B.12: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 4 (continued)*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.3$ | | | |
| 100 | 50 | 0.6518(3.48) | 0.5916(3.46) | 0.03 | 0.5963(3.43) | 0.03 |
| | 100 | 0.6874(2.55) | 0.6555(3.38) | 0.02 | 0.6622(3.63) | 0.02 |
| | 150 | 0.7089(2.07) | 0.6907(3.17) | 0.02 | 0.7005(3.33) | 0.02 |
| | 300 | 0.7441(1.90) | 0.7354(2.87) | 0.02 | 0.7495(2.98) | 0.02 |
| 200 | 50 | 0.5592(2.72) | 0.4463(3.27) | 0.02 | 0.4516(3.21) | 0.02 |
| | 100 | 0.5899(2.79) | 0.4789(2.76) | 0.01 | 0.4843(2.84) | 0.01 |
| | 150 | 0.6032(2.25) | 0.5011(2.43) | 0.01 | 0.5042(2.12) | 0.01 |
| | 300 | 0.6341(1.75) | 0.5363(2.10) | 0.01 | 0.5382(2.11) | 0.01 |
| 500 | 50 | 0.5089(2.30) | 0.3202(2.07) | 0.01 | 0.3279(2.03) | 0.02 |
| | 100 | 0.5330(1.35) | 0.3501(1.58) | 0.01 | 0.3570(1.54) | 0.01 |
| | 150 | 0.5387(1.16) | 0.3597(1.34) | 0.01 | 0.3669(1.33) | 0.01 |
| | 300 | 0.5467(0.87) | 0.3708(1.13) | 0.01 | 0.3769(1.10) | 0.01 |

*Table B.13: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 4 (continued)*

| $n$ | $p$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.8$ | | | |
| 100 | 50 | 0.9628(3.11) | 0.8785(2.30) | 0.02 | 0.8653(1.60) | 0.02 |
| | 100 | 0.9798(2.81) | 0.9184(2.44) | 0.01 | 0.8965(1.79) | 0.01 |
| | 150 | 0.9834(2.49) | 0.9659(3.56) | 0.01 | 0.9378(2.47) | 0.01 |
| | 300 | 0.9863(1.73) | 1.1226(7.94) | 0.01 | 1.1361(8.25) | 0.01 |
| 200 | 50 | 0.9779(2.34) | 0.8458(2.57) | 0.01 | 0.8278(1.44) | 0.01 |
| | 100 | 0.9785(2.06) | 0.8679(2.50) | 0.01 | 0.8490(1.08) | 0.01 |
| | 150 | 0.9864(2.19) | 0.8755(0.96) | 0.01 | 0.8588(0.71) | 0.01 |
| | 300 | 0.9903(1.89) | 0.8937(1.33) | 0.01 | 0.8774(0.59) | 0.01 |
| 500 | 50 | 0.9880(0.79) | 0.8316(1.94) | 0.01 | 0.8129(0.89) | 0.01 |
| | 100 | 0.9964(0.70) | 0.8535(2.54) | 0.01 | 0.8351(0.56) | 0.01 |
| | 150 | 1.0009(0.33) | 0.8651(2.22) | 0.01 | 0.8460(0.49) | 0.01 |
| | 300 | 1.0050(0.24) | 0.8836(1.78) | 0.01 | 0.8633(0.47) | 0.01 |

*Table B.14: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 5*

| $n$ | $p$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0$ | | | |
| 100 | 50 | 0.3531(2.23) | 0.3165(1.67) | 0.10 | 0.3122(0.93) | 0.10 |
| | 100 | 0.3412(0.93) | 0.3320(1.06) | 0.10 | 0.3279(0.62) | 0.10 |
| | 150 | 0.3990(2.11) | 0.3445(1.01) | 0.10 | 0.3387(0.53) | 0.10 |
| | 300 | 0.4109(0.23) | 0.3555(0.95) | 0.10 | 0.3494(0.37) | 0.10 |
| 200 | 50 | 0.2644(1.23) | 0.2484(1.22) | 0.10 | 0.2439(0.93) | 0.10 |
| | 100 | 0.2745(1.37) | 0.2630(1.57) | 0.10 | 0.2561(0.73) | 0.10 |
| | 150 | 0.2876(1.63) | 0.2726(1.33) | 0.10 | 0.2646(0.51) | 0.10 |
| | 300 | 0.2854(1.09) | 0.2833(0.73) | 0.10 | 0.2749(0.32) | 0.10 |
| 500 | 50 | 0.1950(0.80) | 0.1839(1.38) | 0.10 | 0.1785(0.95) | 0.10 |
| | 100 | 0.2152(0.64) | 0.1920(0.64) | 0.10 | 0.1866(0.61) | 0.10 |
| | 150 | 0.2322(0.57) | 0.1990(0.66) | 0.10 | 0.1942(0.47) | 0.10 |
| | 300 | 0.2625(0.33) | 0.2094(0.54) | 0.10 | 0.2044(0.38) | 0.10 |

*Table B.15: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 5 (continued)*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.3$ | | | |
| 100 | 50 | 0.3299(1.17) | 0.3258(1.17) | 0.10 | 0.3249(0.88) | 0.10 |
| | 100 | 0.3450(0.97) | 0.3429(1.45) | 0.10 | 0.3396(0.59) | 0.10 |
| | 150 | 0.3542(0.77) | 0.3532(1.17) | 0.10 | 0.3495(0.49) | 0.10 |
| | 300 | 0.3652(0.79) | 0.3647(0.93) | 0.09 | 0.3607(0.31) | 0.09 |
| 200 | 50 | 0.2807(1.19) | 0.2591(1.56) | 0.10 | 0.2566(0.92) | 0.10 |
| | 100 | 0.2943(0.73) | 0.2714(1.04) | 0.10 | 0.2685(0.57) | 0.10 |
| | 150 | 0.3039(0.67) | 0.2852(1.58) | 0.09 | 0.2786(0.50) | 0.10 |
| | 300 | 0.3198(0.36) | 0.2964(0.72) | 0.07 | 0.2901(0.35) | 0.07 |
| 500 | 50 | 0.2293(1.08) | 0.1972(1.88) | 0.10 | 0.1936(0.89) | 0.10 |
| | 100 | 0.2476(0.74) | 0.2057(0.61) | 0.10 | 0.2043(0.60) | 0.09 |
| | 150 | 0.2576(0.63) | 0.2140(0.58) | 0.10 | 0.2122(0.55) | 0.10 |
| | 300 | 0.2728(0.42) | 0.2246(0.38) | 0.07 | 0.2223(0.33) | 0.08 |

*Table B.16: The Average (standard error in %) of Frobenius Norm-based IRSE for Scenario 5 (continued)*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.8$ | | | |
| 100 | 50 | 0.4619(1.91) | 0.4087(0.69) | 0.10 | 0.4114(0.56) | 0.10 |
| | 100 | 0.4757(1.82) | 0.4252(0.56) | 0.10 | 0.4276(0.41) | 0.10 |
| | 150 | 0.4829(1.85) | 0.4330(0.38) | 0.06 | 0.4357(0.33) | 0.06 |
| | 300 | 0.4841(1.94) | 0.4399(0.32) | 0.04 | 0.4424(0.18) | 0.03 |
| 200 | 50 | 0.4787(0.44) | 0.3809(1.12) | 0.10 | 0.3773(0.58) | 0.10 |
| | 100 | 0.4864(0.31) | 0.3927(0.49) | 0.04 | 0.3917(0.40) | 0.04 |
| | 150 | 0.4925(0.26) | 0.4005(0.45) | 0.03 | 0.3993(0.34) | 0.03 |
| | 300 | 0.4931(0.17) | 0.4083(0.26) | 0.02 | 0.4075(0.23) | 0.01 |
| 500 | 50 | 0.4613(0.43) | 0.3700(0.47) | 0.08 | 0.3692(0.43) | 0.08 |
| | 100 | 0.4698(0.29) | 0.3782(0.33) | 0.03 | 0.3770(0.30) | 0.03 |
| | 150 | 0.4758(0.22) | 0.3853(0.30) | 0.03 | 0.3837(0.28) | 0.02 |
| | 300 | 0.4768(0.16) | 0.3869(0.20) | 0.01 | 0.3851(0.18) | 0.01 |

*Table B.17: The Average SEN, SPE and ACC for Scenario 5 ($\rho = 0$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.4023 | 0.5510 | 0.5693 | 0.8623 | 0.9707 | 0.9708 | 0.7831 | 0.8985 | 0.9017 | 0.10 | 0.10 |
| | 100 | 0.2884 | 0.4884 | 0.5112 | 0.9841 | 0.9860 | 0.9860 | 0.9229 | 0.9422 | 0.9442 | 0.10 | 0.10 |
| | 150 | 0.3499 | 0.4502 | 0.4729 | 0.8826 | 0.9903 | 0.9906 | 0.8511 | 0.9584 | 0.9600 | 0.10 | 0.10 |
| | 300 | 0.1601 | 0.3840 | 0.4145 | 0.9996 | 0.9960 | 0.9957 | 0.9746 | 0.9778 | 0.9783 | 0.10 | 0.10 |
| 200 | 50 | 0.4546 | 0.7438 | 0.7678 | 0.8971 | 0.9710 | 0.9690 | 0.8210 | 0.9319 | 0.9344 | 0.10 | 0.10 |
| | 100 | 0.3963 | 0.6946 | 0.7182 | 0.9334 | 0.9840 | 0.9830 | 0.8861 | 0.9585 | 0.9597 | 0.10 | 0.10 |
| | 150 | 0.3755 | 0.6764 | 0.6911 | 0.9324 | 0.9876 | 0.9890 | 0.8995 | 0.9692 | 0.9714 | 0.10 | 0.10 |
| | 300 | 0.3344 | 0.6220 | 0.6449 | 0.9716 | 0.9944 | 0.9945 | 0.9526 | 0.9833 | 0.9841 | 0.10 | 0.10 |
| 500 | 50 | 0.5612 | 0.9022 | 0.9073 | 0.9144 | 0.9620 | 0.9635 | 0.8536 | 0.9517 | 0.9539 | 0.10 | 0.10 |
| | 100 | 0.4909 | 0.8752 | 0.8817 | 0.9226 | 0.9794 | 0.9811 | 0.8847 | 0.9703 | 0.9723 | 0.10 | 0.10 |
| | 150 | 0.4548 | 0.8578 | 0.8643 | 0.9214 | 0.9862 | 0.9875 | 0.8938 | 0.9787 | 0.9802 | 0.10 | 0.10 |
| | 300 | 0.4027 | 0.8203 | 0.8275 | 0.9212 | 0.9930 | 0.9938 | 0.9058 | 0.9879 | 0.9889 | 0.10 | 0.10 |

*Table B.18: The Average SEN, SPE and ACC for Scenario 5 ($\rho = 0.3$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.3465 | 0.5466 | 0.5687 | 0.9601 | 0.9619 | 0.9605 | 0.8546 | 0.8905 | 0.8931 | 0.10 | 0.10 |
| | 100 | 0.2800 | 0.4947 | 0.5143 | 0.9976 | 0.9804 | 0.9802 | 0.9345 | 0.9377 | 0.9392 | 0.10 | 0.10 |
| | 150 | 0.2583 | 0.4653 | 0.4889 | 0.9986 | 0.9860 | 0.9852 | 0.9549 | 0.9552 | 0.9559 | 0.10 | 0.10 |
| | 300 | 0.2223 | 0.3984 | 0.4193 | 0.9992 | 0.9936 | 0.9936 | 0.9760 | 0.9759 | 0.9765 | 0.09 | 0.09 |
| 200 | 50 | 0.4842 | 0.7550 | 0.7680 | 0.9901 | 0.9515 | 0.9522 | 0.9031 | 0.9177 | 0.9205 | 0.10 | 0.10 |
| | 100 | 0.4322 | 0.6689 | 0.7175 | 0.9966 | 0.9843 | 0.9752 | 0.9469 | 0.9565 | 0.9525 | 0.10 | 0.10 |
| | 150 | 0.4149 | 0.6633 | 0.6969 | 0.9976 | 0.9847 | 0.9808 | 0.9631 | 0.9657 | 0.9640 | 0.09 | 0.10 |
| | 300 | 0.3719 | 0.6091 | 0.6441 | 0.9990 | 0.9926 | 0.9908 | 0.9803 | 0.9812 | 0.9805 | 0.07 | 0.07 |
| 500 | 50 | 0.6473 | 0.9028 | 0.9068 | 0.9880 | 0.9390 | 0.9415 | 0.9294 | 0.9328 | 0.9355 | 0.10 | 0.10 |
| | 100 | 0.6145 | 0.8657 | 0.8787 | 0.9966 | 0.9704 | 0.9671 | 0.9630 | 0.9612 | 0.9593 | 0.10 | 0.09 |
| | 150 | 0.5970 | 0.8451 | 0.8639 | 0.9981 | 0.9811 | 0.9764 | 0.9744 | 0.9730 | 0.9698 | 0.10 | 0.10 |
| | 300 | 0.5605 | 0.8120 | 0.8255 | 0.9991 | 0.9892 | 0.9883 | 0.9860 | 0.9839 | 0.9834 | 0.07 | 0.08 |

*Table B.19: The Average SEN, SPE and ACC for Scenario 5 ($\rho = 0.8$)*

| $n$ | $p$ | SEN | | | SPE | | | ACC | | | Sig. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | ${}_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | $\text{DAC}_2$ | $\text{DAC}_1$ | $\text{DAC}_2$ |
| 100 | 50 | 0.3833 | 0.5765 | 0.6101 | 0.8805 | 0.8972 | 0.8858 | 0.7950 | 0.8420 | 0.8384 | 0.10 | 0.10 |
| | 100 | 0.2988 | 0.5250 | 0.5195 | 0.9439 | 0.9366 | 0.9470 | 0.8871 | 0.9004 | 0.9094 | 0.10 | 0.10 |
| | 150 | 0.2841 | 0.4609 | 0.4713 | 0.9430 | 0.9636 | 0.9651 | 0.9040 | 0.9339 | 0.9360 | 0.06 | 0.06 |
| | 300 | 0.2301 | 0.3535 | 0.3871 | 0.9717 | 0.9873 | 0.9844 | 0.9496 | 0.9685 | 0.9666 | 0.04 | 0.03 |
| 200 | 50 | 0.4036 | 0.7544 | 0.7542 | 0.9776 | 0.8645 | 0.8758 | 0.8789 | 0.8456 | 0.8549 | 0.10 | 0.10 |
| | 100 | 0.3471 | 0.6666 | 0.6896 | 0.9967 | 0.9451 | 0.9352 | 0.9396 | 0.9205 | 0.9136 | 0.04 | 0.04 |
| | 150 | 0.3342 | 0.6243 | 0.6237 | 0.9979 | 0.9634 | 0.9660 | 0.9586 | 0.9434 | 0.9457 | 0.03 | 0.03 |
| | 300 | 0.3096 | 0.5409 | 0.5669 | 0.9992 | 0.9859 | 0.9821 | 0.9786 | 0.9726 | 0.9697 | 0.02 | 0.01 |
| 500 | 50 | 0.5676 | 0.8829 | 0.8794 | 0.9757 | 0.8670 | 0.8696 | 0.9055 | 0.8697 | 0.8713 | 0.08 | 0.08 |
| | 100 | 0.5253 | 0.8144 | 0.8217 | 0.9940 | 0.9509 | 0.9450 | 0.9528 | 0.9389 | 0.9342 | 0.03 | 0.03 |
| | 150 | 0.5123 | 0.7839 | 0.7928 | 0.9968 | 0.9687 | 0.9636 | 0.9682 | 0.9577 | 0.9535 | 0.03 | 0.02 |
| | 300 | 0.4905 | 0.7426 | 0.7526 | 0.9986 | 0.9820 | 0.9785 | 0.9835 | 0.9749 | 0.9718 | 0.01 | 0.01 |

*Table B.20: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 5*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0$ | | | |
| 100 | 50 | 1.2162(11.74) | 0.8466(5.58) | 0.10 | 0.8525(4.35) | 0.10 |
| | 100 | 1.0491(10.92) | 0.9348(4.75) | 0.10 | 0.9399(3.69) | 0.10 |
| | 150 | 1.8601(29.22) | 0.9826(3.90) | 0.10 | 0.9861(2.78) | 0.10 |
| | 300 | 1.1458(2.16) | 1.0229(3.11) | 0.10 | 1.0219(2.68) | 0.10 |
| 200 | 50 | 0.8304(8.13) | 0.6727(4.18) | 0.10 | 0.6686(3.90) | 0.10 |
| | 100 | 0.9486(14.28) | 0.7485(4.24) | 0.10 | 0.7439(3.74) | 0.10 |
| | 150 | 1.1196(21.36) | 0.7878(3.13) | 0.10 | 0.7823(2.96) | 0.10 |
| | 300 | 1.0941(20.72) | 0.8352(3.06) | 0.10 | 0.8268(2.35) | 0.10 |
| 500 | 50 | 0.6042(3.16) | 0.4904(3.50) | 0.10 | 0.4886(3.01) | 0.10 |
| | 100 | 0.7708(2.91) | 0.5407(2.96) | 0.10 | 0.5381(2.85) | 0.10 |
| | 150 | 0.9308(2.97) | 0.5795(2.75) | 0.10 | 0.5779(2.61) | 0.10 |
| | 300 | 1.3403(3.26) | 0.6186(2.22) | 0.10 | 0.6147(2.08) | 0.10 |

*Table B.21: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 5 (continued)*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.3$ | | | |
| 100 | 50 | 0.9350(7.80) | 0.8843(4.18) | 0.10 | 0.8935(3.94) | 0.10 |
| | 100 | 0.9888(3.71) | 0.9802(4.53) | 0.10 | 0.9835(3.47) | 0.10 |
| | 150 | 1.0266(2.78) | 1.0203(3.62) | 0.10 | 1.0222(2.91) | 0.10 |
| | 300 | 1.0671(5.54) | 1.0569(2.59) | 0.09 | 1.0581(2.34) | 0.09 |
| 200 | 50 | 0.7887(3.92) | 0.7079(4.50) | 0.10 | 0.7096(4.03) | 0.10 |
| | 100 | 0.8632(3.29) | 0.7823(3.58) | 0.10 | 0.7816(3.16) | 0.10 |
| | 150 | 0.9086(2.72) | 0.8351(3.54) | 0.09 | 0.8322(3.11) | 0.10 |
| | 300 | 0.9545(1.98) | 0.8814(3.18) | 0.07 | 0.8735(2.50) | 0.07 |
| 500 | 50 | 0.6627(3.29) | 0.5355(4.25) | 0.10 | 0.5373(3.64) | 0.10 |
| | 100 | 0.7451(2.77) | 0.5898(2.60) | 0.10 | 0.5928(2.56) | 0.09 |
| | 150 | 0.7896(2.23) | 0.6284(2.71) | 0.10 | 0.6316(2.74) | 0.10 |
| | 300 | 0.8361(1.97) | 0.6737(2.47) | 0.07 | 0.6733(2.29) | 0.08 |

*Table B.22: The Average (standard error in %) of Spectral Norm-based IRSE for Scenario 5 (continued)*

| $n$ | $p$ | $_{\text{st}}\text{NCM}_1$ | $\text{DAC}_1$ | Sig. | $\text{DAC}_2$ | Sig. |
|---|---|---|---|---|---|---|
| | | | $\rho = 0.8$ | | | |
| 100 | 50 | 1.1946(4.58) | 1.0905(3.03) | 0.10 | 1.0966(2.78) | 0.10 |
| | 100 | 1.2773(4.40) | 1.1765(2.42) | 0.10 | 1.1826(2.22) | 0.10 |
| | 150 | 1.3022(4.41) | 1.2054(1.76) | 0.06 | 1.2109(1.63) | 0.06 |
| | 300 | 1.3283(10.18) | 1.2244(1.33) | 0.04 | 1.2304(1.08) | 0.03 |
| 200 | 50 | 1.2322(1.28) | 1.0299(4.85) | 0.10 | 1.0174(2.95) | 0.10 |
| | 100 | 1.2963(0.96) | 1.1032(2.41) | 0.04 | 1.0953(2.18) | 0.04 |
| | 150 | 1.3130(0.57) | 1.1431(3.23) | 0.03 | 1.1322(1.56) | 0.03 |
| | 300 | 1.3186(0.47) | 1.1670(1.22) | 0.02 | 1.1583(1.27) | 0.01 |
| 500 | 50 | 1.1894(1.46) | 0.9990(2.01) | 0.08 | 0.9921(2.04) | 0.08 |
| | 100 | 1.2546(1.04) | 1.0667(1.60) | 0.03 | 1.0555(1.70) | 0.03 |
| | 150 | 1.2763(0.69) | 1.1048(1.37) | 0.03 | 1.0917(1.35) | 0.02 |
| | 300 | 1.2839(0.62) | 1.1243(1.18) | 0.01 | 1.1089(1.06) | 0.01 |

# Appendix C

# Proofs and Results of Chapter 5

## C.1 Technical Details for AR($p$), MA($q$) and ARMA($p, q$) Processes

To prove Theorem 5.1, we need the following lemma

**Lemma C.1.** *If multivariate variable* $\mathbf{x} \sim N_p(\mu, \Sigma)$*, then the entropy of* $\mathbf{x}$*, denoted as* $h(\mathbf{x})$*, is*

$$h(\mathbf{x}) = \frac{1}{2} \log\left((2\pi e)^p |\Sigma|\right).$$

*Proof.* By the definition of continuous entropy, we have

$$\begin{aligned} h(\mathbf{x}) &= -\int_{\mathbb{R}^p} f(\mathbf{x}) \log\left(f(\mathbf{x})\right) d\mathbf{x}, \\ &= \frac{p}{2}\log(2\pi) + \frac{1}{2}\log(|\Sigma|) + \frac{1}{2}\mathbf{E}\left[(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right], \\ &= \frac{1}{2} \log\left((2\pi e)^p |\Sigma|\right), \end{aligned}$$

this completes the proof. □

Let $\Sigma_m, \Sigma_{11;ms}, \Sigma_{22;ms}$ represent the auto-covariance matrices of vectors $\mathbf{x}^{(m+1)}$, $\mathbf{x}^{(m+1-s)}$ and $\mathbf{x}^{(s)}$ respectively, $\Sigma_{12;ms}$ is the covariance matrix between $\mathbf{x}^{(m+1-s)}$ and $\mathbf{x}^{(s)}$, i.e.,

$$\Sigma_m = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_m \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_m & \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}, \Sigma_{11;ms} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-s} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-s-1} \\ \gamma_2 & \gamma_1 & \cdots & \gamma_{m-s-2} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{m-s} & \gamma_{m-s-1} & \cdots & \gamma_0 \end{bmatrix},$$

and

$$\Sigma_{22;ms} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{s-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{s-2} \\ \gamma_2 & \gamma_1 & \cdots & \gamma_{s-3} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{s-1} & \gamma_{s-2} & \cdots & \gamma_0 \end{bmatrix}, \Sigma_{12;ms} = \begin{bmatrix} \gamma_{m-s+1} & \gamma_{m-s+2} & \cdots & \gamma_m \\ \gamma_{m-s} & \gamma_{m-s+1} & \cdots & \gamma_{m-1} \\ \gamma_{m-s-1} & \gamma_{m-s} & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_1 & \gamma_2 & \cdots & \gamma_s \end{bmatrix}.$$

For simplicity, we have

$$\Sigma_m = \begin{bmatrix} \Sigma_{11;ms} & \Sigma_{12;ms} \\ \Sigma_{21;ms} & \Sigma_{22;ms} \end{bmatrix},$$

we also notice that $R_m = \gamma_0^{-1}\Sigma_m$, $R_{11;ms} = \gamma_0^{-1}\Sigma_{11;ms}$, $R_{12;ms} = \gamma_0^{-1}\Sigma_{12;ms}$, $R_{22;s} = \gamma_0^{-1}\Sigma_{22;ms}$. Based on these facts, we now prove Proposition 5.1.

*Proof of Proposition 5.1.* Since $\varepsilon_i$ is the Gaussian white noise, the distribution of $\mathbf{x}^{(m+1)}$, $\mathbf{x}^{(m+1-s)}$ and $\mathbf{x}^{(s)}$ are multivariate Gaussian with covariance matrices $\Sigma_m, \Sigma_{11;ms}, \Sigma_{22;ms}$ respectively. By Lemma C.1, we can obtain the following results: $h\left(\mathbf{x}^{(m+1)}\right) = 2^{-1}\log\left((2\pi e)^{m+1}\left|\Sigma_m\right|\right)$, $h\left(\mathbf{x}^{(s)}\right) = 2^{-1}\log\left((2\pi e)^{s}\left|\Sigma_{22;ms}\right|\right)$ and $h\left(\mathbf{x}^{(m+1-s)}\right) = 2^{-1}\log\left((2\pi e)^{m+1-s}\left|\Sigma_{11;ms}\right|\right)$. The relative entropy can be expressed as

$$\begin{aligned} \mathcal{I}_s\left(\mathbf{x}^{(m+1)}\right) = \mathrm{RlEn}_s &= \int_{\mathbb{R}^{m+1}} f\left(\mathbf{x}^{(m+1)}\right)\log\left(\frac{f\left(\mathbf{x}^{(m+1)}\right)}{g\left(\mathbf{x}^{(m+1-s)}\right)g\left(\mathbf{x}^{(s)}\right)}\right)d\mathbf{x}^{(m+1)}, \\ &= \frac{1}{2}\log\left(\frac{\left|R_{11;ms}\right|\left|R_{22;ms}\right|}{\left|R_m\right|}\right), \quad 1 \le s \le m, \quad m \ge 1. \end{aligned}$$

By the Yule-Walker equations,

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_p \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_2 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \\ \phi_p \end{bmatrix},$$

the autocorrelation $\rho_i, i = 1, 2, \ldots, \infty$ are totally determined by coefficient $\phi_k, k = 1, \ldots, p$. Hence, $\mathcal{I}_s(\mathbf{x}^{(m+1)})$ is a function of $\phi_k, k = 1, \ldots, p$. As $\{x_i\}$ and $\{\varepsilon_i\}$ are independent, the autocorrelation function of $\{x_i\}$ is independent of $\sigma$ which implies $\mathcal{I}_s(\mathbf{x}^{(m+1)})$ is independent of $\sigma$ as well which completes the proof. □

*Proof of Proposition 5.3.* First, when $m = p$, we prove that

$$\mathcal{I}_1\left(\mathbf{x}^{(m+1)}\right) = -\frac{1}{2}\log\left(1 - \sum_{k=1}^{p}\phi_k\rho_k\right).$$

Next, we prove it still holds when $m > p$.

Note that the $\mathcal{I}_s(\mathbf{x}^{(m+1)})$ relates to the determinant of block matrix, we introduce a general result of determinant of block matrix below. For any block matrix

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

if matrix $A$ is invertible, we have $|M| = |D - CA^{-1}B|\,|A|$. Applying this result to relative entropy when $s = 1$, we have

$$\begin{aligned}\mathcal{I}_1\left(\mathbf{x}^{(m+1)}\right) &= \frac{1}{2}\log\left(\frac{|R_{11;p1}|\,|R_{22;p1}|}{|R_p|}\right), \\ &= \frac{1}{2}\log\left(\frac{1}{1 - R_{21;p1}R_{11;p1}^{-1}R_{12;p1}}\right),\end{aligned}$$

where $R_{21;p1} = (\rho_p, \ldots, \rho_1)$, $R_{12;p1} = R_{21;p1}^T$ and

$$R_{11;p1} = \Psi_p = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_2 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix}.$$

Let $\boldsymbol{\rho}_p = (\rho_1, \ldots, \rho_p)^T = PR_{21;p1}^T$, where $P$ is permutation matrix, i.e.,

$$P = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix},$$

then $R_{21:p1}R_{11:p1}^{-1}R_{12:p1} = \boldsymbol{\rho}_p^T(PR_{11:p1}^{-1}P)\boldsymbol{\rho}_p$. The term $(PR_{11:p1}^{-1}P)$ means that we apply the same permutation $P$ twice to $R_{11:p1}^{-1}$, so $(PR_{11:p1}^{-1}P) = R_{11:p1}^{-1}$. Denote $\Phi_p = (\phi_1, \ldots, \phi_p)^T$, then the matrix form of Yule-Walker equation is

$\boldsymbol{\rho}_p = R_{11:p1}\Phi_p$, finally we have

$$R_{21:p1}R_{11:p1}^{-1}R_{12:p1} = \boldsymbol{\rho}_p^T R_{11:p1}^{-1}\boldsymbol{\rho}_p = \boldsymbol{\rho}_p^T\Phi_p = \sum_{k=1}^{p}\rho_k\phi_k,$$

which completes the proof when $m = p$.

When $m > p$, denote $\boldsymbol{\rho}_{m-p} = (\rho_{p+1}, \ldots, \rho_m)^T$, then $R_{21:m1} = (\rho_m, \ldots, \rho_1) = P(\boldsymbol{\rho}_p^T, \boldsymbol{\rho}_{m-p}^T)^T$. Similarly, we still have

$$\begin{aligned} R_{21:m1}R_{11:m1}^{-1}R_{12:m1} &= (\boldsymbol{\rho}_p^T, \boldsymbol{\rho}_{m-p}^T)PR_{11:m1}^{-1}P(\boldsymbol{\rho}_p^T, \boldsymbol{\rho}_{m-p}^T)^T \\ &= (\boldsymbol{\rho}_p^T, \boldsymbol{\rho}_{m-p}^T)R_{11:m1}^{-1}(\boldsymbol{\rho}_p^T, \boldsymbol{\rho}_{m-p}^T)^T. \end{aligned}$$

Now $R_{11:m1}$ is $m \times m$ size symmetric matrix. We divide it into four blocks, i.e.,

$$R_{11:m1} = \begin{bmatrix} \Psi_p & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix},$$

where

$$\Psi_{12} = \Psi_{21}^T = \begin{bmatrix} \rho_p & \cdots & \rho_{m-1} \\ \rho_{p-1} & \cdots & \rho_{m-2} \\ \vdots & \vdots & \vdots \\ \rho_1 & \cdots & \rho_{m-p} \end{bmatrix}, \Psi_{22} = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{m-p-1} \\ \rho_1 & 1 & \cdots & \rho_{m-p-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{m-p-1} & \rho_{m-p-2} & \cdots & 1 \end{bmatrix}.$$

Since AR($p$) is stationary, then $R_{11:m1}$ is invertible. Let $S = \left(\Psi_{22} - \Psi_{21}\Psi_p^{-1}\Psi_{12}\right)^{-1}$, the inverse of $R_{11:m1}$ is

$$R_{11:m1}^{-1} = \begin{bmatrix} \Psi_p^{-1} + \Psi_p^{-1}\Psi_{12}S\Psi_{21}\Psi_p^{-1} & -\Psi_p^{-1}\Psi_{12}S \\ -S\Psi_{21}\Psi_p^{-1} & S \end{bmatrix},$$

therefore

$$\begin{aligned} R_{21:m1}R_{11:m1}^{-1}R_{12:m1} &= (\boldsymbol{\rho}_p^T, \boldsymbol{\rho}_{m-p}^T)R_{11:m1}^{-1}(\boldsymbol{\rho}_p^T, \boldsymbol{\rho}_{m-p}^T)^T, \\ &= \boldsymbol{\rho}_p^T\Psi_p^{-1}\boldsymbol{\rho}_p + \boldsymbol{\rho}_p^T\Psi_p^{-1}\Psi_{12}S\Psi_{21}\Psi_p^{-1}\boldsymbol{\rho}_p - \\ &\quad \boldsymbol{\rho}_{m-p}^T S\Psi_{21}\Psi_p^{-1}\boldsymbol{\rho}_p - \boldsymbol{\rho}_p^T\Psi_p^{-1}\Psi_{12}S\boldsymbol{\rho}_{m-p} + \boldsymbol{\rho}_{m-p}^T S\boldsymbol{\rho}_{m-p}, \\ &= \boldsymbol{\rho}_p^T\Psi_p^{-1}\boldsymbol{\rho}_p + \left(\boldsymbol{\rho}_p^T\Psi_p^{-1}\Psi_{12} - \boldsymbol{\rho}_{m-p}^T\right)S\left(\Psi_{21}\Psi_p^{-1}\boldsymbol{\rho}_p - \boldsymbol{\rho}_{m-p}\right), \\ &= \boldsymbol{\rho}_p^T\Psi_p^{-1}\boldsymbol{\rho}_p + \left(\Phi_p^T\Psi_{12} - \boldsymbol{\rho}_{m-p}^T\right)S\left(\Psi_{21}\Phi_p - \boldsymbol{\rho}_{m-p}\right). \end{aligned}$$

By Yule-Walker equation, it is easy to verify that $\Phi_p^T\Psi_{12} - \boldsymbol{\rho}_{m-p}^T = 0$. Combining

the previous result $\boldsymbol{\rho}_p^T \Psi_p^{-1} \boldsymbol{\rho}_p = \sum_{k=1}^{p} \rho_k \phi_k$, we finally complete the whole proof. □

*Proof of Proposition 5.4.* Since $\{x_i\}$ is stationary moving average process and $\varepsilon_i, i = 1, 2, \ldots$ are i.i.d., the auto-covariance function of $\{x_i\}$ can be expressed as

$$\gamma_\tau = \begin{cases} \sigma^2 \sum_{j=0}^{q-|\tau|} \theta_j \theta_{j+|\tau|} & \text{if } |\tau| \leq q. \\ 0 & \text{if } |\tau| > q. \end{cases}$$

$$\rho_\tau = \begin{cases} \sum_{j=0}^{q-|\tau|} \theta_j \theta_{j+|\tau|} / \sum_{j=0}^{q} \theta_j^2 & \text{if } |\tau| \leq q. \\ 0 & \text{if } |\tau| > q. \end{cases}$$

where $\theta_0 = 1$. By the similar discussion in Proposition 5.1's proof, equation (5.10) immediately holds. Furthermore, using the block matrix inversion manipulation in the proof of Proposition 5.3, one can verify that when $s = 1$ and $m \geq q_1$ $\mathcal{I}_s(\mathbf{x}^{(m+1)}) = -2^{-1} \log(1 - R^{(1)}_{12;q_1 1}(R^{(1)}_{11;q_1 1})^{-1} R^{(1)}_{21;q_1 1})$. This completes the proof. □

*Proof of Proposition 5.5.* Using the characteristic polynomial, the ARMA$(p, q)$ process can be expressed as

$$\phi(L) x_i = \theta(L) \varepsilon_i, \quad \text{or } x_i = \psi(L) \varepsilon_i,$$

where $\psi(L) = \phi(L)/\theta(L)$. If ARMA$(p, q)$ process is stationary, by Wold representation (Wold, 1948), we have $\psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$. Then $\gamma_0 = \sigma^2 \sum_{j=0}^{\infty} \psi_j < \infty$. It is easy to verify that

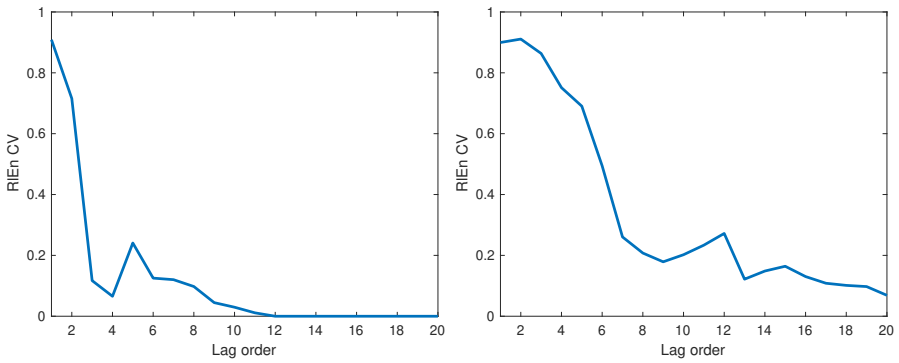
$$\begin{aligned} &\rho_\tau - \phi_1 \rho_{\tau-1} - \cdots - \phi_p \rho_{\tau-p} = 0, \quad \text{for } \tau \geq \max(p, q+1), \\ &\rho_\tau - \phi_1 \rho_{\tau-1} - \cdots - \phi_p \rho_{\tau-p} = c_\tau / \sum_{j=0}^{\infty} \psi_j, \quad \text{for } 0 \leq \tau < \max(p, q+1), \end{aligned}$$

where $c_\tau = \theta_\tau \psi_0 + \theta_{\tau+1} \psi_1 + \cdots + \theta_q \psi_{q-\tau}$. By the similar discussion in Proposition 5.1's proof, Proposition 5.5 immediately holds which completes the proof. □

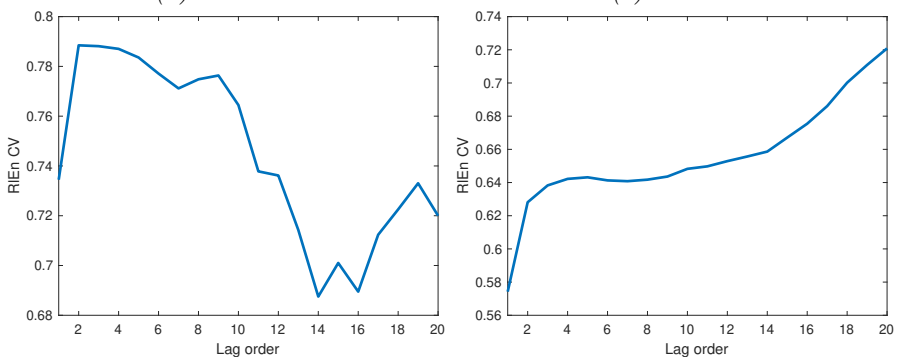
## C.2 Lag Order Selection and Proof

Figure C.1 shows the results of the first time series in Case 2.

*Proof of Theorem 5.1.* The proof is based on the proofs of Vieu (1995) and Shao (1997). First, we construct a new equation $\sigma_\lambda^2(m) = \hat{\sigma}_e^2(m)[1 + \lambda_n(m, \hat{h}_m)]$, where $\lambda_n(m, \hat{h}_m) = v(m, \hat{h}_m) \log(n)/n$. We can regard $\lambda_n(m, \hat{h}_m)$ as a penalty part in $\sigma_\lambda^2(m)$, when $\lambda_n(m, \hat{h}_m) \to 0$, minimizing $BIC(m)$ is equivalent to minimizing $\sigma_\lambda^2(m)$ based on the fact that $\log(1+x) \approx x$ as $x \to 0$. This proof skill is frequently adopted in discussion of BIC or AIC consistency, see, for example, Shibata (1981,




*(a)* $h = 0.05$ *(b)* $h = 0.1$




*(c)* $h = 0.2$ *(d)* $h = 0.3$

*Figure C.1: Relative Entropy against Lag Order for Different Bandwidths*

p. 46), Vieu (1995, p. 314) and Shao (1997, p. 232). Therefore, the sketch of our proof can be summarized into the following two steps: (1) We need to discuss the consistency of lag order selected via $\hat{\sigma}_e^2(m)$; (2) We extend the result to the penalty version $\sigma_\lambda^2(m)$ with controlling $\lambda_n(m, \hat{h}_m) \to 0$ in an appropriate rate. This proof is very similar to that in lag order selection (Vieu, 1995) except the definitions of $\hat{\sigma}_e^2(m)$ and $\lambda_n(m, \hat{h}_m)$. Next, we introduce the conditions used in our proof.

(C11) The time series $\{X_i\}_{i\in\mathbb{N}}$ is $\alpha$-mixing, the mixing coefficient $\alpha(n)$ satisfies: $\exists\ s > 0, \exists\ 0 < t < 1, \forall\ n \geq 1, \alpha(n) \leq st^n$.

(C12) For each $1 \leq m < M$, there exists the nonlinear autoregression functions such that

$$x_{i+m} = \mathfrak{F}\left(\mathbf{X}_i^{(m)}\right) + \varepsilon_{i,m},$$

where $\mathbf{X}_i^{(m)}$ is independent of $\varepsilon_{i,m}$ and $\varepsilon_{i,m}, i = 1, \ldots, n$ are noise with mean zero.

(C13) The unknown function $\mathfrak{F}$ has second-order continuous derivation.

(C14) $\forall\ q \geq 1, \exists\ \mathbb{M}_q$ such that for any $i, \mathrm{E}|X_i|^q \leq \mathbb{M}_q < \infty$.

(C15) Given $m$, for some $0 < \gamma_m < \infty$ and some $0 < \eta_m < 1/(4+m)$, the bandwidth satisfies:

$$h_* \in \mathcal{H}_{n,m} = \left[\gamma_m^{-1} n^{-\eta_m - 1/(4+m)}, \gamma_m n^{\eta_m - 1/(4+m)}\right].$$

(C16) $\lambda_n(m, \hat{h}_m)$ is of order $\log(n)/(n(h_*)^m)$.

(C17) $m$ could be arbitrary large but has an upper bound $M$.

Conditions (C11)–(C16) are quoted from Vieu (1995, pp. 310–311) with appropriate adjustments for our circumstance. Condition (C17) controls the upper bound $M$ to coordinate the proof of RlEn in Appendix C.3. Given lag order $m, 1 \leq m \leq M$ and the underlying lag order $m_0$, we define the distance between $\mathfrak{F}(\mathbf{X}_i^{(m_0)})$ and $\hat{\mathfrak{F}}(\mathbf{X}_i^{(m)}, h_*)$ by

$$\sigma_0^2(m, h_*) = \frac{1}{N - \max(m, m_0)} \sum_{i=1}^{\max(m,m_0)} \left[\mathfrak{F}\left(\mathbf{X}_i^{(m_0)}\right) - \hat{\mathfrak{F}}\left(\mathbf{X}_i^{(m)}, h_*\right)\right]^2.$$

Furthermore, let $\sigma_0^2(m) = \inf_{h_* \in \mathcal{H}_n} \sigma_0^2(m, h_*)$, we have

**Lemma C.2.** *Given Conditions (C11)–(C15), Assumption 1 and Assumption 2, the nonlinear autoregression process (5.14) has underlying lag order $m_0$ and $m_0 \in \{1, \ldots, M\}$, then we have*

*(a)* $\sigma_0^2(m_0) \to 0$, *a.e.,*

*(b) For* $1 \leq m < m_0$, *there exists real positive constant* $c_m > 0$ *such that*

$$\hat{\sigma}_e^2(m) - \hat{\sigma}_e^2(m_0) \geq c_m, \quad a.s.,$$

*(c) For* $m_0 < m \leq M$, $\exists\ c_0 > 0$ *s.t.* $\hat{\sigma}_e^2(m_0) - c_0 \geq 0$, *a.s. and*

$$\frac{\hat{\sigma}_e^2(m) - c_0}{\hat{\sigma}_e^2(m_0) - c_0} \to +\infty, \quad a.s.$$

*Proof of Lemma C.2.* This proof employs the same techniques used in Lemma 1, Lemma 2 and Theorem 3 in Vieu (1995). It can be regarded as a special case of Vieu (1995) except the upper bound $M$. Give $m$, $n = N - m$, the average square predict error of nonlinear autoregression is defined as

$$\sigma^2(m, h_*) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathfrak{F}\left(\mathbf{X}_i^{(m)}\right) - \hat{\mathfrak{F}}\left(\mathbf{X}_i^{(m)}, h_*\right)\right]^2.$$

Under Assumptions 1 and 2, we can rewrite the average square predict error (e.g., Li & Racine, 2007, pp. 83–85) as

$$\sigma^2(m, h_*) = \alpha_{1m} \frac{1}{n h_*^m} + \alpha_{2m} m h_*^4 + o\left(\frac{1}{n h_*^m} + h_*^4\right), \quad a.s., \tag{C.1}$$

where $\alpha_{1m}$ and $\alpha_{2m}$ are constant. We can easily obtain the optimal bandwidth if

we minimize the first two leading terms in equation (C.1), denoted as

$$\hat{h}_m = \left(\frac{4\alpha_{2m}}{\alpha_{1m}} n\right)^{-\frac{1}{4+m}}.$$

Finally, we get

$$\sigma^2(m) = \inf_{h_* \in \mathcal{H}_{n,m}} \sigma^2(m, h_*) = \alpha_{3m} n^{-\frac{4}{4+m}} + o\left(n^{-\frac{4}{4+m}}\right), \quad a.s., \tag{C.2}$$

where $\alpha_{3m}$ is a constant. Especially, when $m = m_0$, then $\sigma^2(m) = \sigma_0^2(m_0)$, immediately, $\sigma_0^2(m_0) \to 0$ holds almost surely. This completes the proof of (a).

*Proof of (b).* For given $m$, $1 \le m < m_0$, let $\hat{h}_{m,cv} = \arg\min_{h_* \in \mathcal{H}_{n,m}} \hat{\sigma}_e^2(m, h_*)$, the bandwidth selected by least square cross-validation is still of order $n^{-1/(4+m)}$ as $\hat{h}_m$ (e.g., Vieu, 1991; Hall *et al.*, 2004), so $\hat{h}_{m,cv} \in \mathcal{H}_{n,m}$, we have

$$\frac{\sigma^2(m, \hat{h}_{m,cv})}{\inf\limits_{h_* \in \mathcal{H}_{n,m}} \sigma^2(m, h_*)} \to 1, \quad a.s. \tag{C.3}$$

In the proof of this property, Vieu (1991) and Vieu (1995) employed the following statement for nonlinear autoregression:

$$\hat{\sigma}_e^2(m, h_*) - \sigma^2(m, h_*) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{i,m}^2 + o\left(\sigma^2(m, h_*)\right), \quad a.s., \tag{C.4}$$

where the operation $o(\cdot)$ is uniform over $h_* \in \mathcal{H}_{n,m}$. Therefore, by equation (C.2) and equation (C.3), $\forall\ m,\ 1 \le m < m_0$, we have $\sigma^2(m, \hat{h}_{m,cv}) = o(n^{-4/(4+m)})$ almost surely. Then, minimizing equation (C.4), we obtain for any $1 \le m < m_0$

$$\hat{\sigma}_e^2(m) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{i,m}^2 + o(1), \quad a.s. \tag{C.5}$$

Let $\delta_m = \mathrm{Var}(\varepsilon_{i,m})$ and $c_m = \delta_m - \delta_{m_0}$, by equation (C.5), we have $\hat{\sigma}_e^2(m) - \hat{\sigma}_e^2(m_0) \to c_m$, a.s., and $c_m > 0$ because from (C15) and Assumption 2, we know $c_m \ge \mathrm{Var}\left[\mathrm{E}(x_{i+m}|\mathbf{X}_i^{(m_0)}) - \mathrm{E}(x_{i+m}|\mathbf{X}_i^{(m)})\right]$, because $1 \le m < m_0$, so $\mathrm{Var}\left[\mathrm{E}(x_{i+m}|\mathbf{X}_i^{(m_0)}) - \mathrm{E}(x_{i+m}|\mathbf{X}_i^{(m)})\right] > 0$ which completes the proof (b).

*Proof of (c).* For $m_0 < m \le M$, replacing $h_*$ in equation (C.4) with $\hat{h}_{m,cv}$, we obtain

$$\hat{\sigma}_e^2(m) - \sigma^2(m, \hat{h}_{m,cv}) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{i,m}^2 + o\left(n^{-4/(4+m)}\right), \quad a.s.$$

We also note that (C.3) implies

$$\sigma^2(m, \hat{h}_{m,cv}) = \inf_{h_* \in \mathcal{H}_{n,m}} \sigma^2(m, h_*) + o\left(n^{-4/(4+m)}\right), \quad a.s.,$$

therefore

$$\hat{\sigma}_e^2(m) = \sigma_0^2(m) + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i,m}^2 + o\left(n^{-4/(4+m)}\right), \quad a.s. \tag{C.6}$$

Like the discussion in Vieu (1995), by Bernstein's inequality for $\alpha$-mixing, for example, see the Theorem 3.1 in Roussas & Ioannides (1988), we have

$$\frac{1}{N-m_0}\sum_{i=1}^{N-m_0} \varepsilon_{i,m_0}^2 - \mathrm{Var}(\varepsilon_{i,m_0}) = o\left(n^{-4/(4+m_0)}\right), \quad a.s. \tag{C.7}$$

Combining (C.6) and (C.7), we get

$$\frac{|\hat{\sigma}_e^2(m) - c_0|}{|\hat{\sigma}_e^2(m_0) - c_0|} = \frac{\sigma_0^2(m)}{\sigma_0^2(m_0)} + o\left(\frac{\sigma_0^2(m)}{\sigma_0^2(m_0)}\right), \quad a.s.,$$

where $c_0 = \mathrm{Var}(\varepsilon_{i,m_0})$. By the fact (C.2), we have $\sigma_0^2(m)/\sigma_0^2(m_0) \to \infty$ almost surely. Immediately, by (C.6), we can conclude $\hat{\sigma}_e^2(m_0) - c_0 \geq 0$, a.s., because $\sigma_0^2(m)$ is positive. This completes the proof. □

Let $\bar{m} = \arg\min \hat{\sigma}_e^2(m)$, based on Lemma C.2, we immediately have

$$\sigma_0^2(\bar{m})/\sigma_0^2(m_0) \to 1, \quad a.s.$$

Moreover, we add the penalty part to $\hat{\sigma}_e^2(m)$, i.e., $\sigma_\lambda^2(m) = \hat{\sigma}_e^2(m)[1 + \lambda_n(m, \hat{h}_m)]$, where $\lambda_n(m, \hat{h}_m) = v(m, \hat{h}_m)\log(n)/n$. Based on Lemma 5.1, Condition (C16) makes sure that the penalty part is not arbitrary large compared with 0. Based on Lemma C.2, previous discussion and Condition (C17), we have

$$\max_{m=1,\ldots,M} \frac{|\sigma_\lambda^2(m) - \hat{\sigma}_e^2(m)|}{\sigma_0^2(m)} = \frac{\log(n)}{n^{\frac{4}{4+m}}} \leq \frac{\log(n)}{n^{\frac{4}{4+O\left(\sqrt{\log(n)}\right)}}} = o(1), \quad a.s. \tag{C.8}$$

Let $\hat{m} = \arg\min \sigma_\lambda^2(m)$, we have $\sigma_0^2(\hat{m}_\lambda)/\sigma_0^2(m_0) \to 1$ almost surely. Because the previous results are almost surely convergence, so

$$P(\hat{m} = m_0) \to 1,$$

holds as well which completes the whole proof. □

Note: Vieu (1995) claimed $M = O(\log(n))$, however, our result shows that $M$ is at least of order $O(\sqrt{\log(n)})$, see equation (C.8). Furthermore, if one want to control $M$ to tend to infinity not as fast as $O(\sqrt{\log(n)})$, the order of $M$ could

be $O(\log(\log(n)))$. However, in order to keep $M$ consistent in the proof of RlEn theory, we sacrifice the relaxation of $m$ to infinity discussed above.

## C.3 Consistency of RlEn

*Proof of Lemma 5.2.* Under Assumptions 1 and 2, we can obtain the uniform rates of convergence for multivariate kernel density estimator (e.g., Li & Racine, 2007, pp. 30–32). For $\mathbf{z} \in \mathbb{I}^m$, it follows that

$$\sup_{\mathbf{z}\in\mathbb{I}^m} |\hat{g}(\mathbf{z}) - g(\mathbf{z})| = O_p\left(\frac{(\log n)^{1/2}}{n^{1/2}h^{m/2}} + mh^2\right), \tag{C.9}$$

and

$$\sup_{\mathbf{z}\in\mathbb{I}^m} \mathrm{E}\left\{[\hat{g}(\mathbf{z}) - g(\mathbf{z})]^2\right\} = O\left(n^{-1}h^{-m} + mh^4\right). \tag{C.10}$$

Using (C.9) and (C.10), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)\right|^3 \\
&\quad\le \sup_{\mathbf{x}_i^{(m)}\in\mathbb{I}^m}\left|\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)\right|\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)\right]^2\right\}, \\
&\quad= O_p\left(\frac{(\log n)^{1/2}}{n^{3/2}h^{3m/2}} + m^2h^6\right).
\end{aligned} \tag{C.11}$$

Then by (C.11) and the inequality $\left|\log(1+x) - x + \frac{1}{2}x^2\right| \le |x|^3$, obviously we have

$$\begin{aligned}
&\left|\hat{I}_{nm}(\hat{g}, g) - \hat{W}_{1S}(m) + \frac{1}{2}\hat{W}_{2S}(m)\right| \\
&\quad\le \frac{1}{n}\sum_{i\in S_n(m)}\left|\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right|^{3,} \\
&\quad= O_p\left(\frac{(\log n)^{1/2}}{n^{3/2}h^{3m/2}} + m^2h^6\right),
\end{aligned} \tag{C.12}$$

where

$$\hat{W}_{1S}(m) = \frac{1}{n}\sum_{i\in S_n(m)}\left[\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right],$$

and

$$\hat{W}_{2S}(m) = \sum_{i\in S_n(m)}\left[\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right]^2.$$

Note that the definitions of $\hat{W}_1(m)$ and $\hat{W}_2(m)$ include the summation of $n$ observations. The difference is

$$\begin{aligned}
&\left[\hat{W}_{1S}(m) - \frac{1}{2}\hat{W}_{2S}(m)\right] - \left[\hat{W}_1(m) - \frac{1}{2}\hat{W}_2(m)\right] \\
&= O_p\left[\frac{1}{n}\sum_{i=1}^{n} P\left(i \notin S_n(m)\right)\right], \qquad \text{(C.13)} \\
&= O_p\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left|\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right|^3\right], \qquad \text{(C.14)} \\
&= O_p\left(\frac{(\log n)^{1/2}}{n^{3/2}h^{3m/2}} + m^2h^6\right). \qquad \text{(C.15)}
\end{aligned}$$

Step (C.13) to step (C.14) is based on the fact that

$$\begin{aligned}
P\left(i \notin S_n(m)\right) &= P\left(\hat{g}\left(\mathbf{x}_i^{(m)}\right) \leq 0\right), \\
&\leq P\left[\left|\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)\right| > g\left(\mathbf{x}_i^{(m)}\right)\right], \\
&\leq \mathrm{E}\left|\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right|^3.
\end{aligned}$$

Combining equations (C.12) and (C.15), we complete the proof of this lemma. □

*Proof of Lemma 5.3.* Firstly, we give a similar result for univariate kernel, then we extend this result to multivariate kernel. For any $x, y \in \mathbb{I}$, denote $\gamma_{n1}(x,y) = \int_0^1 [K_h^J(x^*, y) - \int_0^1 K_h^J(x^*, y^*) g_1(y^*) dy^*] dx^* / g_1(x)$. The numerator of $\gamma_{n1}(x,y)$ includes two terms. The first term can be expanded as

$$\begin{aligned}
&\int_0^1 K_h^J\left(x^* - y\right) \mathrm{d}x^* \\
&\quad = \int_0^h K_h^J\left(x^* - y\right) \mathrm{d}x^* + \int_{1-h}^1 K_h^J\left(x^* - y\right) \mathrm{d}x^* + \int_h^{1-h} K_h\left(x^* - y\right) \mathrm{d}x^*,
\end{aligned} \qquad \text{(C.16)}$$

when $n \to \infty$, then $h \to 0$ such that $y/h \to \infty$ and $(1-y)/h \to \infty$ for any $y \in (0,1)$, see APPENDIX A in Hong & White (2005). Using $K(\cdot)$ having bounded support $[-1,1]$ and change of variable, when $n$ is sufficient large, the first and second terms in equation (C.16) are zero, and the third term is 1. When

$n$ is sufficiently large, the term

$$\begin{aligned}
&\int_0^1 \int_0^1 K_h^J(x^*, y^*) g_1(y^*) dy^* dx^* \\
&= \int_0^1 g_1(y^*) \int_0^1 K_h^J(x^*, y^*) dx^* dy^* \\
&= \int_0^1 g_1(y^*) \cdot 1 dy^* = 1,
\end{aligned} \tag{C.17}$$

as well. Therefore, when $n$ is sufficiently large, $\gamma_{n1}(x, y) = 0$ with probability 1. Recalling that $\mathcal{K}_h^{(m)}(\cdot)$ is multiplicative kernel, we can easily extend this result to multivariate case. It follows that for sufficiently large $n$,

$$P\left[\gamma_{nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right) = 0\right] = 1,$$

this completes the proof. □

*Proof of Lemma 5.4.* Let

$$\phi_{nm}(\mathbf{z}_1, \mathbf{z}_2) = h^m D_{nm}(\mathbf{z}_1, \mathbf{z}_2) = h^m \left[A_{nm}^2(\mathbf{z}_1, \mathbf{z}_2) + A_{nm}^2(\mathbf{z}_2, \mathbf{z}_1)\right],$$

then we have $\hat{\phi}_n(m) = h^m \hat{D}_n(m)$ and

$$\begin{aligned}
\phi_{nm0} &= \int_{\mathbb{I}^m} \int_{\mathbb{I}^m} \phi_{nm}(\mathbf{z}_1, \mathbf{z}_2) g(\mathbf{z}_1) g(\mathbf{z}_2) \, \mathrm{d}\mathbf{z}_1 \, \mathrm{d}\mathbf{z}_2 \\
&= 2h^m \mathrm{E} A_{nm}^2(\mathbf{z}_1, \mathbf{z}_2).
\end{aligned}$$

We note that $g(\mathbf{z})$ is bounded away from zero by Assumption 2, then it follows that

$$\mathrm{E}\phi_{nm}^2 \leq h^{2m} C \int_{\mathbb{I}^m} \int_{\mathbb{I}^m} A_{nm}^4(\mathbf{z}_1, \mathbf{z}_2) \, \mathrm{d}\mathbf{z}_1 \, \mathrm{d}\mathbf{z}_2 = O(1),$$

because $\mathrm{E}A_{nm}^2(\mathbf{z}_1, \mathbf{z}_2) = O(h^{-m})$, Jensen's inequality and Cauchy-Schwarz inequality. Using the same way, one can also verify that $\mathrm{E}\phi_{nm1}^2(\mathbf{x}_i^{(m)}) - \phi_{nm0}^2 = O(1)$, so $h^m D_{nm}(\mathbf{z}_1, \mathbf{z}_2)$ satisfies the conditions in Lemma C.3, immediately have the result (5.28). □

*Proof of Lemma 5.5.* Let

$$\begin{aligned}
\phi_{nm}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) &= h^m \tilde{H}_{2nm}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3), \\
&= h^m \left[A_{nm}(\mathbf{z}_1, \mathbf{z}_2) A_{nm}(\mathbf{z}_1, \mathbf{z}_3) + A_{nm}(\mathbf{z}_2, \mathbf{z}_3) A_{nm}(\mathbf{z}_2, \mathbf{z}_1)\right. \\
&\quad \left. + A_{nm}(\mathbf{z}_3, \mathbf{z}_1) A_{nm}(\mathbf{z}_3, \mathbf{z}_2)\right],
\end{aligned}$$

then we have

$$\hat{\phi}_n(m) = h^m \binom{n}{3}^{-1} \sum_{k=3}^{n} \sum_{j=2}^{k-1} \sum_{i=1}^{j-1} \tilde{H}_{2nm}\left(\mathbf{x}_k^{(m)}, \mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right) = h^m \tilde{H}_{2n}(m),$$

and $\phi_{nm2}(\mathbf{z}_1, \mathbf{z}_2) = h^m H_{2nm}(\mathbf{z}_1, \mathbf{z}_2)$ based on the fact $\int_{\mathbb{I}^m} A_{nm}(\mathbf{z}_1, \mathbf{z}_2) g(\mathbf{z}_2)\, \mathrm{d}\mathbf{z}_2 = 0$. Furthermore, we can easily verify that $\mathrm{E}\phi_{nm}^2(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}, \mathbf{x}_k^{(m)}) = O(1)$, then by Lemma C.4, we immediately obtain equation (5.29) which completes the proof. □

*Proof of Lemma 5.6.* Firstly, $\hat{W}_{22}(m)$ can be expressed as

$$\begin{aligned}\hat{W}_{22}(m) &= \frac{1}{n} \sum_{i=1}^{n} B_{nm}^2\left(\mathbf{x}_i^{(m)}\right), \\ &= \mathrm{E}B_{nm}^2\left(\mathbf{x}_1^{(m)}\right) + \frac{1}{n} \sum_{i=1}^{n} \left[B_{nm}^2\left(\mathbf{x}_i^{(m)}\right) - \mathrm{E}B_{nm}^2\left(\mathbf{x}_1^{(m)}\right)\right].\end{aligned}$$

By Lemma C.5, we have $\mathrm{E}B_{nm}^4(\mathbf{x}_i^{(m)}) = O(h^8)$. When $i \geq m$, $\mathbf{x}_i^{(m)}$ is independent of $\mathbf{x}_1^{(m)}$ and $m$ is bounded by $M$, by Chebyshev's inequality, we immediately have $\hat{W}_{22}(m) = \mathrm{E}B_{nm}^2(\mathbf{x}_1^{(m)}) + O_p(n^{-1/2}h^4)$, this completes the proof. □

*Proof of Lemma 5.7.* Define a new symmetric function

$$\tilde{C}_{nm}(\mathbf{z}_1, \mathbf{z}_2) = A_{nm}(\mathbf{z}_1, \mathbf{z}_2) B_{nm}(\mathbf{z}_1) + A_{nm}(\mathbf{z}_2, \mathbf{z}_1) B_{nm}(\mathbf{z}_2),$$

then

$$\begin{aligned}\hat{W}_{23}(m) &= 2\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\hat{g}\left(\mathbf{x}_i^{(m)}\right) - \bar{g}\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right] \left[\frac{\bar{g}\left(\mathbf{x}_i^{(m)}\right) - g\left(\mathbf{x}_i^{(m)}\right)}{g\left(\mathbf{x}_i^{(m)}\right)}\right], \\ &= \binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \tilde{C}_{nm}\left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right).\end{aligned}$$

Let $\phi_{nm}(\mathbf{z}_1, \mathbf{z}_2) = \tilde{C}_{nm}(\mathbf{z}_1, \mathbf{z}_2)$, then $\phi_{nm0} = 0$ and

$$\begin{aligned}\phi_{nm1}(\mathbf{z}) &= \int_{\mathbb{I}^m} \phi_{nm}(\mathbf{z}, \mathbf{z}_2) g(\mathbf{z}_2)\, \mathrm{d}\mathbf{z}_2 = \int_{\mathbb{I}^m} \phi_{nm}(\mathbf{z}_2, \mathbf{z}) g(\mathbf{z}_2)\, \mathrm{d}\mathbf{z}_2, \\ &= \int_{\mathbb{I}^m} A_{nm}(\mathbf{z}_2, \mathbf{z}) B_{nm}(\mathbf{z}_2) g(\mathbf{z}_2)\, \mathrm{d}\mathbf{z}_2.\end{aligned}$$

As discussion in Lemma 5.4, one can verify that $\mathrm{E}[\phi_{nm}^2(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)})] = O(c_n^2)$,

where $c_n = h^{2-m/2}$. By Lemma C.3, we have

$$\begin{aligned}\hat{W}_{23}(m) = \hat{\phi}_n(m) &= \frac{2}{n}\sum_{i=1}^{n}\phi_{nm1}\left(\mathbf{x}_i^{(m)}\right) + O_p(n^{-1}mh^{2-m/2}),\\ &= 2\hat{C}_n(m) + O_p(n^{-1}mh^{2-m/2}),\end{aligned}$$

which completes the proof. □

*Proof of Lemma 5.8.* Let $\kappa_1 = 2\int_0^1\int_{-1}^{\rho} k_\rho^2(u)\,\mathrm{d}u\,\mathrm{d}\rho - 2\kappa - 1$, then $\mathrm{E}a_n^2(z_{1m}, z_{2m}) = \kappa h^{-1} + \kappa_1 + O(h)$, by equation (C.26) and $\mathbb{H}_0$, we immediately obtain the desired result. □

*Proof of Lemma 5.9.* By equation (C.27), $[\hat{B}_n(m+1) - \hat{B}_n(m) - \hat{b}_n]$ can be expressed as

$$\begin{aligned}&\left[\hat{B}_n(m+1) - \hat{B}_n(m) - \hat{b}_n\right]\\ &\quad= \mathrm{E}B_{nm}(\mathbf{y}_1)b_n(z_{1m}) + \frac{1}{n}\sum_{i=1}^{n}\left[B_{nm}\left(\mathbf{x}_i^{(m)}\right)b_n(x_{i+m}) - \mathrm{E}B_{nm}(\mathbf{y}_1)b_n(z_{1m})\right],\end{aligned}$$

According to Cauchy-Schwarz inequality, $\mathrm{E}[B_{nm}\left(\mathbf{x}_i^{(m)}\right)b_n(x_{i+m})]^2 = O(h^8)$, so the second term

$$n^{-1}\sum_{i=1}^{n}\left[B_{nm}\left(\mathbf{x}_i^{(m)}\right)b_n(x_{i+m}) - \mathrm{E}B_{nm}(\mathbf{y}_1)b_n(z_{1m})\right] = O_p(n^{-1/2}h^4),$$

by Markov inequality and Chebyshev inequality. Furthermore, by equation (C.27) and Lemma C.5, after simple calculations, we have

$$\begin{aligned}&\mathrm{E}B_{n(m+1)}^2(\mathbf{z}_1) - \mathrm{E}B_{n(m)}^2(\mathbf{y}_1) - \mathrm{E}b_n^2(z_{1m})\\ &\quad= 2\mathrm{E}B_{nm}(\mathbf{y}_1)b_n(z_{1m}) + 2\mathrm{E}B_{nm}^2(\mathbf{y}_1)b_n(z_{1m}) + \mathrm{E}B_{nm}(\mathbf{y}_1)b_n^2(z_{1m}) + O(h^8),\\ &\quad= 2\mathrm{E}B_{nm}(\mathbf{y}_1)b_n(z_{1m}) + O(h^6),\end{aligned}$$

which immediately completes the proof. □

*Proof of Lemma 5.10.* Let

$$\begin{aligned}\bar{K}_h^J(z_{1m}) &= \int_0^1 K_h^J(z_{1m}, z)g_1(z)\,\mathrm{d}z,\\ \bar{\mathcal{K}}_h^{(m)}(\mathbf{y}_1) &= \int_{\mathbb{I}^m}\mathcal{K}_h^{(m)}(\mathbf{y}_1, \mathbf{y})g(\mathbf{y})\,\mathrm{d}\mathbf{y},\\ \bar{\mathcal{K}}_h^{(m+1)}(\mathbf{z}_1) &= \int_{\mathbb{I}^{m+1}}\mathcal{K}_h^{(m+1)}(\mathbf{z}_1, \mathbf{z})f(\mathbf{z})\,\mathrm{d}\mathbf{z},\end{aligned}$$

and $\psi_1(z_{1m}) = \bar{K}_h^J(z_{1m})/g_1(x_{i+m})$, $\psi_1(z_{1m}, x_{i+m}) = K_h^J(z_{1m}, x_{i+m})/g_1(x_{i+m})$,

$\psi_m(\mathbf{y}_1) = \bar{\mathcal{K}}_h^{(m)}(\mathbf{y}_1)/g(\mathbf{y}_1)$,$\psi_m(\mathbf{y}_1, \mathbf{x}_i^{(m)}) = \mathcal{K}_h^{(m)}(\mathbf{y}_1, \mathbf{x}_i^{(m)})/g(\mathbf{y}_1)$. Given the definition of multivariate kernel and $\mathbb{H}_0$, we obtain

$$
\begin{aligned}
a_n\left(z_{1m}, x_{i+m}\right) &= \psi_1(z_{1m}, x_{i+m}) - \psi_1(z_{1m}), \\
A_{nm}\left(\mathbf{y}_1, \mathbf{x}_i^{(m)}\right) &= \psi_m\left(\mathbf{y}_1, \mathbf{x}_i^{(m)}\right) - \psi_m\left(\mathbf{y}_1\right), \\
A_{n(m+1)}\left(\mathbf{z}_1, \mathbf{x}_i^{(m+1)}\right) &= \psi_m\left(\mathbf{y}_1, \mathbf{x}_i^{(m)}\right) \psi_1(z_{1m}, x_{i+m}) - \psi_m\left(\mathbf{y}_1\right) \psi_1(z_{1m}).
\end{aligned}
$$

Using equation (C.27) and $f(\mathbf{z}_1) = g(\mathbf{y}_1)g_1(z_{1m})$, we can separately write $\breve{c}(x_{i+m})$, $\breve{C}_m(\mathbf{x}_i^{(m)})$ and $\breve{C}_{m+1}(\mathbf{x}_i^{(m+1)})$ as

$$
\begin{aligned}
\breve{c}(x_{i+m}) &= \int_0^1 \psi_1\left(z_{1m}, x_{i+m}\right) b_n(z_{1m}) g_1(z_{1m}) \, \mathrm{d}z_{1m} \\
&\quad - \int_0^1 \psi_1\left(z_{1m}\right) b_n(z_{1m}) g_1(z_{1m}) \, \mathrm{d}z_{1m}, \\
&= \breve{c}_1(x_{i+m}) - \breve{c}_2,
\end{aligned}
$$

$$
\begin{aligned}
\breve{C}_m\left(\mathbf{x}_i^{(m)}\right) &= \int_{\mathbf{y}_1 \in \mathbb{I}^m} \psi_m\left(\mathbf{y}_1, \mathbf{x}_i^{(m)}\right) B_{nm}(\mathbf{y}_1) g(\mathbf{y}_1) \, \mathrm{d}\mathbf{y}_1 \\
&\quad - \int_{\mathbf{y}_1 \in \mathbb{I}^m} \psi_m\left(\mathbf{y}_1\right) B_{nm}(\mathbf{y}_1) g(\mathbf{y}_1) \, \mathrm{d}\mathbf{y}_1, \\
&= \breve{C}_1\left(\mathbf{x}_i^{(m)}\right) - \breve{C}_2,
\end{aligned}
$$

$$
\begin{aligned}
\breve{C}_{m+1}\left(\mathbf{x}_i^{(m+1)}\right) &= \breve{C}_1\left(\mathbf{x}_i^{(m)}\right) \breve{c}_1(x_{i+m}) - \breve{C}_2 \breve{c}_2 \\
&\quad + \breve{C}_1\left(\mathbf{x}_i^{(m)}\right) \int_0^1 K_h^J\left(z_{1m}, x_{i+m}\right) \, \mathrm{d}z_{1m} \\
&\quad - \breve{C}_2 \int_0^1 \bar{K}_h^J\left(z_{1m}\right) \, \mathrm{d}z_{1m} \\
&\quad + \breve{c}_1(x_{i+m}) \int_{\mathbf{y}_1 \in \mathbb{I}^m} \mathcal{K}_h^{(m)}\left(\mathbf{y}_1, \mathbf{x}_i^{(m)}\right) \, \mathrm{d}\mathbf{y}_1 \\
&\quad - \breve{c}_2 \int_{\mathbf{y}_1 \in \mathbb{I}^m} \bar{\mathcal{K}}_h^{(m)}\left(\mathbf{y}_1\right) \, \mathrm{d}\mathbf{y}_1,
\end{aligned}
$$

then we have

$$
\begin{aligned}
&\breve{C}_{m+1}\left(\mathbf{x}_i^{(m+1)}\right)-\breve{C}_m\left(\mathbf{x}_i^{(m)}\right)-\breve{c}(x_{i+m})\\
&\quad=\breve{C}_1\left(\mathbf{x}_i^{(m)}\right)\breve{c}_1(x_{i+m})-\breve{C}_2\breve{c}_2\\
&\qquad+\breve{C}_1\left(\mathbf{x}_i^{(m)}\right)\left[\int_0^1 K_h^J\left(z_{1m},x_{i+m}\right)\,\mathrm{d}z_{1m}-1\right]\\
&\qquad-\breve{C}_2\left[\int_0^1 \bar{K}_h^J\left(z_{1m}\right)\,\mathrm{d}z_{1m}-1\right]\\
&\qquad+\breve{c}_1(x_{i+m})\left[\int_{\mathbf{y}_1\in\mathbb{I}^m}\mathcal{K}_h^{(m)}\left(\mathbf{y}_1,\mathbf{x}_i^{(m)}\right)\,\mathrm{d}\mathbf{y}_1-1\right]\\
&\qquad-\breve{c}_2\left[\int_{\mathbf{y}_1\in\mathbb{I}^m}\bar{\mathcal{K}}_h^{(m)}\left(\mathbf{y}_1\right)\,\mathrm{d}\mathbf{y}_1-1\right]\\
&\quad=\breve{C}_1\left(\mathbf{x}_i^{(m)}\right)\breve{c}_1(x_{i+m})-\breve{C}_2\breve{c}_2+\sum_{s=1}^{4}\delta_s.
\end{aligned}
$$

Firstly, for terms $\delta_s, s=1,2,3,4$, we prove $\delta_1=0,\delta_2=0,a.e.$, for univariate kernel when $n$ is sufficient large, then we extend this result to multivariate kernel. For any $y\in\mathbb{I}$, by equations (C.16) and (C.17) in Lemma 5.3, we have $\int_0^1 K_h^J\left(x,y\right)\,\mathrm{d}x=1$ and $\int_0^1 \bar{K}_h^J(x)dx=1$ almost surely. Therefore, when $n$ is sufficiently large, $\delta_1=0,\delta_2=0$ almost everywhere. Recalling that $\mathcal{K}_h^{(m)}(\cdot)$ is multiplicative kernel, we can easily extend this result to multivariate case. It follows that for sufficiently large $n$, $\sum_{s=1}^4\delta_s=0$ almost surely. Therefore, $\hat{C}_n(m+1)-\hat{C}_n(m)-\hat{c}_n=n^{-1}\sum_{i=1}^n\breve{C}_1(\mathbf{x}_i^{(m)})\breve{c}_1(x_{i+m})-\breve{C}_2\breve{c}_2$. We also notice that $\mathrm{E}[\breve{C}_1(\mathbf{x}_i^{(m)})\breve{c}_1(x_{i+m})]=\breve{C}_2\breve{c}_2$ and $E[\breve{C}_1(\mathbf{x}_i^{(m)})\breve{c}_1(x_{i+m})]^2=O(h^8)$ because of Lemma C.5. Hence, by Chebyshev inequality, $\hat{C}_n(m+1)-\hat{C}_n(m)-\hat{c}_n=O_p(n^{-1/2}h^4)$, this completes the proof. □

*Proof of Lemma 5.11.* By Lemma C.6 and Lemma C.7, we have $\mathrm{E}[T_{1n0}(m+1)]=\mathrm{E}[T_{1n0}(m)]=O(mn^{-1}h^2)$, $\mathrm{E}[T_{2n0}(m)]=c_2(\tau^m-1)+O(mn^{-1}h)$, $\mathrm{E}[T_{2n0}(m+1)]=c_1(\tau^{m+1}-1)+O(mn^{-1}h)$ where $c_2=[(2n-m)(m-1)]/[(n-m)(n-m+1)]$ and $c_1=[(2n-m-1)m]/[(n-m-1)(n-m)]$. Using the similar change of variable in Lemma C.6 and Lemma C.7, one can verify that $\mathrm{E}[T_{1n0}^2(m)]=O(m^2n^{-2}h^{-m})$, $\mathrm{E}[T_{1n0}^2(m+1)]=O(m^2n^{-2}h^{-m-1})$, $\mathrm{E}[T_{2n0}^2(m)]=O(m^2n^{-2}h^{-m})$ and $\mathrm{E}[T_{2n0}^2(m+1)]=O(m^2n^{-2}h^{-m-1})$. Hence, by Chebyshev inequality, we have

$$
\begin{aligned}
T_{1n0}(m)&=O(mn^{-1}h^2)+O_p(mn^{-1}h^{-\frac{m}{2}}),\\
T_{2n0}(m)&=c_2(\tau^m-1)+O(mn^{-1}h)+O_p(mn^{-1}h^{-\frac{m}{2}}),\\
T_{1n0}(m+1)&=O(mn^{-1}h^2)+O_p(mn^{-1}h^{-\frac{m+1}{2}}),\\
T_{2n0}(m+1)&=c_1(\tau^{m+1}-1)+O(mn^{-1}h)+O_p(mn^{-1}h^{-\frac{m+1}{2}}),
\end{aligned}
$$

which immediately completes the proof comparing with order $n^{-1/2}h^{-(m+1)/2}$. □

*Proof of Theorem 5.5.* Following the THEOREMs A.6 – A.9 in Hong & White (2005), one can extend their theory to multivariate U-statistics with bounded $m$. Hong & White (2005) discussed the pair variables $Z_{jt} = (X_t, X_{t-j})^T$ and $j = o(n)$. THEOREM A.6 constructs a new $2j$-dependent process to show the dependent part of U-statistics is negligible, then they employed a martingale difference sequence in THEOREM A.7 so that the U-statistics can be projected on it. THEOREM A.7 implies that one can apply central limit theorem to martingale difference sequence according to Brown (1971)'s theorem if two conditions in THEOREM A.9 are satisfied. Finally, by Slusky's theorem, Brown's theorem and THEOREMs A.6 – A.9, the central limit theorem of U-statistics is completed. Analogue to Hong & White (2005)'s idea but more tedious than that, for bounded $m$, (5.32) holds as well. □

## C.4 Lemmas for The Second and Third Order U-statistics

**Lemma C.3.** *Let $\mathbf{Z}_i^{(m)} = (X_i, \ldots, X_{i+m-1})^T$, $m < M$ and $\{X_t\}$ is i.i.d. with CDF $G_1(\cdot)$. Consider a second-order U-statistics*

$$\hat{\phi}_n(m) = \binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \phi_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right),$$

*where $\phi_{nm}(\cdot,\cdot)$ is a kernel function such that $\phi_{nm}(\mathbf{z}_1, \mathbf{z}_2) = \phi_{nm}(\mathbf{z}_2, \mathbf{z}_1)$. Let*

$$\phi_{nm0} = \int_{\mathbb{I}^m} \int_{\mathbb{I}^m} \phi_{nm}(\mathbf{z}_1, \mathbf{z}_2)\, \mathrm{d}G_m(\mathbf{z}_1)\, \mathrm{d}G_m(\mathbf{z}_2),$$

*and $\phi_{nm1}(\mathbf{z}) = \int_{\mathbb{I}^m} \phi_{nm}(\mathbf{z}, \mathbf{z}_1)\, \mathrm{d}G_m(\mathbf{z}_1)$, where $G_m(\mathbf{z}) = G_1(x_1)G_1(x_2)\cdots G_1(x_m)$ and $\mathbf{z} = (x_1, \ldots, x_m)^T$. Suppose $E\phi_{nm}^2(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}) - \phi_{nm0}^2 = O(c_n^2)$ holds, then we have*

$$\hat{\phi}_n(m) = \phi_{nm0} + \frac{2}{n}\sum_{i=1}^{n}\left[\phi_{nm1}\left(\mathbf{Z}_i^{(m)}\right) - \phi_{nm0}\right] + O_p(mn^{-1}c_n).$$

*If in addition $E\phi_{nm1}^2(\mathbf{Z}_i^{(m)}) - \phi_{nm0}^2 \le C$ and $c_n = O(n^{1/2})$, then $\hat{\phi}_n(m) = \phi_{nm0} + O_p(m^{1/2}n^{-1/2})$.*

*Proof of Lemma C.3.* Lemma C.3 is the version of LEMMA B.1 in Hong & White (2005). For $m$-consecutive variables, we still use the same notations as Hong & White (2005)'s. Denote $\tilde{\phi}_{nm}(\mathbf{z}_1, \mathbf{z}_2) = \phi_{nm}(\mathbf{z}_1, \mathbf{z}_2) - \phi_{nm1}(\mathbf{z}_1) - \phi_{nm1}(\mathbf{z}_2) + \phi_{nm0}$,

then $\forall\ \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{I}^m$, we have

$$\int_{\mathbb{I}^m} \tilde{\phi}_{nm}\left(\mathbf{z}_1, \mathbf{z}_2\right)\, \mathrm{d}G_m(\mathbf{z}_2) = \int_{\mathbb{I}^m} \tilde{\phi}_{nm}\left(\mathbf{z}_1, \mathbf{z}_2\right)\, \mathrm{d}G_m(\mathbf{z}_1) = 0. \tag{C.18}$$

Using $\tilde{\phi}_{nm}(\mathbf{z}_1, \mathbf{z}_2)$, we can reshape $\hat{\phi}_n(m)$ after simple computations,

$$\begin{aligned}
\hat{\phi}_n(m) &= \phi_{nm0} + \frac{2}{n}\sum_{i=1}^{n}\left[\phi_{nm1}\left(\mathbf{Z}_i^{(m)}\right) - \phi_{nm0}\right] \\
&\quad + \binom{n}{2}^{-1}\sum_{j=2}^{n}\sum_{i=1}^{j-1}\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right), \\
&= \phi_{nm0} + \frac{2}{n}\sum_{i=1}^{n}\left[\phi_{nm1}\left(\mathbf{Z}_i^{(m)}\right) - \phi_{nm0}\right] + \tilde{\phi}_n(m).
\end{aligned} \tag{C.19}$$

Note that, if $j - i \geq m$, then $\mathbf{Z}_j^{(m)}$ and $\mathbf{Z}_i^{(m)}$ have no overlap variable. By this fact, we divide $\tilde{\phi}_n(m)$ into two parts,

$$\begin{aligned}
\tilde{\phi}_n(m) &= \binom{n}{2}^{-1}\sum_{j=1+m}^{n}\sum_{i=1}^{j-m}\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right) \\
&\quad + \binom{n}{2}^{-1}\sum_{j=2}^{n}\sum_{i=1\vee(j-m+1)}^{j-1}\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right), \\
&= \tilde{\phi}_{n1}(m) + \tilde{\phi}_{n2}(m).
\end{aligned}$$

The double summation of $\tilde{\phi}_{n1}(m)$ includes $(n-m)(n-m+1)/2$ terms and there are $(2n-m)(m-1)/2$ summation terms in $\tilde{\phi}_{n2}(m)$. One can easily verify

$$\mathrm{E}\tilde{\phi}_{nm}^2\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right) = \mathrm{E}\phi_{nm}^2\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right) - \phi_{nm0}^2 = O(c_n^2).$$

Hence, using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathrm{E}\left|\tilde{\phi}_{n2}(m)\right| &= \mathrm{E}\left|\binom{n}{2}^{-1}\sum_{j=2}^{n}\sum_{i=1\vee(j-m+1)}^{j-1}\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right)\right|, \\
&= O(mn^{-1})\mathrm{E}\left|\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right)\right|, \\
&\leq O(mn^{-1})\sqrt{\mathrm{E}\tilde{\phi}_{nm}^2\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right)}, \\
&= O(mn^{-1}c_n).
\end{aligned}$$

By Markov's inequality, we have $\tilde{\phi}_{n2}(m) = O_p(mn^{-1}c_n)$. We also notice that the

$\mathbf{Z}_i^{(m)}$ and $\mathbf{Z}_j^{(m)}$ in the first term $\tilde{\phi}_{n1}(m)$ are independent, hence we have

$$\begin{aligned} \mathrm{E}\tilde{\phi}_{n1}^2(m) = \binom{n}{2}^{-2} \sum_{j=1+m}^{n} \sum_{i=1}^{j-m} \sum_{t=1+m}^{n} \sum_{s=1}^{t-m} \\ \mathrm{E}\left[\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right) \tilde{\phi}_{nm}\left(\mathbf{Z}_s^{(m)}, \mathbf{Z}_t^{(m)}\right)\right] \times \mathbb{1}\left(i, j \in S_{ij}\right), \end{aligned} \tag{C.20}$$

where $S_{ij} = [s-m+1, s+m-1] \cup [t-m+1, t+m-1]$. If at least one of $i, j \notin S_{ij}$, by equation (C.18), $\mathrm{E}[\tilde{\phi}_{nm}(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)})\tilde{\phi}_{nm}(\mathbf{Z}_s^{(m)}, \mathbf{Z}_t^{(m)})] = 0$. The number of pair $(s,t)$, $t-s \geq m$ is of order $O(n^2)$, for each given $(s,t)$, if $\mathbf{Z}_i^{(m)}$ has at least one overlap variable with $\mathbf{Z}_s^{(m)}$ or $\mathbf{Z}_t^{(m)}$, and $\mathbf{Z}_j^{(m)}$ has at least one overlap variable with $\mathbf{Z}_s^{(m)}$ or $\mathbf{Z}_t^{(m)}$ as well, then the expectation is nonzero. The indices of $i, j$ have at most $O(m)$ and $O(m)$ choices respectively given $m < M$. So the number of four summation terms in equation (C.20) is of order $O(n^2m^2) = O(n^2)O(m)O(m)$, hence $\mathrm{E}\tilde{\phi}_{n1}^2(m) = O(m^2n^{-2}c_n^2)$ by Cauchy-Schwarz inequality and $\mathrm{E}\tilde{\phi}_{nm}^2(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}) = O(c_n^2)$. It follows that $\tilde{\phi}_{n1}(m) = O_p(mn^{-1}c_n)$ by Chebyshev's inequality, and finally $\tilde{\phi}_n(m) = O_p(mn^{-1}c_n)$.

Next, using a similar way, we discuss the order of the second term in equation (C.19). Note that $\phi_{nm1}(\mathbf{Z}_i^{(m)}) - \phi_{nm0}$ is an $m$-dependence process with mean 0 and

$$\mathrm{E}\left\{\left[\phi_{nm1}\left(\mathbf{Z}_i^{(m)}\right) - \phi_{nm0}\right]\left[\phi_{nm1}\left(\mathbf{Z}_j^{(m)}\right) - \phi_{nm0}\right]\right\} = 0,$$

if $j - i \geq m$. The number of nonzero terms in $\mathrm{E}\{\sum_{i=1}^{n}[\phi_{nm1}(\mathbf{Z}_i^{(m)}) - \phi_{nm0}]\}^2$ is of order $O(nm)$. By these facts, $\mathrm{E}[\phi_{nm1}(\mathbf{Z}_i^{(m)}) - \phi_{nm0}]^2 \leq C$ and Chebyshev's inequality, we have $2n^{-1}\sum_{i=1}^{n}[\phi_{nm1}(\mathbf{Z}_i^{(m)}) - \phi_{nm0}] = O_p(m^{1/2}n^{-1/2})$. This completes the proof. □

**Lemma C.4.** *Let $\mathbf{Z}_i^{(m)} = (X_i, \ldots, X_{i+m-1})^T$, $m < M$ and $\{X_t\}$ is i.i.d. with CDF $G_1(\cdot)$. Consider a third-order U-statistics*

$$\hat{\phi}_n(m) = \binom{n}{3}^{-1} \sum_{k=3}^{n} \sum_{j=2}^{k-1} \sum_{i=1}^{j-1} \phi_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}\right), \tag{C.21}$$

*where $\phi_{nm}(\cdot,\cdot,\cdot)$ is a kernel function in its argument and $\forall\ \mathbf{z}_1 \in \mathbb{I}^m$*

$$\int_{\mathbb{I}^m}\int_{\mathbb{I}^m} \phi_{nm}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)\ \mathrm{d}G_m(\mathbf{z}_2)\,\mathrm{d}G_m(\mathbf{z}_3) = 0, \tag{C.22}$$

*holds, where $G_m(\mathbf{z}) = G_1(x_1)G_1(x_2)\cdots G_1(x_m)$ and $\mathbf{z} = (x_1, \ldots, x_m)^T$. Let*

$$\phi_{nm2}(\mathbf{z}_1, \mathbf{z}_2) = \int_{\mathbb{I}^m} \phi_{nm}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)\ \mathrm{d}G_m(\mathbf{z}_3).$$

*Suppose* $E\phi_{nm}^2(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}) = O(c_n^2)$*, then we have*

$$\hat{\phi}_n(m) = 3\binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \phi_{nm2}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right) + O_p(m^{3/2} n^{-3/2} c_n).$$

*Proof of Lemma C.4.* Lemma C.4 is the version of LEMMA B.2 in Hong & White (2005). For $m$-consecutive variables, we still use the same notations as Hong & White (2005)'s. As in Lemma C.3, we construct a new symmetric third-order $U$-statistics

$$\tilde{\phi}_{nm}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = \phi_{nm}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) - \phi_{nm2}(\mathbf{z}_1, \mathbf{z}_2) - \phi_{nm2}(\mathbf{z}_2, \mathbf{z}_3) - \phi_{nm2}(\mathbf{z}_3, \mathbf{z}_1),$$

and it is easy to verify

$$\int_{\mathbb{I}^m} \tilde{\phi}_{nm}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)\, \mathrm{d}G_m(\mathbf{z}_3) = 0, \quad \forall\ \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{I}^m, \tag{C.23}$$

given equation (C.22). We can rewrite equation (C.21) using $\tilde{\phi}_{nm}(\cdot,\cdot,\cdot)$ as

$$\begin{aligned}
\hat{\phi}_n(m) &= 3\binom{n}{2}^{-1} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \phi_{nm2}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}\right) \\
&\quad + \binom{n}{3}^{-1} \sum_{k=3}^{n} \sum_{j=2}^{k-1} \sum_{i=1}^{j-1} \tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}\right), \\
&= 3\binom{n}{2}^{-1} \hat{\phi}_{n2}(m) + \binom{n}{3}^{-1} \tilde{\phi}_n(m).
\end{aligned}$$

Let

$$\tilde{\phi}_{n1}(m) = \sum_{k=1+2m}^{n} \sum_{j=1+m}^{k-m} \sum_{i=1}^{j-m} \tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}\right), \tag{C.24}$$

then $\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}$ in (C.24) are mutually independent and (C.24) includes $\binom{n-2m+2}{3}$ terms. Then $\tilde{\phi}_{n2}(m) = \tilde{\phi}_n(m) - \tilde{\phi}_{n1}(m)$ includes $\binom{n}{3} - \binom{n-2m+2}{3} = O(mn^2)$ terms by $m < M$. Using Cauchy-Schwarz inequality, we can verify $\mathrm{E}\tilde{\phi}_{nm}^2(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}) = O(c_n^2)$ as well. Hence, we have

$$\begin{aligned}
\mathrm{E}\left|\tilde{\phi}_{n2}(m)\right| &= O(mn^2)\mathrm{E}\left|\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}\right)\right|, \\
&\leq O(mn^2)\sqrt{\mathrm{E}\tilde{\phi}_{nm}^2\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}\right)}, \\
&\qquad \text{(by Cauchy-Schwarz inequality )}, \\
&= O(mn^2 c_n).
\end{aligned}$$

For $\tilde{\phi}_{n1}(m)$, we have

$$\begin{aligned} \mathrm{E}\tilde{\phi}_{n1}^2(m) = & \sum_{k=1+2m}^{n} \sum_{j=1+m}^{k-m} \sum_{i=1}^{j-m} \sum_{r=1+2m}^{n} \sum_{t=1+m}^{r-m} \sum_{s=1}^{t-m} \\ & \mathrm{E}\left[\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}\right) \tilde{\phi}_{nm}\left(\mathbf{Z}_s^{(m)}, \mathbf{Z}_t^{(m)}, \mathbf{Z}_r^{(m)}\right)\right] \\ & \times \mathbb{1}\left(i, j, k \in S_{ijk}\right), \end{aligned} \tag{C.25}$$

where

$$S_{ijk} = [s-m+1, s+m-1] \cup [t-m+1, t+m-1] \cup [r-1].$$

If at least one of $i, j, k \notin S_{ijk}$, then by equation (C.23),

$$\mathrm{E}\left[\tilde{\phi}_{nm}\left(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}\right) \tilde{\phi}_{nm}\left(\mathbf{Z}_s^{(m)}, \mathbf{Z}_t^{(m)}, \mathbf{Z}_r^{(m)}\right)\right] = 0.$$

The number of triplet $(s,t,r)$, $t-s \geq m$ and $r-t \geq m$ is of order $O(n^3)$, for each given triplet $(s,t,r)$, if each of $\mathbf{Z}_i^{(m)}$, $\mathbf{Z}_j^{(m)}$, $\mathbf{Z}_k^{(m)}$ has at least one overlap variable with $\mathbf{Z}_s^{(m)}$, $\mathbf{Z}_t^{(m)}$ or $\mathbf{Z}_r^{(m)}$, then the expectation is nonzero. The indices of $i, j, k$ have at most $O(m)$, $O(m)$ and $O(m)$ choices respectively given $m < M$. So the number of six summation terms in equation (C.25) is of order $O(n^3m^3) = O(n^3)O(m)O(m)O(m)$, hence $\mathrm{E}\tilde{\phi}_{n1}^2(m) = O(m^3n^3c_n^2)$ by Cauchy-Schwarz inequality and $\mathrm{E}\tilde{\phi}_{nm}^2(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)}, \mathbf{Z}_k^{(m)}) = O(c_n^2)$. Finally, it follows that $\binom{n}{3}^{-1}\tilde{\phi}_n(m) = O_p(m^{3/2}n^{-3/2}c_n)$ by Chebyshev's inequality and Markov's inequality. This completes the proof. □

**Lemma C.5.** *Given Assumptions 1 and 2, $\forall\ \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{I}^m$, if $\mathbf{z}_1, \mathbf{z}_2$ are independent, then $EA_{nm}^2(\mathbf{z}_1, \mathbf{z}_2)$ is of order $O(h^{-m})$. Furthermore, for $1 \leq m \leq M$, $\max_m \sup_{\mathbf{z}_1 \in \mathbb{I}^m} B_{nm}(\mathbf{z}_1) = O(h^2)$.*

*Proof of Lemma C.5.* To prove this lemma, we investigate the univariate case, i.e.,

$$a_n(x_1, x_2) = \left[K_h^J(x_1, x_2) - \int_0^1 K_h^J(x_1, x) g_1(x) dx\right] / g_1(x_1).$$

For simplicity, we denote $\bar{K}_h^J(x_1) = \int_0^1 K_h^J(x_1, x) g_1(x) dx$, then

$$a_n(x_1, x_2) = \left[K_h^J(x_1, x_2) - \bar{K}_h^J(x_1)\right] / g_1(x_1),$$

$$\int_0^1 a_n(x_1, x_2) g_1(x_2)\, \mathrm{d}x_2 = 0,$$

and

$$\begin{aligned}\mathrm{E}a_n(x_1,x_2) &= \int_0^1\int_0^1 a_n(x_1,x_2)g_1(x_1)g_1(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2,\\ &= \int_0^1\int_0^1 \left[K_h^J(x_1,x_2)-\bar{K}_h^J(x_1)\right]g_1(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 0.\end{aligned}$$

We have

$$\begin{aligned}a_n^2(x_1,x_2) &= \left[\frac{K_h^J(x_1,x_2)-\bar{K}_h^J(x_1)}{g_1(x_1)}\right]^2,\\ &= \frac{\left[K_h^J(x_1,x_2)\right]^2}{g_1^2(x_1)}+\frac{\left[\bar{K}_h^J(x_1)\right]^2}{g_1^2(x_1)}-2\frac{K_h^J(x_1,x_2)\times\bar{K}_h^J(x_1)}{g_1^2(x_1)},\\ &= \psi_1(x_1,x_2)+\psi_2(x_1,x_2)-2\psi_3(x_1,x_2).\end{aligned}$$

Next, we will discuss each $\psi_i(\cdot,\cdot), i=1,2,3$.

$$\begin{aligned}\mathrm{E}\psi_1(x_1,x_2) &= \int_0^1\int_0^h h^{-2}k^2_{(x_1/h)}\left(\frac{x_1-x_2}{h}\right)\frac{g_1(x_2)}{g_1(x_1)}\,\mathrm{d}x_1\,\mathrm{d}x_2\\ &\quad+\int_0^1\int_h^{1-h} h^{-2}K_h^2\left(\frac{x_1-x_2}{h}\right)\frac{g_1(x_2)}{g_1(x_1)}\,\mathrm{d}x_1\,\mathrm{d}x_2\\ &\quad+\int_0^1\int_{1-h}^1 h^{-2}k^2_{([1-x_1]/h)}\left(\frac{x_1-x_2}{h}\right)\frac{g_1(x_2)}{g_1(x_1)}\,\mathrm{d}x_1\,\mathrm{d}x_2,\\ &= \psi_{11}+\psi_{12}+\psi_{13}.\end{aligned}$$

By change of variable, we have

$$\begin{aligned}\psi_{11} &= \int_0^1\int_0^h h^{-2}k^2_{(x_1/h)}\left(\frac{x_1-x_2}{h}\right)\frac{g_1(x_2)}{g_1(x_1)}\,\mathrm{d}x_1\,\mathrm{d}x_2,\\ &= \int_0^h\int_{-1}^{x_1/h} h^{-1}k^2_{(x_1/h)}(u)\,\frac{g_1(x_1-uh)}{g_1(x_1)}\,\mathrm{d}u\,\mathrm{d}x_1,\\ &= \int_0^h\int_{-1}^{x_1/h} h^{-1}k^2_{(x_1/h)}(u)\left[1-uh\frac{g_1'(x_1)}{g_1(x_1)}\right]\,\mathrm{d}u\,\mathrm{d}x_1,\\ &= \int_0^1\int_{-1}^{\rho} k_\rho^2(u)\,\mathrm{d}u\,\mathrm{d}\rho - h\int_0^1\int_{-1}^{\rho} uk_\rho^2(u)\frac{g_1'(\rho h)}{g_1(\rho h)}\,\mathrm{d}u\,\mathrm{d}\rho+O(h^2).\end{aligned}$$

Using the same method, we have

$$\begin{aligned}\psi_{12} &= \int_0^1\int_h^{1-h} h^{-2}K_h^2\left(\frac{x_1-x_2}{h}\right)\frac{g_1(x_2)}{g_1(x_1)}\,\mathrm{d}x_1\,\mathrm{d}x_2,\\ &= \int_h^{1-h}\int_{-1}^{1} h^{-1}K_h^2(u)\left[1-uh\frac{g_1'(x_1)}{g_1(x_1)}+\frac{1}{2}u^2h^2\frac{g_1''(x_1)}{g_1(x_1)}\right]\,\mathrm{d}u\,\mathrm{d}x_1,\\ &= (h^{-1}-2)\int_{-1}^1 K^2(u)\,\mathrm{d}u+O(h),\end{aligned}$$

and

$$\psi_{13} = \int_0^1 \int_{-\rho}^1 k_\rho^2(u)\,\mathrm{d}u\,\mathrm{d}\rho - h\int_0^1 \int_{-\rho}^1 u k_\rho^2(u)\frac{g_1'(\rho h)}{g_1(\rho h)}\,\mathrm{d}u\,\mathrm{d}\rho + O(h^2).$$

Hence,

$$\mathrm{E}\psi_1(x_1,x_2) = (h^{-1}-2)\int_{-1}^1 K^2(u) + 2\int_0^1\int_{-1}^\rho k_\rho^2(u)\,\mathrm{d}u\,\mathrm{d}\rho + O(h),$$

$$\mathrm{E}\psi_2(x_1,x_2) = \int_0^1 \left[\bar{K}_h^J(x_1)\right]^2/g_1(x_1)\,\mathrm{d}x_1,$$

and

$$\begin{aligned}\mathrm{E}\psi_3(x_1,x_2) &= \int_0^1\int_0^1 \frac{K_h^J(x_1,x_2)\times \bar{K}_h^J(x_1)}{g_1^2(x_1)} g_1(x_1)g_1(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2,\\ &= \int_0^1 \left[\bar{K}_h^J(x_1)\right]^2/g_1(x_1)\,\mathrm{d}x_1.\end{aligned}$$

Now we need to expand $\bar{K}_h^J(x_1)$, when $0 \le x_1 \le h$,

$$\begin{aligned}\bar{K}_h^J(x_1) &= g_1(x_1)\int_{-1}^{x_1/h} k_{(x_1/h)}(u)\,du - hg_1'(x_1)\int_{-1}^{x_1/h} u k_{(x_1/h)}(u)\,du\\ &\quad + \frac{1}{2}h^2 g_1''(x_1)\int_{-1}^{x_1/h} u^2 k_{(x_1/h)}(u)\,du,\\ &= g_1(x_1) + O(h^2),\end{aligned}$$

because

$$\begin{aligned}\int_{-1}^{x_1/h} k_{(x_1/h)}(u)\,du &= 1,\\ \int_{-1}^{x_1/h} u k_{(x_1/h)}(u)\,du &= 0.\end{aligned}$$

Note that for Jackknife kernel $k_\rho(u)$,

$$\int_{-1}^\rho k_\rho(u)\,\mathrm{d}u = \int_{-1}^\rho (1+\beta)\frac{K(u)}{\omega_0(\rho)}\,\mathrm{d}u - \int_{-\alpha}^\rho \frac{\beta}{\alpha}\frac{K(u/\alpha)}{\omega_0(\rho/\alpha)}\,\mathrm{d}u,$$

for $h \le x_1 \le 1-h$ and $1-h < x_1 \le 1$, we also have $\bar{K}_h^J(x_1) = g_1(x_1) + O(h^2)$, so $\left[\bar{K}_h^J(x_1)\right]^2/g_1(x_1) = g_1(x_1) + O(h^2)$, hence $\mathrm{E}\psi_2(x_1,x_2) = \mathrm{E}\psi_3(x_1,x_2) = 1 + O(h^2)$.

Finally,

$$\begin{aligned}
\mathrm{E}a_n^2(x_1,x_2) &= \mathrm{E}\psi_1(x_1,x_2)+\mathrm{E}\psi_2(x_1,x_2)-2\mathrm{E}\psi_3(x_1,x_2)\\
&= (h^{-1}-2)\int_{-1}^{1}K^2(u)du\\
&\quad +2\int_0^1\int_{-1}^{\rho}k_\rho^2(u)\,\mathrm{d}u\,\mathrm{d}\rho-1+O(h),\\
&= O(h^{-1}).
\end{aligned}$$

Now, we have obtained the result of univariate case. For multivariate case, let $\mathbf{z}_1=(z_{10},\ldots,z_{1(m-2)},z_{1(m-1)})^T=(\mathbf{y}_1^T,z_{1(m-1)})^T$ and

$$\mathbf{z}_2=(z_{20},\ldots,z_{1(m-2)},z_{1(m-1)})^T=\left(\mathbf{y}_2^T,z_{2(m-1)}\right)^T,$$

by the Assumptions 1 and 2, the definition of multivariate kernel and $\mathbb{H}_0$, we can prove that

$$\begin{aligned}
\mathrm{E}A_{nm}^2(\mathbf{z}_1,\mathbf{z}_2) &= \mathrm{E}A_{n(m-1)}^2(\mathbf{y}_1,\mathbf{y}_2)\mathrm{E}a_n^2(z_{1(m-1)},z_{2(m-1)})\\
&\quad +\mathrm{E}A_{n(m-1)}^2(\mathbf{y}_1,\mathbf{y}_2)\mathrm{E}\left[\frac{\bar{K}_h^J(z_{1(m-1)})}{g_1(z_{1(m-1)})}\right]^2\\
&\quad +\mathrm{E}a_n^2(z_{1(m-1)},z_{2(m-1)})\mathrm{E}\left[\frac{\bar{\mathcal{K}}_h^{(m-1)}(\mathbf{y}_1)}{g_1(\mathbf{y}_1)}\right]^2, \qquad \text{(C.26)}\\
&= \mathrm{E}A_{n(m-1)}^2(\mathbf{y}_1,\mathbf{y}_2)O(h^{-1})\\
&\quad +\mathrm{E}A_{n(m-1)}^2(\mathbf{y}_1,\mathbf{y}_2)\mathrm{E}\big[b_n(z_{1(m-1)})+1\big]^2\\
&\quad +\mathrm{E}a_n^2(z_{1(m-1)},z_{2(m-1)})\mathrm{E}\big[B_{n(m-1)}(\mathbf{y}_1)+1\big]^2,
\end{aligned}$$

where $\bar{K}_h^J(z_1)=\int_0^1 K_h^J(z_1,z_2)g_1(z_2)\,\mathrm{d}z_2$ and

$$\bar{\mathcal{K}}_h^{(m-1)}(\mathbf{y}_1)=\int_{\mathbb{I}^{m-1}}\mathcal{K}_h^{(m-1)}(\mathbf{y}_1,\mathbf{y})g(\mathbf{y})\,\mathrm{d}\mathbf{y}.$$

Iteratively, we can obtain $\mathrm{E}A_{nm}^2(\mathbf{z}_1,\mathbf{z}_2)=O(h^{-m})$ by equations (C.26) and (C.27).

For the last part, we have $b_n(x_1)=\bar{K}_h^J(x_1)/g_1(x_1)-1=O(h^2)$, given $\mathbb{H}_0$, we can also verify that

$$B_{nm}(\mathbf{z}_1)=B_{n(m-1)}(\mathbf{y}_1)b_n(z_{1(m-1)})+B_{n(m-1)}(\mathbf{y}_1)+b_n(z_{1(m-1)}). \qquad \text{(C.27)}$$

This immediately completes the proof. □

**Lemma C.6.** *Given $\mathbb{H}_0$ and $1\le m<M$, we have*

$$EH_{1nm}(\mathbf{z}_1,\mathbf{z}_2)=\begin{cases}0 & \text{if } \mathbf{z}_1,\mathbf{z}_2 \text{ are independent,}\\ O(h^2) & \text{otherwise.}\end{cases}$$

*Proof of Lemma C.6.* When $\mathbf{z}_1$ and $\mathbf{z}_2$ have no overlap variable, i.e., $\mathbf{z}_1$ and $\mathbf{z}_2$ are independent, by the definition (5.21), (5.22) and $\mathbb{H}_0$, we have $\mathrm{E}H_{1nm}(\mathbf{z}_1, \mathbf{z}_2) = 0$. Next, we will prove the order of $\mathrm{E}H_{1nm}(\mathbf{z}_1, \mathbf{z}_2)$ is $O(h^2)$ if $\mathbf{z}_1$ and $\mathbf{z}_2$ have one or more overlap variables. Note that $\hat{H}_{1n}(m)$ is a U-statistics, $\mathbf{z}_1 \neq \mathbf{z}_2$, then $\mathbf{z}_1$ and $\mathbf{z}_2$ have at most $m-1$ overlap variables. By the fact $\tilde{A}_{nm}(\mathbf{z}_1, \mathbf{z}_2) - A_{nm}(\mathbf{z}_1, \mathbf{z}_2) = \gamma_{nm}(\mathbf{z}_1, \mathbf{z}_2)$ and Lemma 5.3, we only need to proof $\mathrm{E}A_{nm}(\mathbf{z}_1, \mathbf{z}_2) = O(h^2)$. First, we prove the order is $O(h^2)$ if $\mathbf{z}_1$ and $\mathbf{z}_2$ sharing $m-1$ overlap variables, then extend the result to the case whence $\mathbf{z}_1$ and $\mathbf{z}_2$ share only 1 overlap variable. Let $\mathbf{z}_1 = (z_1, \ldots, z_m)$ and $\mathbf{z}_2 = (z_2, \ldots, z_{m+1})$, from Lemma C.5, we know $\forall\, \mathbf{z}_1 \in \mathbb{I}^m$, $\bar{K}_h^{(m)} = g(\mathbf{z}_1) + O(h^2)$. Hence,

$$\begin{aligned}
&\mathrm{E}A_{nm}(\mathbf{z}_1, \mathbf{z}_2) \\
&\quad = \int_0^1 \cdots \int_0^1 \frac{K_h^{(m)}(\mathbf{z}_1, \mathbf{z}_2) - \bar{K}_h^{(m)}(\mathbf{z}_1)}{g(\mathbf{z}_1)} g(\mathbf{z}_1) g_{1(z_{m+1})} \,\mathrm{d}z_1 \cdots \mathrm{d}z_{m+1}, \\
&\quad = \int_0^1 \cdots \int_0^1 \left[ K_h^{(m)}(\mathbf{z}_1, \mathbf{z}_2) - g(\mathbf{z}_1) + O(h^2) \right] g_{1(z_{m+1})} \,\mathrm{d}z_1 \cdots \mathrm{d}z_{m+1}, \\
&\quad = \int_0^1 \cdots \int_0^1 K_h^{(m)}(\mathbf{z}_1, \mathbf{z}_2) g_{1(z_{m+1})} \,\mathrm{d}z_1 \cdots \mathrm{d}z_{m+1} - 1 + O(h^2).
\end{aligned} \tag{C.28}$$

We also notice that

$$\begin{aligned}
&\int_0^1 K_h^{(m)}(\mathbf{z}_1, \mathbf{z}_2) g_{1(z_{m+1})} \,\mathrm{d}z_{m1} \\
&\quad = K_h^J(z_1, z_2) \times \cdots \times K_h^J(z_{m-1}, z_m) \int_0^1 K_h^J(z_m, z_{m+1}) g_{1(z_{m+1})} \,\mathrm{d}z_{m1}, \\
&\quad = K_h^J(z_1, z_2) \times \cdots \times K_h^J(z_{m-1}, z_m) \left[ g_1(z_m) + O(h^2) \right],
\end{aligned} \tag{C.29}$$

iteratively substituting (C.29) into integration (C.28), we finally obtain

$$\mathrm{E}A_{nm}(\mathbf{z}_1, \mathbf{z}_2) = O(h^2).$$

Using the similar method, one can easily prove $\mathrm{E}A_{nm}(\mathbf{z}_1, \mathbf{z}_2) = O(h^2)$ if $\mathbf{z}_1, \mathbf{z}_2$ have only one overlap variable. This completes the proof. □

**Lemma C.7.** *Given $\mathbb{H}_0$ and $2 \leq m < M$, we have*

$$EH_{2nm}(\mathbf{z}_1, \mathbf{z}_2) = \begin{cases} 0 & \text{if } \mathbf{z}_1, \mathbf{z}_2 \text{ are independent,} \\ \tau^m - 1 + O(h) & \text{otherwise,} \end{cases}$$

*where $\tau = \int_{-1}^1 \int_{-1}^1 K(u) K(u+v) \,\mathrm{d}u \,\mathrm{d}v$.*

*Proof of Lemma C.7.* Let $\mathbf{z}_0 = (z_{01}, \ldots, z_{0m})^T$, $\mathbf{z}_1 = (z_1, \ldots, z_m)^T$ and

$\mathbf{z}_2 = (z_2, \ldots, z_{m+1})^T$, then we have

$$\begin{aligned} H_{2nm}(\mathbf{z}_1, \mathbf{z}_2) &= \int_{\mathbb{I}^m} A_{nm}(\mathbf{z}_0, \mathbf{z}_1) A_{nm}(\mathbf{z}_0, \mathbf{z}_2) g(\mathbf{z}_0)\, \mathrm{d}\mathbf{z}_0, \\ &= \int_{\mathbb{I}^m} \frac{K_h^{(m)}(\mathbf{z}_0, \mathbf{z}_1) K_h^{(m)}(\mathbf{z}_0, \mathbf{z}_2)}{g(\mathbf{z}_0)}\, \mathrm{d}\mathbf{z}_0 \\ &\quad - \int_{\mathbb{I}^m} \left[ K_h^{(m)}(\mathbf{z}_0, \mathbf{z}_1) + K_h^{(m)}(\mathbf{z}_0, \mathbf{z}_2) \right] \mathrm{d}\mathbf{z}_0 + 1 + O(h^2), \end{aligned}$$

therefore

$$\begin{aligned} &\mathrm{E} H_{2nm}(\mathbf{z}_1, \mathbf{z}_2) \\ &= \int_0^1 \int_{\mathbb{I}^m} \int_{\mathbb{I}^m} \frac{K_h^{(m)}(\mathbf{z}_0, \mathbf{z}_1) K_h^{(m)}(\mathbf{z}_0, \mathbf{z}_2)}{g(\mathbf{z}_0)} g(\mathbf{z}_1) g_1(z_{m+1})\, \mathrm{d}\mathbf{z}_0\, \mathrm{d}\mathbf{z}_1\, \mathrm{d}z_{m+1} - 1 + O(h^2). \end{aligned}$$

By change of variable and the first-order Taylor expansion, the first term can be expressed as $\tau^m + O(h)$. One can also obtain the same result for $\mathbf{z}_1 = (z_1, \ldots, z_m)^T$ and $\mathbf{z}_2 = (z_m, \ldots, z_{2m-1})^T$ using the same discussion. This completes the proof. □

## C.5 Relationship of CoEn and ApEn

We define the multivariate uniform kernel as

$$\mathcal{K}(\mathbf{x}) = 2^{-m} \mathbb{1}\left( \|\mathbf{x}\|_\infty \leq 1 \right), \tag{C.30}$$

where $m$ is the length of $\mathbf{x}$ and $\|\mathbf{x}\|_\infty$ is the maximum norm. In fact, (C.30) is one type of **multiplicative** kernel. If we let bandwidth for each entry of $\mathbf{x}$ be constant $h$, then the scaled multivariate kernel can be expressed as

$$\mathcal{K}_h(\mathbf{x}) = (2h)^{-m} \mathbb{1}\left( \|\mathbf{x}\|_\infty \leq h \right).$$

We still use the notions $\mathbf{x}_i^{(m)}$ and $\mathbf{x}_i^{(m+1)}$ as defined in Section 5.2, therefore the kernel density estimators of both $\mathbf{x}_i^{(m)}$ and $\mathbf{x}_i^{(m+1)}$ can be written as

$$\begin{aligned} \hat{f}\left( \mathbf{x}_i^{(m+1)} \right) &= \frac{1}{n} \sum_{j=1}^{n} \mathcal{K}_h \left( \mathbf{x}_j^{(m+1)} - \mathbf{x}_i^{(m+1)} \right), \\ &= (2h)^{-(m+1)} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\left( \left\| \mathbf{x}_j^{(m+1)} - \mathbf{x}_i^{(m+1)} \right\|_\infty \leq h \right), \\ &= (2h)^{-(m+1)} C_i^{(m+1)}(h), \quad i = 1, \ldots, n, \end{aligned}$$

$$\begin{aligned}\hat{g}\left(\mathbf{x}_i^{(m)}\right) &= \frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h\left(\mathbf{x}_j^{(m)} - \mathbf{x}_i^{(m)}\right), \\ &= (2h)^{-m}\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}\left(\left\|\mathbf{x}_j^{(m)} - \mathbf{x}_i^{(m)}\right\|_\infty \leq h\right), \\ &= (2h)^{-m}C_i^{(m)}(h), \quad i = 1, \ldots, n.\end{aligned}$$

Hence, conditional entropy can be expressed as

$$\begin{aligned}\text{CoEn} &= -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\hat{f}\left(\mathbf{x}_i^{(m+1)}\right)}{\hat{g}\left(\mathbf{x}_i^{(m)}\right)}\right), \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{C_i^{(m+1)}(h)}{(2h)C_i^{(m)}(h)}\right), \\ &= \text{ ApEn } + \log(2h).\end{aligned}$$

Especially, for Gaussian kernel, the relationship of **CoEn** and **ApEn** is

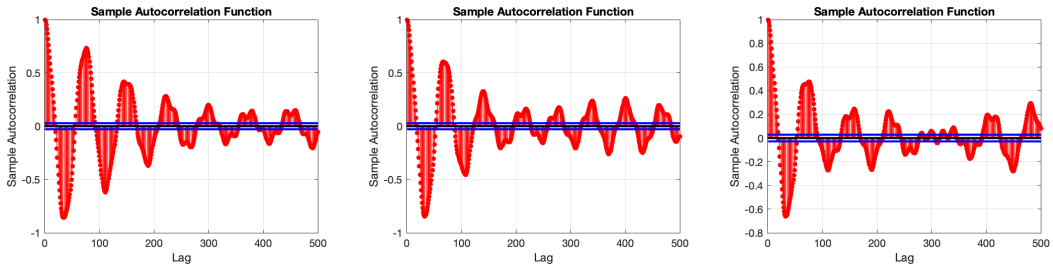
$$\text{CoEn} = \text{ ApEn } + \log\left(\sqrt{2}h\right).$$

## C.6 Seasonal ARIMA Estimation

We divide the real sports time series into three groups according to the change points 16 and 22, i.e., Group 1 indices: 1–15; Group 2 indices: 16–21; Group 3 indices: 22–52. The average of each group denotes as $x_1, x_2, x_3$ respectively. Next we will estimate Processes 1, 2 and 3 based on $x_1, x_2, x_3$ step by step.

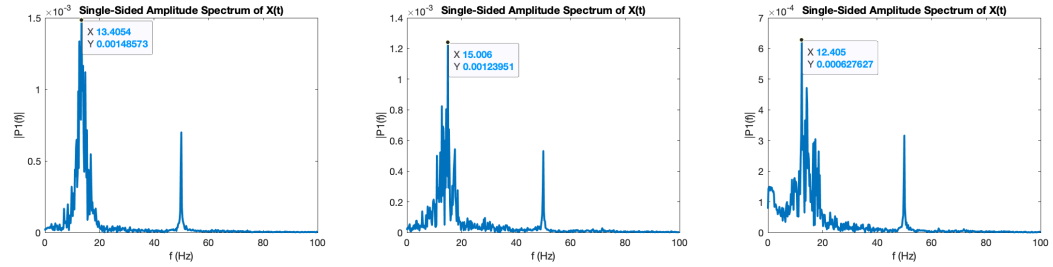
**Degree of Integration** Using the Augmented Dickey-Fuller test (Dickey & Fuller, 1979), we found the degree of integration is 2 for $x_1, x_2, x_3$.

**The Period of Season** We check the graph of sample autocorrelation function, see Figure C.2. In the dataset cleaning step, the dataset is filtered by lower-pass ButterWorth, the cut-off frequency is 20 Hz. Figure C.2 also implies the periodical auto-correlation of $x_1, x_2, x_3$. Discrete Faster Fourier Transformation is used here to convert time series analysis from time domain to frequency domain. For example, Figure C.3(a) demonstrates the result of Discrete Faster Fourier Transformation of $x_1$. Clearly, the largest amplitude is 0.01603, the corresponding frequency is 13.4 Hz, which mean the period is around 75. The seasonality of Process 1 sets to be 75. Using the same way, the seasonalities of Processes 2 and 3 are 67 and 81 respectively. Furthermore, we specify the seasonality order of AR as 1 for simplicity.




*(a) ACF of $x_1$* *(b) ACF of $x_2$* *(c) ACF of $x_3$*

*Figure C.2: Sample Autocorrelation Function*




*(a) $x_1$* *(b) $x_2$* *(c) $x_3$*

*Figure C.3: Single-Sided Amplitude Spectrum Analysis*

**The Choice of Order** Based on the previous analysis, we suggest the following process[1] for Group 1:

$$\phi(L)\Phi(L)(1-L)^D(1-L^s)^{D_s}x_t = c+\theta(L)\varepsilon_t,$$

where $\phi(L) = 1-\phi_1 L-\cdots-\phi_p L^p$ and $\theta(L) = 1+\theta_1 L+\cdots+\theta_q L^q$ represent the AR and MA operator polynomials. $\Phi(L) = 1-\Phi_{p_1}L^{p_1}-\Phi_{p_2}L^{p_2}-\cdots-\Phi_{p_s}L^{p_s}$ is seasonal auto-regressive operator polynomials. $(1-L^s)^{D_s}$ is the so-called Seasonal Difference factor. For Group 1, based on our previous analysis, we let $D = 2$, $s = 75$ and $D_{75} = 1$, the unknown parameters are $\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q,\Phi_{75}$. We can not guarantee each unknown parameter significant at this moment. Therefore, for given pair $(p,q)$, we apply the backward model selection method to choose the significant parameters.

Specifically, we let $p$ and $q$ changes from 0 to 4 respectively, for each combination of $p$ and $q$, the BIC of backward model selection in each step is computed. Then, choose the combination of $p$ and $q$ which has a minimum BIC.

Finally, we summarize the basic steps of SARIMA process as follows:

Step 1: Using the Augmented Dickey-Fuller test to determine $D$;

Step 2: Using Discrete Faster Fourier Transformation to choose the seasonal period;

Step 3: For user-specified order of seasonality of AR and MA polynomial,

[1] https://uk.mathworks.com/help/econ/seasonal-arima-sarima-model.html

choose the lag numbers;

Step 4: For given $(p, q)$, using backward model selection to choose the modal and compute the corresponding BIC;

Step 5: Let $p$ and $q$ change from 0 to $p.\max$ and $q.\max$ respectively, repeat step 4, and choose the combination which has the minimum BIC.

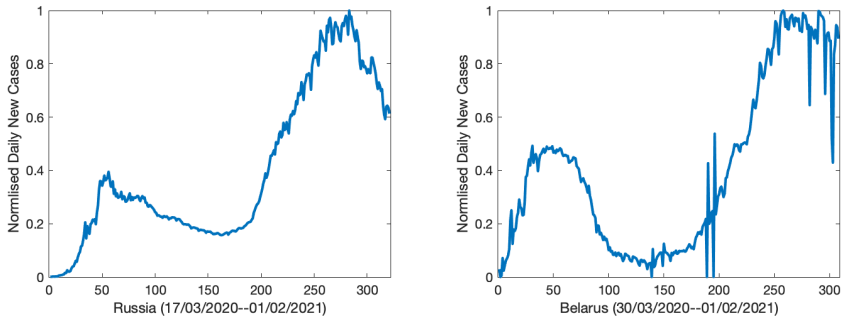
For Group 2 and 3, we use the same procedures to choose the order and parameters of SARIMA process.

Note, the process is not optimal because (1) the range of $p, q$ is from 1 to 4, one can extend this range to 10, 20, etc, but the complexity will increase as well; (2) seasonal autoregressive order is 1, there maybe exist more periods, see Figure C.3(a), if we go further, the complexity and computation consumption will increase significantly.

We also apply the same estimation procedure to $x_2$ and $x_3$, the order $p, q$ are (2, 2) and (2, 1) respectively. We generate 15 time series from Process 1, 6 time series from Process 2 and 31 time series from Process 3. Even in this case, our RlEn method can detect the change points 16 and 22.

# List of publications

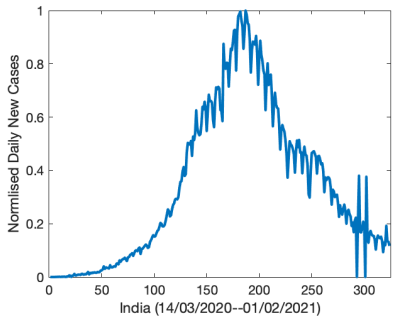
Zhang, Jian & Jie Li (2021). "Factorized Estimation of High-Dimensional Nonparametric Covariance Models". *Scandinavian Journal of Statistics*. DOI: 10.1111/sjos.12529.



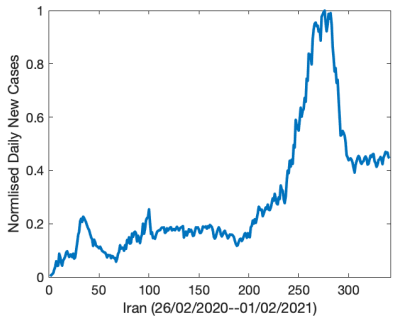

*(a) RlEn=1.886*



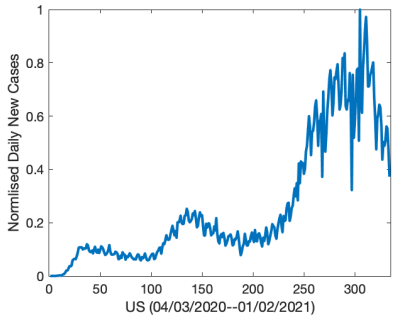

*(b) RlEn=1.578*



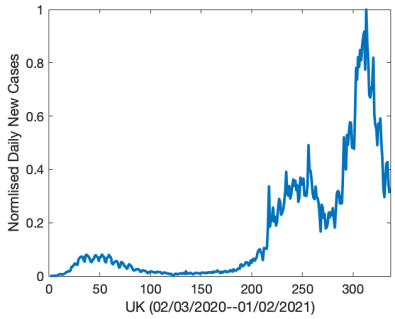

*(c) RlEn=1.393*



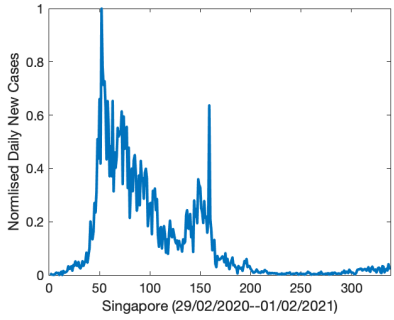

*(d) RlEn=1.359*



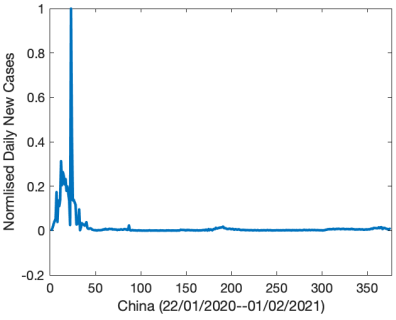

*(e) RlEn=1.055*




*(f) RlEn=0.831*




*(g) RlEn=0.548*




*(h) RlEn=0.103*

*Figure C.4: Normalized Daily New Cases for eight Countries*

# Bibliography


Acharya U, Rajendra, Kannathal N, Ong Wai Sing, Luk Yi Ping & TjiLeng Chua (2004). "Heart Rate Analysis in Normal Subjects of Various Age Groups". *BioMedical Engineering OnLine*, 24. DOI: 10.1186/1475-925X-3-24 (*see p. 4*).

Acharya U., Rajendra, Oliver Faust, N. Kannathal, TjiLeng Chua & Swamy Laxminarayan (2005). "Non-Linear Analysis of EEG Signals at Various Sleep Stages". *Computer Methods and Programs in Biomedicine* **1**, 37–45. DOI: 10.1016/j.cmpb.2005.06.011 (*see p. 4*).

Agre, J. C. & A. A. Rodriquez (1991). "Intermittent Isometric Activity: Its Effect on Muscle Fatigue in Postpolio Subjects". *Archives of Physical Medicine and Rehabilitation* **12**, 971–975 (*see p. 99*).

Ahmed, Amr & Eric P. Xing (2009). "Recovering Time-Varying Networks of Dependencies in Social and Biological Studies". *Proceedings of the National Academy of Sciences of the United States of America* **29**, 11878–11883. DOI: 10.1073/pnas.0901910106 (*see p. 1*).

Ahmed, Mosabber Uddin & Danilo P. Mandic (2012). "Multivariate Multiscale Entropy Analysis". *IEEE Signal Processing Letters* **2**, 91–94. DOI: 10.1109/LSP.2011.2180713 (*see p. 23*).

Amini, Arash A. & Martin J. Wainwright (2008). "High-Dimensional Analysis of Semidefinite Relaxations for Sparse Principal Components". In: *2008 IEEE International Symposium on Information Theory*. Toronto, ON, Canada: IEEE, 2454–2458. DOI: 10.1109/ISIT.2008.4595432 (*see p. 3*).

An, B., J. Guo & Y. Liu (2014). "Hypothesis Testing for Band Size Detection of High-Dimensional Banded Precision Matrices". *Biometrika* **2**, 477–483. DOI: 10.1093/biomet/asu006 (*see p. 42*).

Bai, Zhidong & Jack W. Silverstein (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. New York, NY: Springer New York. 560 pp. DOI: 10.1007/978-1-4419-0661-8 (*see p. 49*).

Benjamini, Yoav, Abba M. Krieger & Daniel Yekutieli (2006). "Adaptive Linear Step-up Procedures That Control the False Discovery Rate". *Biometrika* **3**, 491–507. DOI: 10.1093/biomet/93.3.491 (*see p. 27*).

Bickel, Peter J. & Elizaveta Levina (2008a). "Covariance Regularization by Thresholding". *Annals of Statistics* **6**, 2577–2604. DOI: 10.1214/08-AOS600 (*see pp. 17, 37, 47, 48*).

– (2008b). "Regularized Estimation of Large Covariance Matrices". *Annals of Statistics* **1**, 199–227. DOI: 10.1214/009053607000000758 (*see pp. 3, 62, 63*).

Bien, Jacob & Robert J. Tibshirani (2011). "Sparse Estimation of a Covariance Matrix". *Biometrika* **4**, 807–820. DOI: 10.1093/biomet/asr054 (*see p. 38*).

Biscay, Rolando, Luis M. Rodríguez & Eloísa Díaz-Frances (1997). "Cross-Validation of Covariance Structures Using the Frobenius Matrix Distance as a Discrepancy Function". *Journal of Statistical Computation and Simulation* **3**, 195–215. DOI: 10.1080/00949659708811831 (*see p. 20*).

Braun, J. V., R. K. Braun & H. -G. Muller (2000). "Multiple Changepoint Fitting via Quasilikelihood, with Application to DNA Sequence Segmentation". *Biometrika* **2**, 301–314. DOI: 10.1093/biomet/87.2.301 (*see p. 111*).

Brown, B. M. (1971). "Martingale Central Limit Theorems". *Annals of Mathematical Statistics* **1**, 59–66. DOI: 10.1214/aoms/1177693494 (*see p. 203*).

Buja, Andreas, Trevor Hastie & Robert Tibshirani (1989). "Linear Smoothers and Additive Models". *Annals of Statistics* **2**, 453–510. DOI: 10.1214/aos/1176347115 (*see p. 12*).

Burioka, Naoto, Masanori Miyata, Germaine Cornélissen, Franz Halberg, Takao Takeshima, Daniel T. Kaplan, Hisashi Suyama, Masanori Endo, Yoshihiro Maegaki, Takashi Nomura, Yutaka Tomita, Kenji Nakashima & Eiji Shimizu (2005). "Approximate Entropy in the Electroencephalogram during Wake and Sleep". *Clinical EEG Neuroscience* **1**, 21–24. DOI: 10.1177/155005940503600106 (*see pp. 4, 129*).

Butterworth, S. (1930). "On the Theory of Filter Amplifiers". *Experimental Wireless and the Wireless Engineer*, 536–541 (*see p. 126*).

Cai, Tony & Weidong Liu (2011). "Adaptive Thresholding for Sparse Covariance Matrix Estimation". *Journal of the American Statistical Association* **494**, 672–684. DOI: 10.1198/jasa.2011.tm10560 (*see pp. 17, 33, 38, 48*).

Cao-abad, R. & W. González-Manteiga (1993). "Bootstrap Methods in Regression Smoothing". *Journal of Nonparametric Statistics* **4**, 379–388. DOI: 10.1080/10485259308832566 (*see p. 18*).

Chacón, José E., Tarn Duong & Tarn Duong (2018). *Multivariate Kernel Smoothing and Its Applications*. Chapman and Hall/CRC. ISBN: 978-0-429-48557-2. DOI: 10.1201/9780429485572 (*see p. 18*).

Chamberlain, Gary & Michael Rothschild (1983). "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets". *Econometrica* **5**, 1281–1304. DOI: 10.2307/1912275 (*see p. 2*).

Chaudhuri, Sanjay, Mathias Drton & Thomas S. Richardson (2007). "Estimation of a Covariance Matrix with Zeros". *Biometrika* **1**, 199–216. DOI: 10.1093/biomet/asm007 (*see p. 38*).

Chen, Jia, Degui Li & Oliver B. Linton (2018). "A New Semiparametric Estimation Approach of Large Dynamic Covariance Matrices with Multiple Conditioning Variables". *Social Science Research Network Journal*. DOI: 10.2139/ssrn.3210726 (*see p. 1*).

Chen, Kun, Kung-Sik Chan & Nils Chr. Stenseth (2012). "Reduced Rank Stochastic Regression with a Sparse Singular Value Decomposition: Reduced Rank Stochastic Regression". *Journal of the Royal Statistical Society: Series B* **2**, 203–221. DOI: 10.1111/j.1467-9868.2011.01002.x (*see p. 1*).

Chen, Weiting, Jun Zhuang, Wangxin Yu & Zhizhong Wang (2009). "Measuring Complexity Using FuzzyEn, ApEn, and SampEn". *Medical Engineering & Physics* **1**, 61–68. DOI: 10.1016/j.medengphy.2008.04.005 (*see pp. 21, 23, 26*).

Chen, Xiaohui, Mengyu Xu & Wei Biao Wu (2013). "Covariance and Precision Matrix Estimation for High-Dimensional Time Series". *Annals of Statistics* **6**, 2994–3021. DOI: 10.1214/13-AOS1182 (*see p. 1*).

Chen, Ziqi & Chenlei Leng (2015). "Local Linear Estimation of Covariance Matrices via Cholesky Decomposition". *Statistica Sinica*. DOI: 10.5705/ss.2013.129 (*see p. 21*).

– (2016). "Dynamic Covariance Models". *Journal of the American Statistical Association* **515**, 1196–1207. DOI: 10.1080/01621459.2015.1077712 (*see pp. 1–3, 6, 7, 16–19, 21, 32–35, 37, 38, 41, 47, 49, 54, 55, 61, 62, 64, 77, 79, 135*).

Cheng, Ming-Yen, Jianqing Fan & J. S. Marron (1997). "On Automatic Boundary Corrections". *Annals of Statistics* **4**, 1691–1708. DOI: 10.1214/aos/1031594737 (*see p. 14*).

Chon, Ki H., Christopher G. Scully & Sheng Lu (2009). "Approximate Entropy for All Signals". *IEEE Engineering in Medicine and Biology Magazine* **6**, 18–23. DOI: 10.1109/MEMB.2009.934629 (*see p. 25*).

Clark, R. M. (1977). "Non-Parametric Estimation of a Smooth Regression Function". *Journal of the Royal Statistical Society. Series B* **1**, 107–113. DOI: 10.1111/j.2517-6161.1977.tb01611.x (*see p. 18*).

Clauset, Aaron, M. E. J. Newman & Cristopher Moore (2004). "Finding Community Structure in Very Large Networks". *Physical Review E* **6**, 1–6. DOI: 10.1103/PhysRevE.70.066111 (*see p. 97*).

Cormen, Thomas H., ed. (2009). *Introduction to Algorithms*. 3rd ed. Cambridge, Mass: MIT Press. 1292 pp. ISBN: 978-0-262-03384-8 (*see pp. 9, 65*).

Costa, M., C. -K. Peng, Ary L. Goldberger & Jeffrey M. Hausdorff (2003). "Multiscale Entropy Analysis of Human Gait Dynamics". *Physica A: Statistical Me-*

*chanics and its Applications*. RANDOMNESS AND COMPLEXITY: Proceedings of the International Workshop in Honor of Shlomo Havlin's 60th Birthday **1**, 53–60. DOI: 10.1016/j.physa.2003.08.022 (*see pp. 21, 23*).

Costa, Madalena, Ary L. Goldberger & C.-K. Peng (2005). "Multiscale Entropy Analysis of Biological Signals". *Physical Review E* **2**, 021906. DOI: 10.1103/PhysRevE.71.021906 (*see p. 23*).

Davie, A. M. & A. J. Stothers (2013). "Improved Bound for Complexity of Matrix Multiplication". *Proceedings of the Royal Society of Edinburgh: Section A Mathematics* **2**, 351–369. DOI: 10.1017/S0308210511001648 (*see p. 20*).

Dempster, A. P. (1972). "Covariance Selection". *Biometrics* **1**, 157–175. DOI: 10.2307/2528966 (*see p. 63*).

Dette, Holger & Natalie Neumeyer (2001). "Nonparametric Analysis of Covariance". *Annals of Statistics* **5**, 1361–1400. DOI: 10.1214/aos/1013203458 (*see p. 1*).

Dickey, David A. & Wayne A. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". *Journal of the American Statistical Association* (366a), 427–431. DOI: 10.1080/01621459.1979.10482531 (*see p. 213*).

Donoho, David (1995). "Nonlinear Solution of Linear Inverse Problems by Wavelet-Vaguelette Decomposition". *Applied and Computational Harmonic Analysis*, 101–126. DOI: 10.1006/acha.1995.1008 (*see p. 11*).

Donoho, David L. (1994). "Statistical Estimation and Optimal Recovery". *Annals of Statistics* **1**, 238–270. DOI: 10.1214/aos/1176325367 (*see p. 11*).

Donoho, David L. & Iain M. Johnstone (1994). "Ideal Spatial Adaptation by Wavelet Shrinkage". *Biometrika* **3**, 425–455. DOI: 10.2307/2337118 (*see p. 11*).

– (1995). "Adapting to Unknown Smoothness via Wavelet Shrinkage". *Journal of the American Statistical Association* **432**, 1200–1224. DOI: 10.1080/01621459.1995.10476626 (*see p. 11*).

– (1998). "Minimax Estimation via Wavelet Shrinkage". *Annals of Statistics* **3**, 879–921. DOI: 10.1214/aos/1024691081 (*see p. 12*).

Donoho, David L., Iain M. Johnstone, Gerard Kerkyacharian & Dominique Picard (1995). "Wavelet Shrinkage: Asymptopia?" *Journal of the Royal Statistical Society. Series B* **2**, 301–369. DOI: 10.1111/j.2517-6161.1995.tb02032.x (*see p. 12*).

Douglas, Critchlow E. & Fligner A. Michael (1991). "On Distribution-Free Multiple Comparisons in the One-Way Analysis of Variance". *Communications in Statistics - Theory and Methods* **1**, 127–139. DOI: 10.1080/03610929108830487 (*see p. 88*).

Dunn, Olive Jean (1961). "Multiple Comparisons among Means". *Journal of the American Statistical Association* **293**, 52–64. DOI: 10.1080/01621459.1961.10482090 (*see p. 88*).

Engle, Robert F., Olivier Ledoit & Michael Wolf (2017). "Large Dynamic Covariance Matrices". *Journal of Business & Economic Statistics* **2**, 363–375. DOI: 10.1080/07350015.2017.1345683 (*see pp. 32, 33*).

Enoka, Roger M. & Jacques Duchateau (2008). "Muscle Fatigue: What, Why and How It Influences Muscle Function: Muscle Fatigue". *The Journal of Physiology* **1**, 11–23. DOI: 10.1113/jphysiol.2007.139477 (*see p. 99*).

Fama, Eugene F. & Kenneth R. French (2004). "The Capital Asset Pricing Model: Theory and Evidence". *Journal of Economic Perspectives* **3**, 25–46. DOI: 10.1257/0895330042162430 (*see pp. 2, 34*).

Fan, Jianqing, Theo Gasser, Irène Gijbels, Michael Brockmann & Joachim Engel (1997). "Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency". *Annals of the Institute of Statistical Mathematics* **1**, 79–99. DOI: 10.1023/A:1003162622169 (*see pp. 14, 98*).

Fan, Jianqing & Irène Gijbels (1996). *Local Polynomial Modelling and Its Applications*. 1. CRC Press reprint. Monographs on Statistics and Applied Probability 66. Boca Raton: Chapman & Hall / CRC. 341 pp. ISBN: 978-0-412-98321-4 (*see pp. 11, 12, 14, 18, 21, 64*).

Fan, Jianqing, Yuan Liao & Martina Mincheva (2013). "Large Covariance Estimation by Thresholding Principal Orthogonal Complements". *Journal of the Royal Statistical Society. Series B* **4**, 603–680. DOI: 10.1111/rssb.12016 (*see pp. 2, 3, 17, 32*).

Fan, Jianqing & Qiwei Yao (1998). "Efficient Estimation of Conditional Variance Functions in Stochastic Regression". *Biometrika* **3**, 645–660. DOI: 10.1093/biomet/85.3.645 (*see pp. 12, 15, 71*).

– (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Series in Statistics. New York: Springer-Verlag. 565 pp. ISBN: 978-0-387-26142-3. DOI: 10.1007/978-0-387-69395-8 (*see pp. 18, 50, 107, 113–115, 122*).

Fan, Jianqing, Chunming Zhang & Jian Zhang (2001). "Generalized Likelihood Ratio Statistics and Wilks Phenomenon". *Annals of Statistics* **1**, 153–193. DOI: 10.1214/aos/996986505 (*see pp. 21, 64, 65, 69, 74, 77, 136*).

Forrest, Sarah M., John H. Challis & Samantha L. Winter (2014). "The Effect of Signal Acquisition and Processing Choices on ApEn Values: Towards a ″ Gold Standard ″ for Distinguishing Effort Levels from Isometric Force Records". *Medical Engineering & Physics* **6**, 676–683. DOI: 10.1016/j.medengphy.2014.02.017 (*see pp. 4, 99*).

Fox, Emily B & David B Dunson (2015). "Bayesian Nonparametric Covariance Regression". *Journal of Machine Learning Research* **16**, 2501–2542 (*see pp. 18, 32*).

Freeman, Linton C. (1978). "Centrality in Social Networks Conceptual Clarification". *Social Networks* **3**, 215–239. DOI: 10.1016/0378-8733(78)90021-7 (*see p. 28*).

Friedman, Jerome, Trevor Hastie & Robert Tibshirani (2008). "Sparse Inverse Covariance Estimation with the Graphical Lasso". *Biostatistics* **3**, 432–441. DOI: 10.1093/biostatistics/kxm045 (*see pp. 1, 63*).

Friedman, Jerome H. & Werner Stuetzle (1981). "Projection Pursuit Regression". *Journal of the American Statistical Association* **376**, 817–823. DOI: 10.1080/01621459.1981.10477729 (*see p. 12*).

Fryzlewicz, Piotr (2014). "Wild Binary Segmentation for Multiple Change-Point Detection". *Annals of Statistics* **6**, 2243–2281. DOI: 10.1214/14-AOS1245 (*see pp. 5, 100*).

– (2020). "Detecting Possibly Frequent Change-Points: Wild Binary Segmentation 2 and Steepest-Drop Model Selection". *Journal of the Korean Statistical Society*. DOI: 10.1007/s42952-020-00060-x (*see pp. 5, 100*).

Gasser, Theo & Hans-Georg Müller (1979). "Kernel Estimation of Regression Functions". In: *Smoothing Techniques for Curve Estimation*. Ed. by Th. Gasser & M. Rosenblatt. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 23–68. ISBN: 978-3-540-09706-8. DOI: 10.1007/BFb0098489 (*see p. 44*).

Gasser, Theo, Lothar Sroka & Christine Jennen-steinmetz (1986). "Residual Variance and Residual Pattern in Nonlinear Regression". *Biometrika* **3**, 625–633. DOI: 10.1093/biomet/73.3.625 (*see p. 13*).

González Manteiga, W., M.D. Martnez Miranda & A. Pérez González (2004). "The Choice of Smoothing Parameter in Nonparametric Regression through Wild Bootstrap". *Computational Statistics & Data Analysis* **3**, 487–515. DOI: 10.1016/j.csda.2003.12.007 (*see p. 18*).

Green, P. J. & B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. 1st ed. Monographs on Statistics and Applied Probability 58. London ; New York: Chapman & Hall. 182 pp. ISBN: 978-0-412-30040-0 (*see p. 12*).

Guo, Shaojun, John Leigh Box & Wenyang Zhang (2017). "A Dynamic Structure for High-Dimensional Covariance Matrices and Its Application in Portfolio Allocation". *Journal of the American Statistical Association* **517**, 235–253. DOI: 10.1080/01621459.2015.1129969 (*see p. 2*).

Hall, Peter (1988). "Estimating the Direction in Which a Data Set Is Most Interesting". *Probability Theory and Related Fields* **1**, 51–77. DOI: 10.1007/BF00348752 (*see p. 114*).

Hall, Peter & R. J. Carroll (1989). "Variance Function Estimation in Regression: The Effect of Estimating the Mean". *Journal of the Royal Statistical Society. Series B* **1**, 3–14. DOI: 10.1111/j.2517-6161.1989.tb01744.x (*see p. 13*).

Hall, Peter, Nicholas I. Fisher & Branka Hoffmann (1994). "On the Nonparametric Estimation of Covariance Functions". *Annals of Statistics* **4**, 2115–2134. DOI: 10.1214/aos/1176325774 (*see p. 1*).

Hall, Peter & J. S. Marron (1990). "On Variance Estimation in Nonparametric Regression". *Biometrika* **2**, 415–419. DOI: 10.2307/2336824 (*see pp. 12, 13*).

Hall, Peter, Jeff Racine & Qi Li (2004). "Cross-Validation and the Estimation of Conditional Probability Densities". *Journal of the American Statistical Association* **468**, 1015–1026. DOI: 10.1198/016214504000000548 (*see p. 194*).

Hallac, David, Youngsuk Park, Stephen Boyd & Jure Leskovec (2017). *Network Inference via the Time-Varying Graphical Lasso*. URL: http://arxiv.org/abs/1703.01958 (*see p. 1*).

Härdle, W. & A. Tsybakov (1997). "Local Polynomial Estimators of the Volatility Function in Nonparametric Autoregression". *Journal of Econometrics* **1**, 223–242. DOI: 10.1016/S0304-4076(97)00044-4 (*see p. 14*).

Härdle, Wolfgang (1990). *Applied Nonparametric Regression*. Cambridge University Press. 356 pp. ISBN: 978-0-521-42950-4 (*see pp. 12, 18, 108*).

Hastie, Trevor & Robert Tibshirani (1999). *Generalized Additive Models*. Boca Raton, Fla: Chapman & Hall/CRC. 335 pp. ISBN: 978-0-412-34390-2 (*see pp. 12, 13*).

Hong, Yongmiao & Halbert White (2005). "Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence". *Econometrica* **3**, 837–901. DOI: 10.1111/j.1468-0262.2005.00597.x (*see pp. 7, 21, 24, 25, 30, 109, 111, 113–115, 117, 118, 120, 121, 197, 203, 206*).

Hsu, Wei-Yen (2015). "Assembling A Multi-Feature EEG Classifier for Left – Right Motor Imagery Data Using Wavelet-Based Fuzzy Approximate Entropy for Improved Accuracy". *International Journal of Neural Systems* **08**, 13. DOI: 10.1142/S0129065715500379 (*see p. 26*).

Huang, Jianhua Z., Naiping Liu, Mohsen Pourahmadi & Linxu Liu (2006). "Covariance Matrix Selection and Estimation via Penalised Normal Likelihood". *Biometrika* **1**, 85–98. DOI: 10.1093/biomet/93.1.85 (*see pp. 1, 63*).

Hyndman, Rob J. & George Athanasopoulos (2013). *Forecasting: Principles and Practice*. S.l.: OTexts. 292 pp. ISBN: 978-0-9875071-0-5 (*see p. 127*).

Ihara, Shunsuke (1993). *Information Theory for Continuous Systems*. Singapore ; River Edge, N.J: World Scientific. 308 pp. ISBN: 978-981-02-0985-8 (*see pp. 5, 101*).

Inclán, Carla & George C. Tiao (1994). "Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance". *Journal of the American Statistical Association* **427**, 913–923. DOI: 10.1080/01621459.1994.10476824 (*see p. 111*).

Jiang, Binyan, Ziqi Chen & Chenlei Leng (2020). "Dynamic Linear Discriminant Analysis in High Dimensional Space". *Bernoulli* **2**, 1234–1268. DOI: 10.3150/19-BEJ1154 (*see p. 21*).

John, Rice (1984). "Boundary Modification for Kernel Regression". *Communications in Statistics - Theory and Methods* **7**, 893–900. DOI: 10.1080/03610928408828728 (*see pp. 29, 30, 108*).

Johnstone, Iain M. & Arthur Yu Lu (2009). "On Consistency and Sparsity for Principal Components Analysis in High Dimensions". *Journal of the American Statistical Association* **486**, 682–693. DOI: 10.1198/jasa.2009.0121 (*see pp. 1, 3*).

Jolliffe, I. T. (2002). *Principal Component Analysis*. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag. 518 pp. ISBN: 978-0-387-95442-4. DOI: 10.1007/b98835 (*see p. 49*).

Jones, M. C. (1993). "Simple Boundary Correction for Kernel Density Estimation". *Statistics and Computing* **3**, 135–146. DOI: 10.1007/BF00147776 (*see pp. 29, 108*).

Jones, M. C., J. S. Marron & S. J. Sheather (1996). "A Brief Survey of Bandwidth Selection for Density Estimation". *Journal of the American Statistical Association* **433**, 401–407. DOI: 10.2307/2291420 (*see p. 18*).

Kaffashi, Farhad, Ryan Foglyano, Christopher G. Wilson & Kenneth A. Loparo (2008). "The Effect of Time Delay on Approximate & Sample Entropy Calculations". *Physica D: Nonlinear Phenomena* **23**, 3069–3074. DOI: 10.1016/j.physd.2008.06.005 (*see p. 25*).

Kan, Raymond & Guofu Zhou (2007). "Optimal Portfolio Choice with Parameter Uncertainty". *The Journal of Financial and Quantitative Analysis* **3**, 621–656. DOI: 10.1017/S0022109000004129 (*see p. 63*).

Karunamuni, R. J. & T. Alberts (2005). "On Boundary Correction in Kernel Density Estimation". *Statistical Methodology* **3**, 191–212. DOI: 10.1016/j.stamet.2005.04.001 (*see p. 14*).

Katayama, Keisho, Yasuhide Yoshitake, Kohei Watanabe, Hiroshi Akima & Koji Ishida (2010). "Muscle Deoxygenation during Sustained and Intermittent Isometric Exercise in Hypoxia". *Medicine and Science in Sports and Exercise* **7**, 1269–1278. DOI: 10.1249/MSS.0b013e3181cae12f (*see p. 99*).

Kendall, Maurice G., Alan Stuart, J. Keith Ord, Alan Stuart & Maurice G. Kendall (1973). *Inference and Relationship*. 3. ed. The Advanced Theory of Statistics 2. London: Griffin. 723 pp. ISBN: 978-0-85264-215-3 (*see pp. 66, 74*).

Kiliç, Emrah & Pantelimon Stanica (2013). "The Inverse of Banded Matrices". *Journal of Computational and Applied Mathematics* **1**, 126–135. DOI: 10.1016/j.cam.2012.07.018 (*see p. 44*).

Killick, R., P. Fearnhead & I. A. Eckley (2012). "Optimal Detection of Changepoints With a Linear Computational Cost". *Journal of the American Statistical Association* **500**, 1590–1598. DOI: 10.1080/01621459.2012.737745 (*see pp. 5, 100, 102, 111, 112, 132*).

Kolaczyk, Eric D. (2009). *Statistical Analysis of Network Data*. Springer Series in Statistics. New York, NY: Springer New York. 397 pp. ISBN: 978-0-387-88145-4. DOI: 10.1007/978-0-387-88146-1 (*see p. 28*).

Kolar, Mladen, Le Song, Amr Ahmed & Eric P. Xing (2010). "Estimating Time-Varying Networks". *The Annals of Applied Statistics* **1**, 94–123. DOI: 10.1214/09-AOAS308 (*see p. 1*).

Kooperberg, Charles & Charles J. Stone (1991). "A Study of Logspline Density Estimation". *Computational Statistics & Data Analysis* **3**, 327–347. DOI: 10.1016/0167-9473(91)90115-I (*see p. 12*).

Kooperberg, Charles, Charles J. Stone & Young K. Truong (1995a). "Logspline Estimation of a Possibly Mixed Spectral Distribution". *Journal of Time Series Analysis* **4**, 359–388. DOI: 10.1111/j.1467-9892.1995.tb00240.x (*see p. 12*).

– (1995b). "Rate of Convergence for Logspline Spectral Density Estimation". *Journal of Time Series Analysis* **4**, 389–401. DOI: 10.1111/j.1467-9892.1995.tb00241.x (*see p. 12*).

Kullback, S. & R. A. Leibler (1951). "On Information and Sufficiency". *Annals of Mathematical Statistics* **1**, 79–86. DOI: 10.1214/aoms/1177729694 (*see pp. 5, 100–102*).

Lam, Clifford & Jianqing Fan (2009). "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation". *The Annals of Statistics* (6B). DOI: 10.1214/09-AOS720 (*see p. 1*).

Lam, Clifford & Qiwei Yao (2012). "Factor Modeling for High-Dimensional Time Series: Inference for the Number of Factors". *Annals of Statistics* **2**, 694–726. DOI: 10.1214/12-AOS970 (*see p. 50*).

Lamus, Camilo, Matti S. Hämäläinen, Simona Temereanca, Emery N. Brown & Patrick L. Purdon (2012). "A Spatiotemporal Dynamic Distributed Solution to the MEG Inverse Problem". *Neuroimage* **2**, 894–909. DOI: 10.1016/j.neuroimage.2011.11.020 (*see p. 32*).

Lavielle, Marc (2005). "Using Penalized Contrasts for the Change-Point Problem". *Signal Processing* **8**, 1501–1510. DOI: 10.1016/j.sigpro.2005.01.012 (*see p. 132*).

Ledoit, Olivier & Michael Wolf (2004). "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices". *Journal of Multivariate Analysis* **2**, 365–411. DOI: 10.1016/S0047-259X(03)00096-4 (*see pp. 26, 27, 32, 34, 37, 38, 64*).

Lee, Mihee, Haipeng Shen, Jianhua Z. Huang & J. S. Marron (2010). "Biclustering via Sparse Singular Value Decomposition". *Biometrics* **4**, 1087–1095. DOI: 10.1111/j.1541-0420.2010.01392.x (*see p. 1*).

Li, Qi & Jeffrey Scott Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton, N.J: Princeton University Press. 746 pp. ISBN: 978-0-691-12161-1 (*see pp. 12, 115, 193, 196*).

Li, Y. (2011). "Efficient Semiparametric Regression for Longitudinal Data with Nonparametric Covariance Estimation". *Biometrika* **2**, 355–370. DOI: 10.1093/biomet/asq080 (*see p. 1*).

Lu, Junwei, Mladen Kolar & Han Liu (2017). "Post-Regularization Inference for Time-Varying Nonparanormal Graphical Models". *The Journal of Machine Learning Research* **1**, 7401–7478 (*see p. 1*).

Lu, Sheng, Xinnian Chen, JØrgen K. Kanters, Irene C. Solomon & Ki H. Chon (2008). "Automatic Selection of the Threshold Value $r$ for Approximate Entropy". *IEEE Transactions on Biomedical Engineering* **8**, 1966–1972. DOI: 10.1109/TBME.2008.919870 (*see p. 25*).

Manis, George (2008). "Fast Computation of Approximate Entropy". *Computer Methods and Programs in Biomedicine* **1**, 48–54. DOI: 10.1016/j.cmpb.2008.02.008 (*see p. 25*).

Manis, George, Md Aktaruzzaman & Roberto Sassi (2018). "Low Computational Cost for Sample Entropy". *Entropy* **1** (1), 61. DOI: 10.3390/e20010061 (*see p. 26*).

Meinshausen, Nicolai & Peter Bühlmann (2006). "High-Dimensional Graphs and Variable Selection with the Lasso". *Annals of Statistics* **3**, 1436–1462. DOI: 10.1214/009053606000000281 (*see p. 63*).

Nadaraya, E. A. (1964). "On Estimating Regression". *Theory of Probability and Its Applications* **1**, 141–142. DOI: 10.1137/1109020 (*see pp. 10, 64*).

Nychka, Douglas (1995). "Splines as Local Smoothers". *Annals of Statistics* **4**, 1175–1197. DOI: 10.1214/aos/1176324704 (*see p. 12*).

Page, E. S. (1954). "Continuous Inspection Schemes". *Biometrika* **1-2**, 100–115. DOI: 10.1093/biomet/41.1-2.100 (*see pp. 5, 100, 111*).

Pan, Yu-Hsiang, Yung-Hung Wang, Sheng-Fu Liang & Kuo-Tien Lee (2011). "Fast Computation of Sample Entropy and Approximate Entropy in Biomedicine".

*Computer Methods and Programs in Biomedicine* **3**, 382–396. DOI: 10.1016/j.cmpb.2010.12.003 (*see p. 26*).

Peng, Jie, Pei Wang, Nengfeng Zhou & Ji Zhu (2009). "Partial Correlation Estimation by Joint Sparse Regression Models". *Journal of the American Statistical Association* **486**, 735–746. DOI: 10.1198/jasa.2009.0126 (*see p. 1*).

Pethick, Jamie, Samantha L. Winter & Mark Burnley (2016). "Loss of Knee Extensor Torque Complexity during Fatiguing Isometric Muscle Contractions Occurs Exclusively above the Critical Torque". *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **11**, R1144–R1153. DOI: 10.1152/ajpregu.00019.2016 (*see pp. 4, 99, 129, 131*).

Picard, Franck, Emilie Lebarbier, Mark Hoebeke, Guillem Rigaill, Baba Thiam & Stéphane Robin (2011). "Joint Segmentation, Calling, and Normalization of Multiple CGH Profiles". *Biostatistics (Oxford, England)* **3**, 413–428. DOI: 10.1093/biostatistics/kxq076 (*see p. 111*).

Pincus, S. M. (1991). "Approximate Entropy as a Measure of System Complexity." *Proceedings of the National Academy of Sciences of the United States of America* **6**, 2297–2301. DOI: 10.1073/pnas.88.6.2297 (*see pp. 21, 22, 124, 129, 137*).

Pincus, Steven M. & Wei-Min Huang (1992). "Approximate Entropy: Statistical Properties and Applications". *Communications in Statistics - Theory and Methods* **11**, 3061–3077. DOI: 10.1080/03610929208830963 (*see p. 25*).

Pourahmadi, Mohsen (2013). *High-Dimensional Covariance Estimation*. Hoboken, New Jersey: Wiley. 184 pp. ISBN: 978-1-118-03429-3 (*see p. 1*).

Qiao, Xinghao, Cheng Qian, Gareth M James & Shaojun Guo (2020). "Doubly Functional Graphical Models in High Dimensions". *Biometrika* **2**, 415–431. DOI: 10.1093/biomet/asz072 (*see p. 1*).

Reich, Brian J., Jo Eidsvik, Michele Guindani, Amy J. Nail & Alexandra M. Schmidt (2011). "A Class of Covariate-Dependent Spatiotemporal Covariance Functions for the Analysis of Daily Ozone Concentration". *Ann. Appl. Stat.* **4**, 2425–2447. DOI: 10.1214/11-AOAS482 (*see p. 32*).

Rhea, Christopher K., Tobin A. Silver, S. Lee Hong, Joong Hyun Ryu, Breanna E. Studenka, Charmayne M. L. Hughes & Jeffrey M. Haddad (2011). "Noise and Complexity in Human Postural Control: Interpreting the Different Estimations of Entropy". *PLOS ONE* **3**. DOI: 10.1371/journal.pone.0017696 (*see p. 99*).

Rice, John (1984). "Bandwidth Choice for Nonparametric Regression". *Annals of Statistics* **4**, 1215–1230. DOI: 10.1214/aos/1176346788 (*see pp. 13, 18*).

Richman, Joshua S. & J. Randall Moorman (2000). "Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy". *American Journal of Physiology-Heart and Circulatory Physiology* **6**, H2039–H2049. DOI: 10.1152/ajpheart.2000.278.6.H2039 (*see pp. 21, 22, 129*).

Robinson, P. M. (1991). "Consistent Nonparametric Entropy-Based Testing". *Rev Econ Stud* **3**, 437–453. DOI: 10.2307/2298005 (*see pp. 21, 24, 25, 114*).

Rothman, Adam J. (2012). "Positive Definite Estimators of Large Covariance Matrices". *Biometrika* **3**, 733–740. DOI: 10.1093/biomet/ass025 (*see pp. 1, 38*).

Rothman, Adam J., Elizaveta Levina & Ji Zhu (2009). "Generalized Thresholding of Large Covariance Matrices". *null* **485**, 177–186. DOI: 10.1198/jasa.2009.0101 (*see p. 38*).

– (2010). "A New Approach to Cholesky-Based Covariance Regularization in High Dimensions". *Biometrika* **3**, 539–550. DOI: 10.1093/biomet/asq022 (*see p. 44*).

Roussas, GG & D Ioannides (1988). "Probability Bounds for Sums in Triangular Arrays of Random Variables under Mixing Conditions". *Statistical Theory and Data Analysis II*, 293–308 (*see p. 195*).

Rukhin, Andrew L. (2000). "Approximate Entropy for Testing Randomness". *Journal of Applied Probability* **1**, 88–100. DOI: 10.1239/jap/1014842270 (*see p. 25*).

Ruppert, D., S. J. Sheather & M. P. Wand (1995). "An Effective Bandwidth Selector for Local Least Squares Regression". *Journal of the American Statistical Association* **432**, 1257–1270. DOI: 10.2307/2291516 (*see p. 18*).

Ruppert, David, M. P. Wand, Ulla Holst & Ola HöSJER (1997). "Local Polynomial Variance-Function Estimation". *Technometrics* **3**, 262–273. DOI: 10.1080/00401706.1997.10485117 (*see pp. 12, 14*).

Sabidussi, Gert (1966). "The Centrality Index of a Graph". *Psychometrika* **4**, 581–603. DOI: 10.1007/BF02289527 (*see p. 28*).

Shao, Jun (1997). "An Asymptotic Theory for Linear Model Selection". *Statistica Sinica* **2**, 221–242 (*see pp. 111, 191, 192*).

Shen, Haipeng & Jianhua Z. Huang (2008). "Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation". *Journal of Multivariate Analysis* **6**, 1015–1034. DOI: 10.1016/j.jmva.2007.06.007 (*see pp. 1, 3*).

Shi, Bo, Yudong Zhang, Chaochao Yuan, Shuihua Wang & Peng Li (2017). "Entropy Analysis of Short-Term Heartbeat Interval Time Series during Regular Walking". *Entropy* **10** (10), 568. DOI: 10.3390/e19100568 (*see p. 4*).

Shibata, Ritei (1981). "An Optimal Selection of Regression Variables". *Biometrika* **1**, 45–54. DOI: 10.2307/2335804 (*see pp. 111, 191*).

Taylor, Paul G., Michael Small, Kwee-Yum Lee, Raul Landeo, Damien M. O'Meara & Emma L. Millett (2016). "A Surrogate Technique for Investigating Deterministic Dynamics in Discrete Human Movement". *Motor Control* **4**, 459–470. DOI: 10.1123/mc.2015-0043 (*see p. 99*).

Udhayakumar, Radhagayathri K., Chandan Karmakar & Marimuthu Palaniswami (2017). "Approximate Entropy Profile: A Novel Approach to Comprehend Irregularity of Short-Term HRV Signal". *Nonlinear Dyn* **2**, 823–837. DOI: 10.1007/s11071-016-3278-z (*see p. 25*).

Vieu, Philippe (1991). "Smoothing Techniques in Time Series Analysis". In: *Nonparametric Functional Estimation and Related Topics*. Ed. by George Roussas. NATO ASI Series. Dordrecht: Springer Netherlands, 271–283. ISBN: 978-94-011-3222-0. DOI: 10.1007/978-94-011-3222-0_21 (*see p. 194*).

– (1995). "Order Choice in Nonlinear Autoregressive Models". *Statistics* **4**, 307–328. DOI: 10.1080/02331889508802499 (*see pp. 111, 191–195*).

Wakeman, Daniel G. & Richard N. Henson (2015). "A Multi-Subject, Multi-Modal Human Neuroimaging Dataset". *Scientific Data* **1** (1), 150001. DOI: 10.1038/sdata.2015.1 (*see p. 4*).

Wand, M. P. & M. C. Jones (1995). *Kernel Smoothing*. 1st ed. Monographs on Statistics and Applied Probability 60. London ; New York: Chapman & Hall. 212 pp. ISBN: 978-0-412-55270-0 (*see pp. 11, 14, 18*).

Wang, Hanchao, Bin Peng, Degui Li & Chenlei Leng (2020). "Nonparametric Estimation of Large Covariance Matrices with Conditional Sparsity". *SSRN Journal*. DOI: 10.2139/ssrn.3515624 (*see pp. 1, 2*).

Wang, Jialei & Mladen Kolar (2014). *Inference for Sparse Conditional Precision Matrices*. URL: http://arxiv.org/abs/1412.7638 (*see p. 63*).

Wasserman, Larry (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. New York: Springer. ISBN: 978-0-387-25145-5 (*see pp. 12, 111*).

Watson, Geoffrey S. (1964). "Smooth Regression Analysis". *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)* **4**, 359–372 (*see p. 10*).

Witten, D. M., R. Tibshirani & T. Hastie (2009). "A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis". *Biostatistics* **3**, 515–534. DOI: 10.1093/biostatistics/kxp008 (*see p. 1*).

Wold, Herman O. A. (1948). "On Prediction in Stationary Time Series". *The Annals of Mathematical Statistics* **4**, 558–567. DOI: 10.1214/aoms/1177730151 (*see pp. 107, 191*).

Xu, Lin, Man-Lai Tang & Ziqi Chen (2019). "Analysis of Longitudinal Data by Combining Multiple Dynamic Covariance Models". *Statistics and Its Interface* **3**, 479–487. DOI: 20190612152513 (*see p. 21*).

Yang, Dan, Zongming Ma & Andreas Buja (2014). "A Sparse Singular Value Decomposition Method for High-Dimensional Data". *Journal of Computational and Graphical Statistics* **4**, 923–942. DOI: 10.1080/10618600.2013.858632 (*see p. 1*).

Yang, Jilei & Jie Peng (2018). *Estimating Time-Varying Graphical Models*. URL: http://arxiv.org/abs/1804.03811 (*see p. 1*).

Yentes, Jennifer M., Nathaniel Hunt, Kendra K. Schmid, Jeffrey P. Kaipust, Denise McGrath & Nicholas Stergiou (2013). "The Appropriate Use of Approximate Entropy and Sample Entropy with Short Data Sets". *Ann Biomed Eng* **2**, 349–365. DOI: 10.1007/s10439-012-0668-3 (*see p. 26*).

Yin, Jianxin, Zhi Geng, Runze Li & Hansheng Wang (2010). "Nonparametric Covariance Model". *Stat Sin*, 469–479 (*see pp. 1, 2, 6, 7, 16–19, 21, 32, 35, 38, 43, 47, 61, 68, 135, 137*).

Yu, K & M. C Jones (2004). "Likelihood-Based Local Linear Estimation of the Conditional Variance Function". *Journal of the American Statistical Association* **465**, 139–144. DOI: 10.1198/016214504000000133 (*see pp. 12, 15, 16, 71, 98*).

Yuan, Ming & T. Tony Cai (2010). "A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression". *Annals of Statistics* **6**, 3412–3444. DOI: 10.1214/09-AOS772 (*see pp. 55, 77, 79*).

Yuan, Ming & Y. Lin (2007). "Model Selection and Estimation in the Gaussian Graphical Model". *Biometrika* **1**, 19–35. DOI: 10.1093/biomet/asm018 (*see pp. 1, 63*).

Zhang, Jian & Jie Li (2021). "Factorized Estimation of High-Dimensional Nonparametric Covariance Models". *Scandinavian Journal of Statistics*. DOI: 10.1111/sjos.12529 (*see pp. 50, 51, 53, 59*).

Zhang, Jian & Chao Liu (2015). "On Linearly Constrained Minimum Variance Beamforming". *Journal of Machine Learning Research* **65**, 2099–2145. URL: http://jmlr.org/papers/v16/zhang15b.html (*see pp. 32, 50, 55, 79*).

Zhang, Jian & Li Su (2015). "Temporal Autocorrelation-Based Beamforming With MEG Neuroimaging Data". *Journal of the American Statistical Association* **512**, 1375–1388. DOI: 10.1080/01621459.2015.1054488 (*see p. 32*).

Zhou, Shuheng, John Lafferty & Larry Wasserman (2010). "Time Varying Undirected Graphs". *Machine Learning* **2**, 295–319. DOI: 10.1007/s10994-010-5180-0 (*see pp. 1, 81*).

Zou, Tao, Wei Lan, Hansheng Wang & Chih-Ling Tsai (2017). "Covariance Regression Analysis". *Journal of the American Statistical Association* **517**, 266–281. DOI: 10.1080/01621459.2015.1131699 (*see p. 63*).

Zurek, Sebastian, Przemyslaw Guzik, Sebastian Pawlak, Marcin Kosmider & Jaroslaw Piskorski (2012). "On the Relation between Correlation Dimension, Approximate Entropy and Sample Entropy Parameters, and a Fast Algorithm for Their Calculation". *Physica A: Statistical Mechanics and its Applications* **24**, 6601–6610. DOI: 10.1016/j.physa.2012.07.003 (*see p. 26*).