On the Statistical Dependence – from Nonlinearity to Spatial Structural Estimation

by Yi Li

B.S. in Mathematics, Fudan University
M.S. in Mathematics, Northeastern University

A dissertation submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

June 26, 2018

Dissertation directed by

Adam(Aidong) Ding
Professor of Mathematics
Acknowledgments

First and foremost, I would like to thank my PhD advisor Professor Adam Aidong, Ding, for his uncountable guidance and support over the five years of my PhD study. This work would never have been done without his help. I feel very lucky to have him as my advisor.

I thank Professor Jennifer Dy, who gave me tremendous help and suggestions on my academic study, scientific research and presentation skills. I thank Professor Mikhail Malioutov, for teaching me many high level statistics courses, and providing me kind support for my job applications. I would also like to thank Professor Samuel S. Wu, Professor Dy, and Professor Malioutov for agreeing to be on my dissertation committee.

My sincere thanks also go to my fellow graduate students and program coordinators for continuous support. Moreover, I thank NSF for supporting my thesis work through the grant CCF-1442728.

Last but not least, I am deeply grateful to my parents and grandparents for their and support all the time, no matter I am doing good or bad. And thanks to my lovely daughter, Miranda, for always brightening my day.
Abstract of Dissertation

In this thesis, we discuss several dependence related problems in statistics, which covers three topics I worked on during my PhD study. In the first part, we introduce a new concept of robust-equitability for (nonlinear) dependence measures and identify a robust-equitable copula dependence measure (RCD). We also apply RCD in the application of feature ranking and selection. In the second part, we focus on the estimation of high dimensional spatial covariance matrix. Under the block bandable assumption which arises naturally from spatially correlated data, we propose a rate optimal double tapering estimator for estimating the spatial covariance matrix and demonstrate its advantages over the sample covariance matrix. The third part generalizes the former one in the way that it considers the dependence from both nearby and remote distance grid points, and propose a far-near covariance model which could potentially be used to model teleconnection effect in climate research.
# Table of Contents

Acknowledgments ........................................... 2

Abstract of Dissertation .................................. 3

Table of Contents ........................................ 4

List of Tables ........................................... 6

List of Figures ............................................ 7

Chapter 1 Introduction .................................... 13
  1.1 The Dependence in Statistics ......................... 13
  1.2 Outline and Summary ................................ 15

Chapter 2 The Robust Copula Dependence .................. 18
  2.1 Introduction .......................................... 18
  2.2 A Robust-Equitable Dependence Measure .............. 23
  2.3 Statistical Estimation Errors ......................... 36
  2.4 Experimental Results ................................ 41
  2.5 Conclusions and Discussions ......................... 62

Chapter 3 Estimation of High Dimensional Spatial Covariance Matrices ............. 64
  3.1 Introduction .......................................... 64
  3.2 Block Bandable Spatial Covariance Estimation ....... 66
  3.3 Statistical Error ...................................... 70
  3.4 Experimental Results ................................ 74
Chapter 4  Far-Near Covariance Model

4.1  Introduction ................................................................. 82
4.2  Methodology ................................................................. 84
4.3  Numerical Examples ....................................................... 99
4.4  Conclusion ................................................................. 105

Appendix A  Additional Proofs and Results for Chapter 2

A.1  Proof of Theorem 8 ......................................................... 107
A.2  Proof of Theorem 9 ......................................................... 111
A.3  Proof of Theorem 10 ....................................................... 112
A.4  Proof of Theorem 11 ....................................................... 114
A.5  Selection of Tuning Parameter in the Practical Estimator ......... 118

Appendix B  Additional Proofs and Results for Chapter 3

B.1  Proof of Theorem 13 ....................................................... 128
B.2  Proof of Lemma 16 ......................................................... 130

Appendix C  Additional Proofs and Results for Chapter 4

C.1  Proof of Theorem 20 ....................................................... 131
C.2  Computation of Variance $\tilde{\Sigma}$ of Sample Covariance Elements ....................................................... 132
C.3  Proof of Corollary 21 ....................................................... 133
C.4  Expressions of entries in (4.29) ......................................... 139
C.5  Proof of Theorem 24 ....................................................... 140
C.6  Proof of Theorem 27 ....................................................... 141

Bibliography ................................................................. 143
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Pearson correlation coefficient on three function relationships.</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>The expected values of RCD estimates (with standard deviation in parenthesis) based on 100 simulations, under various functional types, sample sizes, noise types and noise levels.</td>
<td>44</td>
</tr>
<tr>
<td>2.3</td>
<td>The expected values of $\rho$ estimates (with standard deviation in parenthesis) based on 100 simulations.</td>
<td>45</td>
</tr>
<tr>
<td>2.4</td>
<td>The expected values of $CD^2$ estimates (with standard deviation in parenthesis) based on 100 simulations.</td>
<td>45</td>
</tr>
<tr>
<td>2.5</td>
<td>Dependence measure values for synthetic data. Sample sizes $n = 300$ and $n = 10,000$ are considered. Each row corresponds to one type of measure. The MSE is presented in the last row.</td>
<td>50</td>
</tr>
<tr>
<td>2.6</td>
<td>Dependence measure for three features in metabolic reaction network dataset.</td>
<td>52</td>
</tr>
<tr>
<td>2.7</td>
<td>Features that is used in the additive regression model.</td>
<td>53</td>
</tr>
<tr>
<td>2.8</td>
<td>The most frequently selected top five features, with the relative frequency in parenthesis. Asterisk indicates one of the random noise features ($X_9$, ..., $X_{20}$).</td>
<td>55</td>
</tr>
<tr>
<td>2.9</td>
<td>Median of the order of selection for the first eight features in each dependence measure experiment among fifty repeated runs.</td>
<td>56</td>
</tr>
<tr>
<td>2.10</td>
<td>Measures ranked by predictive $MSE$.</td>
<td>58</td>
</tr>
<tr>
<td>A.1</td>
<td>The function relationships used in Figures A.2 - A.7.</td>
<td>119</td>
</tr>
<tr>
<td>A.2</td>
<td>The function relationships used in Figures A.8.</td>
<td>120</td>
</tr>
</tbody>
</table>
## List of Figures

2.1 Left: Bivariate Gaussian with $\rho = 0.75$. Middle: Data with exponential marginal for $X$. Right: The Gaussian copula. The first two distributions have the same copula as in the right figure. ................................. 26

2.2 Left: Additive noise for self-equitability. Right: Mixture noise for robust-equitability. ................................. 29

2.3 Hidden in background noise are 10% (red colored) data on a deterministic curve in (a) and (b), on a narrow strip around the line in (c). ................................. 36

2.4 Seven dependence measure values versus signal levels. ................................. 47

2.5 Six function types, circle, parabola, sine/cosine, twisted curve, two cross lines, and a spiral curve, with 20% uniform noise from 0 to 0.9 (with signal portion from 0.1 to 1). The examples in this figure is of signal level 0.8. ................................. 48

2.6 $(X_1, Y)$ has a nonmonotonic relationship with deterministic signal. $(X_2, Y)$ has linear relationship with uniform additive noise with width $d = 0.2$. $(X_3, Y)$ has background noise with a 0.75 linear signal portion. $(X_4, Y)$ is similar to $(X_3, Y)$ but with a 0.5 signal portion. ................................. 49

2.7 $(X_i, Y), i = 1, 2$ in stock dataset ................................. 51

2.8 Measures for $X_1, X_2$ in stock dataset ................................. 51

2.9 Scatter plots of the features $X_1, \cdots, X_8$ versus $Y$. ................................. 53

2.10 Cross-validated MSE plot with error bar based on 50 repeated runs. ................................. 54

2.11 MSE of mRMR-based Feature Selection on nine real-world datasets ................................. 57

2.12 Degree distribution of the networks based on the two dependence measure. ................................. 59
2.13 Local clustering of the networks based on the two dependence measure (Redder is higher value, and bluer is lower value). ........................................... 60
2.14 Betweenness of the networks based on the two dependence measure (Redder is higher value, and bluer is lower value). ........................................... 60
2.15 Communities of the networks based on the two dependence measure based on the hierarchical clustering algorithm. ........................................... 60
2.16 Visualization of the network structure that are constructed by the remotely highly correlated edges base on their spatial location on map. .............. 61
2.17 Interesting possible teleconnection that is uniquely detected by nonlinear dependence measure. ................................................................. 62

3.1 Left: Example of the spatial covariance matrix. Right: The two-dimensional stations are indexed in a raster scan order to build a one dimensional long vector. ................................................................. 68
3.2 The tapering function $\omega(x)$ which is used in the double tapering estimator. 69
3.3 Relative error (w.r.t sample covariance matrix) for the banding and linear tapering estimator. The comparison is against the decay rate $a$. .............. 76
3.4 Relative error (w.r.t. sample covariance matrix) for the banding and linear tapering estimators. The comparison is on the left: against the dimension $p$; on the right: against the sample size $n$. ......................................................... 77
3.5 Examples of the face data set. The three groups (four images each) from left to right: original images, the reconstructed images by keeping $q = 1000$ principal components based on the sample covariance matrix and based on the double tapered covariance matrix. ......................................................... 78
3.6 Reconstruction error on the testing set as a function of the number of principal component kept. The double tapered covariance matrix leads to lower reconstruction error compared with the sample covariance matrix. .............. 79
3.7 Example of the statistical downscaling task. Left: the climate variable (monthly mean temperature in Kelvin) with a coarse resolution at a single time point (Jan. 2009). Right: the target climate variable with a relatively finer resolution. The goal of statistical downscaling is to learn the target climate variable in finer resolution based on the historical data of the fine and coarse resolution climate data (red is larger and black is smaller).

3.8 Downscaling result of the target climate variable in the testing set. The intensity (red is larger and black is smaller) is the average MSE in each grid point.

3.9 Testing MSE across four seasons: DJF (December, January, February), MAM (March, April, May), JJA (June, July, August), SON (September, October, November). The overall MSE is also presented.

4.1 Example of the adjacent sample covariance elements that are combined linearly to estimate \( \sigma^{\star}_{i\cdot,j\cdot} \), where \( \bar{\sigma} \) is \( \sigma^{\star}_{i\cdot,j\cdot} \) in the middle, and the other four are defined clockwise starting from the top.

4.2 Left: Example of the optimal \( \lambda^* \) for estimating \( \rho \) in the model (4.6). Right: Example of the variance reduction ratio \( \varrho_{rr} \). The color of the lines corresponds to \( \sigma_{e} = 1(\text{red}), 2(\text{green}), 3(\text{blue}) \).

4.3 An example of one-dimensional indexed structures.

4.4 An example of two-dimensional indexed structures.

4.5 Example of the adjacent sample covariance elements that are combined linearly to estimate \( \sigma_{i\cdot,j\cdot} \) in the two dimensional model.

4.6 Comparison of \( \hat{\sigma}_{i\cdot,j\cdot} \) (estimator 1 in red) and \( \hat{\sigma}_{i\cdot,j\cdot}^* \) (estimator 2 in blue). The left panel corresponds to \( a = 0.6 \), while the right panel corresponds to \( a = 0.8 \). The remote signal strength \( \rho \) in x-axis varies from 0.5 to 0.9.
4.7 Empirical power curves against sample size (left column) and signal strength (right column). The near covariance strength $a$ is 0.6, 0.8 with different thickness. The blue curve is our proposed test, and the orange one is test based on the sample covariance matrix. The first row is $p = 100$, the second row is $p = 200$, and the third row is $p = 500$.

4.8 Monthly mean precipitation data.

4.9 Example of the teleconnection detected by sample covariance $\tilde{\sigma}_{i^*,j^*} = 0.58$ (top left). The other three are the scatter plots of the surrounding pairs.

4.10 Monthly mean precipitation data.

4.11 Example of the top teleconnection detected by the proposed far covariance estimation $\bar{\sigma}_{i^*,j^*}$ (top left). The other three are the scatter plots of the surrounding pairs.

A.1 The plot shows the regions $R_1$, $R_2$ and $R_3$. The other two narrow strips neighboring $R_1$ and $R_2$ are for the continuity correction mentioned at the end of the proof.

A.2 The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.25n^{-1/4}$.

A.3 The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.1n^{-1/4}$.

A.4 The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.5n^{-1/4}$.

A.5 Additive noise with $k = 0.25\sqrt{n}$.

A.6 Additive noise with $k = 0.1\sqrt{n}$.
A.7 Additive noise with $k = 0.5\sqrt{n}$. ........................................ 126
A.8 Mixture noise with $k = c\sqrt{n}$, where $c = 0.1, 0.25, 0.5$. .................... 127
Chapter 1

Introduction

The concept of association and dependence plays a significant role in statistics and other scientific fields. The main aim of this dissertation is to discuss this from two perspectives, i.e. nonlinearity and spatial structural estimation. Most of the results and works are related to the application in climate data analysis and accomplished under the support of the CyberSEES \(^1\) project during my PhD study.

1.1 The Dependence in Statistics

The knowledge of the dependence between two or more random variables is crucial in the statistical world. It lays the foundation of many other statistics-related methodologies and applications. This is true because the data that we are analyzing is usually multivariate and inter-related. Having a command of such information will often help us to make further decision.

There have been many research works addressing this topic from various aspects, including the Pearson’s correlation coefficient, which is commonly used in many circumstances, among others. However, with the development and explosion of the form and size of the

\(^1\)https://www.nsf.gov/awardsearch/showAward?AWD_ID=1442728
data nowadays, it is of great importance to re-think this in accord with the characteristics of the data. Certainly, there are many issues can be addressed. Here, we will focus on two particular problems that are related to complex spatial climate data.

The first is the nonlinear dependence measure. This arises naturally in real world problems when the form of the data becomes complex and the linearity assumption is violated in practice. The meaning of nonlinear relationship is broad, ranging from simple V-shaped quadratic relation to the “butterfly effect”-related chaotic series. Unfortunately, even the simple quadratic relationship could not be dealt with by the most commonly used Pearson’s correlation coefficient. Thus, it is of great interest if we could design a nonlinear dependence measure that can resolve this problem and help us detecting interesting (nonlinear) relationships and strong signals from the data. Such problem will be the main topic in Chapter 2.

The second is about how to incorporate spatial structural information in real application. As we know, the multivariate analogue of the Pearson’s correlation coefficient is the correlation matrix, or the covariance matrix for normalized data. Even though the sample covariance (correlation) matrix shares good properties in traditional multivariate analysis, the performance of it becomes problematic and questionable as the dimension of the data becomes higher in the “large p, small n” world, which is the basic setting in the recent high dimensional statistics scenario. Thus, it is vital to find a high dimensional remedy of the sample covariance matrix to fit the current characteristics of the data. In fact, it can be shown that by incorporating the nearby dependence among spatial grids, one can significantly improve the performance of the sample covariance matrix. This will be discussed in Chapter 3. Moreover, in some cases, as a generalization of the model proposed in Chapter 3, we can simultaneously consider the dependence from nearby points and possibly far away points. This will help us to build a more robust model this will be the topic in Chapter 4.
1.2 Outline and Summary

In this section, we outline the general ideas and contributions to each work included in this dissertation.

1.2.1 Feature Selection and RCD

In many applications, not all the features used to represent data samples are important. Often only a few features are relevant for the prediction task. The choice of dependence measures often affect the final result of many feature selection methods. To select features that have complex nonlinear relationships with the response variable, the dependence measure should be equitable, a concept proposed by Reshef et al. [2011]; that is, the dependence measure treats linear and nonlinear relationships equally. Recently, Kinney and Atwal [2014] gave a mathematical definition of self-equitability.

In this chapter, we introduce a new concept of robust-equitability and identify a robust-equitable copula dependence measure, the robust copula dependence (RCD) measure. RCD is based on the $L_1$-distance of the copula density from uniform and we show that it is equitable under both equitability definitions. We also prove theoretically that RCD is much easier to estimate than mutual information. Because of these theoretical properties, the RCD measure has the following advantages compared to existing dependence measures: it is robust to different relationship forms and robust to unequal sample sizes of different features. Experiments on both synthetic and real-world datasets confirm the theoretical analysis, and illustrate the advantage of using the dependence measure RCD for feature selection. Moreover, we also apply RCD in a climate study with a focus on building climate network with this nonlinear dependence measure. We show that some interesting nonlinear relationship are found by using RCD rather than the Pearson’s correlation.
1.2.2 The High Dimensional Covariance Matrix

Spatial covariance matrix estimation is of great significance in many applications in climatology, econometrics and many other fields with complex data structures involving spatial dependencies. High dimensionality brings new challenges to this problem, and no theoretical optimal estimator has been proved for the spatial high-dimensional covariance matrix. Over the past decade, the method of regularization has been introduced to high-dimensional covariance estimation for various structured matrices, to achieve rate optimal estimators.

In this chapter, we aim to bridge the gap in these two research areas. We use a structure of block bandable covariance matrices to incorporate spatial dependence information, and study rate optimal estimation of this type of structured high dimensional covariance matrices. A double tapering estimator is proposed, and is shown to achieve the asymptotic minimax error bound. Numerical studies on both synthetic and real data are conducted showing the improvement of the double tapering estimator over the sample covariance matrix estimator.

1.2.3 The Far-near Covariance Model

Modeling structured data is one of the central topics in the analysis of high dimensional data. In the previous work, we mainly assumes that the data are spatially correlated with the nearby grid points. However, under a distance measure, there are many data contains variable dependencies among both nearby and far away locations, which can be viewed as an extension of the previous model.

In this chapter, we aim to simultaneously model the short range and long range dependencies, and thus propose the far-near covariance model. Such model is able to utilize the short range dependence structure to improve the estimation and detection of the remote dependence signals. Statistical properties are provided for our proposed method. Numerical studies on both synthetic and real climate data are conducted showing the improvement of the far-near covariance model over the methods based on the standard sample covariance
matrix.
Chapter 2

The Robust Copula Dependence

2.1 Introduction

The performance of machine learning algorithms is dependent on the input features representing each data sample. Often not all of these features are useful: some may be irrelevant and some may be redundant. Feature selection is thus needed to help improve the performance of learning tasks. Moreover, feature selection can decrease the computational cost of algorithms, and provide domain experts with an increased understanding of which factors are important.

search and the learning algorithm into a single optimization problem formulation. Wrapper and embedded methods, contrary to filter methods, select features specific to the learning algorithm; thus, they are most likely to be more accurate than filter methods on a particular learning algorithm, but the features they choose may not be appropriate for other algorithms. Another limitation of wrapper methods is that they are computationally expensive because they need to train and test the learning algorithm for each feature subset candidate, which can be prohibitive when working with high-dimensional data.

Filter methods rely on measures based on intrinsic properties of the data. More specifically, they evaluate features based on some dependence measure criterion between features and the target variable and select the subset of features that optimizes this criterion. Let \( d \) be the number of original features. An exhaustive search, which involves \( 2^d \) possible feature subsets is computationally impractical. Thus, one commonly employs heuristic search strategies, such as greedy approaches (e.g., sequential forward/backward search [Pudil et al., 1994]). However, these strategies can lead to local optima. Random search methods, such as genetic algorithms, add some randomness to help escape from local optima. When the dimensionality is very high, one can only afford an individual search. Individual search methods [Guyon and Elisseeff, 2003, He et al., 2005] evaluate each feature individually according to a criterion and then select features, which either satisfies a condition or are top-ranked. The problem with individual search methods is that they ignore feature interaction and dependencies. To account for such interactions and dependencies, Yu and Liu [2004] selects relevant features individually and then add a separate redundancy removal step to account for linear correlation between features; Peng et al. [2005] suggests another way, by maximizing relevance and minimizing redundancy (mRMR) together.

In addition to search strategies, the performance of filter methods depend heavily on the choice of dependence measures. The ability to measure the dependence between random variables is a fundamental problem in statistics and machine learning. One of the simplest
and most common dependence measure is the Pearson correlation coefficient ($\rho_{lin}$). However, this measure only captures linear relationships. Another popular measure is mutual information (MI). MI can capture nonlinear dependencies but is difficult to estimate [Fernandes and Gloor, 2010, Reshef et al., 2011] (see Theorem 3 in Section 8). Kernel-based dependence measures [Gretton et al., 2005a, Fukumizu et al., 2007] (e.g., the Hilbert-Schmidt Independence Criterion (HSIC)) have been introduced as an alternative to MI which does not require explicitly learning joint distributions. However, HSIC depends on the choice of kernels. Hilbert-Schmidt Normalized Information Criterion (HSNIC), also known as normalized conditional cross-covariance operator (NOCCO) [Fukumizu et al., 2007], is kernel-free, meaning it does not depend on the choice of kernels in the limit of infinite data. Even though HSNIC is kernel-free, both HSIC and HSNIC’s values may vary when we use different scales. Poczos et al. [2012] applied Maximum Mean Discrepancy (MMD) after empirical copula transformation to make the kernel-based dependence measure invariant to strictly monotone transformation of the marginal variables. The Copula-MMD (CMMD) can also be written in HSIC formulation after empirical copula transformation. Similarly, Reddi and Póczos [2013] applied HSNIC after empirical copula transformation (CHSNIC). Other dependence measures can also be applied after empirical copula transformation, resulting in measures that are also invariant to strictly monotone transformations. However, they (e.g., CMMD and CHSNIC) may fail to treat non-monotonic relations equally.

Reshef et al. [2011] proposed the concept of equitability, which states that a dependence measure should give equal importance to all relations: linear and nonlinear. For example, we expect a fair dependence measure to treat a perfectly linear relationship and a perfectly sinusoid relationship equally. Kinney and Atwal [2014] mathematically defined equitability by proposing self-equitability – under a nonlinear regression model with additive noise, a dependence measure should be invariant to any deterministic transformation of the marginal variables, under a nonlinear regression model with additive noise (a formal definition is
provided in Definition 1, Section 2.2). A self-equitable dependence measure will treat all forms of relationships equally in the large data limit for the additive noise model. Kinney and Atwal [2014] proved that MI is self-equitable, and recommended its usage.

To choose among the many self-equitable dependence measures, we further propose a new robust-equitability concept such that the measure also treats all forms of relationships equally in the mixture noise model. That is, in a mixture distribution with $p$ proportion of deterministic signal hidden in continuous independent background noise, the measure should reflect the signal strength $p$. The mixture noise model reflects real applications where measurements (features) are often corrupted with noise. For example, sensor data maybe corrupted by noise from hardware and environmental factors. Reshef et al. [2011, 2015b] considered equitability for a statistic. Our robust-equitability, as well as Kinney and Atwal [2014]’s self-equitability, is defined on the population quantity instead. Particularly, in the mixture distribution above, we define a dependence measure as weakly-robust-equitable if it is a monotone transformation of the proportion $p$, and is robust-equitable if it equals to $p$ exactly.

In this chapter, we show that among a class of self-equitable copula-based dependence measures, only robust copula dependence (RCD), defined as the total variational distance (the half of the $L_1$ distance) between copula density and uniform (independence) density, is also weakly-robust-equitable (and robust-equitable). Without referring to the copula density, RCD can be equivalently stated as the total variational distance between the probability distribution and the (independent) product of its marginal distributions, and is equivalent to the Silvey’s Delta measure [Silvey, 1964]. In the literature, Silvey’s Delta (RCD) was only cited as an abstract benchmark. Here, we propose a $k$-nearest-neighbor (KNN)-based estimator for RCD and prove its consistency. Besides the $L_1$ distance RCD, we also investigated properties of the $L_2$ distance between copula density and the uniform density (we call CD$_2$). CD$_2$ is the theoretical value of HSNIC in the large data limit [Fukumizu et al., 2007].
In addition, the robust-equitability study in this chapter provides insights on the difficulty of estimating MI. Some authors studied the convergence of MI estimators by imposing the Hölder condition on the copula density. This Hölder condition, while being a standard condition for density estimations, does not hold for any commonly used copula [Omelka et al., 2009, Segers, 2012]. Under a more realistic Hölder condition on the bounded region of copula density, we provide a theoretical proof that the mutual information (MI)’s minimax risk is infinite. This provides a theoretical explanation on the statistical difficulty of estimating MI observed by practitioners [Fernandes and Gloor, 2010, Reshef et al., 2011]. Moreover, we prove that although both MI and CD$_2$ are self-equitable, they are not robust-equitable. Therefore, MI and CD$_2$ may not rank the features correctly by dependence strength in some cases, even in the large data limit. We confirm this phenomena on both synthetic and real-world datasets. In contrast, RCD is consistently estimable under the same condition. As for kernel-based dependence measures, HSIC and CMMD are neither self-equitable nor robust-equitable, HSNIC and CHSNIC are self-equitable but not robust-equitable and their estimators converge very slowly. Since RCD is the only measure that is both self-equitable and robust-equitable among these measures, it can be very useful for feature selection.

In summary, the contributions of this chapter are: (1) the introduction of the concept of robust-equitability and the discussions about the motivation and rationale of equitability definitions – we discuss the relationship of equitability to Renyi’s theorems, to more copula-based dependence measures and to independence tests; (2) the identification of RCD as a dependence measure that is both self- and robust-equitable and the proposal of a practical consistent estimator for RCD; (3) theoretically proving that non-robust-equitable measures MI and CD$_2$ cannot be consistently estimated and showing that this can lead to incorrect selection of features when sample size is large or when sample sizes are unequal for different features; we show that the difficulty of MI estimation is not due to the unboundedness of its definition, but is intrinsic due its being non-robust-equitable; (4) demonstrating that
the robust-equitable RCD is a better dependence measure for feature selection compared to existing dependence measures through experiments on synthetic and real-world datasets, in terms of robustness to function types, correctness in large sample size and correctness in unequal sample sizes. and finally, (5) the utilization of RCD in complex network study with the application in climate data.

The rest of this chapter is organized as follows. In Section 2.2, we motivate the equitability concepts, discussing different equitability definitions and relationship to copula and Renyi’s theorems. Particularly, we propose the concept of robust-equitability, and define a robust-equitable dependence measure called robust copula dependence. In Section 2.3, we prove MI and CD₂ are not consistently estimable. We also prove RCD can be consistently estimated and provide its estimators based on kernel density estimation (KDE) and k-nearest-neighbors (KNN). In Section 2.4, we provide feature selection experiments on synthetic and real datasets to demonstrate the advantage of RCD compared to existing dependence measures. We end with conclusions and discussions in Section 5.

2.2 A Robust-Equitable Dependence Measure

In this section, we investigate the theoretical properties of different dependence measures. In particular, we would like the dependence measures to have the following characteristics. We would like the dependence measures to rank a feature with less noise as having a stronger dependence with the response variable compared to features with more noise. We do not want measures that prefer a particular type of relationship (e.g., linear). Moreover, we do not want the measures to be too sensitive to sample size (i.e., when different features have unequal sample sizes, the dependence measure should not prefer a feature simply because it has more samples, but should still rank the features based on the strength of the deterministic signal compared to noise). Note that it is becoming more common for databases to have features that have unequal sample sizes due to the prevalence of data collection from heterogeneous
sources. For example, a clinical database may have more samples with clinical features compared to samples with genomic information. In such a setting, we would like to use all the data available to perform feature selection rather than to create equal sample sizes by throwing away samples from the larger set. In this chapter, we formalize these ideas through the recently proposed equitability concept: We want to use dependence measures that reflect the noise level, regardless of relationship type.

### 2.2.1 Self-equitability, Rényi’s Axioms and Copula-based Dependence Measures

Reshef et al. [2011] proposed that an equitable measure should “give similar scores to equally noisy relationships of different types.” Kinney and Atwal [2014] mathematically defined self-equitability through invariance under all nonlinear relationships in the regression model

$$Y = f(X) + \epsilon,$$

where $f$ is a deterministic function, $\epsilon$ is the random noise variable whose distribution may depend on $f(X)$ as long as $\epsilon$ has no additional dependence on $X$.

**Definition 1.** A dependence measure $D[X; Y]$ is self-equitable if and only if $D[X; Y] = D[f(X); Y]$ whenever $f$ is the function in model (2.1).


To understand self-equitability better, we notice that Definition 1 is very similar to Rényi’s Sixth Axiom A6, both are defined through the invariance of the dependence measure under transformations. Rényi [1959] proposed seven axioms for dependence measures $D[X; Y]$. (A1) $D[X; Y]$ is defined for any random variables $X$ and $Y$; (A2) symmetric $D[X; Y] = D[Y; X]$; (A3) $0 \leq D[X; Y] \leq 1$; (A4) $D(X; Y) = 0$ if and only if $X$ and $Y$ are statistically independent; (A5) $D(X; Y) = 1$ if either $X = f(Y)$ or $Y = g(X)$ for some
Borel-measurable functions \( f \) and \( g \); \((A6)\) If \( f \) and \( g \) are Borel-measurable, one-one mappings of the real line into itself then \( D[f(X); g(Y)] = D[X; Y] \); \((A7)\) If the joint distribution of \( X \) and \( Y \) is bivariate Gaussian, with linear correlation coefficient \( \rho \), then \( D[X; Y] = |\rho| \).

For a symmetric dependence measure (satisfying Axiom A2), Axiom A6 can be rewritten as \( D[f(X); Y] = D[X; Y] \) for all Borel-measurable \( f \). Hence self-equitability is a weaker version requiring an extra assumption that \( f \) satisfies the model (2.1).

It is known that Rényi’s maximum correlation coefficient is the only measure that satisfies all seven Rényi’s Axioms. However, Rényi’s maximum correlation coefficient has a number of major drawbacks, e.g., it equals 1 too often and is generally not effectively estimable [Schweizer and Wolff, 1981, Székely and Rizzo, 2009]. Hence enforcing all seven axioms is often considered too strong a constraint on the dependence measure, while some axioms are often considered desirable. For example, HSIC is shown to satisfy the first four axioms by [Gretton et al., 2005a] and [Gretton et al., 2005b].

For comparison, another weakened version of Axiom A6 is to restrict the transformations to monotone functions [Schweizer and Wolff, 1981], but without imposing the regression model (2.1). We may call this version of Axiom (A6*) weak-equitability.

**Definition 2.** A dependence measure \( D[X; Y] \) is weak-equitable if and only if \( D[X; Y] = D[f(X); Y] \) whenever \( f \) is a strictly monotone continuous deterministic function.

The weak-equitable dependence measures treat all monotone (but not all nonlinear) relationships equally, and is a property shared by all copula-based dependence measures (e.g., CMMD and CHSNNIC). Sklar’s theorem ensures that, for any joint distribution function \( F_{X,Y}(x, y) = \Pr(X \leq x, Y \leq y) \), there exists a copula \( C \) – a probability distribution on the unit square \( \mathcal{I}^2 = [0, 1] \times [0, 1] \) – such that

\[
F_{X,Y}(x, y) = C[F_X(x), F_Y(y)] \quad \text{for all } x, y.
\] (2.2)

Here \( F_X(x) = \Pr(X \leq x) \) and \( F_Y(y) = \Pr(Y \leq y) \) are the marginal cumulative distribution functions (CDFs) of \( X \) and \( Y \) respectively. In other words, the copula \( C \) is the joint CDF of
the two copula-transformed, uniformly distributed, variables \( U = F_X(X) \) and \( V = F_Y(Y) \). In this way, the copula decomposition separates the dependence from any marginal effects, and the copula \( C \) captures all the dependence between \( X \) and \( Y \). Figure 2.1 shows the data from two distributions with different marginals but the same dependence structure.

Figure 2.1: Left: Bivariate Gaussian with \( \rho = 0.75 \). Middle: Data with exponential marginal for \( X \). Right: The Gaussian copula. The first two distributions have the same copula as in the right figure.

Table 2.1 shows three simple examples and their respective copula transformations. We can see that the linear correlation \( \rho_{lin} \) prefers the linear relationship in (A). Applying on the copula-transformed variables on the right half of Table 2.1, \( \rho_{lin} \) (now equivalent to Spearman’s \( \rho \)) becomes invariant to monotone transformation in (B), but still cannot capture the non-monotone nonlinear relationship in (C).

<table>
<thead>
<tr>
<th>raw data scale</th>
<th>copula transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

\[ \rho_{lin} = 1 \quad \rho_{lin} = 0.866 \quad \rho_{lin} = 0 \quad \rho_{lin} = 1 \quad \rho_{lin} = 1 \quad \rho_{lin} = 0 \]

Table 2.1: Pearson correlation coefficient on three function relationships.

While copula-based dependence measures treats the monotone functions equally, equitability aims to also treat non-monotone functions equally. However, the original Rényi’s Axiom A6 may be overly strong, and self-equitability aims to treat non-monotone functions equally only under the regression model (2.1).
We first consider some self-equitable copula-based dependence measures and further choose among these measures based on a new equitability definition in the next subsection 2.2.2. Mutual information (MI), the recommended measure in Kinney and Atwal [2014], is self-equitable and is based on copula density $c(u, v)$,

$$\text{MI} = \int_{I^2} \log[c(u, v)]c(u, v)dudv, \quad (2.3)$$

where $I^2$ is the unit square. We now consider a large class of self-equitable copula-based measures. Since the marginal variables $X, Y$ are independent if and only if the corresponding copula distribution is uniform, we measure the dependence between $X, Y$ through the distance between their copula distribution and the uniform distribution. Let the Copula Distance $CD_\alpha$ be the $L_\alpha$ distance between a copula density and the uniform copula density $\pi(u, v) = 1$.

$$CD_\alpha = \int_{I^2} |c(u, v) - 1|^\alpha dudv, \quad \alpha > 0. \quad (2.4)$$

Combining Eq.4 in Fukumizu et al. [2007] and Eq.(2.4) here, $CD_2$ is the theoretical value of HSNIC in the large data limit. Our first result is that, the Copula Distance is self-equitable when $\alpha \geq 1$.

**Lemma 3.** The Copula-Distance $CD_\alpha$ with $\alpha \geq 1$ is self-equitable.

The proof follows from Theorems S3 and S4 of Kinney and Atwal [2014], since $g(x) = |x - 1|^\alpha$ is convex when $\alpha \geq 1$.

**Remark:** Schweizer and Wolff [1981] studied a class of dependence measures that are the $L_\alpha$ distance between a joint copula $C(u, v)$ and the uniform copula $\Pi(u, v) = uv$. The $L_1, L_2$ and $L_\infty$ distance result in, the Wolf’s $\sigma$, Hoeffding’s $\Phi^2$ and Wolf’s $\kappa$ respectively. Schweizer and Wolff [1981] showed that these measures satisfies a modified set of Rényi’s Axioms, including Axiom (A6*) weak-equitability. In contrast to the Copula-Distance $CD_\alpha$ ($L_\alpha$ distance based on copula densities), these measures are based on the cumulative distribution functions and are not self-equitable. Since $C(u, v) = Pr(U \leq u, V \leq v)$ is the cumulative
distribution function, such measures cumulate the deviation from independence from \( u = 0 \) to \( u = 1 \), and do not remain invariant for all nonlinear transformations \( f \) in model (2.1).

### 2.2.2 Robust Equitability

To select among the many self-equitable dependence measures, we want to consider additional equitability conditions. Some self-equitable dependence measures may not perform well in practice. For example, Rényi’s maximum correlation coefficient (Rcor) satisfies the stronger Rényi’s Axiom A6, thus it is also self-equitable. \( Rcor(X; Y) = \sup_{f,g} \rho[f(X); g(Y)] \), where \( \rho \) is the linear correlation coefficient and the supremum is taken over all Borel-measurable functions \( f \) and \( g \). Rcor has a number of major drawbacks, e.g., it equals 1 too often and is generally not effectively estimable [Schweizer and Wolff, 1981, Székely and Rizzo, 2009].

We observe the deficiencies more clearly in another self-equitable measure, the ideal dependence coefficient (IDC): \( IDC(X; Y) = 0 \) if \( X \) and \( Y \) are independent and \( IDC(X; Y) = 1 \) otherwise. IDC satisfies the first six Rényi’s Axioms, and is self-equitable. It equals one for all dependent \( X \) and \( Y \), providing no distinction of the dependence strength. IDC is only an abstract measure, the estimation of IDC is equivalent to testing independence between \( X \) and \( Y \). However, for feature selection, IDC is not helpful at all because it provides no distinction of the dependence strength for different features. It is also hard to estimate. We wish for a new equitability criterion to exclude trivial dependence measures like IDC.

The self-equitability definition focuses on the regression \( Y = f(X) + \epsilon \). However, in practice, this additive noise model does not capture all data types. In some cases, for example in sensor measurements, the deterministic signal is hidden in continuous background noise. Figure 2.2 illustrates these two types of noise. The Left subfigure shows additive noise on a deterministic sinusoidal function. The Right subfigure is the same deterministic signal on a uniform background noise. Mathematically, after the copula transformation, the second mixture noise model is described by a mixture copula: a continuous copula on the unit square.
Figure 2.2: Left: Additive noise for self-equitability. Right: Mixture noise for robust-equitability.

$I^2$ is added to a deterministic signal $C_s$, which is a singular copula. Any copula can always be separated into a singular component and an absolutely continuous component [Nelsen, 2006, page 27]. Independent background noise is represented by taking the absolutely continuous component as the independence copula $\Pi(u,v) = uv$ on $I^2$. Therefore, with $p$ proportion of hidden deterministic relationship, the copula $C = pC_s + (1-p)\Pi$. Here $C_s$ is a singular copula representing the deterministic relationship, so that its support $S$ has Lebesgue measure zero.

The equitability in this mixture noise model means that the dependence measure should give the same value for all types of deterministic signal $C_s$.

**Definition 4.** A dependence measure $D[X;Y]$ is robust-equitable if and only if $D[X;Y] = p$ whenever $(X,Y)$ follows a distribution whose copula is $C = pC_s + (1-p)\Pi$, for a singular copula $C_s$.

Among the self-equitable Copula-Distances ($\alpha \geq 1$), $L_1$ distance is the special case that does reflect the proportion of deterministic relationship in the mixture copula.

$$CD_1 = p \int_S C(du, dv) + \int_{I^2 \setminus S} |(1-p) - 1|dudv = p(1) + p = 2p.$$ 

Therefore we define the scaled version of $CD_1$ as robust copula dependence (RCD)

$$RCD = \frac{1}{2} CD_1 = \frac{1}{2} \int_{I^2} |c(u,v) - 1|dudv. \quad (2.5)$$

**Lemma 5.** The robust copula dependence $RCD$ is robust-equitable.
Mathematically, RCD is the same as Silvey’s Delta [Silvey, 1964]:

\[ \Delta = \int_{\phi > 1} [p(x, y) - p_X(x)p_Y(y)]dxdy, \]

where \( p_X \) and \( p_Y \) are the marginal probability densities for \( X \) and \( Y \), \( p \) is the joint probability density for \( X \) and \( Y \), and \( \phi(x, y) = p(x, y)/[p_X(x)p_Y(y)] \). We write equation (2.5) in terms of the absolutely continuous copula density for ease of understanding. When part of the copula is singular, the RCD in (2.5) can be defined as in Silvey [1964], interpreting \( \phi \) as the Radon-Nikodym derivative of the joint distribution with respect to a dominating probability measure which does cover the possibility of singularity. Alternatively, for a mixture copula \( C \), the RCD can be defined as the limit of \( \lim_{m \to \infty} RCD(C_m) \) for equation (2.5) on any sequence of continuous copulas \( \{C_1, C_2, \ldots\} \) that converges to \( C \). The convergence means that \( \lim_{m \to \infty} \|C_m - C\|_1 := 2 \lim_{m \to \infty} \sup_A |C_m(A) - C(A)| = 0 \), where the supremum is taken over all Borel sets \( A \).

A second way of interpretation is helpful in thinking about why \( CD_\alpha \), when \( \alpha > 1 \), cannot be made robust-equitable and why this leads to statistical difficulties in estimation which we will discuss in detail in the next section.

Roughly speaking, for the mixture copula \( C = pC_s + (1 - p)\Pi \), the copula density for the absolutely continuous component is \( c_c(u, v) = 1 - p \), while we can imagine \( c_s(u, v) \) as an abstract copula density for the singular component such that \( \int_B c_s(u, v)dudv := \int_B C(du, dv) = C_s(B) \) for any subset \( B \subset S \). Since \( S \) has Lebesgue measure zero, \( c_s(u, v) = \infty \) for \( (u, v) \in S \) so that \( c_s \) is not a proper density, but rather an abstract limit of the sequence \( \lim_{m \to \infty} c_{m,s} \).

Here for any convergent sequence of continuous copulas \( \{C_1, C_2, \ldots\} \) above, \( c_{m,s}(u, v) = c_m(u, v) - c_c(u, v) \) is the continuous copula density that approaches the abstract \( c_s(u, v) \). For any open set \( B_\Omega \supset B \), \( C_s(B_\Omega) = \int_{B_\Omega} c_s(u, v)dudv := \lim_{m \to \infty} \int_{B_\Omega} c_{m,s}(u, v)dudv \), and \( C_s(B) = \lim_{B_\Omega \to B} C_s(B_\Omega) \). Hence for any \( \alpha > 1 \), \( \int_B [c_s(u, v)]^\alpha dudv = \int_S [c_s(u, v)]^\alpha - 1 c_s(u, v)dudv = \int_S \infty c_s(u, v)dudv = \infty \). So that \( CD_\alpha = \infty \) whenever \( \alpha > 1 \) and \( p > 0 \). Similarly, \( MI = \infty \) for all \( p > 0 \). They do not distinguish the dependence strength in the mixture distribution according to the signal proportion \( p \), and can not be transformed to be robust-equitable as
they over-emphasize the singular component (high copula density region).

If the dependence measure does not equal \( p \) exactly but is a monotone function of \( p \), then we can scale it to get a robust-equitable version, and call it weakly-robust-equitable.

**Definition 6.** A dependence measure \( D[X; Y] \) is weakly-robust-equitable if and only if \( D[X; Y] \) is a strictly monotone function of \( p \) whenever \( (X, Y) \) follows a distribution whose copula is \( C = pC_s + (1 - p)\Pi \), for a singular copula \( C_s \).

**Lemma 7.** The Copula-Distance \( CD_\alpha \) is weakly-robust-equitable if and only if \( \alpha \leq 1 \).

When \( \alpha > 1 \), since \( CD_\alpha = \infty \) whenever \( p > 0 \), those are not weakly-robust-equitable. When \( \alpha < 1 \), \( \int_S [c_s(u, v)]^\alpha dudv = \int_B [c_s(u, v)]^{\alpha - 1} c_s(u, v) dudv = 0 \) so that the contribution from the singular region \( S \) is zero. In these cases,

\[
CD_\alpha = 0 + \int_{\mathbb{I}^2 \setminus S} |(1 - p) - 1|^\alpha dudv = p^\alpha
\]

is weakly-robust-equitable. And \( CD_1 \) is weakly-robust-equitable from Lemma 5.

The self-equitability for additive noise model requires that \( \alpha \geq 1 \), while the weakly-robust-equitability in the mixture noise model requires that \( \alpha \leq 1 \). Hence only \( \alpha = 1 \) satisfies the equitability condition in both noise models, and resulting in the robust-equitable \( RCD \).

**\( RCD \), other equitability definitions and Rényi’s Axiom**

Having introduced our measure \( RCD \) and the robust-equitability definition, we can further compare them to the other equitability definitions in literature. Reshef et al. [2011] considers equitability as the ability of a statistic \( \hat{D} \) to approximately reflect the nonlinear \( R^2 \) over different relationships. They proposed a statistic MIC and demonstrate numerically such equitability through simulated examples. Kinney and Atwal [2014] propose to formalize such concept for the population parameter \( D[X; Y] \) such that a \( R^2 \)-equitable measure equals \( g(R^2[f(X), Y]) \) in the additive noise model (2.1) \( Y = f(X) + \epsilon \), and showed that no non-trivial dependence measure can satisfy this \( R^2 \)-equitability. The self-equitability definition is
proposed as an alternative. Murrell et al. [2014] pointed out that such impossibility results are due to the non-identifiability due to the specification of $\epsilon$ term, allowing $\epsilon$ to possibly depend on $f(X)$. Under such specification, for example, a noiseless parabola can be realized as a noisy version of a noiseless linear relationship Murrell et al. [2014]. Reshef et al. [2015b] propose another formal equitability framework through interpretable intervals of a statistic under additive homoscedastic noises for both $X$ and $Y$.

Our robust-equitability definition shares some common characteristics with both Kinney and Atwal [2014]’s and Reshef et al. [2015b]’s approach respectively. Similar to Kinney and Atwal [2014], our robust-equitability focuses on the population quantity $D[X; Y]$ instead of a statistic $\hat{D}[X; Y]$. This allows proof of theoretical equitability properties for specific dependence measures, as the statistical estimation error $\hat{D} - D$ can be kept as a separate issue. Robust-equitability does have implications on the statistical estimation error bounds which will be discussed in the next section. Self-equitability and robust-equitability focus on the invariance of $D[X; Y]$ as in Rényi’s Axiom A6, but for different noise models. Reshef et al. [2015b] and our robust-equitability definition each focuses on a noise model to avoid the non-identifiability issue in Kinney and Atwal [2014]’s model: additive homoscedastic noise and mixture uniform noise respectively. Under each model, there is a clearly identifiable quantity of interest: the nonlinear $R^2$ and the mixture proportion $p$ respectively.

Furthermore, our RCD satisfies the first five Rényi’s Axioms. Particularly, $RCD = 0$ if and only if $X$ and $Y$ are statistically independent, $RCD = 1$ if $X$ and $Y$ are related through a deterministic relationship. And $RCD$ is symmetric in that $RCD[X; Y] = RCD[Y; X]$. Notice that this symmetric property is an appropriate requirement for feature selection with the filtering method mRMR Peng et al. [2005]. Given a dataset $X \in \mathbb{R}^{n \times d}$, where $n$ is the number of samples, $d$ is the number of features, and target variable $Y \in \mathbb{R}^{n \times 1}$, let $X_i$ denote the $i$-th feature, mRMR finds from the $d$-dimensional feature space, $\mathbb{R}^d$, a subspace of $m$
features that optimally characterize $Y$, by solving the following optimization problem:

$$\max_S \frac{1}{|S|} \sum_{X_i \in S} D[X_i; Y] - \frac{1}{|S|^2} \sum_{X_i, X_j \in S} D[X_i; X_j],$$

where $S$ is the optimal feature set, the first term maximizes relevance and the second term minimizes redundancy. Notice that this method only requires bivariate dependence between different features $D[X_i; X_j]$ in addition to the bivariate dependence between each feature and response variable $D[X_i; Y]$. The feature selection results of mRMR with an asymmetric dependence measure would depend on how the features are ordered, which is an undesirable characteristic.

RCD satisfies self-equitability, which is a weakened version of the sixth Rényi’s Axiom. The Rényi’s Axiom A7 requires the dependence measure to agree with the natural quantity of $|\rho|$ for bivariate Gaussian distributions (which corresponds to a linear regression model). Our RCD does not satisfy that, but instead agree with $|\rho|$ for the mixture noise setting with a linear deterministic relationship, since in that case $p = |\rho|$. Our robust-equitability definition requires the measure equals $p$ exactly, which provides an easy interpretation in that it is an equitable extension of Pearson’s correlation $|\rho|$ to all forms of hidden nonlinear deterministic relationships. It is not essential to require the exact equality to $p$, as equaling to a monotone function of $p$ (weakly-robust-equitability) would enable a robust-equitable version of the measure through a transformation. However, precisely equaling to $p$ is nice due to the above easier interpretation. Notice that $R^2 = \rho^2$ for the additive noise regression model with a linear relationship (bivariate Gaussian distribution). Hence the nonlinear $R^2$ can be similarly considered as an equitable extension of Pearson’s correlation in the additive noise model to all forms of nonlinear regression relationships. However, unlike $p$, $R^2$ does not satisfy the symmetric property since regressing $Y$ on $X$ and regressing $X$ on $Y$ do not give the same value.

For discrete random variables, equation (2.5) corresponds to the Kolmogorov dependence measure in the pattern recognition literature [Vilmansen, 1972, 1973, Ekdahl and Koski,
2006] and also known as the Mortara dependence index [Bagnato et al., 2013]. In the discrete case, the measure has a maximum value less than 1. In contrast, \( RCD = 1 \) when \( X \) and \( Y \) are deterministically related. For continuous random variables \( X \) and \( Y \), Silvey’s Delta has been cited only as an abstract concept and no practical estimator was used in the literature for data analysis. The new name, \textit{Robust Copula Dependence} (RCD), emphasizes the fact that it is a robust-equitable copula-based dependence measure.

\section{2.2.3 Testing Independence Versus Estimation Errors of Dependence Measures}

In practice, feature selection is based on an estimator \( \hat{D}(X; Y) \) on the data set, since the exact value of the dependence measure \( D(X; Y) \) is unknown to the user. Hence the feature selection results are affected by the estimation error \( \hat{D} - D \). The estimation error also needs to be studied.

An estimator \( \hat{D}[X; Y] \) is often used to test the independence between \( X \) and \( Y \). Some studies compare different dependence measures \( D[X; Y] \)s by the power of independence testing using their corresponding estimators \( \hat{D}[X; Y] \)s Reshef et al. [2011], Simon and Tibshirani [2011], Reshef et al. [2015a]. However, while independence testing is related to the estimation of dependence measures \( D[X; Y] \), the power of the independence test is not a proper way of comparing dependence measures \( D[X; Y] \). In fact, as mentioned in Section 2.2.2, the independence test corresponds to an estimation of the trivial measure IDC: \( IDC(X; Y) = 0 \) if \( X \) and \( Y \) are independent and \( IDC(X; Y) = 1 \) otherwise. Thus the power comparison is comparing the performance of \( \hat{D}[X; Y] \) estimating \( IDC(X; Y) \) rather than estimating its corresponding parameter \( D[X; Y] \). Also, good independence test statistics may not have corresponding interpretable dependence measures [Sun and Zhao, 2014].

As Reshef et al. [2015b] pointed out, the estimator \( \hat{D}[X; Y] \) for equitable \( D[X; Y] \) is most powerful at testing the hypothesis if the signal strength exceeding a threshold \( D[X; Y] \geq D_0 \),
rather than being most powerful at testing independence $D[X; Y] = 0$. Besides simply testing for independence, dependence measures serve another important purpose: ranking the strength of the dependence relationships. For example, in the World Health Organization (WHO) data set in Reshef et al. [2011] the vast majority of the hundreds of the variables show dependence with other variables. So the independence tests do not provide much information there. To achieve sparse representation (feature selection), it is important to pick out the strongest dependence relationships. Equitable dependence measure ensures that the signal strength is reflected, rather than the functional form.

The robust-equitability definitions also have implications on the estimation errors. Some self-equitable measures can equal one too often, the extreme case being the IDC. When the dependence measure $D[X; Y]$ equals one for too many type of distributions, it does not distinguish dependence strength among them, and also makes its estimation difficult. Robust-equitability excludes such measures. In the next section, we theoretically show that robust-equitable RCD is intrinsically much easier to estimate than non-robust-equitable (but only self-equitable) MI and $CD_2$. This is due to the instability in the theoretical values of MI and $CD_2$. The following examples illustrate the difference between self-equitability (Figure 2.3: a versus b) and robust-equitability (Figure 2.3: a versus c). The self-equitable measures (RCD, MI), unlike $\rho$, have the same value in Figures 2.3(a) and 2.3(b) where each has 10% deterministic data on two different curves. In Figure 2.3(c), 10% of the data fall around (rather than exactly on) the line in a very small strip of area $0.1/exp(10) = 4.5 \times 10^{-6}$, which is very close to the distribution in Figure 2.3(a). The robust-equitable RCD values are very similar (differ only in $10^{-6}$ order) in Figures 2.3(a) and 2.3(c), but the MI values changes from $\infty$ to 1. Since these two almost indistinguishable distributions result in very different theoretical MI values ($\infty$ and 1), no estimator can do well.

In Section 2.3, we study the estimation errors theoretically, and provide a practical estimator for RCD.
Figure 2.3: Hidden in background noise are 10% (red colored) data on a deterministic curve in (a) and (b), on a narrow strip around the line in (c).

2.3 Statistical Estimation Errors

We analyze the statistical estimation error $\hat{D} - D$ in this section.

2.3.1 The inconsistency results on estimation of mutual information and $CD_2$

We theoretically show that MI and $CD_2$ are much harder to estimate compared to the robust-equitable RCD. We formally quantify the estimation difficulty through the minimax convergence rate over a family $\mathcal{C}$.

Denote $\mathbf{z} = (u, v)$. Let $\mathcal{C}$ be the family of continuous copulas with the density satisfying the following Hölder condition on the region where $c(\mathbf{z})$ is bounded above by some constant $M > 1$, denoted as $A_M$:

$$|c(\mathbf{z}_1) - c(\mathbf{z}_2)| \leq M_1 \|\mathbf{z}_1 - \mathbf{z}_2\|_1,$$

(2.7) for a constant $M_1$ and for all $\mathbf{z}_1 \in A_M$, $\mathbf{z}_2 \in A_M$, and $\|\cdot\|_1$ denotes the $l_1$ norm.

The estimation of MI has been studied extensively in the literature. Over all distributions, even discrete ones, no uniform rate of convergence is possible for MI [Antos and Kontoyiannis, 2001, Paninski, 2003]. On the other hand, many estimators were shown to converge to MI for
every distribution. These results are not contradictory, but rather common phenomena for many parameters. The first result is about the uniform convergence over all distributions, while the second result is about the pointwise convergence for each distribution. The first restriction is too strong, while the second restriction is too weak. The difficulty of estimating a parameter needs to be studied for uniform convergence over a properly chosen family.

As MI is defined through the copula density, it is natural to consider the families generally used in density estimation literature. Starting from Farrell [1972], it is standard to study the minimax rate of convergence for density estimation over the class of functions whose $m$-th derivatives satisfy the Hölder condition. Since the minimax convergence rate usually is achieved by the kernel density estimator, it is also the optimal convergence rate of density estimation under those Hölder classes. Generally, with the Hölder condition imposed on the $m$-th derivatives, the optimal rate of convergence for two-dimensional kernel density estimator is $n^{-(m+1)/(2m+4)}$ [Silverman, 1986, Scott, 1992].

Therefore, when studying the convergence of MI estimators, it is very tempting to impose the Hölder condition on the copula density. In fact, imposing the Hölder condition (2.7) on the whole $I^2$, Liu et al. [2012] showed that the kernel density estimation (KDE) based MI estimator converges at the parametric rate of $n^{-1/2}$. Pál et al. [2010] also considered similar Hölder condition when they studied the convergence of $k$-nearest-neighbor (KNN) based MI estimator. However, such a condition is usually too strong for copula density, thus these results could not fully reflect the true difficulty of MI estimation. When $c(u, v)$ is unbounded, the Hölder condition (2.7) can not hold for the region where $c(u, v)$ is big. Hence imposing this Hölder condition (2.7) on the whole $I^2$ would exclude many commonly used continuous copula densities (e.g., Gaussian, student-T etc.) since their densities are unbounded [Omelka et al., 2009, Segers, 2012]. Therefore, we impose it only on the region where the copula density is small. Specifically, we assume that the Hölder condition holds only on the region $A_M = \{(u, v) : c(u, v) < M\}$ for a constant $M > 1$. Then this condition
is satisfied by all common continuous copulas in the book by Nelsen [2006]. For example, all Gaussian copulas satisfy the Hölder condition (2.7) on $A_M$ for some constants $M > 1$ and $M_1 > 0$. But no Gaussian copulas, except the independence copula $\Pi$, satisfy the Hölder condition (2.7) over the whole $I^2$.

**Theorem 8.** Let $\hat{MI}$ be any estimator of the mutual information $MI$ based on the observations $Z_1 = (U_1, V_1), \ldots, Z_n = (U_n, V_n)$ from a copula distribution $C \in \mathcal{C}$. And let $\hat{CD}_\alpha$ be any estimator of the $CD_\alpha$ in equation (2.4). Then

$$
\sup_{C \in \mathcal{C}} E[|\hat{MI}(C) - MI(C)|] = \infty, \text{ and}
$$

$$
\sup_{C \in \mathcal{C}} E[|\hat{CD}_\alpha(C) - CD_\alpha(C)|] = \infty, \text{ for any } \alpha > 1.
$$

The detailed proof is provided in Appendix A. This theorem states that MI and CD cannot be consistently estimated over the family $\mathcal{C}$. This result does not depend on the estimation method used, as it reflects the theoretical instability of these quantities. There are many estimators for MI: kernel density estimation (KDE) Moon et al. [1995], the $k$-nearest-neighbor (KNN) Kraskov et al. [2004], maximum likelihood estimation of density ratio [Suzuki et al., 2009]. However, practitioners are often frustrated by the unreliability of these estimation [Fernandes and Gloor, 2010, Reshef et al., 2011]. This theorem provides a theoretical explanation.

Notice that the inconsistency results over this family $\mathcal{C}$ is not due to the unboundedness of MI and CD. They can be transformed into correlation measures with values between 0 and 1 [Joe, 1989]: $MIcon = \sqrt{1 - e^{-2MI}}$ and $\phi_{cor} = \sqrt{CD_2/(1 + CD_2)}$. The $MIcon$ is known as the Linfoot correlation in the literature [Speed, 2011]. We use the name $MIcon$ to indicate it as the scaled version of MI. The next theorem showed that $MIcon$ cannot be consistently estimated over the family $\mathcal{C}$ also.

**Theorem 9.** Let $\hat{MIcon}$ be any estimator of $MIcon$ based on the observations $Z_1 = (U_1, V_1), \ldots, Z_n = (U_n, V_n)$ from a copula distribution $C \in \mathcal{C}$. Then

$$
\sup_{C \in \mathcal{C}} E[|\hat{MIcon}(C) - MIcon(C)|] \geq a_2 > 0,
$$

38
for a positive constant $a_2$.

The detailed proof is provided in Appendix B.

The estimation difficulty of these dependence measures is due to their lack of smoothness related to being not weakly-robust-equitable. Reshef et al. [2015a, Section 4] proved a similar lack of smoothness of MI and $M I c o r$, while their proposed statistic $M I C$ may be considered a smoothed version which shows equitable behavior under their framework. Our results are stronger in that: (a) our results establish the statistical difficulty of estimation via minimax rate, and (b) our results apply to a broader class of dependence measures.

### 2.3.2 The consistent estimation of RCD

The equitability definitions and error analysis above assume a bivariate dependence measure. In this section, we will state the estimation results for a general $d$-dimensional RCD in equation (2.5): $RCD = \frac{1}{2} \int_{\mathbb{R}^d} |c(z) - 1|dz$ for $d$-dimensional $z$. That is, for $z = (z_1, ..., z_d)$, the copula transformation changes each dimension to uniformly distributed variables $u_j = F_{u_j}(z_j)$, $j = 1, ..., d$. Then $\int |c(z) - 1|dz := \int ... \int |c(u_1, ..., u_d) - 1|du_1...du_d$. The robust-equitability definition can be easily changed to the $d$-dimensional mixture copula with a singular component and the $d$-dimensional uniform distribution. And, in the $d$-dimensional case, the calculation above equation (2.5) still holds so that RCD is robust-equitable. Notice that other equitability definitions such as self-equitability is only defined for the bivariate case, and the filtering feature selection method mRMR also uses only the bivariate dependence.

Mathematically, MI (and $M I c o r$) is unstable because it overweighs the region with large density $c(z)$ values. From equation (2.3), MI is the expectation of $\log[c(z)]$ under the true copula distribution $c(z)$. In contrast, RCD in (2.5) takes the expectation at the independence case $\Pi$ instead. Even if $c(z)$ cannot be consistently estimated in the region $A_M$, its error contribution to $\hat{RCD}$ can be bounded. The following theorem, which is proved in Appendix C, shows the result for the KDE estimator for RCD.
Theorem 10. Let the KDE estimator of the d-dimensional copula density based on observations \(Z_1, \ldots, Z_n\) be

\[
\hat{c}_{kde}(Z) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{Z_i - Z}{h}\right). \tag{2.10}
\]

We assume the following conditions:

- The bandwidth \(h \to 0\) and \(nh^d \to \infty\).
- The kernel \(K\) is non-negative and has a compact support in, \(B_0 = \{Z : \|Z\|_2 \leq 1\}\), the d-dimensional unit ball centered at 0.
- The kernel \(K\) is bounded. \(M_K = \max_{s \in B_0} K(s), \int_{B_0} K(s)ds = 1, \mu_2^2 = \int_{B_0} K^2(s)ds < \infty\).

Then the plugged-in estimator \(\hat{RCD} = RCD(\hat{c}_{kde})\) has a risk bound

\[
\sup_{C \in \mathcal{C}} E[|\hat{RCD} - RCD|] \leq M_1 h + \frac{\sqrt{2} \mu_2}{\sqrt{nh^d}} + O\left(\frac{1}{nh^d}\right). \tag{2.11}
\]

In addition to the KDE based RCD estimator, we can estimate RCD consistently by plugging in the KNN estimator [Loftsgaarden and Quesenberry, 1965] of the copula density: \(\hat{c}(Z) = k/n/A_{r(k,n)}\) using copula based observations \(Z_1, Z_2, \ldots, Z_n\). Here \(r(k,n)\) is the distance from (d-dimensional) \(Z\) to the \(k\)-th closest of \(Z_1, Z_2, \ldots, Z_n\), and \(A_r\) is the volume of the d-dimensional hyper-ball with radius \(r\). Then \(\hat{RCD} = RCD(\hat{c}) = \sum_{\hat{c}(Z_i) > 1} [1 - 1/\hat{c}(Z_i)]/n\).

Theorem 11. Assuming \(c\) in \(\mathcal{C}\) has bounded continuous second order derivative in \(A_M\), \(k \to \infty\) and \((k/n) \to 0\) when \(n \to \infty\). Then the plugged-in estimator \(\hat{RCD} = RCD(\hat{c})\) has a risk bound

\[
\sup_{C \in \mathcal{C}} E[|\hat{RCD} - RCD|] \leq \tilde{c}_1\left(\frac{k}{n\epsilon}\right)^2 + \frac{\tilde{c}_2}{\sqrt{k}} + 2\epsilon, \tag{2.12}
\]

for some finite constants \(\tilde{c}_1\) and \(\tilde{c}_2\), and \(\epsilon = \epsilon(n)\) is any sequence converging to 0 slower than \(k/n\).
Here, the extra technical assumption on the second order derivative allows a simpler proof (provided in Appendix D) by citing formulas in Mack and Rosenblatt [1979]. Without it, RCD can still be estimated consistently as in Theorem 10. The error bound (2.12) is minimized by \( \epsilon = (k/n)^{2/(d+2)} \) and \( k = \left( \frac{n}{4(d+6)} \right) \). Hence, in the bivariate \( (d = 2) \) case, we have \( k = O(\sqrt{n}) \). Simulations in Appendix E suggests a practical estimator with \( k = 0.25\sqrt{n} \).

When RCD is estimated well under a sample size, further increasing the sample size does not change its estimation value much. In contrast, the estimated MI and CD\(_2\) values can continuously change by a large margin as sample size increases, altering the ranking of features, sometimes to the wrong order.

**Computational Complexity of KNN-based RCD Estimator** The computation of KNN-based RCD is dominated by empirical ranking for each dimension and \( k \)-nearest neighbor search for each sample. The ranking can be solved by mergesort, which costs \( O(dn \log n) \) for \( d \) dimensions [Cormen, 2009]. The \( k \)-nearest neighbor search can be solved using k-d tree construction, which has \( O(kdn \log n) \) complexity when \( d \) is small (no larger than 20) [Bentley, 1975]. However, the complexity can increase to \( O(kdn^2) \) if \( d \) becomes large. Therefore, the overall complexity is \( O(kdn \log n) \) for low dimension data and \( O(kdn^2) \) for high dimension data.

### 2.4 Experimental Results

In this section, we empirically verify the properties of RCD in our theoretical analysis.

We first check the estimation errors for RCD in synthetic experiments with additive noise and mixture noise respectively. For each type of noise, we simulate data with several different relationships so as to show the effect of self-equitability and robust-equitability respectively. In particular, we compare the RCD estimator with an MI estimator based on the same density estimation. Due to the non-robust-equitability of MI, in the mixture noise cases, the
MI estimator varies widely with the sample sizes. In contrast, RCD converges as sample sizes increases. Therefore, MI may provide misleading ranking of features with unequal sample sizes. Also, the ranking between relationships with the two different noise types are greatly affected by the sample sizes under MI, while ranking under RCD remains relatively stable.

We then conduct several synthetic experiments to illustrate the properties in feature selection, and then show that similar patterns exist on real-world datasets. In section 2.4.3, we show that: (1) Non-self-equitable dependence measures may provide misleading ranking under additive noise; (2) Non-robust-equitable dependence measures may provide misleading ranking under mixture noise when features have unequal sample sizes; (3) The ranking by non-robust-equitable dependence measures between the two types of noises are sensitive to sample size. Section 2.4.4 shows that similar behavior occurs in three real data examples. This confirms that the advantages of self-equitable and robust-equitable dependence measures are not just theoretical, but are real in some practical situations.

Furthermore, we compare the performance of feature selection by the filter method mRMR [Peng et al., 2005], using various dependence measures as measures of relevance and redundancy. We conducted the mRMR method first on synthetic examples in section 2.4.5, to illustrate why non-self-equitability or non-robust-equitability could lead to misleading results. We then did the mRMS on nine benchmark datasets in section 2.4.6, for nonlinear predictions, from the UCI Data Repository [Lichman, 2013]. Notice that the feature selection performance on a particular data set is affected by the type of existing relationships and the type of predictors used. For example, the Pearson’s correlation with a linear regression predictor should do best if linear relationship is dominant in a data set. For a fair comparison, we measure the performance by the 10-fold cross-validated MSE of spline regression, a general nonlinear predictor [Friedman, 1991], using the selected features. Self-equitability and robust-equitability lead to equitable and robust feature selection. Hence RCD should provide stable performance across different types of data, not necessarily best in

42
each situation. However, over many data sets with different types of nonlinear relationships, robust-equitable RCD would provide best average performance as confirmed on these nine benchmark datasets.

Lastly, we apply RCD in a climate study with a focus on building climate network with this nonlinear dependence measure. We show that some interesting nonlinear relationship related to teleconnections can be found by using RCD which can not be detected efficiently by the Pearson’s correlation.

There are some parameters to be set for computing various dependence measures. For kernel based measures, we follow the settings used by Fukumizu et al. [2007]. For HSNIC, we set the regularization parameter $\epsilon_n = 10^{-5}n^{-3.1}$ to satisfy the convergence guarantee given by Theorem 5 from Fukumizu et al. [2007]. As discussed in the previous section, we set $k = 0.25\sqrt{n}$ for the k-NN estimator of MI, RCD and CD$_2$.

### 2.4.1 Estimation Errors and Equitability

In this section, we study the estimation errors of our RCD estimates through synthetic experiments, and examine the equitability effect in ranking features. We first generate data from four different functional types: linear, square root, cubic, and quadratic (cases A, C, D, F in Table A.1 of Appendix E) with two sample sizes of $n = 1000$ and $n = 10000$ respectively. Also, data with two different noise paradigms and three noise levels are tested with three measures, RCD, (non-self-equitable) $\rho$ and (non-robust-equitable) CD$_2$.

As we can see from Table 2.2, the standard deviation of the RCD estimates are small, and the expected value of RCD converges to the true values as sample size increases. The expected values of RCD was similar for the different functional relationships. They are closer to the true values under the mixture noise than under the additive noise. For either noise type, the estimates are very accurate for sample size $n = 10,000$. Although there are some random estimation errors, the RCD estimates would not miss-rank features with moderate
Table 2.2: The expected values of RCD estimates (with standard deviation in parenthesis) based on 100 simulations, under various functional types, sample sizes, noise types and noise levels.

difference in true RCD values. That is, it never ranks features with real RCD = 0.2 as more dependent than features with real RCD = 0.4, under samples \( n = 1000 \) or \( n = 10,000 \). So the RCD estimates can be used to provide reliable dependence ranking that do not change dramatically under sample sizes \( n = 1000 \) versus \( n = 10,000 \).

In contrast, the non-self-equitable Pearson’s \( \rho \) values depend on the functional relationship. Under the mixture noise, the estimated \( \rho \) values are close to the mixture proportion only for the linear relationship. Hence its ranking of features is heavily influenced by the functional relationships. Particularly, it fails to detect the quadratic relationship in the last column of Table-2.3. Even when 80% of data follows the deterministic quadratic relationship, it is still ranked as less dependent than the features with other functional relationships (even if the other features has only 40% deterministic mixture proportion).

For the non-robust-equitable \( \text{CD}_2 \) in Table 2.4, its value changes dramatically under sample sizes \( n = 1000 \) versus \( n = 10,000 \), especially under mixture noise. This demonstrates that sample size affects the ranking by \( \text{CD}_2 \) in contrast to the ranking by the robust-equitable RCD. If the two features have unequal sample sizes, then the feature with larger sample size has a built-in preference by \( \text{CD}_2 \). For each of the functional types, \( \text{CD}_2 \) ranks a feature with mixture proportion (of deterministic data) 0.6 but large sample size \( n = 10000 \) as more
Table 2.3: The expected values of $\rho$ estimates (with standard deviation in parenthesis) based on 100 simulations.

<table>
<thead>
<tr>
<th>type</th>
<th>level</th>
<th>Linear 1k</th>
<th>Linear 10k</th>
<th>Square Root 1k</th>
<th>Square Root 10k</th>
<th>Cubic 1k</th>
<th>Cubic 10k</th>
<th>Quadratic 1k</th>
<th>Quadratic 10k</th>
</tr>
</thead>
<tbody>
<tr>
<td>add</td>
<td>0.4</td>
<td>0.77(0.01)</td>
<td>0.77(0.00)</td>
<td>0.77(0.01)</td>
<td>0.77(0.00)</td>
<td>0.75(0.01)</td>
<td>0.75(0.00)</td>
<td>0.00(0.04)</td>
<td>0.00(0.01)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.90(0.00)</td>
<td>0.90(0.00)</td>
<td>0.90(0.01)</td>
<td>0.90(0.00)</td>
<td>0.87(0.01)</td>
<td>0.87(0.00)</td>
<td>0.00(0.04)</td>
<td>0.00(0.01)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.98(0.00)</td>
<td>0.98(0.00)</td>
<td>0.96(0.00)</td>
<td>0.96(0.00)</td>
<td>0.91(0.00)</td>
<td>0.91(0.00)</td>
<td>0.00(0.04)</td>
<td>0.00(0.01)</td>
</tr>
<tr>
<td>mix</td>
<td>0.4</td>
<td>0.40(0.03)</td>
<td>0.40(0.01)</td>
<td>0.33(0.03)</td>
<td>0.33(0.01)</td>
<td>0.34(0.03)</td>
<td>0.33(0.01)</td>
<td>-0.01(0.03)</td>
<td>0.00(0.01)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.60(0.03)</td>
<td>0.60(0.01)</td>
<td>0.52(0.03)</td>
<td>0.51(0.01)</td>
<td>0.51(0.03)</td>
<td>0.5(0.01)</td>
<td>0.00(0.03)</td>
<td>0.00(0.01)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.80(0.02)</td>
<td>0.80(0.01)</td>
<td>0.73(0.02)</td>
<td>0.72(0.01)</td>
<td>0.69(0.03)</td>
<td>0.69(0.01)</td>
<td>0.00(0.04)</td>
<td>0.00(0.01)</td>
</tr>
</tbody>
</table>

Table 2.4: The expected values of $CD_2$ estimates (with standard deviation in parenthesis) based on 100 simulations.

<table>
<thead>
<tr>
<th>type</th>
<th>level</th>
<th>Linear 1k</th>
<th>Linear 10k</th>
<th>Square Root 1k</th>
<th>Square Root 10k</th>
<th>Cubic 1k</th>
<th>Cubic 10k</th>
<th>Quadratic 1k</th>
<th>Quadratic 10k</th>
</tr>
</thead>
<tbody>
<tr>
<td>add</td>
<td>0.4</td>
<td>0.94(0.04)</td>
<td>1.23(0.02)</td>
<td>0.96(0.04)</td>
<td>1.28(0.02)</td>
<td>1.08(0.04)</td>
<td>1.45(0.02)</td>
<td>0.88(0.03)</td>
<td>1.19(0.01)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>1.7(0.04)</td>
<td>2.25(0.02)</td>
<td>1.81(0.05)</td>
<td>2.46(0.02)</td>
<td>2.19(0.06)</td>
<td>3.11(0.03)</td>
<td>1.6(0.04)</td>
<td>2.27(0.02)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>3.62(0.06)</td>
<td>5.06(0.03)</td>
<td>3.72(0.07)</td>
<td>5.47(0.03)</td>
<td>4.35(0.10)</td>
<td>7.52(0.05)</td>
<td>2.93(0.06)</td>
<td>4.96(0.03)</td>
</tr>
<tr>
<td>mix</td>
<td>0.4</td>
<td>1.29(0.09)</td>
<td>2.62(0.07)</td>
<td>1.30(0.09)</td>
<td>2.69(0.072)</td>
<td>1.24(0.10)</td>
<td>2.59(0.06)</td>
<td>0.83(0.06)</td>
<td>1.71(0.04)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>2.70(0.15)</td>
<td>5.78(0.10)</td>
<td>2.73(0.15)</td>
<td>5.89(0.09)</td>
<td>2.66(0.17)</td>
<td>5.76(0.10)</td>
<td>1.69(0.08)</td>
<td>3.63(0.06)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>4.75(0.15)</td>
<td>10.24(0.11)</td>
<td>4.79(0.17)</td>
<td>10.34(0.11)</td>
<td>4.75(0.16)</td>
<td>10.24(0.11)</td>
<td>2.87(0.08)</td>
<td>6.27(0.07)</td>
</tr>
</tbody>
</table>

dependent than a strongly dependent feature with mixture proportion 0.8 but smaller sample size $n = 1000$. Also, for features with different noise types, their ranking are inconsistent when sample size changes. For example, for the linear relationship under sample size $n = 1000$, $CD_2$ ranks the feature with RCD = 0.6 mixture noise as less dependent than the feature with RCD = 0.8 additive noise. But when sample size is increased to $n = 10000$, the ranking between these two features reverses. While it is not necessary for other dependence measures to rank features across different noise types in the same order as RCD, the stability of the ranking under different sample sizes is desirable. The non-robust-equitable dependence measures may not provide consistent ranking.

In summary, we observe three advantages of RCD for feature selection, in comparison to other dependence measures. (1) Non-equitable measures such as $\rho$ may prefer certain
functional relationships (say, linear), while RCD treat them equitably. (2) Self-equitable but non-robust-equitable measure such as $CD_2$ prefer features with larger sample size. (3) Self-equitable but non-robust-equitable measure such as $CD_2$ does not provide stable ranking among features when sample size changes.

In the following subsection, we first explore the equitability property of RCD versus more dependence measures. Then we will further explore these feature selection issues among all those dependence measures.

### 2.4.2 Multivariate Equitability Analysis

We now perform equitability analysis on simulated multivariate data. Following the framework of Reshef et al. [2014], we generate noisy data from various functional forms, and plot the estimated dependence measure values against the signal level in Figure 2.4. Our robust-equitability definition extends the natural signal level to higher dimensional case in the mixture noise model. Here we study three dimensional ($d = 3$) cases of six different nonlinear function relationships, generated with different portion of uniform noise from 0 to 0.9 (with signal portion from 0.1 to 1). The sample size $n = 1000$ is used in this experiment. Figure 2.5 plots the six three dimensional different nonlinear function relationships at signal level 0.8 (0.2 proportion of uniform noise).

According to Reshef et al. [2014], a measure is more equitable if the length of a band, when cutting each plot in Figure 2.4 with a horizontal line, is smaller. That is, the measure with smaller bandwidth could capture the dependence purely based on the noise level, and is robust to different (linear and nonlinear) relationships. On the other hand, if the band is very wide, it will give the same score for data with a wide range of noise levels, and hence could not identify strong relationships correctly. Figure 2.4 shows that RCD is the most equitable, since it has the narrowest bandwidth and lies on the signal level line. The self-equitable mutual information is second best, has relatively narrower bandwidth compared
to other measures except RCD.

2.4.3 Synthetic Datasets I: Ranking of Features

To compare the performance of each dependence measure in feature selection, we consider four features $X_1, X_2, X_3, X_4$ and target variable $Y$ as shown in Figure 2.6. $Y$ has a non-monotonic but deterministic relationship with $X_1$ and a linear relationship with $X_2$ plus some additive noise. In addition, $Y$ has linear relationships with both $X_3$ and $X_4$ corrupted by increasing level of continuous background noise. For each feature $X_i$, we calculate its dependence measure with $Y$ for different sample sizes $n = 300$ and 10,000. Results are presented in Table 2.5.

Since $X_1$ has a deterministic relationship with $Y$, it should be ranked as more dependent than other features. $X_3$ and $X_4$ has mixture noise with the mixture proportions of 0.75 and 0.5 respectively. We can see that the values learned by RCD are close to those values and correctly ranks $X_3$ as more dependent than $X_4$. The last row of Table 2.5 reports the 10-fold cross-validated mean-squared-error (MSE) of a nonlinear predictor using each feature. Here
Figure 2.5: Six function types, circle, parabola, sine/cosine, twisted curve, two cross lines, and a spiral curve, with 20% uniform noise from 0 to 0.9 (with signal portion from 0.1 to 1). The examples in this figure is of signal level 0.8.

RCD results are consistent with the MSE results, providing higher scores for those with lower MSE values (more predictive of $Y$). Now, we inspect the other dependence measures and observe the three issues mentioned in section 2.4.1.

**Self-equitability.** We expect the self-equitable measures to treat linear and nonlinear models equally (i.e., they should prefer $X_1$ over $X_2$ because $X_1$ is purely deterministic while $X_2$ has some noise). As we can see from Table 2.5, Pearson correlation coefficient $\rho_{lin}$, and kernel-based measures prefer $X_2$ more than $X_1$. On the other hand, self-equitable measures MI, $CD_2$ and RCD were able to rank the features correctly. Although HSNIC and CHSNIC have the same value as $CD_2$ in the large data limit, empirically they behave similarly to other non-self-equitable kernel-based measures due to slow convergence of their estimators [Reddi and Póczos, 2013].
Figure 2.6: $(X_1, Y)$ has a nonmonotonic relationship with deterministic signal. $(X_2, Y)$ has linear relationship with uniform additive noise with width $d = 0.2$. $(X_3, Y)$ has background noise with a 0.75 linear signal portion. $(X_4, Y)$ is similar to $(X_3, Y)$ but with a 0.5 signal portion.

**Selection Correctness in Unequal Sample Sizes.** In real applications, some features may have some missing measurements, resulting in unequal number of samples among the various features. In this setting, we still want to compare feature relevance. An ideal dependence measure should not be influenced greatly by unequal sample sizes. Let us take a closer look at MI and $CD^2$ and on how they rank features $X_3$ and $X_4$. Note that $X_3$ has a stronger signal-to-noise proportion than $X_4$ ($p = 0.5$ versus $p = 0.75$); thus, ideally, one would like the measure to rank $X_3$ higher than $X_4$ as empirically confirmed by the MSE results. The ranking provided by MI and $CD^2$ is correct when both features have the same sample size, $n = 10,000$. However, if the stronger feature $X_3$ has missing measurements so that $n = 300$ for $X_3$, then $X_3$ is ranked lower than $X_4$ by $CD^2$, which would mislead feature selection algorithms. MI will make the same mistake if the sample size for $X_4$ further increases.

**Selection Stability in Different Sample Sizes.** Ideally, a measure should not vary too much as the sample size changes. However, we observe that MI, $CD^2$, and HSNIC’s ranking of features $X_2$, $X_3$ and $X_4$ is affected when the sample size is increased. With fixed sample size $n = 300$, MI ranks $X_2$ as having higher deterministic relationship with $Y$ compared to $X_4$. However, when the sample size is increased to $n = 10,000$, it reverses the ranking of these features. This is due to its non-robust-equitability and resulting estimation difficulty,
as proved in Theorems 8 and 9. Additionally, similar phenomenon appears for CHSNIC and HSNIc on features \( X_2 \) and \( X_3 \). We observe that when \( n = 300 \), they rank \( X_2 \) as having higher dependence with \( Y \) compared to \( X_3 \). However, when \( n = 10,000 \), these rankings are reversed. These inconsistencies may mislead feature selection algorithms.

### Real Datasets I: Ranking Features

In this subsection, we verify that the equitability properties in subsection 2.4.3 are also observed on real data.

**Self-equitability.** Consider the stock dataset from StatLib\(^1\). This dataset provides daily stock prices for ten aerospace companies. Our task is to determine the relative relevance of the stock price of the first two companies \( (X_1, X_2) \) in predicting that of the fifth company \( Y \).

The scatter plots of \( Y \) against \( X_1, X_2 \) are presented in Figure 2.7. Ideally, self-equitable

---

\(^1\)http://lib.stat.cmu.edu/
measures should prefer $X_1$ over $X_2$ because the MSE associated with $X_1$ is lower even though it has a more complex function form. As we can see from Table 2.8, self-equitable measures MI, $CD_2$, and RCD all correctly select $X_1$. While measures that are not self-equitable fail to select the right feature.

**Selection Correctness in Unequal Sample Sizes.** Consider the KEGG metabolic reaction network dataset [Lichman, 2013]. Our task is to select the most relevant features in predicting target variable ‘Characteristic path length’ ($Y$). The ‘Average shortest path’ ($X_1$), ‘Eccentricity’ ($X_2$) and ‘Closeness centrality’ ($X_3$) are used as candidate features. Observe the ranking of $X_1$ and $X_3$ from Table 2.6, when they have equal sample sizes (either 1000 or 20,000), MI, $CD_2$ and RCD all rank $X_1$ as being more relevant than $X_3$. The MSE values also confirmed that $X_1$ is more predictive of $Y$ than $X_3$. However, if there are missing measurements of $X_1$, then we may need to compare $X_1$ with 1000 samples and $X_3$ with 20,000 samples. The feature $X_3$ with less signal strength but larger sample size is given higher ranking by MI and $CD_2$, degrading the performance of feature selection algorithms. In contrast,
RCD correctly identifies $X_1$ as being more relevant than $X_3$ even with the unequal sample size.

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>1k</td>
<td>20k</td>
<td>1k</td>
</tr>
<tr>
<td>MI</td>
<td>3.39</td>
<td>3.95</td>
<td>3.23</td>
</tr>
<tr>
<td>CD$_2$</td>
<td>12.05</td>
<td>31.65</td>
<td>10.67</td>
</tr>
<tr>
<td>RCD</td>
<td>0.85</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>MSE</td>
<td>0.030</td>
<td>0.028</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Table 2.6: Dependence measure for three features in metabolic reaction network dataset

Selection Stability in Different Sample Sizes. Ideally, a measure should not vary too much as the sample size changes. However, in Table 2.6, CD$_2$’s ranking of features $X_2$ and $X_3$ is affected by the increase in sample size. If we fix sample size $n = 1000$, CD$_2$ ranks $X_2$ as more relevant than $X_3$ in predicting $Y$, agreeing with the MSE ranking. However, when the sample size increases to $n = 20,000$, CD$_2$ will prefer $X_3$. CD$_2$ will select the feature $X_3$ when the sample size is large even though it is less relevant to $Y$. RCD has the same ranking under both sample sizes.

2.4.5 Synthetic Datasets II: mRMR Feature Selection

In this part, we investigate the performance of feature selection for each dependence measure with mRMR [Peng et al., 2005]. We generate data from the following additive regression model $Y = 1.5 \cos(3\pi X_1) + (1 - 2|2X_2 - 1|)^2 + \epsilon$, where $X_1$ and $X_2$ are uniformly distributed on [0,1], and $\epsilon \sim N(0,0.05)$. We consider the feature selection from twenty features. The first two features are $X_1$ and $X_2$. The next six features are noisy versions of $X_1$ and $X_2$, with some as mixtures. They are also related to $Y$. A good feature selection method should select $X_1$ and $X_2$ before these more noisy features $X_3$ to $X_8$. The rest 12 features $X_9$ to $X_{20}$ are simply random noise not related to $X_1$, $X_2$ or $Y$. The features are listed in Table 2.7. And we plot the first eight features versus $Y$ in one simulation run in Figure 2.9.
Figure 2.9: Scatter plots of the features $X_1, \cdots, X_8$ versus $Y$.

Table 2.7: Features that is used in the additive regression model.

<table>
<thead>
<tr>
<th>Feature $X_i$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$X_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$X_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>25% $X_2$ and 75% $Y$ with additive noise $U(-0.75, 0.75)$</td>
</tr>
<tr>
<td>$X_4$</td>
<td>20% $X_2$, 20% $Y$ and 60% background noise</td>
</tr>
<tr>
<td>$X_5$</td>
<td>$X_2$ with additive noise $U(-0.05, 0.05)$</td>
</tr>
<tr>
<td>$X_6$</td>
<td>50% $X_2$ and 50% background noise</td>
</tr>
<tr>
<td>$X_7$</td>
<td>$X_1$ with additive noise $U(-0.2, 0.2)$</td>
</tr>
<tr>
<td>$X_8$</td>
<td>50% $X_1$ and 50% background noise</td>
</tr>
<tr>
<td>$X_9,...,X_{20}$</td>
<td>pure random noise</td>
</tr>
</tbody>
</table>

We generate data sets with size $n = 1000$, and select the features with mRMR based on the eight dependence measures. This experiment is repeated 50 times, and record the order of each feature being selected in each data sets. We then use the spline regression model to get cross-validated mean square error (MSE) on top one to ten selected feature sets. The cross-validated MSE averaging over the 50 runs are plotted in Figure 2.10.

In this synthetic example, RCD yields the lowest MSE. We can see why by looking at the feature selection result in more details. Table 2.8 showed the features that are most frequently selected as the top one to five features using mRMR. Inversely, Table 2.9 shows the median order of being selected for each relevant feature $X_1, \ldots, X_8$.

From the tables, RCD correctly selects the two true features $X_1$ and $X_2$ as the top two
features, thus resulting in the lowest MSE curve in Figure 2.10. The non-self-equitable measures, for this data distribution, incorrectly ranks first the feature $X_3$ which has linear relationship with $Y$ in parts of the data (75% mixture). Then Pearson’s correlation $\rho$ does not rank $X_1$ and $X_2$ high because the relationships are nonlinear. Some nonlinear non-self-equitable measures (HSNIC, CMMD, CHSNIC) are able to rank the feature $X_1$ second, but can not select feature $X_2$ which has a nonlinear relationship with $Y$. That is due to $X_2$, which also has some dependence with $X_3$, was penalized when $X_3$ was selected first. The self-equitable measures MI, CD$_2$ and RCD, in contrast, correctly rank $X_1$ first. However, only the robust-equitable RCD ranks $X_2$ second. The non-robust-equitable MI and CD$_2$ incorrectly ranks the noisy $X_4$ second, instead of $X_2$, due to mishandling of the mixture noise.

### 2.4.6 Real Datasets II: mRMR feature selection

Here we used the various dependence measures as measures of relevance and redundancy in the mRMR-based search strategy, and compare the feature selection results on nine real-world datasets. Due to the cubic computational cost of kernel-based measures (HSNIC, CHSNIC), up to 1000 samples are used for each dataset. We compare the results of RCD
Table 2.8: The most frequently selected top five features, with the relative frequency in parenthesis. Asterisk indicates one of the random noise features ($X_9, ..., X_{20}$).

<table>
<thead>
<tr>
<th>Order of selection</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>$X_3(100%)$</td>
<td>*</td>
<td>$X_4(50%)$</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>HSIC</td>
<td>$X_3(100%)$</td>
<td>$X_7(96%)$</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>HSNIC</td>
<td>$X_3(100%)$</td>
<td>$X_1(100%)$</td>
<td>$X_4(78%)$</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>CMMD</td>
<td>$X_3(100%)$</td>
<td>$X_1(100%)$</td>
<td>*</td>
<td>$X_4(52%)$</td>
<td>*</td>
</tr>
<tr>
<td>CHSNIC</td>
<td>$X_3(100%)$</td>
<td>$X_1(100%)$</td>
<td>$X_4(98%)$</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>MI</td>
<td>$X_1(100%)$</td>
<td>$X_4(78%)$</td>
<td>$X_3(90%)$</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>CD$_2$</td>
<td>$X_1(100%)$</td>
<td>$X_4(80%)$</td>
<td>$X_3(80%)$</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>RCD</td>
<td>$X_1(100%)$</td>
<td>$X_2(82%)$</td>
<td>$X_3(90%)$</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

versus other dependence measures by showing plots of 10-fold cross-validated MSE using spline regressor with the features selected by these measures versus the number of selected features in Figure 2.11.

We used the Kruskal-Wallis test to compare the MSE between different measures. Table 2.10 lists the top dependence measures in order of their MSE. For each data set, we only include the dependence measures resulting in MSE equivalent to the best MSE (p-value $> 0.05$ for Kruskal-Wallis test) in the table. We can see that RCD performs the best in 8 out of 9 data sets. Most other measures are worse off in more than half of the data sets. Only CHSNIC and CMMD are close in performance. In particular, CMMD is among the best measures in 5 data sets. CHSNIC performs best in 6 data sets and actually beats RCD in one data set *whitewine*. RCD in fact find the best top feature in every data set including *whitewine*. However, the best second feature in *whitewine* was not selected by RCD. Overall, RCD has the best performance in mRMR-based feature selection compared to competing dependence measures.
Table 2.9: Median of the order of selection for the first eight features in each dependence measure experiment among fifty repeated runs.

### 2.4.7 Real Datasets III: Climate Network Analysis

Ever since the identification of certain types of the small world network models Watts and Strogatz [1998], the complex network has been a mature field with a wide range of applications, ranging from the modeling of social network, the structure of the World Wide Web (WWW), to the gene network. Recently, the complex network theory has been brought to the climate community as the climate network Donges et al. [2009] Steinhaeuser et al. [2010] Steinhaeuser et al. [2011] Steinhaeuser et al. [2012], which aims to view and model the climate data from the complex network point of view.

Most of the current research in climate network is based the Pearson’s correlation coefficient ($\text{cor}$), which is an excellent tools for detecting linear signals. However, as is known to us all, the climate system involves a great amount of nonlinearity. Thus, we attempt to apply a cutting edge nonlinear dependence measure, the robust copula dependence measure (RCD) Chang et al. [2016] Ding et al. [2017], to help constructing the climate network. This could potentially enable us to include more nonlinear features or information based on our
data, which may be helpful for us in understanding more about the nature.

On the other hand, teleconnection Choi et al. [2015] Tsonis et al. [2008] has been a traditional but still on-going research topic in climate science. Detecting relatively weak but informative information of the teleconnection from the highly correlated nearest-neighbour effect is a challenging problem. The usage of nonlinear dependence measure in climate network could help to discover novel and interesting teleconnection phenomenon.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Top Dependence Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>abalone</td>
<td>RCD, CHSNIC, HSIC</td>
</tr>
<tr>
<td>bodyfat</td>
<td>RCD, CMMD, CHSNIC</td>
</tr>
<tr>
<td>building</td>
<td>RCD, CHSNIC, Pearson, CMMD, CD₂</td>
</tr>
<tr>
<td>chemical</td>
<td>RCD, CHSNIC</td>
</tr>
<tr>
<td>housing</td>
<td>RCD, CMMD</td>
</tr>
<tr>
<td>metabolic</td>
<td>RCD, CHSNIC, CD₂, CMMD, HSNI, Pearson</td>
</tr>
<tr>
<td>protein</td>
<td>RCD, CMMD</td>
</tr>
<tr>
<td>stock</td>
<td>RCD, CD₂</td>
</tr>
<tr>
<td>whitewine</td>
<td>CHSNIC</td>
</tr>
</tbody>
</table>

Table 2.10: Measures ranked by predictive $MSE$

Climate Data

We consider the reanalysis monthly mean temperature data (1948 - 2016) from the National Oceanic & Atmospheric Administration (NOAA) \(^2\). The resolution of this data is $5° \times 5°$. We normalize the raw data (temperature) by removing the historical long range mean for each month. Namely, for a given climate variable $x(y, m)$ at year $y$ and month $m \in \{1, 2, \cdots, 12\}$, the anomaly data $\tilde{x}(y, m) = x(y, m) - \langle x(y, m) \rangle_y$, where $\langle \cdot \rangle_y$ is the average over $y$.

The Climate Network

A network, or graph, is a combination of vertexes and edges which is usually denoted by $G = (V, E)$, where $V$ is the collection of all the vertexes and $E$ is the set of all the edges with the endpoints from $V$. Moreover, the climate network models the climate data in the way that the vertexes are the geographical locations while the edges represent the possible relationship between each pair. In our current situation, the vertexes are the spatial grid points from the reanalysis model. While $\text{cor}$ is often used in describing the relationships along each edge, we apply RCD as a nonlinear proxy instead.

\(^2\)https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html
Figure 2.12: Degree distribution of the networks based on the two dependence measure.

To analyze the climate network, several statistics can be used to summarize the characteristics of the intrinsic structure. Here, we focus on four major aspects: the degree distribution, clustering coefficient, betweenness, and community detection. The networks are constructed based on the top 0.5 percentile of the edge weight, although other testing based method is also possible.

The degree of a vertex is the total number of edges correspond to this vertex. The overall distribution of the degree can provide us a general information about the connectivity of the network. As we can see from Figure 2.12, the two network based on Pearson’s correlation coefficient and RCD have similar degree distribution.

Local clustering effect is a way trying to delineate, for a given vertex, whether the connected points are also connected. It is defined as the fraction of connected triples through each vertex that are closed. As we can see from both plots in Figure 2.13, the clustering effect are more likely to occur around the equator and the poles.

As another important summary statistics, the betweenness is a way to measure the centrality of a network. A vertex with high betweenness plays a significant role to bridge two sub-network together. It is the number of geodesic paths that pass through the corresponding vertex. Figure 2.14 shows the relatively different results based on cor and RCD. As we can see from the plots, the nonlinear RCD based network indicates the grid points around the
Figure 2.13: Local clustering of the networks based on the two dependence measure (Redder is higher value, and bluer is lower value).

Figure 2.14: Betweenness of the networks based on the two dependence measure (Redder is higher value, and bluer is lower value).

America continent has higher betweenness.

Community detection is a traditional way to find homogeneous regions based on the relationship/distance between each pair of the vertexes. Figure 2.15 presents the hierarchical clustering results based on the two distance matrices (here, we define $1 - |\text{cor}(x, y)|$ and $1 - RCD(x, y)$ as a semi-distance). Results shows that both of the communities are distributed according to their latitude.

Figure 2.15: Communities of the networks based on the two dependence measure based on the hierarchical clustering algorithm.
Figure 2.16: Visualization of the network structure that are constructed by the remotely highly correlated edges base on their spatial location on map.

**Teleconnection**

It is of great interest and significance to study the teleconnection, the dependence relationship between faraway locations, of the earth system. The application of the nonlinear dependence measure in the climate network could cast light on finding novel information from climate data set. To this end, we first filter out the neighbourhood points within five grid units, and carefully examine the networks constructed based on the top 100 dependence scores.

Figure 2.16 shows that the basic structure of the two networks. The vertexes are geographical grids and the edges are the cords. While most of the relationships are around the equator, we notice that the nonlinear dependence RCD detect an arc with long latitude distance. This arc is further analyzed in Figure 2.17.

Figure 2.17 is the scatter plot that is uniquely identified with the nonlinear dependence measure RCD, which is between the grid point (65°W, 60°N) and (30°E, 65°S). As we can see from the Figure 2.17, the anomaly data in these two positions exhibit a bilinear relationship, which cannot be detected by linear dependence measure like Pearson’s correlation coefficient. This bilinear relationship of the anomaly data indicates some possible interesting finding: if the air temperature of one place is in its average value, the other place is more likely in its extreme case, and vise versa. This may help to guide the further research in studying the climate extremes.
2.5 Conclusions and Discussions

As the data size explodes, researchers are studying increasingly complex relationships among features. Restricting the focus on simple linear relationship can miss very informative features. Therefore, how to measure the dependence strength equitably for various functional relationships has attracted recent interest from researchers [Reshef et al., 2011, Kinney and Atwal, 2014, Murrell et al., 2014, Reshef et al., 2015b]. This chapter provides a theoretical treatment of various equitability definitions, including our proposal of the robust-equitability concept. The robust copula dependence (RCD) is proven to be both self-equitable and robust-equitable. Theoretically we show that RCD is intrinsically easier to estimate than some other self-equitable dependence measures (such as MI and CD$_2$). Particularly, through minimax rate of convergence, we provide a theoretical explanation for the difficulty of accurately estimating MI which is noted by practitioners. A practical estimator is provided for RCD, which enables its usage in feature selection.

Through theoretical and empirical studies, we have shown that RCD does better in rank-
ing the features according to deterministic signal strengths compared to other dependence measures. The non-self-equitable measures may prefer noisy features with certain types of relationships (e.g., monotonic) over less noisy features with more complex relationships. Self-equitable but non-robust-equitable measures (such as MI and CD$_2$) overcome this deficiency but have estimation problems, leading to non-robust feature selection particularly when comparing features with unequal sample sizes. RCD can be used in feature selection to overcome these limitations. Moreover, we also apply RCD in building climate network with this nonlinear dependence measure. Results show that RCD can be used as an nonlinear alternative to construct complex network and to recover nonlinear network dynamics.

Using nonlinear dependence measures, rather than the Pearson’s correlation, in high-dimensional data analysis (e.g., independent component analysis) is becoming more popular to deal with possibly non-Gaussian noises. The equitability properties of RCD makes it an ideal choice of dependence measure in such applications. Replacing measures such as MI by RCD may lead to more robust results as shown in the examples of Section 2.4. The RCD estimation in high-dimension case however, similar to MI, may be inaccurate as it involves high-dimensional density estimation. The improvement on nonparametric RCD estimation remains an ongoing research effort, and can lead to wider applications.
Chapter 3

Estimation of High Dimensional Spatial Covariance Matrices

3.1 Introduction

Investigation of the covariance structure is one of the central topics in high dimensional statistics Pourahmadi [2014], Cai et al. [2016], Fan et al. [2016]. The covariance matrix, as well as its inverse, plays a significant role in many applications, including classification problem in discriminant analysis Anderson [2003], portfolio selection Markowitz [1952], Gaussian graphical models Wang et al. [2016] and dimensionality reduction (e.g., principal component analysis). However, it has been shown that the sample covariance matrix has inferior performance in the high dimensional setting Johnstone [2001]. On the other hand, regularization methods, including banding, tapering, and thresholding the sample covariance matrix, have been successfully applied in this area as the improvement against the inferior performance of the sample covariance matrix estimator in estimating covariance in high dimensions Bickel and Levina [2008a,b]. For the high dimensional covariance estimation problem Cai et al. [2016], Fan et al. [2016], theoretical asymptotic rate optimality has been proven for various classes of structured matrices, such as bandable matrices Cai et al. [2010], Cai and Yuan
[2012] and Toeplitz matrices Xiao and Wu [2012], Cai et al. [2013]. Those structures however do not describe spatial dependence and the theoretical results of rate optimal estimation for high-dimension spatial covariance matrix is yet to be proven.

Spatial data are encountered in a wide range of disciplines. With the increasing complex data model being investigated, for example in climate science Benestad et al. [2008], high dimensional spatial covariance matrix estimation is a crucial element of spatial data analysis. The spatial covariance matrix is useful in both supervised regression problem Cressie [1993], Christensen [2002], Montero et al. [2015] and unsupervised problem Benestad et al. [2015], as well as other areas like spatial econometrics LeSage and Pace [2009].

For high dimensional covariance matrix estimation, there are two types of regularized estimators based on different matrix specifications. One specification does not depend on the index structure, for which the thresholding Bickel and Levina [2008b] estimator is used. The other specification is based on the index structure (i.e., some feature dependency structure is assumed), which corresponds to banding and tapering Bickel and Levina [2008a] estimators. As is noted in Bickel and Levina [2008b], in the application of spatial data, banding or tapering the covariance matrix can be applied as long as there is an appropriate metric (for example, some underlying distance) on variable indexes. The tapering estimator has been studied in spatial statistics; while most literature focus on computational issues, see Kaufman [2008], Shaby and Ruppert [2012] and the references therein, the theoretical rate optimality is not established, as there was not a good class of structured matrices describing the spatial covariance matrix. Hence the theoretical optimality results on bandable matrices and Toeplitz matrices (which were developed for time series data) do not apply.

As noted in Zhu and Liu [2009], the theoretical treatment of spatial matrix estimation is more difficult than that of time series data (partially) because there is no natural ordering (or indexing) of the observation. In this chapter, we show that spatial covariance matrices can exhibit a block bandable structure under the default ordering method in Zhu and Liu...
Bandable covariance matrix is a class of matrices that has large values when the elements are close to its diagonal and decays gradually to its upper-right and lower-left corners, which is applicable to covariance matrices of some time series. A block bandable matrix has many blocks of bandable submatrices, while the blocks also decay further from the diagonal in the pattern of a bandable matrix. We establish the minimax risk bound for block bandable matrix estimation under the Frobenius norm, which is a commonly used norm in covariance estimation problems Lam and Fan [2009], Ravikumar et al. [2011]. We show that a double-tapering estimator achieves the minimax rate, which indicates that it preforms best in the worst possible case allowed in the problem, thus is indeed theoretically rate optimal. Numerical performance of the double-tapering estimator is shown through simulation studies and two applications to image compression and climate downscaling.

3.2 Block Bandable Spatial Covariance Estimation

In this section, we first define the class of block bandable matrices, and show its connection with the covariance matrices of two-dimensional spatial data. Then we describe the regularization estimator based on this structure of covariance matrices.

3.2.1 Block Bandable Spatial Covariance

Spatial data is recorded based on its geographical location, e.g., latitude and longitude. One of the main challenge for this kind of data in \( \mathbb{R}^d \) with \( d \geq 2 \) is that, they are not ordered as the case in \( \mathbb{R}^1 \) such as in time series data. As a consequence, the simple structures like Toeplitz matrices Xiao and Wu [2012], Cai et al. [2013] are not applicable for spatial data whose dimension is higher than one Zhu and Liu [2009].

We consider the case that the spatial dependence decays with respect to their spatial distances, which can be viewed as a type of generalization of the parametric isotropic covariance models Montero et al. [2015], Sherman [2011]. In this case, when we index the
spatial observations in a regular scheme, the resulting covariance matrices exhibits a certain trackable structure.

For example, we consider a grid of $10 \times 10$ spatial stations as shown in the right panel of Figure 3.1 where we index the spatial stations by raster scan. If the dependence decays exponentially with the underlying spatial distance, the covariance structure is shown as the left panel in Figure 3.1 (whiter pixels correspond to higher values, and redder corresponds to lower values).

Generally speaking, if we have $p = p_1 \times p_2$ spatial locations with $p_1$ rows and $p_2$ columns, and index them by raster scan order from 1 to $p$, the resulting $p \times p$ dimensional covariance matrix has $p_1 \times p_1$ blocks of sub-matrices. Each sub-matrix is of $p_2 \times p_2$ dimensions and is a bandable matrix Bickel and Levina [2008a], Cai et al. [2010]. In this way, the spatial observatory station in the $i^{th}$ row and $j^{th}$ column corresponds to the $[p_2(i - 1) + j]^{th}$ indexed element used for constructing the covariance matrix. Moreover, the values on the sub-diagonals of the blocks closer to the main diagonal tend to be greater than the values on the sub-diagonals of those blocks more apart from the main diagonal. We term this kind of matrices the block bandable matrices, of which the class is formally defined next. Thus, the block bandable matrices can be used to model the dependency structure of two-dimensional data that decays according to its spatial distance. It is worth noting that matrices that exhibit block structure plays a significant role in spatial modeling, such as the circulate embedding of random fields Lord et al. [2014].

Let $\Sigma_{p \times p}$ be the covariance matrix of dimensions $p \times p$ with $p = p_1p_2$. It consists of block sub-matrices $\Sigma^{st}$ with $1 \leq s \leq p_1$, and $1 \leq t \leq p_1$. Here $\Sigma^{st} = \{\sigma^{st}_{ij}\}_{1 \leq i,j \leq p_2}$ with $\sigma^{st}_{ij}$ being the element at the $i^{th}$ row and $j^{th}$ column of the block sub-matrix. We then focus the theoretical study on the following class of block bandable matrices:

\footnote{Such an ordering method is termed default ordering in Zhu and Liu [2009], which shows relatively better results than other ordering methods discussed therein.}
\[ \mathcal{B}_{\alpha,\beta}(M, \varepsilon) = \{ \Sigma : |\sigma_{ij}^{st}| \leq M(|i-j|^{-\alpha} \wedge |s-t|^{-\beta}), \forall 1 \leq i \neq j \leq p_2, 1 \leq s \neq t \leq p_1, \quad \varepsilon \leq \lambda_{\min}(\Sigma^{st}), \lambda_{\max}(\Sigma^{st}) \leq \frac{1}{\varepsilon}, \forall s, t \}, \] (3.1)

where \( \lambda_{\min}(A) \) and \( \lambda_{\max}(A) \) correspond to the minimal and maximal eigenvalue of a matrix \( A \). The \( \wedge \) denotes the minimum operator throughout the paper. The first condition reflects the bandable structure in both directions. The covariance decays, in the tail, at exponential rates bounded by \( \alpha \) and \( \beta \) respectively along the latitude and along the longitude. The second condition imposes the boundedness of the eigenvalues to ensure non-singularity. This class is an extension of the classes studied by Cai et al. [2010], Bickel and Levina [2008a], Cai and Yuan [2012], Xue and Zou [2013] which are suited for covariances among locations in an equidistant one-dimensional grid, while our class \( \mathcal{B}_{\alpha,\beta}(M, \varepsilon) \) is the corresponding class for the two-dimensional grid.

Figure 3.1: Left: Example of the spatial covariance matrix. Right: The two-dimensional stations are indexed in a raster scan order to build a one dimensional long vector.

### 3.2.2 Double Tapering Estimation

We assume that the observed normalized data \( \{X_i\}_{i=1}^n \) is drawn from a \( p \) dimensional normal distribution with zero mean and covariance matrix \( \Sigma \). The (maximal likelihood estimator)
sample covariance matrix is $\tilde{\Sigma}$. We define the double tapering covariance matrix $\hat{\Sigma}$, consisting of sub-block matrices $\hat{\Sigma}_{st}^{st}$, as

$$
\hat{\Sigma}_{st}^{st} = \{\omega\left(\frac{|i-j|}{k}\right)\tilde{\sigma}_{ij}^{st}\}_{1 \leq i,j \leq p^2}, \quad \hat{\Sigma} = \{\omega\left(\frac{|s-t|}{l}\right)\hat{\Sigma}_{st}^{st}\}_{1 \leq s,t \leq p^2},
$$

(3.2)

where $\omega(x)$ is a tapering function which is non-increasing and equals to one near zero, and zero when $x$ is large. The two-step tapering estimator corresponds to regularization in both dimensions, i.e. inside each sub-block matrix and across all sub-block matrices. Thus, we name it the double tapering estimator for the spatial block bandable matrix. For example, banding function could be defined as $\omega(x) = 1_{[0,\frac{1}{4}]}(x)$, while linear tapering function could be defined as follow,

$$
\omega(x) = \begin{cases} 
1, & \text{for } x < 0.5, \\
2 - 2x, & \text{for } 0.5 \leq x \leq 1, \\
0, & \text{for } x > 1. 
\end{cases}
$$

(3.3)

Other possible tapering and banding function $\omega(\cdot)$ are plotted in Figure 3.2.

![Figure 3.2: The tapering function $\omega(x)$ which is used in the double tapering estimator.](image)

---

2The tapering function $\omega(\cdot)$ is similar to the kernel function, but is not restricted to the unit integration.
3.3 Statistical Error

In this section, we first study the upper bound of the asymptotic minimax risk for the proposed double tapering estimator. Then the asymptotic minimax lower bound over the class of block bandable covariance matrices is derived, which shows the rate optimality of our proposed method. We consider the high-dimensional asymptotic case of all $n$, $p_1$ and $p_2$ increases to infinity.

3.3.1 Upper Bound

To derive the upper bound of the tapering estimator, we will adopt the so-called norm compression (in)equality (NCI) as a tool to deal with our block bandable matrix. NCI has been studied with various matrix norms King [2003], Audenaert [2006], which is of independent interest. Here we will use the Frobenius norm, the square-root of the sum of squares of every element in the matrix, denoted as $\| \cdot \|_F$. Let matrix $A$ be partitioned into $d_1 \times d_2$ blocks of sub-matrices $A_{ij}$ as:

$$
A = \begin{bmatrix}
A_{11} & A_{12} & \cdots & A_{1d_2} \\
A_{21} & A_{22} & \cdots & A_{2d_2} \\
& & \cdots & \\
A_{d_11} & A_{d_12} & \cdots & A_{d_1d_2}
\end{bmatrix}.
$$  \hfill (3.4)

If we replace each $A_{ij}$ with its (Frobenius) norm, then we get a “compressed” version of $A$, denoted by $N(A)$:

$$
N(A) = \begin{bmatrix}
\|A_{11}\| & \|A_{12}\| & \cdots & \|A_{1d_2}\| \\
\|A_{21}\| & \|A_{22}\| & \cdots & \|A_{2d_2}\| \\
& & \cdots & \\
\|A_{d_11}\| & \|A_{d_12}\| & \cdots & \|A_{d_1d_2}\|
\end{bmatrix}.
$$  \hfill (3.5)
The NCIs establish the relationship between $\|A\|$ and $\|N(A)\|$, which becomes an equality for the Frobenius norm.

**Lemma 12** (NCI for Frobenius norm). *Let $A$ be a matrix that could be partitioned as in (3.4), and $N(A)$ be its compressed matrix as defined in (3.5) with Frobenius norm. We have*

$$\|A\|_F = \|N(A)\|_F.$$  

(3.6)

**Proof.** Since $\|A\|^2_F$ is the sum of squares of every elements in the matrix while $\|A_{i_1i_2}\|^2_F$ is the sum of squares of every elements in $A_{i_1i_2}$, the $i_1^{th}$ and $i_2^{th}$ sub-matrix of the partitioned $A$, so obviously

$$\|A\|^2_F = \sum_{i_1=1}^{d_1} \sum_{i_2=1}^{d_2} \|A_{i_1i_2}\|^2_F = \|N(A)\|^2_F.$$  

(3.7)

This implies that the risk of the block bandable covariance matrix under Frobenius norm could be broken down into the sum of the risk corresponding to each individual small block. Based on this fact, we first derive the element-wise error and combine the results to get the risk upper bound under the Frobenius norm, which leads to the following theorem (proof in Appendix B.1).

**Theorem 13** (Upper Bound). *Let the double tapering estimator $\hat{\Sigma}_k$ of the covariance matrix $\Sigma_{p \times p}$ be defined in (3.2). Then we have the following upper bound for the asymptotic minimax risk over the block bandable matrices class $B_{\alpha, \beta}$ with $\alpha, \beta > 1$, as $n \to \infty$, $p_1 \to \infty$ and $p_2 \to \infty$,

$$\sup_{B_{\alpha, \beta}} \mathbb{E} \frac{1}{n} \|\hat{\Sigma} - \Sigma\|_F^2 \leq Cn^{-\frac{(2\alpha - 1)(2\beta - 1)}{4\alpha\beta - 1}} p_1^{\frac{2\beta - 1}{2\alpha\beta - 1}} p_2^{\frac{2\alpha - 1}{2\alpha\beta - 1}} \wedge \frac{p}{n}.$$  

(3.8)

**3.3.2 Lower bound**

In this part, we discuss the optimality, i.e. the minimax lower bound for the estimation over the class $B_{\alpha, \beta}$ of block bandable matrices. The proof applies a similar approach as in Cai
et al. [2010]. To show the above upper bound is optimal, we only need to show that the risk is also bounded below by $Ckl/n$ for such

$$l = n^{2\alpha-1}p_1^{-\frac{1}{4\alpha-1}}p_2^{\frac{2\alpha}{4\alpha-1}} \land \frac{p_1}{2}, \quad k = n^{2\beta-1}p_1^{-\frac{1}{4\alpha-1}}p_2^{\frac{2\beta}{4\alpha-1}} \land \frac{p_2}{2}.$$  

(3.9)

In this chapter, $C$ denotes a general positive constant. Thus the $C$ here can differ from the constant $C$ in other places including in Theorem 13.

We first describe a version of Assouad’s Lemma, a standard tool to get the minimax lower bound, whose proof is in Yu [1997], Tsybakov [2008], van der Vaart [1998]. Let $\Omega = \{\omega = (\omega_1, \cdots, \omega_N) : \omega_j \in \{0, 1\}\}$ be the set of binary sequences of length $N$. Let $\mathcal{P} = \{P_\omega : \omega \in \Omega\}$ be a set of $2^N$ distributions indexed by the elements of $\Omega$. Let $H(\omega, \nu) = \sum_{j=1}^{N} 1_{\omega_j \neq \nu_j}$ be the Hamming distance between $\omega, \nu \in \Omega$.

**Lemma 14** (Assouad van der Vaart [1998]). Let $\mathcal{P} = \{P_\omega : \omega \in \Omega\}$ be a set of distributions indexed by $\Omega$ and let $\theta(P)$ be a parameter. For any $m > 0$ and any metric $d$,

$$\max_{\omega \in \Omega} 2^m E_\omega [d^m(\hat{\theta}, \theta(P_\omega)))] \geq \frac{N}{2} \min_{H(\omega, \nu) \geq 1} d^m(\theta(P_\omega), \theta(P_\nu)) \cdot \min_{H(\omega, \nu) = 1} \|P_\omega \land P_\nu\|. \quad (3.10)$$

Here, $\|P \land Q\| := \lambda(p \land q)$ is the affinity (or the total variational distance) of two probability measures $P$ and $Q$ with density $p$ and $q$ with respect to common dominating measure $\lambda$ Pollard [2002].

To establish the minimax lower bound with Lemma 14, we construct a finite collection of normal distributions with the covariance matrices as a subset of $B_{\alpha, \beta}$, and separately bound the Frobenius norm and the affinity on this subset. The $\sigma_{ij}^{st}$ and $\sigma_{ji}^{ts}$ in the covariance matrix represent the symmetric elements and have the same value. Therefore, we consider the following class of covariance matrices where we perturb each pair by a fixed amount $\tau$.

$$\mathcal{B}_2 = \{\Sigma(\theta) : \Sigma(\theta) = I_p + (\theta_{ij}^{st} - \frac{1}{\sqrt{n}} 1_{1 \leq |i-j| \leq k, 1 \leq |s-t| \leq l})_{p \times p} \land \theta_{ji}^{ts} = 0 \text{ or } 1, \forall 1 \leq s, t \leq p_1, 1 \leq k, l \leq p_2\}, \quad (3.11)$$
where the binary $\theta_{st}^{ij}$ indicates whether the $(i, j)^{th}$ element in the $(s, t)^{th}$ sub-block matrix is perturbed, and $0 < \tau < M$. As $n \to \infty$ and for $k, l$ in (3.9), it is easy to see that $B_2 \subset B_{\alpha, \beta}$, allowing us to establish the asymptotic minimax rate. Thus, in terms of Lemma 14, we have $\theta \in \Omega = \{0, 1\}^N$ with $N \asymp klp_1p_2$. Here and in the following, the $\asymp$ denote that the asymptotic order is the same. That is, $\xi_n \asymp \eta_n$ if there exist a positive constant $\epsilon$ such that $\epsilon \leq \xi_n / \eta_n \leq 1 / \epsilon$. Since any two elements $\theta$ and $\theta'$ in $\Omega$ differs in exactly $H(\theta, \theta')$ components, we have

$$
\|\Sigma(\theta) - \Sigma(\theta')\|_F^2 = \frac{\tau^2}{n} \sum_{s,t,i,j} |\theta_{st}^{ij} - (\theta')_{st}^{ij}|^2 = \frac{\tau^2}{n} H(\theta, \theta')
$$

(3.12)

To bound the affinity, we first relate it to Kullback-Leibler (KL) divergence by the following Lemma (see Tsybakov [2008] for proof).

**Lemma 15.** Let $KL(P, Q)$ be the KL divergence between two probability measures $P$ and $Q$, we have

$$
\|P \wedge Q\| \geq \frac{1}{2} e^{-KL(P, Q)}.
$$

(3.13)

Using this, we have the following lower bound for the affinity (proof in Appendix B.2).

**Lemma 16.** Let $P_\theta \sim N(0, \Sigma(\theta))$ be the joint distribution of i.i.d sample $\{X_i\}_{i=1}^n$ with $\Sigma(\theta) \in B_2$. Then

$$
\min_{H(\theta, \theta')} \|P_\theta \wedge P_{\theta'}\| \geq C_1.
$$

(3.14)

Using Frobenius norm as the metric in Lemma 14 and $m = 2$, combining equations (3.12) and (3.14) with $C = C_1 \tau^2$, we get the lower bound on risk.

**Theorem 17.** The asymptotic minimax risk for estimating the covariance matrix $\Sigma$ over $B_2$ under the Frobenius norm satisfies, as $n \to \infty$ and for $k, l$ in (3.9),

$$
\inf_{\hat{\Sigma}} \sup_{\Sigma} \frac{1}{p} \left\| \hat{\Sigma} - \Sigma \right\|_F^2 \geq C \frac{kl}{n},
$$

(3.15)

for some $C > 0$. 73
Combining Theorem 13 and Theorem 17, we have the exact optimal minimax risk rate.

**Corollary 18.** The minimax risk for estimating the covariance matrix $\Sigma$ under Frobenius norm satisfies

$$\inf_{\Sigma} \sup_{\mathcal{B}_{\alpha,\beta}} \mathbb{E} \frac{1}{p} \| \Sigma - \hat{\Sigma} \|^2_F \asymp n^{-\frac{(2\alpha-1)(2\beta-1)}{4\alpha\beta-1}} p_1^{\frac{2\beta-1}{4\alpha\beta-1}} p_2^{\frac{2\alpha-1}{4\alpha\beta-1}} \wedge \frac{p}{n}. \tag{3.16}$$

And this optimal minimax risk rate is achieved by the double tapering estimator.

The tapering in each direction should be adjusted according to the decay rate in that direction to achieve the optimal rate. The double tapering estimator is then rate optimal over the class $\mathcal{B}_{\alpha,\beta}$ when $\alpha, \beta > 1$. For the simple case of $\alpha = \beta$ (the decay rates are the same along the latitude as along the longitude), the optimal minimax risk rate is $O(n^{-\frac{(2\alpha+1)^2}{4(\alpha+1)^2-1}} p_1 \frac{2\alpha+1}{4(\alpha+1)^2-1})$, which only depends on the total size of the grid $p = p_1 p_2$ but not on the specific values of $p_1$ and $p_2$.

### 3.4 Experimental Results

In this section, we study the numerical performance of our estimators. We first apply our method on synthetic data to demonstrate its performance under various decay rates, feature dimensions and the sample size. Then, we test it with real data in an application of image compression using principal component analysis (PCA) and a climate application with statistical downscaling.

#### 3.4.1 Synthetic Data

The simulation of block bandable covariance matrices is based on the class $\mathcal{B}_{\alpha,\beta}$ in (3.1). Specifically, let the true covariance matrix $\Sigma$ have the form

$$\sigma_{i,j}^{st} = \begin{cases} M(|i-j|^{-\alpha} \wedge |s-t|^{-\beta}), & i \neq j \text{ or } s \neq t \\ 2.5, & i = j, s = t \end{cases} \tag{3.17}$$
where $M = 1, \alpha = \beta = a$. The multivariate Gaussian random data is generated with mean zero and covariance matrix $\Sigma$. The rate of the parameters $k$ and $l$ are decided by the result in equation (3.9). Each of the simulation study is based on the average results from 100 replications.

In particular, we are interested in the relative performance of the regularized estimators (banding and tapering) under the Frobenius norm against the sample covariance matrix. Several factors are considered in our simulation study including the decay rate ($a$), dimension ($p$) and sample size ($n$). We focus on the square cases where $p_1 = p_2$, thus $p = p_1p_2 = p_1^2$.

**Decay Rate.** We consider the decay rate $a = 1.1, 1.2, 1.3, 1.4$ and 1.5. The error for each setting is measured by the Frobenius norm of the difference between the (regularized) estimators and the true covariance matrix. The relative error, i.e. the ratio of the error for regularized estimators with respect to the sample covariance matrix is considered. Thus, smaller relative error (less than one) implies better performance than the sample covariance matrix.

Result of this comparison is presented in Figure 3.3. The sample size is set to be $n = 1000$, while the dimension is $p = 3600$. The red bar corresponds to the sample covariance matrix to itself, which is always one, while the blue and green bar correspond to the (double) banding and (double) linear tapering estimators, which reduce the relative error of the sample covariance matrix. Thus, both the banding and linear tapering estimator significantly improve the performance of the sample covariance matrix.

**Dimension.** We now compare cases where $p_1 = p_2 = 10, 20, 30, 40, 50, 60$, thus $p = 100, 400, \ldots, 3600$. Result of this comparison is presented in Figure 3.4. The sample size is set to be $n = 1000$, while the decay rate is $a = 1.5$. The red dashed line corresponds to the sample covariance matrix to itself, which is one, while the blue and green lines correspond to the banding and linear tapering estimators respectively. As we can see from the plot, both
Figure 3.3: Relative error (w.r.t sample covariance matrix) for the banding and linear tapering estimator. The comparison is against the decay rate $a$.

the banding and linear tapering estimator improve the performance of the sample covariance matrix more as the dimension increases.

**Sample Size.** Similar to the previous settings, we consider sample size $n$ ranging from 250 to 3000. Result of this comparison is presented in Figure 3.4. The dimension is set to be $p = 3600$, while the decay rate is $a = 1.5$. The red dashed line corresponds to the sample covariance matrix to itself, while the blue and green lines correspond to the banding and linear tapering estimators respectively, which greatly reduce the relative error of the sample covariance matrix.

A more comprehensive simulation study results are reported in the supplemental materials. The overall result shows the effectiveness of our double tapering estimator.

### 3.4.2 Application on Image Compression

Image compression aims to reduce the size of the image data in order to store or transmit data efficiently. Principal component analysis is a classical way to perform dimensionality reduction and to compress images. In this part, we compare the performance of the image
Figure 3.4: Relative error (w.r.t. sample covariance matrix) for the banding and linear tapering estimators. The comparison is on the left: against the dimension $p$; on the right: against the sample size $n$.

compression and reconstruction results using principal component analysis (PCA) based on the sample covariance matrix versus the results using PCA based on the proposed double tapered covariance estimator.

As an example, we consider the Olivetti face data\(^3\) which contains 400 face images of 40 persons and each image is of size 64 by 64. We use 40% of the dataset for training, and the rest 60% dataset for testing.

In the training period, we estimate the mean of the features and the principal subspace spanned by the top $q$ principal components based on the training data. For the proposed double tapering estimator, the parameters $k = 85$ and $l = 9$ are chosen via five-fold cross-validation within the training data. In the testing period, we compressed the testing images by projecting the data onto the trained principal subspace. The compressed images are then reconstructed with the top $q$ principal components and the feature mean. The reconstruction error is based on the Frobenius norm between the reconstructed images and the true testing

\(^3\)This dataset contains a set of face images taken at AT&T Laboratories Cambridge, which is available via the following link [scikit-learn.org/stable/datasets/olivetti_faces.html](http://scikit-learn.org/stable/datasets/olivetti_faces.html)
3.4.3 Application on Climate Data

In this part, we apply our proposed spatial covariance estimation in a real climate task, i.e. statistical downscaling [Benestad et al., 2008, 2015]. It is generally believed that the Global Climate Models (GCMs) provide coarse resolution outputs which preclude their application to accurately assess the effects of climate change on finer regional scale events. Statistical downscaling are methods that use statistical models to infer the local-scale climate informa-
Figure 3.6: Reconstruction error on the testing set as a function of the number of principal component kept. The double tapered covariance matrix leads to lower reconstruction error compared with the sample covariance matrix.

In this experiment, we consider the reanalysis monthly mean temperature data from the National Oceanic & Atmospheric Administration (NOAA) \(^4\) as the coarse resolution data, which includes grid points from the world atlas, while the finer resolution temperature data is from the University of Idaho Gridded Surface Meteorological Data (UofI METDATA) \(^5\), which contain the data of contiguous united states.

We use the common time range, years 1979-2013, from the two data sets for training and testing. We choose years 1979-2008 data as the training set, and years 2009-2013 as the testing set. We normalize the data by removing the historical long range mean for each month. The spatial covariance matrix is obtained by the sample covariance between each grid point. Moreover, the double-tappered spatial covariance can be derived based on previous

\(^4\)https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html
\(^5\)http://metdata.northwestknowledge.net/
discussion. Since the features can be highly correlated in this task, we apply hierarchical clustering method to cluster the data and using the center of each cluster as the derived features for the benchmark predictive model (10-fold cross-validated LASSO). As we can see from Figure 3.8 and 3.9, result shows that the predict performance (testing MSE) based on the double-tappered spatial covariance matrix is better than the one with the sample covariance for both season-wise and overall comparison.

Figure 3.7: Example of the statistical downscaling task. Left: the climate variable (monthly mean temperature in Kelvin) with a coarse resolution at a single time point (Jan. 2009). Right: the target climate variable with a relatively finer resolution. The goal of statistical downscaling is to learn the target climate variable in finer resolution based on the historical data of the fine and coarse resolution climate data (red is larger and black is smaller).

Figure 3.8: Downscaling result of the target climate variable in the testing set. The intensity (red is larger and black is smaller) is the average MSE in each grid point.
3.5 Conclusions

With the large amount of spatial data becoming available, high-dimensional spatial covariance matrix estimation is becoming more important for data analysis. Here we provide the first minimax rate optimality results of tapering estimators in this high-dimensional setting. To represent the spatial structure, we propose the block bandable matrices structure to incorporate spatial information. Accordingly, a double tapering estimator is proposed and is shown to achieve the optimal minimax rate over a class of block bandable covariance matrices. Numerical study confirms the performance of the proposed estimator.
Chapter 4

Far-Near Covariance Model

4.1 Introduction

Covariance matrix plays a significant role in statistics and machine learning, especially in the study of multivariate data, and has wide applications ranging from discriminant analysis [Anderson, 2003], portfolio allocation [Markowitz, 1952], gene expression analysis, and Gaussian graphical models [Wang et al., 2016]. In recent years, there have been many developments to improve estimation accuracy over the sample covariance matrix in the high dimensional setting [Pourahmadi, 2014, Cai et al., 2016, Fan et al., 2016]. For instance, Cai et al. [2010], Cai and Yuan [2012] discuss the estimation of bandable covariance matrices, where the true covariance has high values concentrated near the diagonal. Xiao and Wu [2012], Cai et al. [2013] study the Toeplitz matrices, which is closely related to stationary time series. In order to facilitate the analysis of the theoretical and empirical properties of the data, most of the aforementioned works imposes certain structural condition for the data, which result in various sparse covariance matrix models. These methods for sparse covariance matrices are useful in spatial statistics. For example, regularization methods like tapering have been studied in spatial statistics, with the emphasize on the computational issue, see Kaufman [2008], Shaby and Ruppert [2012] and the references therein. However,
the structural conditions studied in these sparse covariance matrix models do not use the spatial distance, thus not fully taken the spatial information into account.

The covariance structure has been studied in spatial statistics, especially in climatology. For example, the spatial covariance matrix is useful in both supervised regression problem [Cressie, 1993, Christensen, 2002, Montero et al., 2015] and unsupervised problem [Benestad et al., 2015], as well as other areas like spatial econometrics [LeSage and Pace, 2009]. Traditional covariance model use some parametric form to model that the covariance decreases as the spatial distance between location increases. However, such spatial covariance model may not model enough the complexity of the covariance in real data. For instance, the term “butterfly effect” originated from the climate studies: the climate variable in one location may have strong dependence on certain far away locations. The sparse strong correlations between far away locations are called teleconnections and are attracting interest in climate studies [Magrin et al., 2007, Choi et al., 2015, Tsonis et al., 2008]. Also, there has been an emerging research interests in the climate community to apply complex network [Watts and Strogatz, 1998] in analyzing climate data [Scarsoglio et al., 2013]. Such climate network analysis [Donges et al., 2009, Steinhaeuser et al., 2010, 2011, 2012] needs to consider both short range and long range correlations. However, the climate network analysis currently is built on sample covariance matrix which does not account for spatial distance. There has not been literature on modelling the dense short range dependence and the sparse long range dependence.

In this paper, we propose a far-near covariance model (FNCM) that simultaneously model the short range (near) and long range (far) dependency from the data. A key insight is that, due to the near covariance, any far covariance between a pair of far apart locations will also propagate to the neighborhoods around the pair of locations. Taking such information into account, we can derive an estimator of the long range dependence signal that is more accurate than the simple sample covariance estimator.
Since the long range dependence signals are often sparse, the detection of such teleconnections can use the statistical tests on the high dimensional covariance matrix [Cai, 2017]. Particularly, the $\tau$-coherence [Donoho, 2006] can be used to test existence of non-zero correlated pairs of locations at least $\tau$ distance units apart. The statistical properties of the $\tau$-coherence, the largest entries that is at $\tau$ away from the main diagonal, of the high-dimensional covariance matrix is studied in [Jiang, 2004, Cai and Jiang, 2011]. Combining our proposed far covariance estimator from the FNCM with the $\tau$-coherence testing, we can get an improved test for the existence of sparse teleconnections. We apply our proposed methods in the climate study with the application of teleconnection discovery. By utilizing near covariance information, we can reduce spurious teleconnection detection.

The rest of the paper is organized as follows. Section 4.2 describes the proposed far-near covariance model, and then derives an optimal linear unbiased estimator for the far covariance signal strength. Estimators of other model parameters are discussed, and the far covariance estimator are used for testing of teleconnections. Statistical properties of the proposed procedure is derived in the high-dimensional setting. Section 4.3 illustrate the performance of the proposed procedures using numerical examples. The improvement in estimation accuracy and test power are demonstrated on synthetic simulated data. We also apply the procedure for teleconnection detection on a climate data set.

4.2 Methodology

In this part, we first introduce our far-near covariance model as a factor analysis model for locations along an one-dimensional line. Then we show that, by utilizing the near covariance, the far covariance can be estimated and detected better than by methods using the sample covariance directly. We then extend the model to include two-dimensional spatial distances.
4.2.1 The Far-Near Covariance Model

Suppose that we have \( p \) locations, \( i = 1, 2, ..., p \), and the observed response variables are Gaussian random variables \( X_1, \cdots, X_p \). Assuming the responses have mean \( 0 \) and covariance \( \Sigma \), the information on the correlations among different locations are contained in this \( p \times p \) covariance matrix \( \Sigma \). To consider spatial effect, the distances among the locations should be available given the indices, \( i = 1, 2, ..., p \). For simplicity of presentation, we first focus on the one-dimensional equidistant case where the distance between two locations indexed by \( i \) and \( j \) is proportional to the indices difference \( |i - j| \). The model with two-dimensional locations will be presented afterwards in section 4.2.5.

We will model the far-near covariance using a factor analysis model. The factor analysis is a widely used tool for multivariate analysis [Anderson, 2003]. Many high dimensional statistics problems, such as covariance estimation and spiked PCA, can be written in terms of the factor analysis model. Specifically, we model the response variables \( X_1, \cdots, X_p \) as driven by latent factors \( f_1, \cdots, f_p \) and are observed with additive noises. Let \( X_{p \times 1} = (X_1, \cdots, X_p)^\top \), \( f = (f_1, \cdots, f_p)^\top \) and \( \epsilon_{p \times 1} = (\epsilon_1, \cdots, \epsilon_p)^\top \) be respectively the vectors of observed variables, latent factors and random noises. Then

\[
X = \mu + Af + \epsilon, \tag{4.1}
\]

where \( \mu \) is the mean, and \( A_{p \times p} \) is the matrix of factor loadings. The main idea is to impose the structure on the factor loadings \( A_{p \times p} \) to represent and to clearly delineate the short range and the long range dependence. The factors are standardized to have mean \( \mathbb{E}[f] = 0_p \) and variance \( \mathbb{V}[f] = I_p \), where \( 0_p \) denotes the \( p \)-dimensional zero vector and \( I_p \) denotes the \( p \times p \) identity matrix. The factors are independent of the random noise \( \epsilon \). The noise \( \epsilon \) is assumed to have i.i.d. entries with mean zero and common variance \( \sigma_\epsilon^2 \), i.e., \( \mathbb{V}[\epsilon] = \sigma_\epsilon^2 I_p \). For simplicity of presentation, we assume that \( X \) is centered to have mean zero, thus \( \mu \) is a zero vector.
Given the factor analysis model (4.1), the true covariance $\Sigma$ of $X$ is

$$\Sigma = AA^\top + \mathbb{V}[\epsilon] = AA^\top + \sigma^2 I_p. \quad (4.2)$$

The factor loading matrix $A$ introduces a structure on the covariance matrix $\Sigma$, and we should estimate $\Sigma$ using this structural information. We now consider how to model both far (long range) and near (short range) covariance using the factor loading matrix $A$.

First we consider modelling the very short range of dependence, where the factor $f_i$ affects not only the $i$-th location but the neighboring $(i - 1)$-th and $(i + 1)$-th locations. That is, the model in (4.1) at the $i$-th location can be rewritten as

$$X_i = af_{i-1} + f_i + af_{i+1} + \epsilon_i, \forall i \in \{1, \cdots, p\}, \quad (4.3)$$

where $a$ is a parameter indicating the strength of the neighbourhood effect. For convenience of presentation, we denote $f_j = 0$ for $j$ values outside the index range $\{1, 2, \cdots, p\}$. For example, $f_0 = f_{p+1} = 0$.

Now we consider modelling the long range dependence. The long range dependence usually are sparse and occur only at some specific pair $(i^*, j^*)$ with $|i^* - j^*| > 2$. A natural way of considering the long range dependence is to let the two locations share a common factor. For the correlated two locations, instead of the two independent factors $f_{i^*}$ and $f_{j^*}$ (each of mean zero and variance one), we may consider the two dependent factors $f_{i^*}$ and $\tilde{f}_{j^*}$ (each of mean zero and variance one) with covariance $\rho$. Then

$$\text{Cov}(f_{i^*}, \tilde{f}_{j^*} - \rho f_{i^*}) = \text{Cov}(f_{i^*}, \tilde{f}_{j^*}) - \rho \text{Cov}(f_{i^*}, f_{i^*}) = \rho - \rho(1) = 0.$$

That is, we can consider $\rho f_{i^*}$ as the common factor for the two locations, and recover the factor analysis model by denoting $f_{j^*} = (\tilde{f}_{j^*} - \rho f_{i^*})/\sqrt{1 - \rho^2}$ which also has mean zero and variance one, and are independent of the other factors. Thus the latent factor at the $j^*$-th location is in fact $\tilde{f}_{j^*} = \rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*}$. More precisely, the far-near covariance model at the $j^*$-th location is

$$X_{j^*} = af_{j^*-1} + [\rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*}] + af_{j^*+1} + \epsilon_{j^*}. \quad (4.4)$$
Since the latent factor at the \( j^* \)-th location also affect its neighbor, say the \((j^* - 1)\)-th location, through near covariance, we have

\[
X_{j^*-1} = af_{j^*-2} + f_{j^*-1} + a[\rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*}] + \epsilon_{j^*-1}.
\]  

(4.5)

Therefore, the Far-Near Covariance model will have the factor analysis model form as

\[
\begin{align*}
X_i &= af_{i-1} + f_i + af_{i+1} + \epsilon_i, \quad \forall |i - j^*| > 1, \\
X_{j^*-1} &= af_{j^*-2} + f_{j^*-1} + a[\rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*}] + \epsilon_{j^*-1}, \\
X_{j^*} &= af_{j^*-1} + [\rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*}] + af_{j^*+1} + \epsilon_{j^*}, \\
X_{j^*+1} &= a[\rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*}] + f_{j^*+1} + af_{j^*+2} + \epsilon_{j^*+1}.
\end{align*}
\]  

(4.6)

When \( \rho = 0 \), there is no long range dependence signal, and (4.6) reduces back to (4.3).

With both far and near dependence explicitly modeled for the factor loading matrix \( A \) in (4.1), we have the structured covariance matrix \( \Sigma \) from equation (4.2) that contains the effects of both types of dependence. A main objective of the statistical analysis is to detect the sparse long range dependence signal, i.e., to distinguish

\[
H_0 : \quad \text{The model comes from (4.3)},
\]

\[
H_1 : \quad \text{The model comes from (4.6)}.
\]  

(4.7)

Such information on the long range dependence is crucial in various applications, for example, market efficiency in financial market, the existence of teleconnection in climatology, and long memory of time series, etc. We now use a specific example to illustrate the structured covariance matrix \( \Sigma \) resulting from this FNCM model.

**Example 19.** Consider \( p = 10 \) locations with their indices differences indicating the distances between locations. Besides the near dependence, the two locations at indices \( i^* = 2 \) and \( j^* = 9 \) are also dependent (teleconnected). Then the load matrix can be represented as \( A = B + E_{i^*,j^*} \) where \( B \) is a tri-diagonal matrix and \( E_{i^*,j^*} \) is only supported on a submatrix.
centered at \((i^*, j^*)\):

\[
\begin{pmatrix}
X_1 \\
X_2 \\
X_3 \\
X_4 \\
X_5 \\
X_6 \\
X_7 \\
X_8 \\
X_9 \\
X_{10}
\end{pmatrix} = \begin{pmatrix}
1 & a \\
a & 1 & a \\
a & 1 & a \\
a & 1 & a \\
a & 1 & a \\
\rho & a & \sqrt{1-\rho^2} \\
\rho & a & \sqrt{1-\rho^2} \\
\rho & a & \sqrt{1-\rho^2} \\
\rho & a & \sqrt{1-\rho^2} \\
\rho & a & \sqrt{1-\rho^2}
\end{pmatrix} \begin{pmatrix}
f_1 \\
f_2 \\
f_3 \\
f_4 \\
f_5 \\
f_6 \\
f_7 \\
f_8 \\
f_9 \\
f_{10}
\end{pmatrix} + \begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\epsilon_6 \\
\epsilon_7 \\
\epsilon_8 \\
\epsilon_9 \\
\epsilon_{10}
\end{pmatrix} \tag{4.8}
\]

Note that \(\rho\) is the portion of teleconnected factor \(f_2\) contributing to the observed variable \(X_9\) at the 9-th location, and \(f_2\) also contribute to nearby variables \(X_8\) and \(X_{10}\). Given equation (4.8), the covariance matrix \(\Sigma\) for this example is

\[
\begin{pmatrix}
a^2 + 1 & 2a & a^2 & a^2\rho & a\rho & a^2\rho \\
2a & 2a^2 + 1 & 2a & a^2 & a\rho & \rho & a\rho \\
a^2 & 2a & 2a^2 + 1 & 2a & a^2 & a^2\rho & a\rho & a^2\rho \\
a^2 & 2a & 2a^2 + 1 & 2a & a^2 & a^2\rho & a\rho & a^2\rho \\
a^2 & 2a & 2a^2 + 1 & 2a & a^2 & a^2\rho & a\rho & a^2\rho \\
a^2 & 2a & 2a^2 + 1 & 2a & a^2 & a^2\rho & a\rho & a^2\rho \\
a^2 & 2a & 2a^2 + 1 & 2a & a^2 & a^2\rho & a\rho & a^2\rho \\
a^2 & 2a & 2a^2 + 1 & 2a & a^2 & a^2\rho & a\rho & a^2\rho \\
\epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 & \epsilon_7 & \epsilon_8 & \epsilon_9 & \epsilon_{10}
\end{pmatrix} + \sigma_e^2 I_p. \tag{4.9}
\]

The near covariance induces a band of non-zero entries around the main diagonal of \(\Sigma\), while the far covariance induces non-zero off-diagonal entries away from this band. Note that the long range dependence between \((i^*, j^*)\) in (4.8) has been propagated to their neighbours due to the short range dependency. For example, the covariance between \(X_1\) and \(X_9\) is \(a\rho\), which is weaker than the major teleconnection pair \((i^*, j^*)\) but stronger than the one with \(X_1\)
and $X_s$. Due to the interaction between far and near covariance, the teleconnection does not appear only at a pair of isolated points in $\Sigma$, but also at small squared regions caused by the short range dependence.

As we can see from equation (4.9), the covariances at nearby locations around the major teleconnection pair $(i^*, j^*)$ also contains information on the far covariance strength $\rho$ under the FNCM. Next we discuss how to utilize the information at surrounding locations to improve the estimation for $\rho$.

**4.2.2 Estimation of the Far Covariance $\rho$**

As mentioned in the introduction, the detection and estimation of the far covariance (teleconnection) is a major goal of the statistical analysis. Prior methods are based on sample covariance matrix $\hat{\Sigma}$. That is, the sample covariance $\hat{\sigma}_{i^*,j^*}$ for the pair of $(i^*, j^*)$-th locations is used to estimate $\rho$. As we just observed in (4.9), under the FNCM, there is a local structure on the covariance in the neighborhood of $\sigma_{i^*,j^*}$. Thus we will combine the entries adjacent to $\hat{\sigma}_{i^*,j^*}$ in the sample covariance matrix with $\hat{\sigma}_{i^*,j^*}$ through a linear combination. The resulting weighted estimator, by using the additional information propagated from the neighbourhood, will have smaller variance than $\hat{\sigma}_{i^*,j^*}$, and be a more accurate estimator for $\sigma_{i^*,j^*} = \rho$. We derive the optimal weights for such type of linear estimators here. Then this improved estimator will also be used later to improve the statistical power for detecting teleconnection.

In this section, we derive the optimal weights assuming that, $a$, the strength of neighborhood effect is known. In practice, $a$ will be first estimated by $\hat{a}$ as described in Section 4.2.3. We also assume that the locations of teleconnected pair, $(i^*, j^*)$, are known. Later we can use this estimator to test for non-zero $\rho$ at each far $(i, j)$ pair to detect existence of teleconnection and its location.

We put the sample far covariance $\hat{\sigma}_{i^*,j^*}$ and its four adjacent elements (as shown in Fig-
Figure 4.1: Example of the adjacent sample covariance elements that are combined linearly to estimate $\sigma_{i^*,j^*}$, where $\bar{\sigma}_1$ is $\hat{\sigma}_{i^*,j^*}$ in the middle, and the other four are defined clockwise starting from the top.

We focus on the unbiased linear estimators, that is, $\lambda^\top \mu = 1$, where $\mu = (1, a, a, a, a)^\top$ due to (4.9). Let $\tilde{\Sigma} := cov(\bar{\sigma})$ denote the 5 by 5 covariance matrix of the selected sample covariance elements in $\bar{\sigma}$. Then the variance of the linear estimator $\bar{\sigma}_{i^*,j^*}$ is $\lambda^\top \tilde{\Sigma} \lambda$. The best unbiased linear estimator should have smallest variance, thus its weights are obtained by solving the following mean-variance quadratic optimization problem:

$$\min_{\lambda \in \mathbb{R}^5} \frac{1}{2} \lambda^\top \tilde{\Sigma} \lambda, \quad s.t. \quad \lambda^\top \mu = 1. \quad (4.10)$$

This is a standard quadratic programming problem and the following Theorem is a direct consequence of the above optimization problem. The detailed proofs can be found in Appendix C.1.

**Theorem 20.** The best unbiased linear estimator of $\sigma_{i^*,j^*}$ (based on $\bar{\sigma}$) is

$$\bar{\sigma}_{i^*,j^*}^* = \lambda^*\top \bar{\sigma}, \quad (4.11)$$

where $\lambda^*$ is the solution of the mean-variance problem (4.10),

$$\lambda^* = \frac{\Sigma^{-1} \mu}{\mu^\top \Sigma^{-1} \mu}. \quad (4.12)$$
where $\Sigma$ is the covariance matrix of $\tilde{\sigma}$.

The optimal weights formula (4.12) can be made more explicit for the FNCM (4.6) through calculation of $\Sigma$. In fact, the covariance $\Sigma$ of the five sample covariance elements in $\bar{\sigma}$ has the following form:

$$
\Sigma = \begin{bmatrix}
d_1 & r_1 & r_1 & r_1 \\
r_1 & d_2 & s_1 & s_1 \\
r_1 & s_1 & d_2 & s_1 \\
r_1 & s_2 & s_1 & d_2 \\
r_1 & s_1 & s_2 & s_1 \\
\end{bmatrix},
$$

(4.13)

where $d_1, d_2, r_1, s_1, s_2$ are some functions of $a, \rho$ and $\sigma_\epsilon$, whose detailed formulas are provided in equation (C.10) of Appendix C.3. Hence the optimal weights $\lambda^*$ in (4.12) should be a function of $d_1, d_2, r_1, s_1, s_2$, which is summarized in the Corollary 21.

**Corollary 21.** Assume that (4.13) holds. Then $\lambda^* = (1 - 4a\lambda^*_2, \lambda^*_2, \lambda^*_2, \lambda^*_2, \lambda^*_2)^\top$ where

$$
\lambda^*_2 = \frac{ad_1 - r_1}{d_2 + 2s_1 + s_2 - 8ar_1 + 4a^2d_1}.
$$

(4.14)

Moreover, we have

**Corollary 22.** Let $X$ follows the FNCM (4.6). Then $\lambda^* = (1 - 4a\lambda^*_2, \lambda^*_2, \lambda^*_2, \lambda^*_2, \lambda^*_2)^\top$ where

$$
\lambda^*_2 = \frac{4a^5 - a + 4a^3\sigma^2_\epsilon + a\sigma^4_\epsilon}{16a^6 - 10a^4 + a^2 + 1 + (16a^4 - 3a^2 + 2)\sigma^2_\epsilon + (4a^2 + 1)\sigma^4_\epsilon}.
$$

(4.15)

We plot the optimal weights $\lambda^*$, namely, $\lambda^*_1 := 1 - 4a\lambda^*_2$ (solid) and $\lambda^*_2$ (dashed) in the left panel of Figure 4.2 as a function of $a$, where thicker line corresponds to higher noise level $\sigma_\epsilon$. Naturally for stronger near covariance strength $a$, our estimator $\bar{\sigma}^*_i, j^*$ (Theorem 20) will put more weights on adjacent sample covariance elements. The relationship between the optimal weights and the additive noise level $\sigma_\epsilon$ is more complex, depending on the strength of near covariance strength $a$. 

91
Due to the optimization, our estimator \( \tilde{\sigma}_{i^*j^*} \) has smaller variance than the standard estimator of the sample covariance \( \hat{\sigma}_{i^*j^*} \). We denote the variance reduction ratio as

\[
\varrho_{rr} = \frac{\lambda^{*\top} \tilde{\Sigma} \lambda^*}{\lambda_0^{\top} \Sigma \lambda_0},
\]

where \( \lambda_0 = (1, 0, 0, 0, 0)^\top \) corresponds to the weights for the sample covariance estimator \( \hat{\sigma}_{i^*j^*} \). The right panel of Figure 4.2 plots the reduction ratio \( \varrho_{rr} \) against the near covariance strength \( a \), where different colors represent different levels of \( \sigma_\epsilon \). As we can see from the right panel of Figure 4.2, the reduction becomes more effective as the nearby relationship becomes stronger (bigger \( a \) value).

### 4.2.3 Estimation of Other Parameters

The optimal weight expression (4.15) involves two other parameters: near covariance strength \( a \) and noise level \( \sigma_\epsilon \). Up to now we have assumed that they are known. In practice, these parameters should also be estimated.
We first consider the estimation of the near covariance strength $a$ under the FNCM (4.6). Notice from equation (4.9), the sub-diagonal elements of the covariance matrix $\Sigma$ are determined by $a$: $\sigma_{i,j} = 2a$, for $|i - j| = 1$ and $\sigma_{i,j} = a^2$, for $|i - j| = 2$, $\forall i, j = 1, \cdots, p$. Let us consider two quantities based on the corresponding sub-diagonal elements of the sample covariance matrix $\hat{\Sigma}$:

$$\bar{k}_1 = \frac{\sum_{i,j=1}^{p} \hat{\sigma}_{ij}}{p - 1}, \quad \bar{k}_2 = \frac{\sum_{i,j=1}^{p} \hat{\sigma}_{ij}}{p - 2}. \quad (4.17)$$

Then $\bar{k}_1$ and $\bar{k}_2$ are unbiased estimators for $2a$ and $a^2$ respectively. Although we can use either one of them to estimate $a$, a natural scale-invariant estimator of $a$ is based on the ratio between these two quantities:

$$\hat{a} = \frac{2\bar{k}_2}{\bar{k}_1}. \quad (4.18)$$

**Remark 23.** We have assumed that the data is appropriately centered and scaled so that the factor analysis model (4.6) holds where each latent factor $f_i$ has mean zero and variance one. In practice, the raw data may not be measured at that scale and may differ from the model by a constant multiple which is unknown to the user. One common method is to apply statistical procedures on normalized data so that each $X_i$ has mean zero and variance one. Then the sample covariance matrix $\hat{\Sigma}$ of the normalized data becomes the sample correlation matrix of the raw data. The scale-invariant estimator $\hat{a}$ in (4.18) would give the same value using the sample covariance matrix of the raw data or the normalized data. Hence we can derive its statistical distribution assuming the correct scale is already applied so that (4.6) holds.

**Theorem 24.** Let $X_{n \times p}$ be the data matrix for $p$ dimensional multivariate Gaussian distribution with mean zero and covariance $\Sigma$, and $X_{n \times p}$ follows the far-near covariance model (4.6) where $a > 0$ and $n$ is the sample size. Then, for $\hat{a}$ defined in (4.18), as $n \to \infty$, almost surely

$$|\hat{a} - a| = O_p\left(\frac{1}{\sqrt{np}}\right). \quad (4.19)$$
The proof of Theorem 24 is provided in Appendix C.5. Theorem 24 states that the near covariance parameter is well estimated by $\hat{a}$, particularly in the high-dimensional case where $p$ is also large.

Furthermore we estimate $\sigma_\epsilon$ with

$$\hat{\sigma}_\epsilon = \sqrt{k_0 - 2\hat{a}^2 - 1},$$

(4.20)

where $\bar{k}_0 = \frac{1}{p-2} \sum_{i=2}^{p-1} \hat{\sigma}_{ii}$. Similar arguments as in the proof of Theorem 24 shows that $|\hat{\sigma}_\epsilon - \sigma_\epsilon| = O_p\left(\frac{1}{\sqrt{np}}\right)$ almost surely as $n \to \infty$.

**Remark 25.** By plug-in $\hat{a}$ and $\hat{\sigma}_\epsilon$ into (4.15), we get the estimated optimal weights for estimating the far covariance $\rho$. As the theory here indicated, $\hat{a}$ and $\hat{\sigma}_\epsilon$ are very well estimated in the high dimensional case. Therefore the resulting far covariance estimator in (4.11) will achieve variance reduction as considered in the last section 4.2.2.

**Remark 26.** The covariance among nearby locations are abundant in many applications such as climate data. While the near covariance may not follow (4.3) perfectly in those applications, $\hat{a}$ can be considered a first order approximation to the strength of the abundant near covariance. The weighted estimator in (4.11), compared to the simple sample covariance, likely will still improve the estimation accuracy for the sparse far covariance $\rho$.

### 4.2.4 Test for Detecting Teleconnection

We now consider the detection of the teleconnection. Knowing the location of the teleconnection, $(i^*, j^*)$, this is the hypothesis testing of (4.7): distinguish between model (4.3) and model (4.6). When $\rho = 0$, model (4.6) reduces to (4.3), thus testing (4.7) is equivalent to distinguish between $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$. The natural test is to reject $H_0$ when an estimator $\hat{\rho}$ exceeds a cutoff value corresponding to specified type I error rate. The estimator $\hat{\rho}$ can be taken as the sample covariance $\hat{\sigma}_{i^*,j^*}$ or our proposed far covariance estimator $\bar{\sigma}_{i^*,j^*}$. Since $\bar{\sigma}_{i^*,j^*}$ has smaller variance, the test based on it should distinguish (4.7) better.
More generally, the location \((i^*, j^*)\) is not known prior to the data analyst. Then the objective is to detect existence of teleconnection at any pair of far away locations. More strictly, we define the teleconnection as dependence between two locations with distance greater or equal to a threshold \(\tau\). Then the existence of any teleconnection is mathematically expressed as non-zero \(\tau\)-coherence of the covariance matrix \(\Sigma\).

For the \(p \times p\) matrix \(\Sigma = \{\sigma_{i,j}\}\), the \(\tau\)-coherence is

\[
L = \max_{1 \leq i < j \leq p, |i-j| \geq \tau} |\sigma_{i,j}|.
\]  

(4.21)

Then detecting of existence of any teleconnection is equivalent to statistically testing

\[
H_0 : L = 0, \quad \text{versus} \quad H_1 : L > 0.
\]  

(4.22)

The natural test for (4.22) is based on sample covariance of data \(X_{nxp}\), and such tests have been studied in the literature. Generally, the data matrix \(X_{nxp}\) is first normalized so that at every location \((i = 1, ..., p)\), \(X_i\) has unit variance. That is, we instead work with the sample correlation matrix \(R = \{\hat{\rho}_{i,j}\}_{i,j=1}^{p}\) where each element is the Pearson correlation coefficient

\[
\hat{\rho}_{i,j} = \frac{\hat{\sigma}_{i,j}}{\sqrt{\hat{\sigma}_{i,i} \hat{\sigma}_{j,j}}}. 
\]  

(4.23)

The theoretical distribution of the \(\tau\)-coherence of \(R\),

\[
\hat{L} = \max_{1 \leq i < j \leq p, |i-j| \geq \tau} |\hat{\rho}_{i,j}|,
\]  

(4.24)

is derived in Cai and Jiang [2011], Xiao and Wu [2013], and the test can be conducted accordingly for (4.22) using \(\hat{L}\).

Since our far covariance estimator \(\bar{\sigma}_{i,j}^*\) is more accurate than the sample covariance \(\hat{\sigma}_{i,j}\) under FNCM, we propose to test (4.22) based on \(\bar{\sigma}_{i,j}^*\). Plug \(\bar{\sigma}_{i,j}^*\) into (4.23), we have a new estimator \(\hat{\rho}_{i,j}\) of correlation for \(|i - j| \geq \tau\). Then the \(\tau\)-coherence of \(R = \{\hat{\rho}_{i,j}\}_{i,j=1}^{p}\) is

\[
\bar{L} = \max_{1 \leq i < j \leq p, |i-j| \geq \tau} |\hat{\rho}_{ij}|.
\]  

(4.25)

The asymptotic behavior of \(L\) can be derived similarly to that of \(\hat{L}\) as in Jiang [2004].
Theorem 27. Let $X_{n \times p}$ be the data matrix for $p$ dimensional multivariate Gaussian distribution with mean zero and covariance $\Sigma$, and $X_{n \times p}$ follows the far-near covariance model (4.6) with $\rho = 0$ where $a > 0$ and $n$ is the sample size. Assuming $\frac{n}{p} \to \gamma \in (0, \infty)$, then

$$\lim_{n \to \infty} \sqrt{\frac{n}{\log n}} \bar{L} = 2\sqrt{\varrho_{rr}} \quad a.s.$$  (4.26)

The proof of Theorem 27 can be found in Appendix C.6. Theorem 27 and Theorem 20 together implies that $\bar{L}$ of (4.25) has a smaller limit compared to that of $\hat{L}$ in (4.24). The asymptotic limit of the latter corresponding to $\varrho_{rr} = 1$ without variance reduction. Thus the test based on $\bar{L}$ will have a lower cut-off than the test based on $\hat{L}$. As $\hat{L}$ is based on the better estimator $\bar{\sigma}_{i,j}^*$, there is a better separation between its null and alternative distribution. Thus the test based on $\hat{L}$ should have higher power for detecting the existence of any teleconnection. This power improvement is confirmed by the numerical studies in Section 4.3 later.

4.2.5 Extension to Two Dimension Models

In the above, we have assumed that the differences in the indices are proportional to the distances between corresponding locations. That is, the variables are indexed according to a one-dimensional structure as plotted in Figure 4.3. This type of structure is suitable when the variables $X_1, ..., X_p$ represents data from a time series or equidistant locations on some one-dimensional manifold. For applications such as climate data analysis, we need to consider the case that $X_1, ..., X_p$ correspond to two dimensional spatial locations. We now consider building the FNCM model on a two-dimensional equidistant grid, for example, as plotted in Figure 4.4.
Generally speaking, if we have \( p = p_x \times p_y \) spatial locations with \( p_x \) rows and \( p_y \) columns, we can index them by row (raster scan order) from 1 to \( p \). Then the near covariance model of (4.3) becomes

\[
X_i = a f_{i-p_y} + a f_{i-1} + f_i + a f_{i+1} + a f_{i+p_y} + \epsilon_i, \quad \forall i \in \{1, \cdots, p\},
\]

(4.27)

For the far covariance at a pair \((i^*, j^*)\) corresponding to locations apart by at least \( \tau \), we can similarly model the latent factor at \( j^*-\)th location by \( \rho f_{i^*} + \sqrt{1-\rho^2} f_{j^*} \).

\[
\begin{align*}
X_{j^*} &= a f_{j^*-p_y} + a f_{j^*-1} + [\rho f_{i^*} + \sqrt{1-\rho^2} f_{j^*}] + a f_{j^*+1} + a f_{j^*+p_y} + \epsilon_{j^*}, \\
X_{j^*-1} &= a f_{j^*-1-p_y} + a f_{j^*-2} + f_{j^*-1} + a[\rho f_{i^*} + \sqrt{1-\rho^2} f_{j^*}] + a f_{j^*-1+p_y} + \epsilon_{j^*-1}, \\
X_{j^*+1} &= a f_{j^*+1-p_y} + a[\rho f_{i^*} + \sqrt{1-\rho^2} f_{j^*}] + f_{j^*+1} + a f_{j^*+2} + a f_{j^*+1+p_y} + \epsilon_{j^*+1}, \\
X_{j^*-p_y} &= a f_{j^*-2-p_y} + a f_{j^*-p_y-1} + f_{j^*-p_y} + a f_{j^*-p_y+1} + a[\rho f_{i^*} + \sqrt{1-\rho^2} f_{j^*}] + \epsilon_{j^*-p_y}, \\
X_{j^*+p_y} &= a[\rho f_{i^*} + \sqrt{1-\rho^2} f_{j^*}] + a f_{j^*+p_y-1} + f_{j^*+p_y} + a f_{j^*+p_y+1} + \epsilon_{j^*+p_y}.
\end{align*}
\]

(4.28)

Then based on the two-dimensional FNCM (4.28), we can derive the best unbiased linear estimator for \( \rho \) as before. Here, instead of the 4 adjacent elements, there are now 8 adjacent sample covariance elements in the two-dimensional model that we should use in the estimation. As we labeled in Figure 4.5, we combine the sample covariance \( \hat{\sigma}_{i^*, j^*} \) with its 8 adjacent
Figure 4.5: Example of the adjacent sample covariance elements that are combined linearly to estimate $\sigma_{i^*,j^*}$ in the two dimensional model.

elements in the two dimensional model as

$$
\mathbf{\tilde{\Sigma}}_{9\times 1} := (\hat{\sigma}_{i^*,j^*}, \hat{\sigma}_{i^*,j^*-1}, \hat{\sigma}_{i^*,j^*+1}, \hat{\sigma}_{i^*,j^*-1}, \hat{\sigma}_{i^*,j^*+py}, \hat{\sigma}_{i^*,j^*-py}, \hat{\sigma}_{i^*,j^*+py}, \hat{\sigma}_{i^*,j^*-py})^\top.
$$

We now find the corresponding versions of Corollary 21 and 22 in the two-dimensional FNCM (4.28). The structure of the covariance of the nine sample covariance elements in $\mathbf{\tilde{\Sigma}}_{9\times 1}$ under the model (4.28) has the following form:

$$
\mathbf{\tilde{\Sigma}} =
\begin{bmatrix}
    d_1 & r_1 & r_1 & r_1 & r_1 & r_1 & r_1 & r_1 & r_1 \\
    r_1 & d_2 & s_1 & s_2 & s_1 & s_3 & s_1 & s_3 & s_1 \\
    r_1 & s_1 & d_2 & s_1 & s_2 & s_1 & s_3 & s_1 & s_3 \\
    r_1 & s_2 & s_1 & d_2 & s_1 & s_3 & s_1 & s_3 & s_1 \\
    r_1 & s_1 & s_2 & s_1 & d_2 & s_1 & s_3 & s_1 & s_3 \\
    r_1 & s_3 & s_1 & s_3 & s_1 & d_2 & s_1 & s_2 & s_1 \\
    r_1 & s_1 & s_3 & s_1 & s_3 & s_1 & d_2 & s_1 & s_2 \\
    r_1 & s_3 & s_1 & s_3 & s_1 & s_2 & s_1 & d_2 & s_1 \\
    r_1 & s_1 & s_3 & s_1 & s_3 & s_1 & s_2 & s_1 & d_2 \\
\end{bmatrix}, \quad (4.29)
$$
where \( d_1, d_2, r_1, s_1, s_2, s_3 \) are some functions of \( a, \rho \) and \( \sigma_\varepsilon \) whose formulas are given in (C.11) in Appendix C.4.

**Corollary 28.** Assume that (4.29) holds. Then \( \lambda^* = (1 - 8a\lambda_2^*, \lambda_2^*, \lambda_2^*, \lambda_2^*, \lambda_2^*, \lambda_2^*)^T \) where

\[
\lambda_2^* = \frac{ad_1 - r_1}{d_2 + 4s_1 + s_2 + 2s_3 - 16ar_1 + 8a^2d_1}.
\]

(4.30)

Furthermore, plug-in expressions of \( d_1, d_2, r_1, s_1, s_2, s_3 \) from (C.11), we can have the following explicit formula.

**Corollary 29.** Let \( X \) follows the FNCM (4.28). Then \( \lambda^* = (1 - 8a\lambda_2^*, \lambda_2^*, \lambda_2^*, \lambda_2^*, \lambda_2^*, \lambda_2^*)^T \) where

\[
\lambda_2^* = \frac{16a^5 - 2a + 4a^3\sigma_\varepsilon^2 + a\sigma_\varepsilon^4}{32a^6 - 10a^4 + a^2 + 1 + (32a^4 - 7a^2 + 2)\sigma_\varepsilon^2 + (8a^2 + 1)\sigma_\varepsilon^4}.
\]

(4.31)

Correspondingly, the test for detecting teleconnection in the 2-dimensional FNCM can be built upon statistics using \( \tilde{\sigma}_{i,j}^* \) with these optimal weights.

### 4.3 Numerical Examples

In this section, we check the performance of our proposed far-near covariance model analysis methods with numerical examples. We first use simulations to study the estimation accuracy of far covariance and to study the power for detecting existence of teleconnection. Then, we apply our analysis on a climate data set.

#### 4.3.1 Synthetic Simulation

We generate data according to the far-near covariance model (4.6). Then we apply the far covariance signal estimator \( \tilde{\sigma}_{i,j}^* \) in (4.11) and the sample covariance estimator \( \hat{\sigma}_{i,j}^* \) on the generated data sets to compare their estimation accuracy. We then apply the \( \tau \)-coherence tests based on proposed statistic in (4.25) and on the statistic in (4.24) to compare their power to detect teleconnection.
We first simulate data with \( n = 1000, p = 100, \sigma_\epsilon = 2 \) and various far covariance signal strengths of \( a = 0.5, 0.6, \cdots, 0.9 \). The positions \((i^*, j^*)\) of the teleconnected pair is set to be the 23% and 77% quantiles of 1, \( \cdots, p \) in the simulation. Figure 4.6 shows the performances of \( \hat{\sigma}_{i^*, j^*} \) and \( \bar{\sigma}_{i^*, j^*} \) based on \( N = 100 \) replications. We can see from the boxplots that both

![Boxplots showing performances of estimators](image)

Figure 4.6: Comparison of \( \hat{\sigma}_{i^*, j^*} \) (estimator 1 in red) and \( \bar{\sigma}_{i^*, j^*} \) (estimator 2 in blue). The left panel corresponds to \( a = 0.6 \), while the right panel corresponds to \( a = 0.8 \). The remote signal strength \( \rho \) in x-axis varies from 0.5 to 0.9.

estimators are unbiased, but \( \bar{\sigma}_{i^*, j^*} \) (in blue) has smaller variance than \( \hat{\sigma}_{i^*, j^*} \) (in red). This agrees with the theory from previous sections.

Since the \( \bar{\sigma}_{i^*, j^*} \) has smaller variance, the test statistic based on it should have more separated distributions under null hypothesis \( H_0 \) and alternative hypothesis \( H_1 \) of equation (4.7). Thus the test based on \( \bar{\sigma}_{i^*, j^*} \) should have higher power.

For testing the existence of teleconnection, the position \((i^*, j^*)\) is assumed unknown, and we apply the \( \tau \)-coherence statistic in (4.25) and (4.21) on the simulated data. We calculate their power on simulated data sets with varying values of sample size \( n \), dimension \( p \), near covariance strength \( a \) and far covariance strength \( \rho \). Figure 4.7 plots the power curves of these two tests. The left column shows the power curves against the sample size \( n \), while
the right panel shows the power curves against the signal strength $\rho$. As we can see from Figure 4.7, the $\tau$-coherence test based on the proposed estimators (in blue) indeed has higher power than the $\tau$-coherence test based on the sample covariance matrix (in orange) in most cases.

4.3.2 Teleconnection in Precipitation Data

We now apply our model to analyze climate data. Climate network analysis is an active recent research topic, which aims to discover complex structures in data through analyzing dependence in climate variables across various spatial locations. There often are abundant dependence for climate variables at nearby locations. The long range dependence is relatively rare and such teleconnections reveal interesting patterns in the climate study. Previous literature on climate networks and teleconnections built their analysis methods on the sample covariances. We now apply our model for detecting the teleconnections in the climate data, and compare with the results based on the sample covariances.

We study the climate data from the National Oceanic & Atmospheric Administration (NOAA) \(^1\). It is the CPC Merged Analysis of Precipitation (CMAP) data and it contains the monthly mean precipitation data on $5^\circ$ by $5^\circ$ grids from year 1979 to 2016. We remove missing values near the poles. Here we consider teleconnections as dependence between precipitation at locations that are at least $\tau = 5$ grid points apart. This data set contains totally more than $10^6$ pairs of locations that are at least 5 grid points apart.

We first check for the existence of teleconnections. Based on the sample covariances, at significant level $\alpha = 0.01$, there are indeed teleconnections if $\hat{L} \geq 0.49$. There are many highly correlated pairs of far apart locations in this data set. In fact, the magnitude of sample correlations exceeds the cut-off 0.49 for more than $5 \times 10^4$ far apart pairs. Therefore, there is strong indication that some teleconnections exists. Both tests based on $\hat{L}$ in (4.24)

\(^1\)https://www.esrl.noaa.gov/psd/data/gridded/tables/precipitation.html
Figure 4.7: Empirical power curves against sample size (left column) and signal strength (right column). The near covariance strength $\alpha$ is 0.6, 0.8 with different thickness. The blue curve is our proposed test, and the orange one is test based on the sample covariance matrix. The first row is $p = 100$, the second row is $p = 200$, and the third row is $p = 500$. 
and $\bar{L}$ in (4.25) have p-value less than $10^{-3}$.

Knowing there are teleconnections, the next step is to identify the location of the teleconnection and estimate the dependence strength. As the number of locations is big in the data set, there is a high probability that large sample covariance values can occur spuriously at some pairs of far apart locations that are not truly correlated. By utilizing the neighboring sample covariance information, our estimator can reduce chances of such false positive signals for teleconnection. For example, Figure 4.8 shows an apparently teleconnected pair of locations based on a relatively high sample correlation value of 0.58 (this ranked in top one percent of the sample correlation matrix among the far apart pairs).

![Precipitation](image)

**Figure 4.8:** Monthly mean precipitation data.

Figure 4.9 presents the scatter plot of the monthly mean precipitations at this pair of locations, in the top left panel. The scatter plots at three adjacent pairs are presented in the rest three panels of Figure 4.9. Even though there is a relatively high correlation between these two locations, their nearby pairs have very low correlations indicating that a real precipitation teleconnection between these two locations is unlikely. Our proposed estimator takes the surrounding correlations into consideration and yields a much smaller estimated far covariance strength $\bar{\sigma}_{r^*j^*} = 0.08$ which ranks below top ten percent among the far apart pairs.

In contrast, we illustrate what a genuine teleconnection looks like by checking the top teleconnected pair ranked by our proposed estimator. The pair has one location in South
Figure 4.9: Example of the teleconnection detected by sample covariance $\hat{\sigma}_{r,s} = 0.58$ (top left). The other three are the scatter plots of the surrounding pairs.
America and one location in South Africa as shown in Figure 4.10. Similar to Figure 4.8, Figure 4.10 shows the scatter plots monthly mean precipitations at this pair and three adjacent pairs. We can observe that these adjacent pairs also exhibit high correlation, which may serve as a confirmation that there are indeed some common factors strongly affecting the precipitations at both locations. This teleconnection for precipitation between South Africa and South American has indeed been studied in the literature [Cook et al., 2004, Grimm and Reason, 2015].

4.4 Conclusion

In this work, we present a far-near covariance model to delineate short range and long range dependence in spatial data. The near covariance structural information were used to improve estimation for the far covariance strength. We show theoretically and numerically that the proposed methods improve upon the statistical analysis methods based on the sample covariance matrix. The proposed model can be very useful in analyzing data with abundant short range dependence and some long range dependence, such as teleconnection detections and network analysis for climate data.
Figure 4.11: Example of the top teleconnection detected by the proposed far covariance estimation $\tilde{\sigma}_{i,j}$ (top left). The other three are the scatter plots of the surrounding pairs.
Appendix A

Additional Proofs and Results for Chapter 2

A.1 Proof of Theorem 8

For simplicity, we focus on the bivariate case \((X, Y)\) are each one-dimensional variables). The extension of the proof to the multivariate case is straightforward. We first work on mutual information, then show the similar arguments on the copula distances. To prove the theorem, we use Le Cam [1973]'s method to find the lower bound on the minimax risk of the estimating mutual information \(MI\). To do this, we will use a more convenient form of Le Cam’s method developed by Donoho and Liu [1991]. Define the module of continuity of a functional \(T\) over the class \(F\) with respect to Hellinger distance as in equation (1.1) of Donoho and Liu [1991]:

\[
    w(\varepsilon) = \sup \{|T(F_1) - T(F_2)| : F_i \in F, H(F_1, F_2) \leq \varepsilon\}.
\] (A.1)

Here \(H(F_1, F_2)\) denotes the Hellinger distance between \(F_1\) and \(F_2\). Then the minimax rate of convergence for estimating \(T(F)\) over the class \(F\) is bounded below by \(w(n^{-1/2})\).

We now look for a pair of density functions \(c_1(u, v)\) and \(c_2(u, v)\) on the unit square for
distributions that are close in Hellinger distance but far away in their mutual information. This provides a lower bound on the module of continuity for mutual information MI over the class $\mathcal{C}$, and hence leads to a lower bound on the minimax risk. We outline the proof next.

We first divide the unit square into three disjoint regions $R_1$, $R_2$ and $R_3$ with $R_1 \cup R_2 \cup R_3 = [0, 1] \times [0, 1]$. The first density function $c_1(u, v)$ puts probability masses $\delta, a$ and $1-a-\delta$ respectively on the regions $R_1, R_2$ and $R_3$ each uniformly. The $a$ is an arbitrary small fixed value, for example, $a = 0.01$. For now, we take $\delta$ to be another small fixed value. The area of the region is chosen so that $c_1(u, v) = M$ on region $R_2$ and $c_1(u, v) = M^*$ on region $R_1$ for a very big $M^*$. The second density function $c_2(u, v)$, compared to $c_1(u, v)$, moves a small probability mass $\varepsilon$ from $R_1$ to $R_2$. We will see that the Hellinger distance between $c_1$ and $c_2$ is of the same order as $\varepsilon$, but the change in MI is unbounded for big $M^*$. Hence module of continuity $w(\varepsilon)$ is unbounded for mutual information MI. Therefore the MI can not be consistently estimated over the class $\mathcal{C}$.

Specifically, the region $R_1$ is chosen to be a narrow strip immediately above the diagonal, $R_1 = \{(u, v) : -\delta_1 < u - v < 0\}$; and $R_2$ is chosen to be a narrow strip immediately below the diagonal, $R_2 = \{(u, v) : 0 \leq u - v < \delta_2\}$. The remaining region is $R_3 = [0, 1] \times [0, 1] \setminus (R_1 \cup R_2)$. The values of $\delta_1$ and $\delta_2$ are chosen so that the areas of regions $R_1$ and $R_2$ are $\delta/M^*$ and $a/M$ respectively. Then clearly $c_1(u, v) = M^*$ on $R_1$; $c_1(u, v) = M$ on $R_2$; $c_1(u, v) = (1-a-\delta)/(1-a/M-\delta/M^*)$ on $R_3$. And $c_2(u, v) = M^* - \varepsilon(M^*/\delta)$ on $R_1$; $c_2(u, v) = M + \varepsilon(M/a)$ on $R_2$; $c_2(u, v) = c_1(u, v)$ on $R_3$. See the Figure A.1.

Then we have

$$2H^2(c_1, c_2) = \int (\sqrt{c_2(u, v)} - \sqrt{c_1(u, v)})^2 du dv$$

$$= (\sqrt{M^* - \varepsilon(M^*/\delta)} - \sqrt{M^*})^2 \delta/M^* + (\sqrt{M + \varepsilon(M/a)} - \sqrt{M})^2 a/M$$

$$= \delta(\sqrt{1 - \varepsilon/\delta} - 1)^2 + a(\sqrt{1 + \varepsilon/a} - 1)^2$$

$$= \delta(\varepsilon/2\delta)^2 + a(\varepsilon/2a)^2 + o(\varepsilon^2)$$

$$= \varepsilon^2 \left(\frac{1}{4\delta^2} + \frac{1}{4a} \right) + o(\varepsilon^2).$$
Figure A.1: The plot shows the regions $R_1$, $R_2$ and $R_3$. The other two narrow strips neighboring $R_1$ and $R_2$ are for the continuity correction mentioned at the end of the proof.

Hence the Hellinger distance is of the same order as $\varepsilon$:

$$H(c_1, c_2) = \varepsilon \sqrt{\frac{1}{8\delta} + \frac{1}{8a} + o(\varepsilon)}.$$  

On the other hand, the difference in the mutual information is

$$\text{MI}(c_1) - \text{MI}(c_2)$$

$$= \delta \log(M^*) + a \log(M) - (\delta - \varepsilon) \log[M^* - \varepsilon(M^*/\delta)] - (a + \varepsilon) \log[M + \varepsilon(M/a)] \quad (A.2)$$

$$= \varepsilon \log(M^*) - \varepsilon \log(M) - (\delta - \varepsilon) \log(1 - \varepsilon/\delta) - (a + \varepsilon) \log(1 + \varepsilon/a).$$

Here $M$, $\delta$ and $a$ are fixed constants. Hence when $M^* \to \infty$, this difference in MI also goes to $\infty$. For example, if we let $M^* = e^{1/(\varepsilon)^2}$, then the module of continuity $w(\varepsilon) \geq O(1/\varepsilon)$. That means, the rate of convergence is at least $O(w(n^{-1/2})) = O(n^{1/2}) \to \infty$. In other words, MI can not be consistently estimated.

Now, let us consider the CD$_{\alpha} = \int_{I^2} |c(u, v) - 1|^\alpha dudv$, for $\alpha > 1$, where $I^2$ is the unit square.
\[\begin{align*}
\text{CD}_\alpha(c_1) - \text{CD}_\alpha(c_2) & = |M^*-1|^\alpha \delta/M^* + |M-1|^\alpha a/M - |M^* - 1 - \varepsilon(M^*/\delta)|^\alpha \delta/M^* + |M - 1 + \varepsilon(M/a)|^\alpha a/M \\
& = |M^* - 1|^\alpha - |M^* - 1 - \varepsilon(M^*/\delta)|^\alpha \delta/M^* + |M-1|^\alpha - |M - 1 + \varepsilon(M/a)|^\alpha a/M \\
& = \alpha [(M^* - 1)^{\alpha-1} \delta/M^* - (M - 1)^{\alpha-1} M/a] \varepsilon + o(\varepsilon^2).
\end{align*}\]

(A.3)

Again, \(M, \delta\) and \(a\) are fixed constants. Hence when \(M^* \to \infty\), this difference in \(\text{CD}_\alpha, \alpha > 1\) also goes to \(\infty\). For example, if we let \(M^* = (\varepsilon^{-2} + M^n) \frac{1}{\alpha-1} + 1\), then the module of continuity \(w(\varepsilon) \geq O(1/\varepsilon)\). Note that \(\alpha > 1\) is essential here. That means, the rate of convergence is at least \(O(w(n^{-1/2})) = O(n^{1/2}) \to \infty\). In other words, \(\text{CD}_\alpha, \alpha > 1\) can not be consistently estimated.

The above outlines the main idea of the proof, ignoring some mathematical subtleties. One is that the example densities \(c_1\) and \(c_2\) are only piecewise continuous on the three regions, but not truly continuous as required for the class \(\mathcal{C}\). This can be easily remedied by connecting the three pieces linearly. Specifically we set the densities \(c_i(u, v) = M, i = 1, 2\), on the boundary between \(R_1\) and \(R_3\), \(\{(u, v) : u - v = -\delta_1\}\), and on the boundary between \(R_2\) and \(R_3\), \(\{(u, v) : u - v = \delta_2\}\). Then we use two narrow strips within \(R_3\), \(\{(u, v) : -\delta_3 \leq u - v \leq -\delta_1\}\) and \(\{(u, v) : \delta_2 \leq u - v \leq \delta_4\}\) to connect the constant \(c_i(u, v)\) values on the rest of region \(R_3\) with the boundary value \(c_i(u, v) = M\) continuously through linear (in \(u - v\)) \(c_i(u, v)\)'s on the two strips that satisfies the Hölder condition (2.7) of the main text. By the Hölder condition, the connection can be made with strips of width at most \((M-1+a+\delta)/M_1\). This continuity modification does not affect the calculation of the difference \(\text{MI}(c_1) - \text{MI}(c_2)\) or \(\text{CD}_\alpha(c_1) - \text{CD}_\alpha(c_2)\) above as \(c_1\) and \(c_2\) only differ on regions \(R_1\) and \(R_2\). Within regions \(R_1\) and \(R_2\), the densities \(c_1\) and \(c_2\) can be further similarly connected continuously linearly in \(u - v\). As there is no Hölder condition on \(A_M^\alpha\), the connection within \(R_1\) and \(R_2\) can be as steep as we want. Clearly the order obtained through above calculations will not change if we make these connections very steep so that their effect is negligible.
Another technical subtlety is that the \( c_1 \) and \( c_2 \) defined above are only densities on the unit square but not copula densities which require uniform marginal distributions. However, it is clear that the marginal densities for \( c_i \)'s are uniform over the interval \((\delta_3, 1 - \delta_4)\) and linear in the rest of interval near the two end points 0 and 1. The copulas densities \( c^*_i \)'s corresponding to \( c_i \)'s can be calculated directly through Sklar’s decomposition (1) in the main text. It is easy to see that the order for the module of continuity \( w(\varepsilon) \) remains the same for using the corresponding copula densities \( c^*_i \)'s.

### A.2 Proof of Theorem 9

The proof is almost the same as the proof for MI, but need some modification of the pair of least favorable \( c_1 \) and \( c_2 \) above. The small difference in Hellinger distance of \( c_1 \) and \( c_2 \) can lead to unbounded difference in \( MI(c_1) \) and \( MI(c_2) \) since \( MI \) is unbounded. After the transformation \( MI_{cor} = \sqrt{1 - e^{-2MI}} \) is bounded. The difference between \( MI_{cor}(c_1) \) and \( MI_{cor}(c_2) \) in the above example is actually small since the \( MI \) are big for both \( c_1 \) and \( c_2 \) (leading to corresponding \( MI_{cor} \)s close to zero). However, \( MI_{cor} \) is also very hard to estimate over the class \( \mathcal{C} \). To see this, we follow the same reasoning above but modify the example of \( c_1 \) and \( c_2 \). First, we notice that for any pair of densities \( c_1 \) and \( c_2 \),

\[
|MI_{cor}(c_1) - MI_{cor}(c_2)| = |\sqrt{1 - e^{-2MI(c_1)}} - \sqrt{1 - e^{-2MI(c_2)}}|
\]

\[
= \left| \frac{|1 - e^{-2MI(c_1)}| - |1 - e^{-2MI(c_2)}|}{\sqrt{1 - e^{-2MI(c_1)}} + \sqrt{1 - e^{-2MI(c_2)}}} \right|
\]

\[
\geq \frac{1}{2} |e^{-2MI(c_1)} - e^{-2MI(c_2)}|
\]

\[
= \frac{1}{2} e^{-2MI(c_1)} |1 - e^{-2[MI(c_1) - MI(c_2)]}|.
\]

For the difference \( MI_{cor}(c_1) - MI_{cor}(c_2) \) to be the same order of the difference \( MI(c_1) - MI(c_2) \), we need to set \( MI(c_1) \) at constant order when \( \varepsilon \to 0 \).

Therefore, we modify the above \( c_1 \) to have probability mass \( \delta = 2\varepsilon \) in region \( R_1 \), varying
with the $\varepsilon$ value instead of fixed as before. And we set $M^* = e^{1/\varepsilon}$, leading to

\[
MI(c_1) = \delta \log(M^*) + a \log(M) + (1 - a - \delta) \log[(1 - a - \delta)/(1 - a/M - \delta/M^*)]
\]

\[
= 2 + a \log(M) + (1 - a - 2\varepsilon) \log[(1 - a - 2\varepsilon)/(1 - a/M - 2\varepsilon e^{-1/\varepsilon})],
\]

which converges to a fixed constant $a_1 = 2 + a \log(M) + (1 - a) \log[(1 - a)/(1 - a/M)]$ as $\varepsilon \to 0$. Using (A.2), recall that $\delta = 2\varepsilon$ and $M^* = e^{1/\varepsilon}$, we have

\[
MI(c_1) - MI(c_2)
\]

\[
= \varepsilon \log(M^*) - \varepsilon \log(M) - (\delta - \varepsilon) \log(1 - \varepsilon/\delta) - (a + \varepsilon) \log(1 + \varepsilon/a)
\]

\[
= 1 - \varepsilon \log(M) - \varepsilon \log(1/2) - (a + \varepsilon) \log(1 + \varepsilon/a),
\]

which converges to 1 as $\varepsilon \to 0$. Hence we have

\[
\lim_{\varepsilon \to 0} w(\varepsilon) \geq \lim_{\varepsilon \to 0} \frac{1}{2} e^{-2MI(c_1)} |1 - e^{-2[MI(c_1) - MI(c_2)]}| = \frac{1}{2} e^{-2a_1} (1 - e^{-2(1)}),
\]

a positive constant $a_2 = e^{-2a_1} (1 - e^{-2})/2$. Therefore, $MI_{cor}$ can not be estimated consistently over the class $\mathcal{C}$ either.

**A.3 Proof of Theorem 10**

The first two terms in (2.12) corresponds to bias and standard deviation of kernel density estimation when the copula density is bounded. When the copula density is unbounded, the kernel density estimation $\hat{c}(Z)$ is not consistent. However, a smaller order $O(\frac{1}{nh^d})$ term bounds the overall error contribution to $\hat{RCD}$ resulting from $\hat{c}(Z)$ in the unbounded copula density region.

Let $M_2 = \frac{M + 1}{2}$, $A_{M_2} = \{Z|c(Z) \leq M_2\}$, $T_1(c) = \int_{A_{M_2}} (1 - c(Z))_+ dZ$, $T_2(c) = \int_{A_{M_2}} (1 - c(Z))_+ dZ$, $RCD = T_1(c) + T_2(c)$, and $\overline{RCD} = T_1(\hat{c}) + T_2(\hat{c})$

Firstly, we consider the region $A_{M_2}$ with bounded copula density. Here we calculate the bias and variance of the kernel density estimator using standard methods first.
\[ \tilde{c}_n(Z) = E[\hat{c}_{kde}(Z)] = \frac{1}{h^d} \int K\left( \frac{z - Z}{h} \right) c(z) dz = \int K(s) c(Z + sh) ds. \]

Hence
\[
|Bias(Z)| = |\int K(s) c(Z + sh) ds - c(Z)| \leq \int_{B_0} K(s)|c(Z + sh) - c(Z)| ds \\
\leq \int_{B_0} K(s) M_1 h ds \\
= M_1 h. \tag{A.4}
\]

\[
|Var(Z)| = \frac{1}{n} Var\left[ \frac{1}{n h^d} K\left( \frac{Z - Z}{h} \right) \right] \leq \frac{1}{n} E\left[ \frac{1}{n h^d} K^2\left( \frac{Z - Z}{h} \right) \right] \\
= \frac{1}{n h^d} \int_{B_0} K^2(s)c(Z + sh) ds \\
\leq \frac{1}{n h^d} \int_{B_0} K^2(s)|c(Z) + M_1 h| ds \\
= \frac{\mu^2}{n h^d}[c(Z) + M_1 h]. \tag{A.5}
\]

Hence the integrated mean square error of the density estimator \( \hat{c}_n(Z) \) over regions \( A_{M_2} \) is
\[
IMSE(Z) = \int_{A_{M_2}} [Bias^2(Z) + Var(Z)] dZ \\
\leq \int_{A_{M_2}} [M_1^2 h^2 + \frac{\mu^2}{n h^d}[c(Z) + M_1 h]] dZ \\
\leq M_1^2 h^2 + \frac{\mu^2}{n h^d}[1 + M_1 h]. \tag{A.6}
\]

Hence the error of \( \overline{RCD} \) on \( A_{M_2} \) is bounded by
\[
E|T_1(\hat{c}) - T_1(c)| \leq E \int_{A_{M_2}} |(1 - \hat{c}_n(Z))_+ - (1 - c(Z))_+| dZ \\
\leq E \int_{A_{M_2}} |\hat{c}_n(Z) - c(Z)| dZ \\
\leq \sqrt{E \int_{A_{M_2}} (\hat{c}_n(Z) - c(Z))^2 dZ} \\
\leq \sqrt{d^2 M_1^2 h^2 + \frac{2\mu^2}{n h^d}} \\
\leq dM_1 h + \sqrt{2\mu_2}\left(\frac{1}{nh^d}\right)^{1/2}.
\]

Now we consider the region \( A_{M_2}^c \) with unbounded copula density. For \( Z \in A_{M_2}^c, \hat{c}(Z) \) does not have a finite variance bound in (A.5). But we can bound the variance by the expectation
\[
\tilde{c}_n(Z) = E[\hat{c}_n(Z)]. \]

Let \( M_3 = \frac{M_2 + 1}{2} \), when \( h \) small, \( Z \in A_{M_2}^c \) implies \( Z + sh \in A_{M_3}^c \). Hence
\[
|Var(Z)| \leq \frac{1}{n h^d} \int_{B_0} K^2(s) c(Z + sh) ds \leq \frac{M_K}{nh^d} \int_{B_0} K(s) c(Z + sh) ds = \frac{M_K}{nh^d} \tilde{c}_n(Z)
\]

113
Using Chebyshev’s inequality,

\[
E[1\{\hat{c}_n(Z) < 1\}] = P(\hat{c}_n(Z) < 1) \leq P(\mid c_n(Z) - \hat{c}_n(Z)\mid > \hat{c}_n(Z) - 1)
\]

\[
\leq \frac{\text{Var}[\hat{c}_n(Z)\mid \hat{c}_n(Z) - 1]}{\hat{c}_n(Z) - 1} \leq \frac{M_K}{n h^d (\hat{c}_n(Z) - 1) \leq \frac{M_K M_4}{n h^d}}
\]

where \(M_4 = \frac{M_3}{(M_3 - 1)^2}\).

Hence the error of \(\widehat{RCD}\) on \(A_{M_2}^c\) is bounded by

\[
E[T_2(\hat{c}) - T_2(c)] = E[T_2(\hat{c})] \leq \int_{A_{M_2}^c} E[1\{\hat{c}_n(Z) < 1\}] dZ \leq \frac{M_K M_4}{n h^d}
\]

Combining the above results:

\[
E[\mid \widehat{RCD} - RCD\mid] \leq M_1 h + \frac{\sqrt{2}\mu_2}{\sqrt{n} h^{d/2}} + \frac{M_K M_4}{n h^d}.
\] (A.7)

This finishes the proof.

Note that we can use any \(L_p\) norm \((1 \leq p \leq \infty)\) in the Hölder condition: equation (2.7). The kernel \(K\) is then assumed to have support in the unit ball \(B_0\) corresponding to that \(L_p\) norm. The proof remains exactly the same. We in fact will use \(L_\infty\) norm in our estimator for computational simplicity. In that case, the unit ball \(B_0 = \{Z : \|Z\|_\infty \leq 1\}\) is in fact the \(d\)-dimensional cube.

### A.4 Proof of Theorem 11

Here for \(\widehat{RCD} = RCD(\hat{c})\) we use the k-NN estimator [Loftsgaarden and Quesenberry, 1965]

of the copula density

\[
\hat{c}_{knn}(Z) = \frac{k(n)}{A_{r(k(n), n)Z}},
\] (A.8)

where \(Z_1, Z_2, \cdots, Z_n\) are the copula based observations, \(r(k(n), n)\) is the distance from \(Z\) to the \(k^{th}\) closest of \(Z_1, Z_2, \cdots, Z_n\) and \(A_{r(k(n), n)Z}\) is the volume of the \(d\)-dimensional hyper-ball with radius \(r(k(n), n)\).
In the following, without ambiguity, we denote \( r(k(n), n) \) by \( r \), and \( k(n) \) by \( k \). Hence the volume \( A_r, Z \) is \( v_d \cdot r^d \), where \( v_d \) is the volume of the \( d \)-dimensional unit ball \( \mathbb{B}_0 \). And \( \hat{c}_{knn}(Z) = k/(v_dnr^d) \). For \( l_2 \) norm, \( v_d = \pi^{d/2}/\Gamma[(d + 2)/2] \) where \( \Gamma(\cdot) \) denotes the Gamma function.

Moore and Yackel [1977a] showed that, for bounded densities, there is equivalence between the consistency of the KDE density estimator and the consistency of the k-NN estimator. To cite the results of [Moore and Yackel, 1977a], we assume a slightly stronger version of the Hölder condition than (2.7). That is, we assume that \( c \) also has bounded continuous second order derivative in \( A_M \). Let \( Q(Z) = tr[\frac{\partial^2 c(Z)}{\partial Z^2}] \) denote the trace of the Hessian matrix of copula density \( c(Z) \). For the \( d \)-dimensional vector \( Z = (z_1, ..., z_d) \), the Hessian matrix \( \frac{\partial^2 c(Z)}{\partial Z^2} \) has entries

\[
\left[ \frac{\partial^2 c(Z)}{\partial Z^2} \right]_{ij} = \frac{\partial^2 c(Z)}{\partial z_i \partial z_j}.
\]

Then we rewrite the error bound in Theorem 11 explicitly as

\[
\sup_{C \in \mathcal{C}} E[|\hat{RCD} - RCD|] \leq 2\bar{Q}(\frac{k}{n\epsilon})^{\frac{d}{2}} + \frac{2M}{\sqrt{k}} + 2\epsilon,
\]

where \( \bar{Q} = \frac{1}{2(d+2)\pi} \Gamma^{2/d}(\frac{d+2}{2}) \sup_{Z \in A_M} Q(Z) \), and \( \epsilon = \epsilon(n) \) is any sequence converging to 0 slower than \( k/n \). We suppress the \( n \) from the notation in \( \epsilon \) without ambiguity as in \( k \) and \( r \) above.

We shall use the following asymptotic results on k-NN density estimator in Mack and Rosenblatt [1979]. Denote \( \tilde{Q}(Z) = \frac{1}{2(d+2)\pi} \Gamma^{2/d}(\frac{d+2}{2})Q(Z) \). Then

\[
\begin{align*}
Bias[\hat{c}_{knn}(Z)] &= \hat{Q}(Z)^{2d/k} + O\left(\frac{c(Z)}{k}\right), \\
Var[\hat{c}_{knn}(Z)] &= \frac{c^2(Z)}{k} + o\left(\frac{1}{k}\right).
\end{align*}
\]

These expressions provide control on the error contribution of \( \hat{c}(Z) \) to \( \hat{RCD} \) when \( c(Z) \) is bounded both from above and from below. Similar to the proof of KDE-based \( \hat{RCD} \), we prove that the error contribution to \( \hat{RCD} \) from the big copula density region is of a smaller order \( O(1/k) \). Different from the KDE, the k-NN density estimator also does not have finite
bias bound in (A.9) when the copula density \( c(Z) \) is not bounded below. Therefore, we also need to control the error contribution to \( \widehat{RCD} \) from the small \((< \epsilon)\) copula density region separately.

As before, let \( M_2 \) be a constant between 1 and \( M \), say, \( M_2 = \frac{M+1}{2} \). We now separate the three regions by copula density: \( A^c_{M_2} = \{ Z : c(Z) > M_2 \} \) (big), \( A_{M_2,\epsilon} = \{ Z : \epsilon \leq c(Z) \leq M_2 \} \) (middle) and \( A_\epsilon = \{ Z : c(Z) < \epsilon \} \) (small). Then we can separate \( RCD \) into three components \( RCD = T_1(c) + T_2(c) + T_3(c) \): \( T_1(c) = \int_{A^c_{M_2}}[1 - c(Z)]_+dZ \), \( T_2(c) = \int_{A_{M_2,\epsilon}}[1 - c(Z)]_+dZ \) and \( T_3(c) = \int_{A_\epsilon}[1 - c(Z)]_+dZ \).

Firstly, we look at the error bound on \( A^c_{M_2} \), the region of big copula density. Similar to the KDE, the error in \( \hat{c}_{knm}(Z) \) can be arbitrarily large for \( Z \in A^c_{M_2} \). However, the error only leads to the error in \( \widehat{RCD} \) if \( \hat{c}_{knm}(Z) < 1 \). From equation (A.8), \( \hat{c}_{knm}(Z) < 1 \) if and only if

\[
r > \left( \frac{k}{n\bar{r}} \right)^{1/d} \equiv \bar{r}.
\]

This occurs when at most \( k-1 \) of observations \( Z_1, Z_2, \cdots, Z_n \) fall into the ball \( B(Z; \bar{r}) \) which is centered at \( Z \) with radius \( \bar{r} \).

Let \( \tilde{N}(Z) \) denotes the number of observations falling into \( B(Z; \bar{r}) \). Then \( \tilde{N}(Z) \) follows a binomial distribution with mean \( n\bar{p} \), where \( \bar{p} = \int_{B(Z;\bar{r})} c(z)dz \). Since \( k/n \to 0 \), \( \bar{r} \to 0 \). Hence \( M_1\bar{r} < (M_2 - 1)/2 \) when \( n \) is large enough. Then the whole ball \( B(Z; \bar{r}) \) is contained in \( A^c_{M_3} \) with \( M_3 = (M_2 + 1)/2 \) as before. Hence \( \bar{p} = \int_{B(Z;\bar{r})} c(z)dz \geq M_3v_d\bar{r}^d = M_3k/n \). Using Chebyshev’s inequality,

\[
Pr[\hat{c}_{knm}(Z) < 1] = E[1 \{ \tilde{N}(Z) < k \}] \leq \frac{\text{Var}[\tilde{N}(Z)]}{E[\tilde{N}(Z)] - k} = \frac{n\bar{p}(1-\bar{p})}{(n\bar{p} - k)^2} \\
\leq \frac{1}{n\bar{p}(1-\bar{p})^2} \leq \frac{1}{M_3k[1 - 1/M_3]^2} = O(\frac{1}{k}).
\]

Hence

\[
E[T_1(c) - T_1(\hat{c}_{knm})] = \int_{A^c_{M_2}} E[1 \{ \hat{c}_{knm}(Z) < 1 \}]dZ \leq \frac{1}{M_3k[1 - 1/M_3]^2} = O(\frac{1}{k}).
\]
Secondly, we look at the error bound on $A_{M_2, \epsilon}$, the region of middle copula density. Using (A.9), for $Z \in A_{M_2, \epsilon}$, the mean squared error of $\hat{c}_{knn}(Z)$ is

$$E[(\hat{c}_n(Z) - c(Z))^2] = bias^2(Z) + Var(Z)$$

$$= \frac{\tilde{Q}(Z)}{c(Z)^{2/d}} \left(\frac{k}{n}\right)^{2/d} + c^2(Z) + o\left(\frac{k}{n}\right)^{4/d} + \frac{1}{k}$$

$$\leq \frac{\tilde{Q}^2}{c(Z)^{2/d}} \left(\frac{k}{n}\right)^{4/d} + \frac{M_2}{k} + o\left(\frac{k}{n}\right)^{4/d} + \frac{1}{k}.$$

Hence

$$E|\hat{c}_{knn}(Z) - c(Z)| \leq \sqrt{E[(\hat{c}_n(Z) - c(Z))^2]} \leq \sqrt{2}\tilde{Q}\left(\frac{k}{n}\epsilon\right)^{2/d} + \frac{M_2}{\sqrt{k}}[1 + o(1)].$$

We get

$$E[T_2(\hat{c}_{knn}) - T_2(c)] \leq E[\int_{A_{M_2, \epsilon}} |\hat{c}_{knn}(Z) - c(Z)|dZ] \leq \sqrt{2}\tilde{Q}\left(\frac{k}{n}\epsilon\right)^{2/d} + \frac{M_2}{\sqrt{k}}[1 + o(1)]. \quad (A.10)$$

Thirdly, we look at the error bound on $A_\epsilon$, the region of small copula density. From equation (A.8), $\hat{c}_{knn}(Z) \geq 2\epsilon$ if and only if

$$r \leq \left(\frac{k}{2n\epsilon v_d}\right)^{1/d} \equiv r^*.$$

This occurs when at least $k$ of observations $Z_1, Z_2, \cdots, Z_n$ fall into the ball $B(Z; r^*)$. Since $k/(n\epsilon) \to 0$, $r^* \to 0$.

Let $N^*(Z)$ denotes the number of observations falling into $B(Z; r^*)$. Then $N(Z)$ follows a binomial distribution with mean $np^*$, where $p^* = \int_{B(Z, r^*)} c(z)dz$.

Using Taylor expansion, we have (from last line page 228 in Biau et al. [2011])

$$\int_{B(Z, r)} c(z)dz = c(Z)v_dr_d + \tilde{Q}(Z)v_dr_d^{d+2} + o(r^{d+2}).$$

Therefore, using $r^* \to 0$, we have $p^* = c(Z)v_dr_d^{d+2}[1 + o(1)] \leq \epsilon v_d(r^*)^d[1 + o(1)]$ which converges to $k/(2n)$. Hence for $n$ big, $p^* < 0.6k/n$. Using Chebyshev’s inequality,

$$Pr[\hat{c}_{knn}(Z) > 2\epsilon] = E[I\{N^*(Z) < k\}] \leq \frac{Var[N^*(Z)]}{k - E[N^*(Z)]^2} = \frac{np^*(1-p^*)}{(k-np^*)^2}$$

$$\leq \frac{0.6k}{(0.4k)^2} \leq \frac{3}{k} = O\left(\frac{1}{k}\right).$$
Since $c(Z) \leq \epsilon$, if $\hat{c}_{knn}(Z) \leq 2\epsilon$, then $|\hat{c}_{knn}(Z) - c(Z)| \leq 2\epsilon$. Hence

$$E[T_3(c) - T_3(\hat{c}_{knn})] \leq \int_{A_\epsilon} E[|\hat{c}_{knn}(Z) - c(Z)|] dZ \leq \int_{A_\epsilon} \{2\epsilon + Pr[\hat{c}_{knn}(Z) > 2\epsilon]\} dZ$$

$$< 2\epsilon + \frac{3}{k} = O(\epsilon + \frac{1}{k}).$$

Finally, when combining the three parts, the terms $O(1/k) = o(1/\sqrt{k}) < (2 - \sqrt{2})/\sqrt{k}$. Hence we arrive at

$$\sup_{c \in \mathcal{C}} E[|\hat{RCD} - RCD|] \leq 2[\hat{Q}(\frac{k}{n\epsilon})^{2/d} + M_2 \sqrt{\frac{\log k}{k}} + \epsilon],$$

(A.11)

which finished the proof.

Note that we can use other $l_p$ norms, which changes the $v_d$ in the proof to the volume of the unit ball under the corresponding norm. The rate does not change.

We can also prove the consistency under Hölder condition without assuming continuous second derivatives. However, that involve tedious derivation of bias and variance bounds similar to (A.9) for k-NN density estimators. We provide the simple proof here by citing (A.9) from Mack and Rosenblatt [1979].

To minimize the error bound in (2.12), we get $\epsilon = (k/n)^{2/(d+2)}$ and $k = n^{4/(d+6)}$. So in bivariate ($d = 2$) case, we take $k = O(n^{1/2})$. Taking $k$ below the $n^{4/(d+6)}$ rate will make the $O(1/\sqrt{k})$ term dominant in the error bound. In that case, the asymptotic results on the k-NN density estimation states that $\sqrt{k}[\hat{c}(Z) - c(Z)]/c(Z)$ converge to a standard Gaussian distribution. Then $\sqrt{k}[\hat{RCD} - RCD]$ converges to an integral of a Gaussian process.

### A.5 Selection of Tuning Parameter in the Practical Estimator

For a practical estimator for $\hat{RCD}$, we need to decide the bandwidth in KDE-based estimator or the number of neighbors $k$ in KNN-based estimator. Theorem 4 and Theorem 5 provides the rates. For bivariate case, $h = O(n^{-1/4})$ and $k = O(\sqrt{n})$. To decide the constant coefficient, we used empirical simulations.
First, for KDE estimators, we tested $\hat{RCD}$ on nine functions (listed in Table A.1) with various levels of additive noises. Four sample sizes of $n = 10^2, 10^3, 10^4$ and $10^5$ are used. Figure A.2 plots the simulation results using $h = 0.25n^{-1/4}$. We can see that the performance of $\hat{RCD}$ improves as sample size increases, and gives very accurate estimates for $RCD$ under big sample sizes. For illustration, we showed the plots with bandwidth $h = 0.1n^{-1/4}$ and $h = 0.5n^{-1/4}$ in Figure A.3 and Figure A.4 respectively. Those bandwidth choices are clearly either too small ($h = 0.1n^{-1/4}$ estimator overshoot in several cases when $RCD$ is small) or too big ($h = 0.5n^{-1/4}$ estimator converges slowly when $RCD$ is large). Hence the bandwidth $h = 0.25n^{-1/4}$ is a good choice.

| A | Linear    | $y = x$ |
| B | Quadratic | $y = x^2$ |
| C | Square Root | $y = \sqrt{x}$ |
| D | Cubic     | $y = x^3$ |
| E | Centered Cubic | $y = 4(x - 1/2)^3$ |
| F | Centered Quadratic | $y = 4x(1 - x)$ |
| G | Cosine (Period 1) | $y = [\cos(2\pi x) + 1]/2$ |
| H | Circle    | $(x - 1/2)^2 + y^2 = 1/4$ |
| I | Cross     | $y = \pm(x - 1/2)$ |

Table A.1: The function relationships used in Figures A.2 - A.7.

According to the equivalence results between the KDE and the KNN estimator by [Moore and Yackel, 1977b], the $k$ in the KNN density estimation corresponds to the bandwidth in KDE estimator as $c(z)(2h)^2 = k/n$. As the mean of copula density $c(\alpha)$ is one, $h = 0.25n^{-1/4}$ corresponds to $k = n(2h)^2 = 0.25\sqrt{n}$. The simulation results for KNN-based $\hat{RCD}$ with $k = 0.25\sqrt{n}$, $k = 0.1\sqrt{n}$ and $k = 0.5\sqrt{n}$ are plotted in Figures A.5 - A.7. Similar pattern as in KDE-based estimator are observed. Hence we propose the practical KNN-based $\hat{RCD}$ to use $k = 0.25\sqrt{n}$.
Furthermore, we also checked the KNN-based $\hat{RCD}$ on the mixture noise setting used in definition 2: a proportion ($p$) of deterministic function is hidden in independent continuous noise. Six types of deterministic function are used, as listed in Table A.2. When $n = 5000$, the $\hat{RCD}$ is close to the true value $p$ in the simulations. And compared to the two choices of $k = 0.1\sqrt{n}$ and $k = 0.5\sqrt{n}$, $k = 0.25\sqrt{n}$ provide a good balance of approximating the true values when $RCD$ is small or large.

<table>
<thead>
<tr>
<th>A</th>
<th>Linear</th>
<th>$y = x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Centered Quadratic</td>
<td>$y = 4(x - 1/2)^2$</td>
</tr>
<tr>
<td>C</td>
<td>Cosine</td>
<td>$y = \cos(4\pi x)$</td>
</tr>
<tr>
<td>D</td>
<td>Cross</td>
<td>$y = \pm x1_{{0 \leq x \leq 1}}$</td>
</tr>
<tr>
<td>E</td>
<td>Circle</td>
<td>$(2x - 1)^2 + y^2 = 1$</td>
</tr>
<tr>
<td>F</td>
<td>Cross 2</td>
<td>$y = \pm(x - 1/2)1_{{0 \leq x \leq 1}}$</td>
</tr>
</tbody>
</table>

Table A.2: The function relationships used in Figures A.8.
Figure A.2: The comparison of $RCD$ with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.25n^{-1/4}$. 
Figure A.3: The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.1n^{-1/4}$. 
Figure A.4: The comparison of $RCD$ with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.5n^{-1/4}$. 
Figure A.5: Additive noise with $k = 0.25\sqrt{n}$. 
Figure A.6: Additive noise with $k = 0.1 \sqrt{n}$. 
Figure A.7: Additive noise with $k = 0.5\sqrt{n}$. 
Figure A.8: Mixture noise with $k = c\sqrt{n}$, where $c = 0.1, 0.25, 0.5$. 

127
Appendix B

Additional Proofs and Results for Chapter 3

B.1 Proof of Theorem 13

Let $X_m = (X_m^{(1)}, \cdots, X_m^{(p)})^\top$, $m = 1, 2, \cdots, n$. It is easy to see that $\tilde{\sigma}_{ij}^{st}$, the $(i, j)^{th}$ element of the $(s, t)^{th}$ sub-matrix corresponds to the $(i^*, j^*)^{th}$ element of the global matrix $\tilde{\Sigma}$, where $i^*(s, i) = (s - 1)p_1 + i$ and $j^*(t, j) = (t - 1)p_2 + j$. Therefore, we have $\tilde{\sigma}_{ij}^{st} = \frac{1}{n} \sum_{m=1}^{n} X_m^{(i^*(s, i))} X_m^{(j^*(t, j))}$. In the rest of the proof, we will write $i^*(s, i)$ and $j^*(t, j)$ more compactly as $i^*$ and $j^*$ if their meaning is clear from the text. Now we have $E\tilde{\sigma}_{ij}^{st} = \sigma_{ij}^{st}$, and $Var(\tilde{\sigma}_{ij}^{st}) = \frac{1}{n} Var(X_m^{(i^*)} X_m^{(j^*)}) \leq \frac{1}{n} E[(X_m^{(i^*)} X_m^{(j^*)})^2] \leq \frac{1}{n} \sqrt{E(X_m^{(i^*)})^4} \sqrt{E(X_m^{(j^*)})^4} \leq C.$

For the two-step definition of (2) in the original text, denote the coefficient (weight) of $\tilde{\sigma}_{ij}$ inside the sub-block matrix in the first step as $W = \{w_{ij}^{st}\}$, while denote the weight in the second step among sub-block matrices by $V = \{v_{ij}^{st}\}$. Then we have $\tilde{\Sigma} = V \circ W \circ \tilde{\Sigma}$, where $\circ$ is the Hadamard product. Thus for the tapering estimator in (2) in the original text, the bias of $\hat{\sigma}_{ij}^{st}$ is $(1 - v_{ij}^{st}w_{ij}^{st})\sigma_{ij}^{st}$ and the variance of $\hat{\sigma}_{ij}^{st}$ is $(v_{ij}^{st}w_{ij}^{st})^2 Var(\tilde{\sigma}_{ij}^{st})$. Using the bound
on the variance above, we have the element-wise risk bound

$$E(\hat{\sigma}_{ij}^t - \sigma_{ij}^t)^2 = E(v_{ij}^t w_{ij}^t \hat{\sigma}_{ij}^t - \sigma_{ij}^t)^2 \leq (1 - v_{ij}^t w_{ij}^t)^2 (\sigma_{ij}^t)^2 + (v_{ij}^t w_{ij}^t)^2 \frac{C}{n}. \quad (B.1)$$

Apply this element-wise risk bound with Lemma 1, we get

$$\frac{1}{p} E \left\| \hat{\Sigma} - \Sigma \right\|_F^2 \leq \frac{1}{p} \sum_{1 \leq s, t \leq p_1} \sum_{1 \leq i, j \leq p_2} [(1 - v_{ij}^t w_{ij}^t)^2 (\sigma_{ij}^t)^2 + (v_{ij}^t w_{ij}^t)^2 \frac{C}{n}]$$

Note that when both $|i - j| < \frac{k}{2}$ and $|s - t| < \frac{l}{2}$, $(1 - v_{ij}^t w_{ij}^t)^2 (\sigma_{ij}^t)^2 = 0$ which can be dropped. Otherwise, it can be upper bounded by $(\sigma_{ij}^t)^2$. Likewise, when either $|i - j| > k$ or $|s - t| > l$, $(v_{ij}^t w_{ij}^t)^2 \frac{C}{n} = 0$. Otherwise, it can be upper bounded by $\frac{C}{n}$. Taken together, the above expression can be upper bounded by the sum of $R_1 = \frac{1}{p} \sum_{|s-t| \leq l} \sum_{|i-j| \leq k} \frac{C}{n}$ and $R^* = \frac{1}{p} \sum_{|s-t| \geq \frac{l}{2}} \sum_{|i-j| \geq \frac{k}{2}} (\sigma_{ij}^t)^2$. Furthermore, the second term $R^* \leq R_2 + R_3 + R_4$ where

$$R_2 = \frac{1}{p} \sum_{|s-t| < \frac{l}{2}} \sum_{|i-j| \geq \frac{k}{2}} (\sigma_{ij}^t)^2, \quad R_3 = \frac{1}{p} \sum_{|s-t| \geq \frac{l}{2}} \sum_{|i-j| < \frac{k}{2}} (\sigma_{ij}^t)^2, \quad R_4 = \frac{1}{p} \sum_{|s-t| \geq \frac{l}{2}} \sum_{|i-j| \geq \frac{k}{2}} (\sigma_{ij}^t)^2.$$

Then we have

$$R_1 = \frac{1}{p} \sum_{|s-t| \leq l} \sum_{|i-j| \leq k} \frac{C}{n} \leq \frac{1}{p} C p_1 k p_2 \frac{1}{n} = C \frac{k l}{n}$$

$$R_2 = \frac{1}{p} \sum_{|s-t| < \frac{l}{2}} \sum_{|i-j| \geq \frac{k}{2}} (\sigma_{ij}^t)^2 \leq \frac{C}{p} l p_1 p_2 \frac{1}{k^{2\alpha + 1}} = C l k^{-2\alpha + 1}$$

Similarly, we have $R_3 \leq C k l^{-2\beta + 1}$. Moreover,

$$R_4 = \frac{1}{p} \sum_{|s-t| \geq \frac{l}{2}} \sum_{|i-j| \geq \frac{k}{2}} (\sigma_{ij}^t)^2 \leq \frac{1}{p} C [p(p_1 - \frac{l}{2}) k^{-2\alpha + 1} \land p(p_2 - \frac{k}{2}) l^{-2\beta + 1}]$$

Summing over $R_1$, $R_2$, $R_3$ and $R_4$, the above inequalities imply that for some constant $C > 0$,

$$E \frac{1}{p} \left\| \hat{\Sigma} - \Sigma \right\|_F^2 \leq C \frac{k l}{n} + C \left[ k^{-2\alpha + 1} + \frac{k l}{2} l^{-2\beta + 1} + (p_1 - \frac{l}{2}) k^{-2\alpha + 1} + (p_2 - \frac{k}{2}) l^{-2\beta + 1} \right]$$

$$= C \left( \frac{k l}{n} + p_1 k^{-2\alpha + 1} + p_2 l^{-2\beta + 1} \right)$$

$$\leq C n \frac{(2\alpha - 1)(2\beta - 1)}{4\alpha\beta - 1} p_1^{2\alpha - 1} p_2^{2\beta - 1} \frac{2\alpha - 1}{p_1^{4\alpha\beta - 1} p_2^{4\alpha\beta - 1}} \frac{2\beta}{p_1^{4\alpha\beta - 1} p_2^{4\alpha\beta - 1}}. \quad (B.2)$$

In the last step, the optimal $k$ and $l$ is chosen to make all three terms in the second to last expression to be of the same order. That is,

$$k^* = n \frac{2\beta - 1}{4\alpha\beta - 1} p_1^{2\beta - 1} p_2^{2\alpha - 1}, \quad l^* = n \frac{2\alpha - 1}{4\alpha\beta - 1} p_1^{2\alpha - 1} p_2^{2\beta - 1}. \quad (B.3)$$
In the case of $k^* > p_2$ and $l^* > p_1$, we take $k = p_2$ and $l = p_1$. Thus

$$E \frac{1}{p} \left\| \hat{\Sigma} - \Sigma \right\|_F^2 \leq C \frac{p}{n}. \quad (B.4)$$

The equations (B.2) and (B.4) provide the upper bound in Theorem 13.

### B.2 Proof of Lemma 16

Using the Lemma 15, we only need to bound from above the KL divergence between two perturbations of the distributions within the class.

$$KL(\mathbb{P}_\theta, \mathbb{P}_{\hat{\theta}}) = n \left[ \frac{1}{2} Tr(\Sigma(\hat{\theta})\Sigma^{-1}(\theta)) - \frac{1}{2} \log \det(\Sigma(\hat{\theta})\Sigma^{-1}(\theta)) - \frac{p}{2} \right] \quad (B.5)$$

Let $\Delta = \Sigma(\hat{\theta}) - \Sigma(\theta)$. Then, we have

$$Tr(\Sigma(\hat{\theta})\Sigma^{-1}(\theta)) - p = Tr(\Delta\Sigma^{-1}(\theta)). \quad (B.6)$$

Let the eigenvalues of $\Delta\Sigma^{-1}(\theta)$ be $\{\lambda_i\}_{i=1}^p$. Note that the eigenvalues of $I_p + \Delta\Sigma^{-1}(\theta)$ are $1 + \lambda_i$s, and with Taylor expansion we have

$$\log \det(\Sigma(\hat{\theta})\Sigma^{-1}(\theta)) = \log \det(I + \Delta\Sigma^{-1}(\theta)) = Tr(\Delta\Sigma^{-1}(\theta)) - r_0, \quad (B.7)$$

with $r_0 \leq C_0 \sum_{i=1}^p \lambda_i^2$ for some constant $C_0$.

Using (B.6) and (B.7), the trace term cancels in (B.5), and we have $KL(\mathbb{P}_\theta, \mathbb{P}_{\hat{\theta}}) \leq nr_0/2$. Since $\|\Delta\|_F^2$ is of order $1/n$, $KL(\mathbb{P}_\theta, \mathbb{P}_{\hat{\theta}})$ is bounded by a positive constant $C$. Then (3.13) implies $\min_{H(\theta, \hat{\theta})=1} ||\mathbb{P}_\theta \land \mathbb{P}_{\hat{\theta}}|| \geq C_1$, for some constant $C_1 = 0.5e^{-C} > 0$. 

130
Appendix C

Additional Proofs and Results for Chapter 4

C.1 Proof of Theorem 20

Proof. To solve the quadratic optimization problem, we apply the method of Lagrangian multiplier by converting the constrained problem to an unconstrained problem. Specifically, the Lagrangian of the problem is

\[ L(\lambda, w) = \frac{1}{2} \lambda^\top \tilde{\Sigma} \lambda - w(\lambda^\top \mu - 1), \] (C.1)

by introducing the additional parameter \( w \). By the necessary condition of the optimal solution, we set the first derivative of the Lagrangian in (C.1) to zero and get

\[ \frac{\partial}{\partial \lambda} L(\lambda, w) = \tilde{\Sigma} \lambda - w \mu = 0 \] (C.2)

\[ \frac{\partial}{\partial w} L(\lambda, w) = \lambda^\top \mu - 1 = 0 \] (C.3)

Thus, we have from (C.2)

\[ \lambda = w \tilde{\Sigma}^{-1} \mu. \] (C.4)
Plugging (C.4) into the constrain (C.3), we have

\[ w = \frac{1}{\mu^\top \tilde{\Sigma}^{-1} \mu}. \tag{C.5} \]

Combining (C.4) and (C.5) we have

\[ \lambda = \frac{\tilde{\Sigma}^{-1} \mu}{\mu^\top \tilde{\Sigma}^{-1} \mu}, \tag{C.6} \]

which is the desired optimal weight.

\[ \square \]

### C.2 Computation of Variance \( \tilde{\Sigma} \) of Sample Covariance Elements

The optimal weights in equation (4.12) depends on matrix \( \tilde{\Sigma} \), the covariance structure of the sample covariance matrix elements, as in equations (4.13) and (4.29). Here we derive a general formula which can be used to calculate \( \tilde{\Sigma} \).

Let \( X \) be our \( n \) by \( p \) design matrix with normalized columns, and the rows of \( X \) are iid observations from multivariate Gaussian distribution with covariance \( \Sigma \). Then from the multivariate statistics theory, the cross-product matrix \( M = X^\top X \) follows the Wishart distribution with \( n - 1 \) degrees of freedom, i.e., \( \text{Wish}(\Sigma, n - 1) \). Therefore, the sample covariance matrix \( \hat{\Sigma} = M/(n - 1) \) follows the distribution \( \text{Wish}(\Sigma, n - 1)/(n - 1) \). The following Lemma about the Wishart distribution can then be used to describe \( \hat{\Sigma} \).

**Lemma 30.** Let \( S \) follows the \( p \) dimensional Wishart distribution with a positive definite scale matrix \( \Sigma_{p \times p} \) and the degree of freedom \( m \), i.e. \( S \sim \text{Wish}(\Sigma, m) \). Moreover, let \( \Sigma = CC^\top \) be the Cholesky decomposition of the matrix \( \Sigma \) with a \( p \) by \( p \) matrix \( C \). Then we have

\[
\text{cov}(\text{vec}(S)) = m(C \otimes C)(I_p \otimes I_p + T_p)(C^\top \otimes C^\top),
\]

\[
= m[\Sigma \otimes \Sigma + (C \otimes C)T_p(C^\top \otimes C^\top)], \tag{C.7}
\]

where \( T_p \) is the \( p \) by \( p \) transformation or permutation matrix that transfer the column stacked vector \( \text{vec}(S) \) to row stacked vector \( \text{vec}(S^\top) \), i.e. \( \text{vec}(S^\top) = T_p \text{vec}(S) \).
Proof. Let $\Sigma = CC^\top$ be the Cholesky decomposition of the matrix $\Sigma$ with a $p$ by $p$ matrix $C$. Then, $x_i$ can be written with the linear combination with standardized variables $\{z, z_i\}$, $x_i = Cz_i$ with $E(z_i z_i^\top) = I_p$. Thus,

$$
cov(\text{vec}(S)) = m\text{cov}(\text{vec}(Czz^\top C^\top))
= m\text{cov}((C \otimes C)\text{vec}(zz^\top))
= m(C \otimes C)\text{cov}(z \otimes z)(C^\top \otimes C^\top)
= m(C \otimes C)(I_p \otimes I_p + T_p)(C^\top \otimes C^\top),
$$

where the last equality uses the facts that: $\text{V}(z_k^2) = 2$, $\text{V}(z_k z_l) = 1$, $\text{cov}(z_k z_l, z_k z_l) = 1$, $k \neq l$ and $\text{cov}(z_k z_l, z_p z_q) = 0$ with $k, l, p, q$ in other combinations.

Lemma 30 provides the covariance matrix formula for the elements in the sample covariance matrix $\hat{\Sigma}$. In real applications, the dimension of this covariance matrix $\text{cov}(\text{vec}(\hat{\Sigma}))$ can be very large. However, we only need $\hat{\Sigma}$ which is a small portion of $\text{cov}(\text{vec}(\Sigma))$. Therefore, we can work with individual elements of the matrix in (C.7). Notice that the $i^{th}$ row of $C \otimes C$ in (C.7) can be computed as

$$(C \otimes C)_{i,:} = \text{vec}(C_{j_0, i_0,:}^\top, C_{i_0,:})^\top,$$

where $i = (j_0 - 1)p + i_0$ and $A_{i,:}$ is the $i^{th}$ row of matrix $A$. In this way, each single element of $\text{cov}(\text{vec}(\Sigma))$ can be obtained without calculating the whole matrix.

C.3 Proof of Corollary 21

While Appendix C.2 provides a general formula to compute $\hat{\Sigma}$ in (4.13), here we get explicit expressions for parameters in (4.13) through direct calculation.
Proof. It is enough to show the explicit form of the parameters defined in (4.13). In fact, based on the FNCM (4.6), we have

\[
\hat{\sigma}_{i^*,j^*} = (a(f_i^* - 1 + f_{i^*} + a f_{i^* + 1} + \epsilon_{i^*})(a f_{j^* - 1} + \rho f_{j^*} + \sqrt{1 - \rho^2} f_{j^*} + a f_{j^* + 1} + \epsilon_{j^*})
\]

\[
\hat{\sigma}_{i^*-1,j^*} = (a(f_i^* - 2 + f_{i^*-1} + a f_{i^*} + \epsilon_{i^*-1})(a f_{j^* - 1} + \rho f_{j^*} + \sqrt{1 - \rho^2} f_{j^*} + a f_{j^* + 1} + \epsilon_{j^*})
\]

Thus we calculate \( r_1 \) as

\[
r_1 = \text{cov}(\hat{\sigma}_{i^*,j^*}, \hat{\sigma}_{i^*-1,j^*})
\]

\[
= \text{cov}[(a(f_i^* - 1 + f_{i^*} + a f_{i^* + 1} + \epsilon_{i^*})(a f_{j^* - 1} + \rho f_{j^*} + \sqrt{1 - \rho^2} f_{j^*} + a f_{j^* + 1} + \epsilon_{j^*}),
\]

\[
(a(f_i^* - 2 + f_{i^*-1} + a f_{i^*} + \epsilon_{i^*-1})(a f_{j^* - 1} + \rho f_{j^*} + \sqrt{1 - \rho^2} f_{j^*} + a f_{j^* + 1} + \epsilon_{j^*})].
\]

Notice that the above covariance contains terms that, after taking the multiplicative constants out, have the basic form \( \text{cov}(g_1 g_2, g_3 g_4) \) where \( g_i \in \{f_j\}_{j=1}^p \cup \{\sigma_k^e\}_{k=1}^p \) for \( i = 1, 2, 3, 4 \). Notice that the elements in the \( \{f_j\}_{j=1}^p \cup \{\sigma_k^e\}_{k=1}^p \) are all i.i.d. standard normal random variable. Hence it is easy to see that, \( \text{cov}(g_1 g_2, g_3 g_4) = 0 \) if \( \text{Card}(\{g_1, g_2, g_3, g_4\}) \geq 3 \), where \( \text{Card}(\cdot) \) denotes the cardinality of a set. Thus, after expanding the terms in the above
covariance and dropping the zero terms, we can simplify the expression of $r_1$ as:

$$
\begin{align*}
    r_1 &= \text{cov}(f_1, \epsilon_j) + \text{cov}(a f_{i-1} \epsilon_j, f_{i-1} \epsilon_j) + \text{cov}(a f_{i-1} a f_{j-1}, f_{i-1} a f_{j-1}) \\
    &\quad + \text{cov}(a f_{i-1} \rho f_{i}, f_{i-1} \rho f_{i}) + \text{cov}(a f_{i-1} \sqrt{1 - \rho^2 f_{j}}, f_{i-1} \sqrt{1 - \rho^2 f_{j}})
    \\
    &\quad + \text{cov}(a f_{i-1} a f_{j+1}, f_{i-1} a f_{j+1}) + \text{cov}(f_{i} a f_{j-1}, f_{i} a f_{j-1}) + \text{cov}(f_{i} \rho f_{i}, a f_{i} \rho f_{i}) \\
    &\quad + \text{cov}(f_{i} \sqrt{1 - \rho^2 f_{j}}, a f_{i} \sqrt{1 - \rho^2 f_{j}}) + \text{cov}(f_{i} a f_{j+1}, a f_{i} a f_{j+1})
    \\
    &= \text{acov}(f_1, \epsilon_j) + \text{acov}(f_{i-1} \epsilon_j, f_{i-1} \epsilon_j) + a^3 \text{cov}(f_{i-1} f_{j-1}, f_{i-1} f_{j+1}) \\
    &\quad + a \rho^2 \text{cov}(f_{i-1} f_{i}, f_{i-1} f_{i}) + a(1 - \rho^2) \text{cov}(f_{i-1} f_{j}, f_{i-1} f_{j})
    \\
    &\quad + a^3 \text{cov}(f_{i-1} f_{j+1}, f_{i-1} f_{j+1}) + a^3 \text{cov}(f_{i} f_{j-1}, f_{i} f_{j-1}) + a \rho^2 \text{cov}(f_{i} f_{i}, f_{i} f_{i}) \\
    &\quad + a(1 - \rho^2) \text{cov}(f_{i} f_{j}, f_{i} f_{j}) + a^3 \text{cov}(f_{i} f_{j+1}, f_{i} f_{j+1})
    \\
    &= 2a \sigma^2 + a^3 + a \rho^2 + a(1 - \rho^2) + a^3 + 2a \rho^2 + a(1 - \rho^2) + a^3
    \\
    &= 2a \sigma^2 + (4a^3 + 2a + a \rho^2),
\end{align*}
$$

where the third equality follows the fact that $\text{cov}(g^2, g^2) = 2$ and $\text{cov}(gg', gg') = 1$ for all $g, g' \in \{f_j\}_{j=1}^p \cup \{\frac{1}{\sigma} f_k\}_{k=1}^p$ and $g \neq g'$.

The explicit expressions of $d_1, d_2, s_1, s_2$ can be derived through similar direction calculations. In fact, we have

$$
\begin{align*}
    \hat{\sigma}_{i, j} &= (a f_{i-1} + f_{i} + a f_{i+1} + \epsilon_i)(a f_{j-1} + \rho f_{i} + \sqrt{1 - \rho^2 f_{j}} + a f_{j+1} + \epsilon_j) \\
    \hat{\sigma}_{i-1, j} &= (a f_{i-2} + f_{i-1} + a f_{i} + \epsilon_{i-1})(a f_{j-1} + \rho f_{i} + \sqrt{1 - \rho^2 f_{j}} + a f_{j+1} + \epsilon_j) \\
    \hat{\sigma}_{i, j+1} &= (a f_{i-1} + f_{i} + a f_{i+1} + \epsilon_i)(a f_{j} + \sqrt{1 - \rho^2 f_{j}} + f_{j+1} + a f_{j+2} + \epsilon_{j+1}) \\
    \hat{\sigma}_{i+1, j} &= (a f_{i} + f_{i+1} + a f_{i+2} + \epsilon_{i+1})(a f_{j-1} + \rho f_{i} + \sqrt{1 - \rho^2 f_{j}} + a f_{j+1} + \epsilon_{j+1}) \\
    \hat{\sigma}_{i, j-1} &= (a f_{i-1} + f_{i} + a f_{i+1} + \epsilon_i)(a f_{j-2} + f_{j-1} + a f_{i} + \sqrt{1 - \rho^2 f_{j}} + \epsilon_{j-1})
\end{align*}
$$

And thus,

$$
\begin{align*}
    d_1 &= \text{cov}(\hat{\sigma}_{i, j}, \hat{\sigma}_{i, j}) \\
    &= \text{cov}[(a f_{i-1} + f_{i} + a f_{i+1} + \epsilon_i)(a f_{j-1} + \rho f_{i} + \sqrt{1 - \rho^2 f_{j}} + a f_{j+1} + \epsilon_{j}), \\
    &\quad (a f_{i-1} + f_{i} + a f_{i+1} + \epsilon_i)(a f_{j-1} + \rho f_{i} + \sqrt{1 - \rho^2 f_{j}} + a f_{j+1} + \epsilon_{j})]
\end{align*}
$$

135
\[ d_2 = \text{cov}(\sigma_{i^*,-1,j^*}, \sigma_{i^*,-1,j^*}) \]
\[ = \text{cov}[(a f_{i^*-2} + f_{i^*-1} + a f_{i^*} + \epsilon_{i^*-1})(a f_{j^*-1} + \rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*} + a f_{j^*-1} + \epsilon_{j^*}), \]
\[ (a f_{i^*-2} + f_{i^*-1} + a f_{i^*} + \epsilon_{i^*-1})(a f_{j^*-1} + \rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*} + a f_{j^*-1} + \epsilon_{j^*})] \]
\[ = \text{cov}[a^2 f_{i^*-2} f_{j^*-1} + a \rho f_{i^*-2} f_{i^*} + a \sqrt{1 - \rho^2} f_{i^*-2} f_{j^*} + a^2 f_{i^*-1} f_{j^*-1} + a f_{i^*-2} \epsilon_{i^*} \epsilon_{j^*} \]
\[ + a f_{i^*-1} f_{j^*-1} + \rho f_{i^*-1} f_{i^*} + \sqrt{1 - \rho^2} f_{i^*-1} f_{j^*} + a f_{i^*-1} f_{j^*-1} + f_{i^*-1} \epsilon_{j^*}, \]
\[ + a^2 f_{i^*-1} f_{j^*-1} + a \rho f_{i^*-1} f_{i^*} + a \sqrt{1 - \rho^2} f_{i^*-1} f_{j^*} + a^2 f_{i^*-1} f_{j^*-1} + a f_{i^*-1} \epsilon_{j^*} \]
\[ + a \epsilon_{i^*-1} f_{j^*-1} + \rho \epsilon_{i^*-1} f_{i^*} + \sqrt{1 - \rho^2} \epsilon_{i^*-1} f_{j^*} + a \epsilon_{i^*-1} f_{j^*-1} + \epsilon_{i^*-1} \epsilon_{j^*}, \]
\[ a^2 f_{i^*-2} f_{j^*-1} + a \rho f_{i^*-2} f_{i^*} + a \sqrt{1 - \rho^2} f_{i^*-2} f_{j^*} + a^2 f_{i^*-1} f_{j^*-1} + a f_{i^*-2} \epsilon_{i^*} \epsilon_{j^*} \]
\[ + a f_{i^*-1} f_{j^*-1} + \rho f_{i^*-1} f_{i^*} + \sqrt{1 - \rho^2} f_{i^*-1} f_{j^*} + a f_{i^*-1} f_{j^*-1} + f_{i^*-1} \epsilon_{j^*}, \]
\[ + a^2 f_{i^*-1} f_{j^*-1} + a \rho f_{i^*-1} f_{i^*} + a \sqrt{1 - \rho^2} f_{i^*-1} f_{j^*} + a^2 f_{i^*-1} f_{j^*-1} + a f_{i^*-1} \epsilon_{j^*} \]
\[ + a \epsilon_{i^*-1} f_{j^*-1} + \rho \epsilon_{i^*-1} f_{i^*} + \sqrt{1 - \rho^2} \epsilon_{i^*-1} f_{j^*} + a \epsilon_{i^*-1} f_{j^*-1} + \epsilon_{i^*-1} \epsilon_{j^*}] \]
\[ = (a^4 + a^2 \rho^2 + a^2 (1 - \rho^2) + a^4 + a^2 + \rho^2 + (1 - \rho^2) + a^2 + a^4 + 2a^2 \rho^2 + a^2 (1 - \rho^2) + a^4) \]
\[ + (a^2 + 1 + a^2 + a^2 + 1 + a^2)\sigma_e^2 + \sigma_e^4 \]
\[ = 4a^4 + 4a^2 + a^2\rho^2 + 1 + (4a^2 + 2)\sigma_e^2 + \sigma_e^4. \]

\[ s_1 = \text{cov}(\hat{\sigma}_{i^* -1,j^*}, \hat{\sigma}_{i^*,j^*+1}) \]
\[ = \text{cov}[(af_{i^* -1} + f_{i^* -1} + af_{i^*} + \epsilon_{i^* -1})(af_{j^* -1} + \rho f_{i^*} + \sqrt{1 - \rho^2} f_{j^*} + af_{j^*+1} + \epsilon_{j^*}), \]
\[ (af_{i^* -1} + f_{i^*} + af_{i^*+1} + \epsilon_{i^*})(ap f_{i^*} + a\sqrt{1 - \rho^2} f_{j^*} + f_{j^*+1} + af_{j^*+2} + \epsilon_{j^*+1})] \]
\[ = \text{cov}[a^2 f_{i^* -2}f_{j^* -1} + ap f_{i^* -2}f_{i^*} + a\sqrt{1 - \rho^2} f_{i^* -2}f_{j^*} + a^2 f_{i^* -2}f_{j^*+1} + af_{i^* -2}\epsilon_{j^*}, \]
\[ + af_{i^* -1}f_{j^* -1} + \rho f_{i^* -1}f_{i^*} + \sqrt{1 - \rho^2} f_{i^* -1}f_{j^*} + af_{i^* -1}f_{j^*+1} + f_{i^* -1}\epsilon_{j^*}, \]
\[ + a^2 f_{i^*}f_{j^* -1} + ap f_{i^*}f_{i^*} + a\sqrt{1 - \rho^2} f_{i^*}f_{j^*} + a^2 f_{i^*}f_{j^*+1} + af_{i^*}\epsilon_{j^*}, \]
\[ + a\epsilon_{i^* -1}f_{j^* -1} + \rho \epsilon_{i^* -1}f_{i^*} + \sqrt{1 - \rho^2} \epsilon_{i^* -1}f_{j^*} + a\epsilon_{i^* -1}f_{j^*+1} + \epsilon_{i^* -1}\epsilon_{j^*}, \]
\[ a^2 \rho f_{i^* -1}f_{i^*} + a^2 \sqrt{1 - \rho^2} f_{i^* -1}f_{j^*} + af_{i^* -1}f_{j^*+1} + a^2 f_{i^* -1}f_{j^*+2} + af_{i^* -1}\epsilon_{j^*+1} \]
\[ + ap f_{i^*}f_{i^*} + a\sqrt{1 - \rho^2} f_{i^*}f_{j^*} + f_{i^*}f_{j^*+1} + af_{i^*}f_{j^*+2} + f_{i^*}\epsilon_{j^*+1} \]
\[ + a^2 \rho f_{i^*+1}f_{i^*} + a^2 \sqrt{1 - \rho^2} f_{i^*+1}f_{j^*} + af_{i^*+1}f_{j^*+1} + a^2 f_{i^*+1}f_{j^*+2} + af_{i^*+1}\epsilon_{j^*+1} \]
\[ + a\rho \epsilon_{i^*}f_{i^*} + a\sqrt{1 - \rho^2} \epsilon_{i^*}f_{j^*} + \epsilon_{i^*}f_{j^*+1} + a\epsilon_{i^*}f_{j^*+2} + \epsilon_{i^*}\epsilon_{j^*+1}] \]
\[ = 2a^2 \rho^2 + a^2(1 - \rho^2) + a^2 + 2a^2 \]
\[ = 4a^2 + a^2\rho^2 \]

\[ s_2 = \text{cov}(\hat{\sigma}_{i^*,j^* -1}, \hat{\sigma}_{i^*,j^*+1}) \]
\[ = \text{cov}[(af_{i^* -1} + f_{i^*} + af_{i^*+1} + \epsilon_{i^*})(ap f_{i^*} + a\sqrt{1 - \rho^2} f_{j^*} + f_{j^* -1} + af_{j^* -2} + \epsilon_{j^* -1}), \]
\[ (af_{i^* -1} + f_{i^*} + af_{i^*+1} + \epsilon_{i^*})(ap f_{i^*} + a\sqrt{1 - \rho^2} f_{j^*} + f_{j^* +1} + af_{j^* +2} + \epsilon_{j^* +1})] \]
\[ = \text{cov}[a^2 \rho f_{i^* -1}f_{i^*} + a^2 \sqrt{1 - \rho^2} f_{i^* -1}f_{j^*} + af_{i^* -1}f_{j^* -1} + a^2 f_{i^* -1}f_{j^* -2} + af_{i^* -1}\epsilon_{j^* -1} \]
\[ + ap f_{i^*}f_{i^*} + a\sqrt{1 - \rho^2} f_{i^*}f_{j^*} + f_{i^*}f_{j^* -1} + af_{i^*}f_{j^* -2} + f_{i^*}\epsilon_{j^* -1} \]
\[ + a^2 \rho f_{i^*+1}f_{i^*} + a^2 \sqrt{1 - \rho^2} f_{i^*+1}f_{j^*} + af_{i^*+1}f_{j^* -1} + a^2 f_{i^*+1}f_{j^* -2} + af_{i^*+1}\epsilon_{j^* -1} \]
Thus, we have

\[
\begin{align*}
    d_1 &= 4a^4 + 4a^2 + \rho^2 + 1 + (4a^2 + 2)\sigma_\epsilon^2 + \sigma_\epsilon^4, \\
    d_2 &= 4a^4 + 4a^2 + a^2\rho^2 + 1 + (4a^2 + 2)\sigma_\epsilon^2 + \sigma_\epsilon^4, \\
    r_1 &= 4a^3 + 2a + a\rho^2 + 2a\sigma_\epsilon^2, \\
    s_1 &= 4a^2 + a^2\rho^2, \\
    s_2 &= 2a^4 + a^2 + a^2\rho^2 + a^2\sigma_\epsilon^2.
\end{align*}
\]  
(C.10)

Then, base on Corollary 21, we have the desired result. \qed
C.4 Expressions of entries in (4.29)

By direct calculations similar to those in Appendix C.3, model (4.28) implies

\[
\begin{align*}
  d_1 &= 16a^4 + 8a^2 + \rho^2 + 1 + (4a^2 + 2)\sigma_\epsilon^2 + \sigma_\epsilon^4, \\
  d_2 &= 16a^4 + 8a^2 + a^2\rho^2 + 1 + (4a^2 + 2)\sigma_\epsilon^2 + \sigma_\epsilon^4, \\
  r_1 &= 8a^3 + 2a + a\rho^2 + 2a\sigma_\epsilon^2, \\
  s_1 &= 4a^2 + a^2\rho^2, \\
  s_2 &= 4a^4 + a^2\rho^2 + a^2\sigma_\epsilon^2, \\
  s_3 &= 8a^4 + 2a^2 + a^2\rho^2 + 2a^2\sigma_\epsilon^2.
\end{align*}
\]

For example, we have

\[
d_1 = \text{cov}(\hat{\sigma}_{i^*j^*}, \hat{\sigma}_{i^*j^*})
= \text{cov}[(f_{i^*} + a(f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) + \epsilon_{i^*})(\rho f_{j^*} + \sqrt{1 - \rho^2} f_{j^*})
+ a(f_{j^*+1} + f_{j^*-1} + f_{j^*+p_y} + f_{j^*-p_y}) + \epsilon_{j^*}),
(f_{i^*} + a(f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) + \epsilon_{i^*})(\rho f_{j^*} + \sqrt{1 - \rho^2} f_{j^*})
+ a(f_{j^*+1} + f_{j^*-1} + f_{j^*+p_y} + f_{j^*-p_y}) + \epsilon_{j^*})]
= \text{cov}[\rho f_{i^*} f_{j^*} + \sqrt{1 - \rho^2} f_{i^*} f_{j^*} + a f_{i^*}(f_{j^*+1} + f_{j^*-1} + f_{j^*+p_y} + f_{j^*-p_y}) + f_{i^*} \epsilon_{j^*}
+ \rho a(f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) f_{i^*} + \sqrt{1 - \rho^2} a(f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) f_{j^*}
+ a^2 (f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) (f_{j^*+1} + f_{j^*-1} + f_{j^*+p_y} + f_{j^*-p_y})
+ a (f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) \epsilon_{j^*}
+ \rho \epsilon_{i^*} f_{i^*} + \sqrt{1 - \rho^2} \epsilon_{i^*} f_{j^*} + a \epsilon_{i^*} (f_{j^*+1} + f_{j^*-1} + f_{j^*+p_y} + f_{j^*-p_y}) + \epsilon_{i^*} \epsilon_{j^*},
\rho f_{i^*} f_{j^*} + \sqrt{1 - \rho^2} f_{i^*} f_{j^*} + a f_{i^*} (f_{j^*+1} + f_{j^*-1} + f_{j^*+p_y} + f_{j^*-p_y}) + f_{i^*} \epsilon_{j^*}
+ \rho a (f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) f_{i^*} + \sqrt{1 - \rho^2} a (f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) f_{j^*}
+ a^2 (f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y}) (f_{j^*+1} + f_{j^*-1} + f_{j^*+p_y} + f_{j^*-p_y})]
\]
\[ + a(f_{i^*+1} + f_{i^*-1} + f_{i^*+p_y} + f_{i^*-p_y})e_{j^*} \]
\[ + \rho e_{i^*} f_{i^*} + \sqrt{1 - \rho^2} e_{j^*} f_{j^*} + a e_{i^*} (f_{j^*+1} + f_{j^*-1} + f_{j^*+p_y} + f_{j^*-p_y}) + e_{i^*} e_{j^*}] \]
\[ = (1 + 4a^2 + 4(a^2 + 4a^4) + a^2 \rho^2) + \rho^2 + 1 + (4a^2 + 2)\sigma_e^2 + \sigma_e^4 \]
\[ = 16a^4 + 8a^2 + \rho^2 + 1 + (4a^2 + 2)\sigma_e^2 + \sigma_e^4. \]

Similar results hold for \(d_2, r_1, s_1, s_2, s_3\).

C.5 Proof of Theorem 24

Proof. To get the estimation result of \(a\), we first consider the performance of \(\bar{k}_1\) and \(\bar{k}_2\).

Here, we show that \(\bar{k}_1 = 2a + O_p\left(\frac{1}{\sqrt{np}}\right)\) and \(\bar{k}_2 = a^2 + O_p\left(\frac{1}{\sqrt{np}}\right)\).

Recall that \(\bar{k}_1 = \sum_{p,j=1,|i-j|=1}^p \hat{\sigma}_{ij} \) with \(\hat{\sigma}_{ij} = \frac{1}{n-1} \sum_{k=1}^n x_i^{(k)} x_j^{(k)}\). By the unbiasedness of \(\hat{\sigma}_{ij}\) and the far-near covariance model where \(\sigma_{i,j} = 2a\), for \(|i-j|=1\), we have

\[ E[\hat{\sigma}_{i,i+1}] = \sigma_{i,i+1} = 2a. \] (C.12)

Moreover, since we have \(n\) iid data \((x_i^{(k)})\) for each feature \(x_i\), we have

\[ \text{Var}[\hat{\sigma}_{i,i+1}] = \frac{n}{(n-1)^2} \text{Var}(x_i^{(1)} x_j^{(1)}) = O\left(\frac{1}{n}\right). \]

On the other hand, considering the series \(\hat{\sigma}_{i,i+1}\) with \(i = 1, 2, \cdots\), each two pairs with index differs by two or more becomes independent since they are driven by different hidden factors \((f_i\)'s) according to the factor form for the far-near covariance model. In this way, it forms a stationary 2-dependent sequence [Ferguson, 1996], where the 2-dependent sequence means that each subsequence of that series is independent with the subsequence shifted by two. Thus, by the central limit theorem for \(m\)-dependent sequence [Ferguson, 1996], we have

\[ \sqrt{p}(\bar{k}_1 - 2a) \Rightarrow N(0, O\left(\frac{1}{n}\right)), \] (C.13)

where " \(\Rightarrow\) " means convergence in law. This means that

\[ \bar{k}_1 = 2a + O_p\left(\frac{1}{\sqrt{np}}\right). \] (C.14)
By similar argument, we have $\tilde{k}_2 = a^2 + O_p(\frac{1}{\sqrt{np}})$. Combining this with (C.14) we have $\hat{a} = 2\tilde{k}_2/\tilde{k}_1 = a + O_p(\frac{1}{\sqrt{np}})$.

## C.6 Proof of Theorem 27

**Proof.** We will prove this theorem using strategy in Jiang [2004] which proved similar results for the sample covariance matrix. We define $||A||_\tau = \max_{i,j=1,\ldots,p,|i-j|\geq\tau} |a_{ij}|$ for any square matrix $A_{p\times p} = \{a_{ij}\}$. Assumed that our design matrix $X_{n\times p}$ is normalized. The $(i,j)^{th}$ element of $X^\top X$ is $\sum_{k=1}^n x_{ki}x_{kj}$. Jiang [2004] defined

$$y_{ij}^{(l)} = \sum_{k=1}^l x_{ki}x_{kj}, \quad (C.15)$$

and

$$W_n = \max_{1\leq i,j\leq p,|i-j|\geq\tau} |y_{ij}^{(n)}|. \quad (C.16)$$

The strategy in Jiang [2004] is to show that $n\hat{L}$ is very close to $W_n$, and the asymptotic limit of $W_n/\sqrt{n\log(n)}$ is 2. Therefore, the asymptotic limit of $\hat{L} = ||R_n||_\tau$, where $R_n$ is the sample correlation matrix, is found. Here we are deriving similar results for $\hat{L} = ||\tilde{R}_n||_\tau$ in equation (4.26). Instead of using the cross-product of $x_{ki}x_{kj}$, we utilize the weighted version of the corresponding term. That is, we define

$$\tilde{y}_{ij}^{(l)} = \sum_{k=1}^l \lambda_1 x_{ki}x_{kj} + \lambda_2 x_{k,i-1}x_{kj} + \lambda_3 x_{k,i}x_{k,j+1} + \lambda_4 x_{k,i+1}x_{kj} + \lambda_5 x_{k,i}x_{k,j-1}, \quad (C.17)$$

and

$$\tilde{W}_n = \max_{1\leq i,j\leq p,|i-j|\geq\tau} |\tilde{y}_{ij}^{(n)}| = \max_{1\leq i,j\leq p,|i-j|\geq\tau} |(x_i^\top x_j, x_{i-1}^\top x_j, x_i^\top x_{j+1}, x_{i+1}^\top x_j, x_i^\top x_{j-1})\lambda|. \quad (C.18)$$

Similarly, we wants to show that (a) $n\hat{L}$ is very close to $\tilde{W}_n$ and (b) the asymptotic limit of $W_n/\sqrt{n\log(n)}$ is $2\sqrt{\bar{\rho}_{rr}}$. 

141
First, we note that
\[
|n\bar{L}_n - \bar{W}_n| \leq \max_{1 \leq i,j \leq p, |i-j| \geq \tau} |n\tilde{\rho}_{ij}\lambda - (x_i^\top x_j, x_{i-1}^\top x_j, x_i^\top x_{j+1}, x_{i+1}^\top x_j, x_i^\top x_{j-1})\lambda|
\leq |\lambda_1||n\rho_{ij} - x_i^\top x_j| + |\lambda_2||n\rho_{i-1,j} - x_{i-1}^\top x_j| + |\lambda_3||n\rho_{i,j+1} - x_i^\top x_{j+1}|
+ |\lambda_4||n\rho_{i+1,j} - x_{i+1}^\top x_j| + |\lambda_5||n\rho_{i,j-1} - x_i^\top x_{j-1}|
\leq |\lambda_1||nR_n - X^\top X||_\tau.
\] (C.19)

As summarized in equation (3.2) of Jiang [2004], the Lemma 2.2 and Lemma 2.3 of that paper implies
\[
||nR_n - X^\top X||_\tau \leq 4n^{-\frac{1}{2}}W_n + n^\frac{3}{4}.
\] (C.20)

Since Lemma 3.1 of Jiang [2004] implies that \(4n^{-\frac{1}{2}}W_n|\lambda|_1 = O_p(n^{1/3} \log(n))\), combining (C.19) and (C.20), we have
\[
n\bar{L}_n - \bar{W}_n = O_p(n^{\frac{1}{3}}).
\] (C.21)

Notice that (C.19) only introduces an extra factor \(|\lambda|_1\) compared to the bound \(||nR_n - X^\top X||_\tau\) in Jiang [2004]. This factor \(|\lambda|_1\) can be calculated from (4.12) and is a constant when \(n\) and \(p\) varies, thus the asymptotic rate \(O_p(n^{\frac{1}{3}})\) in (C.21) is the same as that of Jiang [2004].

Second, we derive the asymptotic limit of the asymptotic limit of \(\bar{W}_n\). From Theorem 20, we know that the variance for each of the weighted sum in the summation has the variance \(\varrho_{rr}\). Thus, let \(\tilde{y}^{(l)}_{ij} = \frac{\hat{y}^{(l)}_{ij}}{\sqrt{\varrho_{rr}}}\). Then \(\tilde{y}^{(l)}_{ij}\) is a sum of \(l\) i.i.d random variables with mean zero and variance one, similar to the \(y^{(l)}_{ij}\) in (C.15) used by Jiang [2004]. Replacing \(y^{(l)}_{ij}\) by \(\tilde{y}^{(l)}_{ij}\) in the proof of Lemma 3.1 in Jiang [2004], we have a similar result
\[
\frac{1}{\sqrt{\varrho_{rr}}}\bar{W}_n \sqrt{n \log(n)} \rightarrow 2 \text{ a.s. as } n \rightarrow \infty.
\] (C.22)

Dividing \(\sqrt{n \log(n)}\) on both side of (C.21) and coming with (C.22) \(\frac{W_n}{\sqrt{n \log(n)}} \rightarrow 2\sqrt{\varrho_{rr}}\), we have the desired result of (4.26).
Bibliography


Rasmus Benestad, Deliang Chen, Abdelkader Mezghani, Lijun Fan, and Kajsa Parding. On using principal components to represent stations in empirical-statistical downscaling. *Tellus A*, 67(0), 2015. ISSN 1600-0870. 3.1, 3.4.3, 4.1


Yale Chang, Yi Li, Adam Ding, and Jennifer Dy. A robust-equitable copula dependence measure for feature selection. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*. Citeseer, 2016. 2.4.7


Thomas H Cormen. *Introduction to algorithms*. MIT press, 2009. 2.3.2


Adam Ding, Jennifer Dy, Yi Li, and Yale Chang. A robust-equitable measure for feature ranking and selection. *Journal of Machine Learning Research*, 2017. 2.4.7


Andrew D Fernandes and Gregory B Gloor. Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics*, 26(9):1135–1139, 2010. 2.1, 2.3.1

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007. 2.1, 2.2.1, 2.4

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005a. 2.1, 2.2.1


Christopher King. Inequalities for trace norms of 2 by 2 block matrices. *Communications in Mathematical Physics*, 242(3):531–545, 2003. ISSN 1432-0916. doi: 10.1007/s00220-003-0955-9. 3.3.1

Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014. 1.2.1, 2.1, 2.2.1, 2.2.1, 2.2.1, 2.2.1, 2.2.2, 2.5


J. LeSage and R.K. Pace. *Introduction to Spatial Econometrics*. Chapman-Hall, 2009. 3.1, 4.1

M. Lichman. UCI machine learning repository. [http://archive.ics.uci.edu/ml](http://archive.ics.uci.edu/ml), 2013. 2.4, 2.4.4

Han Liu, John D Lafferty, and Larry A Wasserman. Exponential concentration for mutual information estimation with application to forests. In *NIPS*, pages 2546–2554, 2012. 2.3.1


Gabriel J. Lord, Catherine E. Powell, and Tony Shardlow. *An Introduction to Computational Stochastic PDEs.*. Cambridge University Press, 006 2014. doi: 10.1017/CBO9781139017329. 3.2.1


Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1), 1952. ISSN 1540-6261. 3.1, 4.1


Hanchuan Peng, Fulmi Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005. 2.1, 2.2.2, 2.4, 2.4.5

150


Mohsen Pourahmadi. *High-dimensional Covariance Estimation: with High-Dimensional Data*. Wiley, 2014. 3.1, 4.1


Sashank J Reddi and Barnabás Póczos. Scale invariant conditional dependence measures. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1355–1363, 2013. 2.1, 2.4.3


David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011. 1.2.1, 2.1, 2.2.1, 2.2.2, 2.2.3, 2.3.1, 2.5

Yakir A. Reshef, David N. Reshef, Pardis C. Sabeti, and Michael Mitzenmacher. Theoretical foundations of equitability and the maximal information coefficient, 2014. 2.4.2, 2.4.2


152


Gábor J Székely and Maria L Rizzo. Brownian distance covariance. *The annals of applied statistics*, pages 1236–1265, 2009. 2.2.1, 2.2.2


Wang, Lingxiao, Xiang Ren, and Quanquan Gu. Precision matrix estimation in high dimensional gaussian graphical models with faster rates. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*. Citeseer, 2016. 3.1, 4.1


with weighted $l_1$ penalty. *Journal of Nonparametric Statistics*, 21(7):925–942, 2009. 3.1, 3.2.1, 1