ABSTRACT

ZHOU, WEILIAN. Resampling Methods in Statistical Inference. (Under the direction of Soumendra Lahiri and Eric Chi).

In statistical inference, statisticians are trying to infer the property of the population from the samples. In most cases, we are not only interested in the point estimation but also the uncertainty of the estimation such as bias and variance. However, it is impossible to derive the explicit formula of the interest quantity. Resampling methods, a class of nonparametric methods, play an important role in estimating the interested quantities without approaching the explicit formula. There are two main kinds of resampling methods, the Bootstrap and the Jackknife. For Bootstrap, it estimates the sample distribution by sampling with replacement from the original data. For any statistics with the form of the functional of the population distribution, it is very natural to utilize the bootstrap method to construct a plug-in estimator by replacing the population distribution by the empirical distribution. Unlike the bootstrap method, the Jackknife method computes the parameter by systematically leaving out samples. It is useful in estimating the bias and the variance of the statistics. Since the jackknife method doesn't approximate the population distribution, it is not applied as generally as the bootstrap method. However, the jackknife method enjoys the advantage of efficient computing. It is useful in estimating the bias and the variance of the estimators. In this dissertation, we explore the resampling methods in different situations.

Resampling Methods in Statistical Inference

by
Weilian Zhou

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2021

APPROVED BY:

Rui Song	Min Kang
Soumendra Lahiri Co-chair of Advisory Committee	Eric Chi Co-chair of Advisory Committee

DEDICATION

To my loving parents.

BIOGRAPHY

The author was born in 1993 in Wuhan, Hubei, China. In 2012, he was admitted to the School of Management Science at the University of Science and Technology of China (USTC). There he found his interests in using statistical methods to solve the real-world problem. After receiving a Bachelor degree in statistics in 2016, he attended North Carolina State University to pursue a PhD in statistics. With the valuable instruction and guidance from his advisors Dr. Soumendra Lahiri and Dr. Eric Chi, he focuses his research on resampling methods for statistical inference. He will complete his Ph.D in Sep 2021.

ACKNOWLEDGEMENTS

First of all, I would like to take this opportunity to express my thanks and appreciation to my advisors, Dr. Soumendra Lahiri and Dr. Eric Chi. I feel very fortunate to work under two great people. From Dr. Lahiri, I learned not only how to construct research from 0 to 1 but also how to think about questions in a scientific way. He also encouraged me to try my statistical knowledge in different industries to find my real interests and offered me a lot of help during my hard time finding the job during the COVID period. From Dr. Eric Chi, I learnt a lot of optimization knowledge through working on the convex clustering project under his guide. I still remember that he checked my code line by line and gave me suggestions on the coding style. I feel very lucky to have both as my advisors, without whose guidance, I would not be able to accomplish this work.

I would also like to extend my appreciation to my committee members Dr. Rui Song, Dr.Min Kang for their constant encouragement and wonderful feedback on my written prelim and oral prelim. I would also like to express my thanks to Dr Wenbin Lu, he offered me a lot of help during his course and my job application processes.

I also owe my gratitude to all my supervisors at internships. From those internships, I had the chance to see how my statistical and mathematical ability apply in the different areas. In particular, Dr. Bingzhi Zhang at Sanofi offered me the first chance to work in industry and many insightful suggestions on how to work in industry. Dr. Jianqi Wang at Facebook helped me know how to work in a fast-paced working environment and taught many strategies on coding and machine learning. Matthieu de Rochemonteix at Citadel led me into quantitative finance and tried his best to unblock me during the challenging internship.

Last but not least, I would like to express my deepest gratitude to my great parents Peifang Zhang and Deliang Zhou. Their unconditional support and love is always the

source of my strength. In every crossroads of my life, they gave me valuable advice and respected my decision.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

1.1 Motivation

Resampling is a class of statistical methodology to derive statistical inference by repeatedly resampling the sample data. There are two main kinds of resampling methods, the Bootstrap and the Jackknife. They both treat the sample data as the population and approximate the sampling mechanism from resampling from the sample data sets. They are useful in bias and variance estimation. In this dissertation, we want to investigate the resampling method in different scenarios.

1.2 Overview of the resampling methods

In this section, we give an overview of the resampling methods. The resampling method main consists of the Bootstrap method and the Jackknife method. In the following parts, we introduce some common Bootstrap and Jackknife methods.

1.2.1 iid Bootstrap

Assume $\{X_i\}$ is a sequence of iid random variables with a common distribution F, we have the following observations $\mathscr{X}_n = \{X_1, X_2 \cdots X_n\}$ and the interested statistics is defined as $T_n = T(\mathscr{X}, F)$. The bootstrap aims at finding an approximation to the unknown distribution of T_n or some population quantities such as the standard error of T_n.

In order to mimic the mechanism of sampling from the population, the Bootstrap treats the sample $\mathscr{X}$ as the population. Each time, we draw a random sample $\mathscr{X}^\star = \{X_1^\star, X_2^\star, \cdots X_n^\star\}$ with replacement from $\mathscr{X}$. Conditional on $\mathscr{X}$, $X_i^\star$ are i.i.d random variables with a distribution

$$P\left(X_i^\star = X_j\right) = \frac{1}{n}.$$

Thus the common distribution of $X_i^\star$ is the empirical distribution generated by the original observations, denoted as F_n. The bootstrap version of $T_n^\star$ is defined as

$$T_n^\star \quad = \quad t(\mathscr{X}, F_n).$$

Through B times of replication of the resampling, we can have B copies of the Bootstrap version of the target estimator T_n, $\{T_n^{\star(1)}, T_n^{\star(2)}, \cdots T_n^{\star,(B)}\}$. Let $\hat{L}_n^\star$ denotes the conditional distribution of $T_n^\star$ given $\mathscr{X}_n$, the Bootstrap method advocates $\hat{L}_n^\star$ as an estimator to the true

distribution L_n of T_n. Then the variance of T_n can be treated as a plug-in estimator,

$$\hat{Var}(T_n) = \frac{1}{B}\sum_{i=1}^{B} T_n^{\star(i)2} - \left(\frac{1}{B}\sum_{i=1}^{B} T_n^{\star(i)}\right)^2. \tag{1.1}$$

If T_n is the normalized sample mean $T_n = \sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma$, the validity of the i.i.d Bootstrap method is shown in the following theorem

Theorem 1.2.1. *If $X_1, X_2 \cdots X_n$ are iid with $\sigma^2 = Var(X_1) < \infty$, then*

$$\sup_x \left| P_\star\left(T_n^\star \le x\right) - \Phi(x/\sigma)\right| = o(1), \quad n \to \infty. \tag{1.2}$$

1.2.2 Moving Block Bootstrap

The iid Bootstrap is applicable under the very strong condition that data needs to follow the i.i.d assumption. This is not the case for most situations. The moving block bootstrap is designed to resolve the inference problem in the dependent data situations.

In contrast to resampling a single observation at a time in the i.i.d Bootstrap method, the moving block bootstrap resamples a block of observations every time. The block preserves the dependence relationship in the original observations. In the moving block Bootstrap, we need to define ℓ as the length of the block, then $\mathscr{B}_i = (X_i, X_{i+1} \cdots X_{i+\ell-1})$ is the block starting at X_i where $1 \le i \le n-\ell+1$. Suppose $n = k\ell$, each time we resample with replacement to get $\mathscr{B}_1^\star, \mathscr{B}_2^\star, \cdots, \mathscr{B}_k^\star$ and notice that in each block we have ℓ consecutive observations. Denote the observations in $\mathscr{B}_i^\star$ as $\left(X^\star_{(i-1)\ell+1}, \cdots X^\star_{i\ell}\right)$, then $\mathscr{X}^\star = \left(X_1^\star, X_2^\star, \cdots X_n^\star\right)$ constitute the moving block Bootstrap sample. The moving block Bootstrap version of estimators of the form $\hat{\theta}_n = T(F_n)$ is defined as

$$\theta_n^\star = T\left(F_n^\star\right), \tag{1.3}$$

where $F_n^\star$ is the empirical distribution of $\left(X_1^\star, X_2^\star, \cdots X_n^\star\right)$.

1.2.3 Nonoverlapping Block Bootstrap

Nonoverlapping block Bootstrap is very similar to the moving block Bootstrap. The only difference is on the blocking rule. In the moving block Bootstrap, for the blocks $\mathscr{B}_i$ and $\mathscr{B}_j$, there is an overlap between them if $\left|i-j\right| < k$. However, in the nonoverlapping block bootstrap, there is no overlap between two blocks. The i-th block is defined as $\mathscr{B}_i = \left(X_{(i-1)\ell+1}, X_{(i-1)\ell+2}, \cdots X_{i\ell}\right)$. In total, there are $b = \left[\frac{n}{\ell}\right]$ blocks.

The implementation of the Nonoverlapping block Bootstrap is exactly the same as the moving block Bootstrap. We select the random sample of blocks $\mathscr{B}_1^\star, \mathscr{B}_2^\star \cdots \mathscr{B}_k^\star$ from the $\mathscr{B}_1, \cdots \mathscr{B}_b$. With $m = \ell k$, denote $F_{m,n}^\star$ is the empirical distribution of $\{X_1^\star, X_2^\star \cdots X_m^\star\}$. Then, the nonoverlapping block Bootstrap version of T_n is

$$\theta_n^\star \quad = \quad T_n\left(F_{m,n}\right).$$

1.2.4 The Jackknife

The Jackknife is a resampling technique to estimate the bias and the variance of the statistics. Unlike Bootstrap, the Jackknife utilizes a subset of the data each time. For an estimator of the interested parameter θ of the form $\hat{\theta} = \phi_n(X_1, X_2, \cdots X_n)$, the Jackknife is useful in estimating the standard error of it when we don't know the true distribution of ϕ_n. The standard Jackknife, also known as the delete one Jackknife, systematically deletes one observation from the sample to compute the corresponding pseudo value of ϕ_n. In i-th time, the pseudo value Ps_i is defined as

$$Ps_i \quad = \quad n\phi_n(X_1, X_2, \cdots X_n) - (n-1)\phi_{n-1}(X_1, \cdots X_{i-1}, X_{i+1}, \cdots, X_n).$$

Then the variance of $\hat{\theta}$ is defined as

$$Var\left(\hat{\theta}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(Ps_i - Ps_{(\cdot)}\right)^2, \tag{1.4}$$

Where $Ps_{(\cdot)} = \frac{\sum_{i=1}^{n} Ps_i}{n}$.

More specifically, if the interested parameter θ is the population mean. The Jackknife variance estimation 1.4 can be simplified as

$$Var\left(\hat{\theta}\right) = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2/(n-1).$$

It is the same as the sample variance.

Moreover, the pseudo value in the Jackknife serves as a bias correction in some situations. Suppose the estimated parameter has the following expression of the expectation,

$$\mathbb{E}\left(\hat{\theta}_n\right) = \theta + \frac{a}{n} + \frac{b}{n^2} + \mathcal{O}\left(\frac{1}{n^3}\right). \tag{1.5}$$

Then the delete-1 estimator $\theta_n^{(i)}$ has the following expectation form

$$\mathbb{E}\left(\hat{\theta}_n^{(i)}\right) = \theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + \mathcal{O}\left(\frac{1}{(n-1)^3}\right). \tag{1.6}$$

The i-th pseudo value has the expectation

$$\mathbb{E}(Ps_i) = \theta + \mathcal{O}\left(\frac{1}{n^2}\right). \tag{1.7}$$

From the derivation, we can see that the pseudo value removes the bias of $\mathcal{O}\left(\frac{1}{n}\right)$ order.

The jackknife method enjoys the simplicity of computation compared with Bootstrap. The consistency of the Jackknife variance estimation requires the smoothness of the func-

tionals. The standard jackknife will fail in some cases such as estimating the variance of the sample median. Under such circumstances, a more general Jackknife method should be used, delete-d Jackknife.

Delete-d Jackknife deletes d observations systematically each time computing the pseudo value. Assume without loss of generality that deleting d observations are

$$(X_1, X_2, \cdots, X_d),$$

the corresponding pseudo value is

$$Ps_i = n\phi(X_1, X_2, \cdots, X_n) - (n-d)\phi(X_{d+1}, X_{d+2}, \cdots, X_n). \tag{1.8}$$

Then the variance is defined as the sample variance of the pseudo values. However, when d is relatively large, it is impossible to compute all $\binom{n}{d}$ combinations. A common way to resolve this computation problem is using the random sample and replicate the procedure for a large number of times. It is a reasonable approximation to the true delete-d jackknife variance estimation.

1.3 Structure of the dissertation

In the chapter 2 of this paper, we introduce a new resampling method called the stationary Jackknife. The stationary jackknife is useful in the variance inference for the dependent data. It is a generalization of the moving block jackknife and it allows the length of the deleting block to be a random variable. We will show the consistency of the stationary Jackknife variance estimation, and unlike the moving block Jackknife, it is not sensitive to the expected block length.

In the chapter 3, we focus on the variance estimation of the Lasso estimator. Lasso is very popular in high dimensional settings. Lasso is useful in providing model selection and shrinkage estimation at the same time. This perfectly meets the common assumption of sparsity in the high dimensional case. We investigate the Jackknife method in providing a valid variance inference on the Lasso estimator.

In the chapter 4, we introduce the Markowitz Portfolio Optimization problem. This is a breakthrough mathematical framework for assembling a portfolio of assets such that the expected return is maximized for a given level of risk. We focus on the problem under a high dimensional setting. We utilize the factor model to decrease the dimension and derive the consistency of the optimal risk estimator by applying the structure of the factor model. Then we propose a resampling framework to give the confidence interval of the optimal risk.

CHAPTER

2

STATIONARY JACKKNIFE

2.1 Introduction

The Jackknife(Tukey 1958) is an intriguing nonparametric method for estimating the bias and variance of statistics we are interested in. The role of the jackknife in bias correction and robust confidence interval has been fully explained in (Miller 1974). However, the assumption of independence of the observations is very crucial. It seems that the standard jackknife will give an unreliable estimation if dependence is ignored. In most cases, especially in the applications involving time series, dependence between observations is not negligible. When fitting a parametric model to the time series, it is always very difficult to model all

important features of the observed time series, and the parametric inference approach often suffers from the risk brought by the effect of the parameter estimation or model misspecification. Thus, it is very important to derive the nonparametric estimator for the dependent data. The moving block jackknife (Kunsch 1989), an extension of the standard jackknife, works well for the arbitrary stationary time series with the short range dependence. In this paper, we introduce a new resampling method called the stationary jackknife. This is an extension of the moving block jackknife and makes use blocks of random lengths, similarly as the extension of the moving block bootstrap to the stationary bootstrap(Politis and Romano 1989). This is the reason why we call our method the stationary jackknife.

The stationary jackknife is suitable for variance estimation of statistics from observations generated by weakly dependent stationary time series. The stationary jackknife can be applied in similar situations with the moving block jackknife. The major difference between the stationary jackknife and the moving block lies in the block length. For the stationary jackknife, we delete l_i consecutive observations as the ith block to compute the ith pseudo value. In the settings of the moving block jackknife, $\ell_i = \ell$, the block length is a constant for all pseudo-value computation. Whereas in the stationary jackknife, ℓ_i are not necessarily the same every time and we treat ℓ_i as a random variable with a truncated geometric distribution. Since a variable with the geometric distribution is unbounded, we set an upper bound on the variable to avoid deleting a large segment of the observations. Moreover, the fact that the tail of the geometric distribution decays exponentially helps constrain the block length to within a reasonable range. For the starting point of the block near the end of observations, the last observation serves as a natural cutting point for the block. After deleting the block, we do a smooth transformation on the remaining observations to get the statistic. The stationary jackknife variance estimator is the standardized version of sample variance of statistics obtained in the way mentioned above. One common difficulty in the moving block jackknife is the choice of the optimal order of the block length. The optimal

order of the block length may vary when jackknifing different statistics or the same class of statistics under different data generating distributions. Nonetheless, we observed that the randomness in the block length makes the stationary jackknife method more robust to the expected block length. The simulation results in section 5 illustrate that the stationary jackknife provides a reasonable variance estimation in a wider range of the expected block length compared with the moving block jackknife.

In section 2, we introduce the stationary jackknife and illustrate its differences with the moving block jackknife. In section 3, we prove the consistency of the stationary jackknife variance estimator in the case of the arithmetic mean. We derive the asymptotic bias and variance terms of the stationary jackknife variance estimator. By minimizing the mean square error (MSE), we get the optimal order of the expected block length for the stationary jackknife. In Section 4, we investigate the consistency of the stationary jackknife variance estimator for more general statistics. In Section 5, we compare the stationary jackknife with different resampling methods such as the moving block bootstrap(Kunsch 1989), the moving block jackknife and the stationary bootstrap(Politis and Romano 1989) on the different simulated datasets.

2.2 Definition of Stationary jackknife

In this section, we formalize the definition of the stationary jackknife. One important issue is that we need to define the statistic with a variable length missing block of the observations. The issue can be overcame by focusing on a certain class of statistics we are going to introduce later in this section. This class of statistics is sufficiently general to include many statistics we are interested in.

2.2.1 Estimator defined on the empirical distribution

For observations $\{X_1, X_2 \cdots X_N\}$ from a stationary processes, the empirical k-dimensional marginal distribution is defined as

$$\rho_N^k = (N-k+1)^{-1} \sum_{i=1}^{N-k+1} \delta_{(X_i, X_{i+1} \cdots X_{i+k-1})},$$

where the δ_y is the point mass on $y \in R^k$. For any functional $T(\cdot)$ with the values in R^q defined on the all probability measures on R^k , we consider the statistic T_N of the form

$$T_N(X_1, X_2, \cdots, X_N) = T\left(\rho_N^k\right).$$

For simplicity, we assume $q = 1$. Moreover if we define k-tuple as a block

$$Y_t = (X_t, X_{t+1} \cdots X_{t+k-1}),$$

then we have

$$\begin{aligned} \rho_N^k &= \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}, \\ &= \rho_n, \end{aligned}$$

where $n = N - k + 1$. This is equivalent to the case $k = \ell$. In the later part of the paper, we consider $k = \ell$ for simplicity.

2.2.2 The Stationary Jackknife

Before deriving the formula for the stationary jackknife, we first recap how the moving block jackknife works to produce the pseudo values. In the moving block jackknife, the

length of the block ℓ is fixed over the time. In producing the j-th pseudo value, we first define the corresponding marginal $\tilde{\rho}_n^{(j)}$ after deleting the j-th block from the observations,

$$\tilde{\rho}_n^{(j)} = \frac{1}{n-\ell}\sum_{t=1}^{n}\Big(1-\mathbb{1}_{([j,j+\ell-1])}(t)\Big)\delta_{Y_t}, \quad j=1,2\cdots n-l+1.$$

Then the j-th moving block jackknife replication is represented as the estimator defined on the empirical marginal $\tilde{\rho}_n^{(j)}$:

$$\tilde{T}_n^{(j)} = T\Big(\tilde{\rho}_n^{(j)}\Big).$$

For the stationary jackknife, the length of the deleting blocks will be different each j. We need to define $\{\tilde{L}_1,\tilde{L}_2,\cdots\tilde{L}_m\}$ to represent the block length variables. The corresponding empirical marginal ρ_n is defined as

$$\rho_n^{(j)} = \frac{1}{n-\tilde{L}_j}\sum_{t=1}^{n}\Big(1-\mathbb{1}_{([j,j+\tilde{L}_j-1])}(t)\Big)\delta_{Y_t} \quad j=1,2\cdots n-[2\ell\log n]+1.$$

More specifically, $\{\tilde{L}_1,\tilde{L}_2,\cdots\tilde{L}_m\}$ are independent and each has the same distribution as $\tilde{L}$. $\tilde{L}$ is defined as

$$\tilde{L} = \min\big(L,[2\ell\log n]\big),$$

where L is a random variable with a geometric distribution $L\sim Geo(p)$ and $\ell=\frac{1}{p}$, $[a]$ represents the largest integer not exceed a. Therefore $\tilde{L}$ has a truncated geometric distribution with the pdf

$$P(\tilde{L}_j=k)=\begin{cases} p(1-p)^{k-1} & 1\le k\le[2\ell\log n]-1, \\ (1-p)^{[2\ell\log n]-1} & k=[2\ell\log n]. \end{cases} \tag{2.1}$$

Then the j-th stationary jackknife replication is calculated as

$$T_n^{(j)} = T\left(\rho_n^{(j)}\right) \quad j = 1, 2 \cdots [2\ell \log n] + 1.$$

As we can see that the main difference between the stationary jackknife and the moving block jackknife is that the length of our delete block is no more a constant but a random variable. We set a upper bound $\left[2\ell \log n\right]$ on the block length in order to make sure the length of the missing part is not comparable to the length of the original observations. In total, there are two layers of randomness in the stationary jackknife, one comes from the empirical distribution, the other one comes from the realizations of the truncated geometric distribution.

The stationary jackknife estimator of the variance of T_n is a standardized version of the sample variance of the $T_n^{(j)}$

$$\hat{\sigma}^2_{jack} = \frac{(n-\ell)^2}{n\ell m} \sum_{j=1}^{m} \left(T_n^{(j)} - T_n^{(\cdot)}\right)^2, \tag{2.2}$$

where $m = n - [2\ell \log n] + 1$ and $T_n^{(\cdot)} = \frac{\sum_{j=1}^{m} T_n^{(j)}}{m}$. The estimator can be decomposed into two parts. The first part is sample variance of the stationary jackknife replications

$$\frac{1}{m} \sum_{j=1}^{m} \left(T_n^{(j)} - T_n^{(\cdot)}\right)^2.$$

The second part is the standardizing factor $\frac{(n-\ell)^2}{n\ell}$. The standardizing factor is inversely proportional to the expected block length ℓ. The larger the expected deleting length we choose, the larger the ratio we down weight on the sample variance of the jackknife replications. The formula (2.2) is very similar to the moving block jackknife variance estimator(Kunsch 1989).

2.3 Consistency of the Stationary Jackknife Estimator

In this section, we show the consistency of the stationary jackknife variance estimator for the arithmetic mean under several assumptions. The assumptions are similar to the assumptions for the moving block jackknife. The routine method in proving the consistency of the estimator is to show both the bias and the variance of the estimator converge to zero. We derive the exact formula of the bias and variance in the following theorems. Moreover, we can investigate the optimal expected block length by minimizing the mean squared error of the estimator.

Here we investigate the properties of the stationary jackknife for the arithmetic mean. The arithmetic mean is obviously an element of the estimator defined on the empirical distribution. It corresponds to the population mean $T(F) = \int x F(dx) = \mathbb{E}(x) = \mu$. The functional is linear and allows the explicit calculations of all quantities of interest.

Suppose $\{X_t\}_{t=1,2,\cdots\infty}$ are from a weak stationary processes. Then X_t enjoys the following properties,

$$\begin{aligned} \mathbb{E}(X_t) &= \mu. \\ \mathbb{E}\left(X_t^2\right) &< \infty. \\ \mathbb{E}\left(\left(X_{t+h}-\mu\right)\left(X_t-\mu\right)\right) &= R(h), \quad \forall t. \end{aligned}$$

According to the stationarity of the observations, we can write the the standardized variance of the arithmetic mean, $Var\left(\sqrt{n}\bar{X}\right)$, analytically as

$$Var\left(\sqrt{n}\bar{X}\right) = \sum_{h=-n+1}^{n-1} \frac{(n-h)}{n} R(h). \tag{2.3}$$

We denote σ_{as}^2 as the limit of $Var\left(\sqrt{n}\bar{X}\right)$. When $R(h)$ is absolutely summable, then we

have

$$\sigma_{as}^2 = \sum_{h=-\infty}^{h=\infty} R(h). \tag{2.4}$$

For the j-th stationary jackknife replication $T_n^{(j)}$ in the arithmetic mean case, we have

$$T_n^{(j)} = \frac{\sum_{i=1}^{n} X_i - S_{i,\tilde{L}_i}}{n - \tilde{L}_i},$$

where $S_{i,l} = \sum_{j=i}^{i+l-1} X_j$, the partial sum of $\{X_t\}$.

Before looking at the bias and the variance terms of the stationary jackknife variance estimator, we need to first specify some assumptions to ensure the consistency of the estimator. Here are some assumptions needed for the later theorems,

Assumption 1

$\frac{1}{\ell} + \frac{\ell^2}{N} \to 0$.

Assumption 2

$\sum_{k=-\infty}^{\infty} |k||R(k)| < \infty$.

Remark

The assumption 1 actually represents two conditions, the first one is $\frac{1}{\ell} \to 0$ and the second one is $\frac{\ell^2}{n} \to 0$. This means that the expected block length ℓ goes to infinity as n goes to infinity. The order of ℓ is not comparable to the observation length since we don't want to

delete almost the whole series. Overall, the Assumption 1 is equivalent to

$$\begin{aligned} \ell &= \ell(n) \to \infty, \\ \ell &= o\left(n^{\frac{1}{2}}\right). \end{aligned}$$

Assumption 2 focuses on the weak dependence of the observations. The constraint on the infinite sum of the autocovariance of $\{X_t\}$ avoids the existence of the long dependence. The autocorrelation needs to have the decay rate at least $\frac{1}{k^{2+\delta}}$. The exponential decay rate on the autocorrelation is a sufficient condition for the assumption 2, so the stationary AR, MA, ARMA time series are qualified for the assumption.

2.3.1 Bias of the stationary jackknife variance estimator

Theorem 2.3.1. *Under assumption 1 and 2, the bias of the stationary jackknife estimator* (2.2) *for the arithmetic mean is following*

$$\mathbb{E}\left(n\hat{\sigma}^2_{jack}\right) = \sigma^2_{as} + \mathcal{O}\left(\frac{1}{\ell}\right) + \mathcal{O}\left(\left(\frac{\ell}{n}\right)^{\frac{1}{2}}\right). \tag{2.5}$$

Proof. By the simple algebraic operations, for any constant c the j-th stationary jackknife replication can be expressed as

$$T_n^{(j)} = \bar{X}_n + \frac{1}{n-\tilde{L}_j}\left[\tilde{L}_j\left(\bar{X}_n - c\right) - \sum_{t=j}^{j+\tilde{L}_j-1}\left(X_t - c\right)\right]. \tag{2.6}$$

Now we let

$$c = \hat{\mu}_n, = \frac{\sum_{j=1}^{m} \frac{1}{n-\tilde{L}_j} S_{j,\tilde{L}_j}}{\sum_{j=1}^{m} \frac{\tilde{L}_j}{n-\tilde{L}_j}}.$$

Now we have

$$\begin{aligned} T_n^{(\cdot)} &= \frac{1}{m}\sum_{j=1}^{m} T_n^{(j)}, \\ &= \bar{X}_n + \frac{1}{m}\sum_{j=1}^{m} \frac{1}{n-\tilde{L}_j}\left(\tilde{L}_j\left(\bar{X}_n - \hat{\mu}_n\right)\right). \end{aligned}$$

Moreover

$$\begin{aligned} T_n^{(j)} - T_n^{(\cdot)} &= \left(\frac{\tilde{L}_j}{n-\tilde{L}_j} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_i}{n-\tilde{L}_i}\right)\bar{X}_n \\ &- \left(\frac{S_{j,\tilde{L}_j}}{n-\tilde{L}_j} - \frac{1}{m}\sum_{i=1}^{m}\frac{S_{i,\tilde{L}_i}}{n-\tilde{L}_i}\right). \end{aligned} \tag{2.7}$$

After plugging the formula (2.7) of $T_n^{(j)} - T_n^{(\cdot)}$ into the expression of $n\hat{\sigma}^2_{jack}$, we have

$$\begin{aligned} n\hat{\sigma}^2_{jack} &= \frac{(n-\ell)^2}{m\ell}\sum_{j=1}^{m}[(\frac{\tilde{L}_j}{n-\tilde{L}_j} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_j}{n-\tilde{L}_j})\bar{X}_n - (\frac{S_{j,\tilde{L}_j}}{n-\tilde{L}_j} - \frac{1}{m}\sum_{i=1}^{m}\frac{S_{i,\tilde{L}_i}}{n-\tilde{L}_i})]^2 \\ &= \left(1+\mathcal{O}\left(\frac{\ell\log n}{n}\right)\right)\frac{n^2}{m\ell}\left[\sum_{j=1}^{m}(\frac{\tilde{L}_j}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_j}{n})^2\bar{X}_n^2 + \sum_{j=1}^{m}\frac{S^2_{j,\tilde{L}_j}}{n^2} - \frac{1}{m}(\sum_{j=1}^{m}\frac{S_{j,\tilde{L}_j}}{n})^2\right. \\ &- \left. 2\sum_{j=1}^{n}\left(\frac{\tilde{L}_j}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_i}{n}\right)\left(\frac{S_{j,\tilde{L}_j}}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{S_{i,\tilde{L}_i}}{n}\right)\bar{X}_n\right]. \end{aligned} \tag{2.8}$$

Notice that the second equality is due to the following

$$\begin{aligned} \frac{1}{\left(n-\tilde{L}_j\right)^2} &= \frac{n^2}{\left(n-\tilde{L}_j\right)^2}\frac{1}{n^2} \\ &\leq \frac{n^2}{(n-2\ell\log n)^2}\frac{1}{n^2} \\ &= \left(1+\mathcal{O}\left(\frac{\ell\log n}{n}\right)\right)\frac{1}{n^2}. \end{aligned}$$

Under the assumption 1, we have $\frac{\ell^2}{n} \to 0$. This means

$$\begin{aligned}\mathcal{O}\left(\frac{\ell \log n}{n}\right) &= o\left(\frac{\ell}{n^{\frac{2}{3}}}\right), \\ &= o(1).\end{aligned}$$

So we only consider the leading term in (2.8).

$$\begin{aligned}n\hat{\sigma}^2_{jack} &= \frac{n^2}{ml}[\sum_{j=1}^{m}(\frac{\tilde{L}_j}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_j}{n})^2\bar{X}_n^2 - \frac{1}{m}(\sum_{j=1}^{m}\frac{S_{j,\tilde{L}_j}}{n})^2 + \sum_{j=1}^{m}\frac{S^2_{j,\tilde{L}_j}}{n^2} \\ &- 2\sum_{j=1}^{n}(\frac{\tilde{L}_j}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_i}{n})(\frac{S_{j,\tilde{L}_j}}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{S_{i,\tilde{L}_i}}{n})\bar{X}_n], \\ &= A_1 - A_2 + A_3 - A_4.\end{aligned}$$

We decompose the leading term of $n\hat{\sigma}^2_{jack}$ into four parts and investigate the order of the each part.

For A_1,We have

$$\begin{aligned}\mathbb{E}(A_1) &= \mathbb{E}\frac{n^2}{lm}\sum_{j=1}^{m}\left(\frac{\tilde{L}_j}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_i}{n}\right)^2\bar{X}_n^2, \\ &= \frac{n}{ml}\mathbb{E}n\bar{X}_n^2\mathbb{E}\left(\sum_{j=1}^{m}(\frac{\tilde{L}_j}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_i}{n})^2\right), \\ &= \frac{n}{ml}\mathbb{E}n\bar{X}_n^2\mathbb{E}\left(\sum_{j=1}^{m}\left(\frac{\tilde{L}_j^2}{n^2}\right) - \mathbb{E}\frac{1}{m}\left(\sum_{i=1}^{m}\frac{\tilde{L}_i}{n}\right)^2\right) \\ &= \frac{n}{ml}\left(\sum_{k=-\infty}^{\infty}|R(k)|\right)\mathbb{E}\left(\sum_{j=1}^{m}\frac{\tilde{L}_j^2}{n^2}\right), \\ &= \mathcal{O}\left(\frac{\ell}{n}\right).\end{aligned}$$

During the derivation, we use the following two facts,

$$\begin{aligned}\mathbb{E}n\bar{X_n}^2 &= \sum_{k=-n+1}^{n-1}(1-\frac{|k|}{n})R(k),\\ &\leq \sum_{k=-n+1}^{n-1}|R(k)|,\\ &\leq \sum_{k=-\infty}^{\infty}|R(k)|.\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left(\tilde{L}_j^2\right) &= \sum_{i=1}^{\lceil 2\ell\log n\rceil-1} i^2p(1-p)^{i-1}+\lceil 2\ell\log n\rceil^2\left(1-p\right)^{\lceil 2\ell\log n\rceil-1},\\ &\leq \sum_{i=1}^{\infty} i^2p\left(1-p\right)^{i-1},\\ &= \frac{2-p}{p^2}.\end{aligned}$$

For part A_2, we consider the coefficient of the X_i in the summation $\sum_{i=1}^{m} S_{i,\tilde{L}_i}$. We denote

$$\sum_{i=1}^{m} S_{i,\tilde{L}_i} = \sum_{t=1}^{n} b_t X_t,$$

where b_t is a random variable with the following relationship

$$b_t = \sum_{j=1}^{t}\mathbb{I}_{\left(\tilde{L}_j>t-j\right)}.$$

Suppose $\tilde{L} = \sigma\left(\tilde{L}_1, \tilde{L}_2\cdots\tilde{L}_m\right)$, the sigma algebra generated by the block length variables.

Hence $\mathbb{E}\left(\left(\sum_{j=1}^{m} S_{j,\tilde{L}_j}\right)^2\right)$ can be expressed as

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{j=1}^{m} S_{j,\tilde{L}_j}\right)^2\right) &= \mathbb{E}\left(\mathbb{E}\left(\left(\sum_{j=1}^{m} S_{j,\tilde{L}_j}\right)^2 |\mathscr{L}\right)\right)\\ &= \mathbb{E}\sum_{-n+1}^{n-1} v(|k|)R(k),\end{aligned}$$

where $v(|k|)=\sum_{t=1}^{n-|k|} b_t b_{t+k}$.

Now we focus on the quantity $\mathbb{E}(b_t b_{t+k})$. For any $k \geq 0$, we have

$$\begin{aligned}\mathbb{E}b_t b_{t+k} &= \mathbb{E}\sum_{j=1}^{t+k}\mathbb{1}_{(\tilde{L}_j>t+k-j)}\sum_{j=1}^{t}\mathbb{1}_{(\tilde{L}_j>t-j)},\\ &\leq \mathbb{E}\sum_{j=1}^{t+k}\mathbb{1}_{(\tilde{L}_j>t+k-j)} + E\sum_{j=1}^{t+k} I_{(\tilde{L}_j>t+k-j)} E\sum_{j=1}^{t} I_{(\tilde{L}_j>t-j)}\\ &\leq \sum_{j=1}^{t+k}(1-p)^{t+k-j} + \sum_{j=1}^{t+k}(1-p)^{t+k-j}\sum_{j=1}^{t}(1-p)^{t-j}\\ &\leq \frac{1}{p}+\frac{1}{p^2}.\end{aligned}$$

Consequently, we have

$$v(|k|) \leq \frac{n-|k|}{p}+\frac{n-|k|}{p^2}.$$

Based on the bound of $v(|k|)$, we can derive the bound for A_2,

$$\begin{aligned}|\mathbb{E}(A_2)| &\leq \frac{n^2}{m^2 n\ell}\sum_{k=-n+1}^{n-1}\left(\ell+\ell^2\right)|R(k)|,\\ &= \frac{\ell}{n}\sum_{-\infty}^{\infty}|R(k)| + o\left(\frac{\ell}{n}\right),\\ &= \mathscr{O}\left(\frac{\ell}{n}\right).\end{aligned}$$

For part A_3

$$\begin{aligned}
\mathbb{E}(A_3) &= \frac{n}{m\ell}\sum_{j=1}^{m}\mathbb{E}\left(\mathbb{E}\left(\frac{S^2_{j,\tilde{L}_j}}{n}|\tilde{L}_j\right)\right),\\
&= \frac{n}{\ell}\mathbb{E}\left(\mathbb{E}\left(\frac{S^2_{j,\tilde{L}_j}}{n}|\tilde{L}_j\right)\right),\\
&= \frac{n}{l}\left(\sum_{s=1}^{[2\ell\log n]-1}\sum_{k=-s+1}^{s-1}\frac{(s-|k|)R(k)}{n}p(1-p)^{s-1}\right.\\
&+ \left.\sum_{k=-[2\ell\log n]+1}^{[2\ell\log n]-1}\frac{([2\ell\log n]-|k|)R(k)}{n}(1-p)^{[2\ell\log n]}\right),\\
&= \frac{1}{\ell}\sum_{k=-[2\ell\log n]+1}^{[2\ell\log n]-1}\sum_{s=|k|+1}^{[2\ell\log n]}(s-|k|)R(k)p(1-p)^{s-1}+\mathcal{O}\left(\frac{1}{n\ell}\right),\\
&= \frac{1}{\ell}\sum_{k=-[2\ell\log n]+1}^{[2\ell\log n]-1}\sum_{t=1}^{[2\ell\log n]-|k|}(1-p)^{|k|}R(k)tp(1-p)^{t-1}+\mathcal{O}\left(\frac{1}{n\ell}\right),\\
&= \sum_{k=-[2\ell\log n]+1}^{[2\ell\log n]-1}(1-p)^{|k|+1}-(1-p)^{[2\ell\log n]}R(k)+\mathcal{O}\left(\frac{1}{n^2}\right)+\mathcal{O}\left(\frac{1}{n\ell}\right),\\
&\leq \sum_{k=-[2\ell\log n]+1}^{[2\ell\log n]-1}(1+p(|k|+1))R(k)+\mathcal{O}\left(\frac{1}{n\ell}\right),\\
&= \sigma^2_{\infty}+\mathcal{O}\left(\frac{1}{\ell}\right).
\end{aligned}$$

The derivation is based on repeatedly using the following lemma.

Lemma A. ssume that $p+(np)^{-1}\to 0$ as $n\to\infty$. Let t_n be a sequence of positive number such that $t_n\to\infty$ as $n\to\infty$. L_1 follows a Geometric distribution with the parameter p. Then we have

$$P\left(L_1 > t_n p^{-1}\right) = \mathcal{O}\left(exp(-t_n)\right).$$

Proof. Let $q = 1-p$, then we have

$$\begin{aligned} P\left(L_1 > t_n p^{-1}\right) &= \sum_{i > p^{-1}t_n} p q^{i-1} \\ &\leq q^{\frac{t_n}{p}-1} \\ &= \exp\left(\left[p^{-1}t_n - 1\right]\log q\right) \\ &= \mathcal{O}\left(\exp\left(-p\left(p^{-1}t_n - 1\right)\right)\right) \\ &= \mathcal{O}\left(\exp\left(-t_n\right)\right). \end{aligned}$$

□

For the cross part A_4, from the result of the former part it follows from the Cauchy-Schwarz inequality that

$$\begin{aligned} \mathbb{E}(A_4) &= \mathbb{E}\left(\frac{2n^2}{m\ell}\sum_{j=1}^{n}\left(\frac{\tilde{L}_j}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_i}{n}\right)\left(\frac{S_{j,\tilde{L}_j}}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{S_{i,\tilde{L}_i}}{n}\right)\bar{X}_n\right), \\ &\leq \frac{n^2}{m\ell}\left(\mathbb{E}\sum_{j=1}^{m}\left(\frac{\tilde{L}_j}{n} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_i}{n}\right)^2 \bar{X}_n^2\right)^{\frac{1}{2}} \\ &\quad \left(\mathbb{E}\sum_{j=1}^{m}\frac{S_{j,\tilde{L}_j}^2}{n^2} - \frac{1}{m}\left(\sum_{j=1}^{m}\frac{S_{j,\tilde{L}_j}}{n}\right)^2\right)^{\frac{1}{2}}, \\ &= \mathcal{O}\left(\left(\frac{\ell}{n}\right)^{\frac{1}{2}}\right). \end{aligned}$$

Finally based on the results of the A_1, A_2, A_3 and A_4, we have

$$\begin{aligned} \mathbb{E}\left(n\hat{\sigma}_{jack}^2\right) &= \mathbb{E}\left(A_1 + A_2 + A_3 + A_4\right), \\ &= \sigma_{as}^2 + \mathcal{O}\left(\frac{1}{\ell}\right) + \mathcal{O}\left(\left(\frac{\ell}{n}\right)^{\frac{1}{2}}\right). \end{aligned}$$

□

In order to derive the variance expression of the stationary jackknife estimator $n\hat{\sigma}^2_{jack}$, we need two additional assumptions,

Assumption 3

$\mathbb{E}\left(|X_t|^{6+\delta}\right) < \infty$.

Assumption 4

$\sum_{k=-\infty}^{\infty} k^2 \alpha(k)^{\frac{\delta}{6+\delta}} < \infty$.

The assumption 3 impose a constraint on the moment of X_t that X_t has a finite $6+\delta$-th moment. Assumption 4 allows the series X_t to be weakly dependent. It is sufficient for us to impose the strong mixing condition on X_t. Roughly speaking, we require the mixing coefficient

$$\alpha(k) = \sup_{A \in \mathscr{F}_{-\infty}^{0}, B \in \mathscr{F}_{k}^{+\infty}} |P(A)P(B) - P(A \cap B)| \tag{2.9}$$

to decrease exponentially as $k \to \infty$, where $\mathscr{F}_{T_1}^{T_2}$ is the σ-algebra generated by $\{X_t\}_{t=T_1}^{T_2}$.

With respect to the α-mixing coefficient, we introduce a fundamental inequality for the covariance of mixing random variables.

Lemma L. et X and Y be measurable random variables with respect to two σ-algebras $\mathscr{A}$ and $\mathscr{B}$. Define $\|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}}$, then we have the following inequality

$$|Cov(X,Y)| \leq 8\alpha^{\frac{1}{r}}(\mathscr{A},\mathscr{B})\|X\|_p \|Y\|_q, \tag{2.10}$$

for any $p, q, r \geq 1$ and $\frac{1}{r} + \frac{1}{p} + \frac{1}{q} = 1$.

Proof. See section 1.2.2 in (Doukhan 1994). □

Theorem 2.3.2. *Under assumption 1 to 4, the expression of the variance of stationary jackknife variance estimator $n\hat{\sigma}^2_{jack}$ of the arithmetic mean is*

$$Var\left(n\hat{\sigma}^2_{jack}\right) = \frac{\ell}{4n}\sigma^4_{as} + o\left(\frac{\ell}{n}\right). \tag{2.11}$$

Proof. Without loss of generality, we assume $\mathbb{E}(X_i) = 0$. Then it is obvious that

$$\begin{aligned}
\mathbb{E}\left(S_{j,\tilde{L}_j}\right) &= \mathbb{E}\left(\mathbb{E}\left(S_{j,\tilde{L}_j}|\tilde{L}_j\right)\right), \\
&= \mathbb{E}\left(\tilde{L}_j\right)\mathbb{E}\left(X_j\right), \\
&= 0.
\end{aligned}$$

First we look at $\sum_{j=1}^{m} Cov^2\left(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}\right)$, it can be decomposed into

$$\sum_{j=1}^{[2\ell\log n]} Cov^2(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}),$$

and

$$\sum_{j=\left[2\ell\log n\right]+1}^{m} Cov^2\left(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}\right).$$

When $j \geq \left[2\ell \log n\right]+1$, we have

$$
\begin{aligned}
Cov(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}) &= \mathbb{E}(S_{1,\tilde{L}_1} S_{j,\tilde{L}_j}) = \mathbb{E}(\mathbb{E}(S_{1,\tilde{L}_1} S_{j,\tilde{L}_j} | \tilde{L}_1, \tilde{L}_j)) \\
&= \sum_{k=1}^{[2\ell \log n]-1} \mathbb{E} S_{1,k} S_{j,\tilde{L}_j} p(1-p)^{k-1} + \mathbb{E} S_{1,[2\ell \log n]} S_{j,\tilde{L}_j} (1-p)^{[2\ell \log n]-1} \\
&= \left(\sum_{k=1}^{[2\ell \log n]-[l^{\frac{1}{2}}]} + \sum_{k=\left[2\ell \log n\right]-\left[\ell^{\frac{1}{2}}\right]+1}^{\left[2\ell \log n\right]-1} \right) \mathbb{E} S_{1,k} S_{j,\tilde{L}_j} p(1-p)^{k-1} + \mathscr{O}\left(\frac{\ell \left(\log n\right)^{\frac{1}{2}}}{n^2} \right) \\
&\leq \sum_{k=1}^{[2\ell \log n]-[\ell^{\frac{1}{2}}]} \alpha^{\frac{\delta}{12+2\delta}}(j-k) \| S_{1,k} \|_{\frac{12+2\delta}{6+\delta/2}} \| S_{j,\tilde{L}_j} \|_{\frac{12+2\delta}{6+\delta/2}} p(1-p)^{k-1} \\
&+ \sum_{k=[2\ell \log n]-[\ell^{\frac{1}{2}}]+1}^{[2\ell \log n]-1} \sqrt{\mathbb{E} S_{1,k}^2 \mathbb{E} S_{j,\tilde{L}_j}^2} p(1-p)^{k-1} + \mathscr{O}(\frac{\ell \log^{\frac{1}{2}} n}{n^2}) \\
&\leq \alpha^{\frac{\delta}{12+2\delta}}(j-[2\ell \log n]+[\ell^{\frac{1}{2}}]) C_1 p \ell^{\frac{1}{2}} \left(\sum_{k=1}^{[2\ell \log n]-[\ell^{\frac{1}{2}}]} k(1-p)^{2(k-1)} \right)^{\frac{1}{2}} \left(\left[2\ell \log n\right]-\left[\ell^{\frac{1}{2}}\right] \right)^{\frac{1}{2}} \\
&+ \mathscr{O}\left(\frac{\ell \left(\log n\right)^{\frac{1}{2}}}{n^2} \right) \\
&= \mathscr{O}\left(\ell \log^{\frac{1}{2}} n \right) \alpha^{\frac{\delta}{12+2\delta}}(j-[2\ell \log n]+[l^{\frac{1}{2}}]) + \mathscr{O}\left(\frac{\ell \left(\log n\right)^{\frac{1}{2}}}{n^2} \right).
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
\sum_{j=\left[2\ell \log n\right]+1}^{m} Cov^2\left(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}\right) &\leq \sum_{j=\left[2\ell \log n\right]+1}^{m} \mathscr{O}\left(\ell^2 \log n\right) \alpha^{\frac{\delta}{\delta+6}} \left(j-\left[2\ell \log n\right]+\left[\ell^{\frac{1}{2}}\right] \right) \\
&\leq \sum_{j=[\ell^{\frac{1}{2}}]}^{m} \ell \alpha^{\frac{\delta}{6+\delta}}\left(j\right) \mathscr{O}\left(\ell \log n\right) \\
&\leq \sum_{j=[\ell^{\frac{1}{2}}]}^{m} j^2 \alpha^{\frac{\delta}{6+\delta}}\left(j\right) \mathscr{O}\left(\ell \log n\right) \\
&= \mathscr{O}\left(\ell \log n\right).
\end{aligned}
$$

When $j \le [2\ell \log n]$, we can write the expression of $Cov\left(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}\right)$ as

$$\begin{aligned} Cov\left(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}\right) &= \mathbb{E}\left(S_{1,\tilde{L}_1} S_{j,\tilde{L}_j}\right) \\ &= \sum_{u=-\left[2\ell \log n\right]}^{\left[2\ell \log n\right]} b(u) R\left(j-1+u\right), \end{aligned}$$

where

$$\begin{aligned} b(u) &= \sum_{i=1}^{\left[2\ell \log n\right]-|u|} P\left(\tilde{L} \ge i\right) P\left(\tilde{L} \ge i+|u|\right) \\ &= \left(1-p\right)^{|u|} \sum_{i=1}^{\left[2\ell \log n\right]-|u|} \left(1-p\right)^{2(i-1)} \\ &= \frac{\ell}{2}\left(1-p\right)^{|u|} - \frac{\ell}{2}\left(1-p\right)^{2\left[2\ell \log n\right]+2-|u|} \\ &= \frac{\ell}{2}\left(1-p\right)^{|u|} + \mathcal{O}\left(\frac{\ell}{n^4}\right). \end{aligned} \tag{2.12}$$

In the last step we use the facts that $(1-p)^n = \mathcal{O}\left(exp\left(-\frac{n}{\ell}\right)\right)$ and $\tilde{L}$ is a random variable with the truncated geometric distribution.

From (2.12) we know the coefficient for $R(0)$ in $Cov\left(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}\right)$ is $\frac{\ell}{2}(1-p)^{|j-1|}$, which is denoted as $a_0^{(j)}$. For any $j \ge 1$, we are interested in the coefficient of $R(k), |k| \le \ell$ in the quantity $Cov\left(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}\right)$. The reason is that

$$\sum_{|K|>\ell} R(K) = \mathcal{O}\left(\frac{1}{\ell^2}\right).$$

Denote $a_k^{(j)}$ as the coefficient of $R(k)$ in $Cov\left(S_{1,\tilde{L}_1}, S_{j,\tilde{L}_j}\right)$, then for $|k| \le \ell$, we have a bound for $a_k^{(j)}$, that is

$$a_k^{(j)} \in \left[\left(1-p\right)^{|k|} a_0^{(j)}, \left(1-p\right)^{-|k|} a_0^{(j)}\right].$$

Due to $\lim_{p\to 0}\left(1-p\right)^{\ell}=e^{-1}$, we know $a_k^{(j)}$ is the same order of $a_0^{(j)}$. Thus we have

$$Cov\left(S_{1,\tilde{L}_1},S_{j,\tilde{L}_j}\right) = \mathcal{O}\left(a_0^{(j)}\sigma_{as}^2\right). \tag{2.13}$$

By repeatedly applying the equation (2.12), we have

$$\begin{aligned}\sum_{j=1}^{[2\ell\log n]} Cov^2\left(S_{1,\tilde{L}_1},S_{j,\tilde{L}_j}\right) &= \sum_{j=1}^{[2\ell\log n]} {a_0^{(j)}}^2\sigma_{as}^4\\ &= \frac{\ell^2}{4}\sigma_{as}^4\sum_{j=1}^{[2\ell\log n]}\left(1-p\right)^{2j}\\ &= \frac{\ell^3}{8}\sigma_{as}^4\left(1+o\left(1\right)\right).\end{aligned}$$

Then we need to show that

$$\sum_{j=1}^{m}\left|Cov\left(S_{1,\tilde{L}_1}^2,S_{j,\tilde{L}_j}^2\right)-2Cov^2\left(S_{1,\tilde{L}_1},S_{j,\tilde{L}_j}\right)\right| = o\left(\ell^3\right). \tag{2.14}$$

The key method to prove (2.14) is the following inequality, for $s\le t\le u\le v$,

$$\begin{aligned}&\left|\mathbb{E}(X_sX_tX_uX_v)-\mathbb{E}(X_sX_t)\mathbb{E}(X_uX_v)-\mathbb{E}(X_sX_u)\mathbb{E}(X_tX_v)-\mathbb{E}(X_sX_v)\mathbb{E}(X_tX_u)\right|\\ &\qquad\le C\alpha(\max(t-s,u-t,v-u))^{\frac{\delta}{6+\delta}}.\end{aligned} \tag{2.15}$$

This is followed by a repeated application of Theorem 17.2.3 of (Politis and Romano 1989)

In our case, for a certain maximum gap t between variable, there are at most $3k^2[2\ell\log n]$ sets of indices $\{s,t,u,v\}$ where $\{s,t\}\in\{1,2\cdots[2\ell\log n]\}$ and $\{u,v\}\in\{j,j+1,\cdots j+[2\ell\log n]-$

1} for any j. Now we have

$$\begin{aligned}
&\sum_{j=1}^{m}\left|Cov(S^2_{1,\tilde{L}_1},S^2_{j,\tilde{L}_j})-2Cov^2(S_{1,\tilde{L}_1},S_{j,\tilde{L}_j})\right| \\
&\leq (\sum_{j=1}^{[2\ell\log n]}\sum_{k=0}^{j+[2\ell\log n]}+\sum_{j=[2\ell\log n]+1}^{m}\sum_{k=0}^{j+[2\ell\log n]})3[2\ell\log n]k^2\alpha^{\frac{\delta}{6+\delta}}(k) \\
&\leq \sum_{j=1}^{[2\ell\log n]}\sum_{k=0}^{\infty}3[2\ell\log n]k^2\alpha^{\frac{\delta}{\delta+6}}(k)+\sum_{j=[2\ell\log n]}^{\infty}\sum_{k=j-[2l\ell\log n]}^{j+[2\ell\log n]}3[2\ell\log n]k^2\alpha^{\frac{\delta}{\delta+6}}(k) \\
&\leq \sum_{j=1}^{[2\ell\log n]}\sum_{k=0}^{\infty}3[2\ell\log n]k^2\alpha^{\frac{\delta}{\delta+6}}(k)+\sum_{j=[2\ell\log n]}^{\infty}\sum_{k=j-[2\ell\log n]}^{j+[2\ell\log n]}3[2\ell\log n]k^2\alpha^{\frac{\delta}{\delta+6}}(k) \\
&= \mathcal{O}(\ell\log n)+(\sum_{k=1}^{2[2\ell\log n]}\sum_{j=[2\ell\log n]+1}^{[2\ell\log n]+k}+\sum_{k=[2\ell\log n]+1}^{m}\sum_{j=k-[2\ell\log n]}^{k+[2\ell\log n]})3[2\ell\log n]k^2\alpha^{\frac{\delta}{\delta+6}}(k) \\
&= \mathcal{O}\left(\ell\log n\right)+\mathcal{O}\left(\ell^2\left(\log n\right)^2\right) \\
&= o\left(\ell^3\right).
\end{aligned} \tag{2.16}$$

Above derivation is based on the fact that when $j>[2\ell\log n]+1$ the maximum gap between the indices is at least $j-[2\ell\log n]$ and assumption in theorem 3.2 that $\sum_{k=0}^{\infty}k^2\alpha^{\frac{\delta}{6+\delta}}(k)<\infty$. Then it is obvious that

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{m}Cov\left(S^2_{i,\tilde{L}_i},S^2_{j,\tilde{L}_j}\right)&=\sum_{i=1}^{m}\sum_{j=1}^{m}\Big(Cov(S^2_{i,\tilde{L}_i},S^2_{j,\tilde{L}_j}\Big)-2Cov^2(S_{i,\tilde{L}_i},S_{j,\tilde{L}_j}) \\
&\quad+2Cov^2(S_{i,\tilde{L}_i},S_{j,\tilde{L}_j})) \\
&=\frac{ml^3}{4}\sigma^4_{as}(1+o(1))
\end{aligned} \tag{2.17}$$

Now back to the variance of our jackknife estimator

$$Var(n\hat{\sigma}^2_{jack})=\frac{n^4}{m^2l^2}Var(\sum_{j=1}^{m}(\frac{S_{j,\tilde{L}_j}}{n}+\sum_{i=1}^{n}w_i^{(j)}X_i)^2)(1+o(\frac{l}{n}))$$

Here we denote

$$(\frac{\tilde{L}_j}{n}-\frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{L}_i}{n})\bar{X}_n+\frac{1}{m}\sum_{i=1}^{m}\frac{S_{i,\tilde{L}_i}}{n}=\sum_{i=1}^{n}w_i^{(j)}X_i$$

Notice that although $w_i^{(j)}$ are all random variables but it is easily to verify that $|w_i^{(j)}| = \mathcal{O}(\frac{\ell\log n}{n^2})$.

First from the derivation above, we could naturally get that

$$\begin{aligned}\frac{n^4}{m^2\ell^2}Var(\sum_{j=1}^{m}\frac{S_{j,\tilde{L}_j}^2}{n^2}) &= \frac{1}{m^2\ell^2}\sum_{i=1}^{m}(\sum_{j=1}^{n}Cov^2(S_{i,\tilde{L}_i},S_{j,\tilde{L}_j})+o(\ell^3))\\ &= \frac{1}{m^2l^2}\frac{\ell^3}{4}m\sigma_{as}^2(1+o(1))\\ &= \frac{\ell}{4n}\sigma_{as}^4(1+o(1)).\end{aligned} \tag{2.18}$$

For part $\frac{n^4}{m^2l^2}Var(\sum_{j=1}^{m}(\sum_{i=1}^{n}w_i^{(j)}X_i)^2)$

$$\begin{aligned}\frac{n^4}{m^2\ell^2}Var(\sum_{j=1}^{m}(\sum_{i=1}^{n}w_i^{(j)}X_i)^2) &\le (\frac{2\ell\log n}{n^2})^4\frac{n^4}{m\ell^2}E(\sum_{i=1}^{n}X_i)^4\\ &= \mathcal{O}(\frac{\ell^2\log^4 n}{n^3})\\ &= o(\frac{\ell}{n}).\end{aligned} \tag{2.19}$$

From this we can see that in the expression of the variance of the stationary jackknife variance estimator, the part $Var\left(\sum_{j=1}^{m}\frac{S_{j,\tilde{L}_j}^2}{n^2}\right)$ plays the dominant role, so we can ignore the other parts and then the expression for our variance is

$$Var\left(n\hat{\sigma}_{jack}^2\right) = \frac{\ell}{4n}\sigma_{as}^4+o\left(\frac{\ell}{n}\right).$$

□

Remark: through the expression of bias and variance of stationary jackknife estimator we

can see that when $\ell \to \infty$ and $\ell = o(n)$, both bias and variance goes to zero. This shows the consistency of our estimator in this case.

2.3.2 Rate of Optimal block length

After obtaining the expression of bias and variance, we can obtain the optimal length of block length ℓ in the order of n by minimizing the MSE of our estimator.
By the standard decomposition of the MSE, we can know that

$$MSE(n\hat{\sigma}^2_{jack}) = Bias^2(n\hat{\sigma}^2_{jack}) + Var(n\hat{\sigma}^2_{jack}) = \mathcal{O}(\frac{1}{\ell^2}) + \mathcal{O}(\frac{\ell}{n})$$

Then we can get $\ell = \mathcal{O}(n^{\frac{1}{3}})$ to be the rate of optimal block length.

2.4 Extension to the nonlinear statistics

In this part, we generalize the consistency results to the nonlinear statistics. The corresponding result for the block jackknife has been well developed in (Kunsch 1989). Under the suitable assumptions, the stationary jackknife estimator is a consistent estimator for the nonlinear statistics.

2.4.1 The Stationary Jackknife on nonlinear statistics

For any nonlinear statistics $T_n = T(X_1, X_2, \cdots X_n)$, the corresponding population parameter is $\theta = T(F)$ where F is the marginal distribution of X. The linearization of T_n at F gives

$$\begin{aligned} T_n &= T(F) + n^{-1}\sum_{i=1}^{n} IF(X_i, F) + R_n \\ &= M_n + R_n. \end{aligned}$$

where IF is the influence function that $IF(x,F)=\lim_{\epsilon\to 0}\frac{[T((1-\epsilon)F+\epsilon\delta_x)-T(F)]}{\epsilon}$. Then we have

$$\begin{aligned} T_n^{(j)} &= T(F)+\left(n-\tilde{L}_j\right)^{-1}\sum_{i=1}^{n}\left(1-\mathbb{1}_{\left(j\le i\le j+\tilde{L}_j-1\right)}\right)IF(X_i,F)+R_n^{(j)} \\ &= L_n^{(j)}+R_n^{(j)}. \end{aligned}$$

For large n, the behavior of $\hat{\sigma}^2_{jack}(T_n)$ is similar as the behavior of $\hat{\sigma}^2_{jack}(L_n)$ if $R_n^{(j)}$ is in the small order. One sufficient condition for the reminder $R_n^{(j)}$ is $\max_j \mathbb{E}\left(R_n^{(j)}\right)=\mathcal{O}\left(n^{-2}\right)$.

Under this condition, it is obvious that $\sum_{j=1}^{n-\ell}R_n^{(j)^2}=\mathcal{O}_p\left(n^{-1}\right)$. Moreover we have

$$\sum\left(R_n^{(j)}-R_n^{(\cdot)}\right)^2 \le \sum R_n^{(j)^2},$$

and the standardized factor for the stationary jackknife estimator is $\mathcal{O}\left(\frac{1}{\ell}\right)$. Thus we have

$$\begin{aligned} n\hat{\sigma}^2_{jack}(T_n) &= n\hat{\sigma}^2_{jack}(L_n)+n\hat{\sigma}^2_{jack}(R_n)+2\frac{(n-\ell)^2}{\ell m}\sum_j\left(L_n^{(j)}-L_n^{\cdot}\right)\left(R_n^{(j)}-R_n^{\cdot}\right) \\ &\le n\hat{\sigma}^2_{jack}(L_n)+n\hat{\sigma}^2_{jack}(R_n)+2n\sqrt{\hat{\sigma}^2_{jack}(L_n)\hat{\sigma}^2_{jack}(R_n)} \\ &= n\hat{\sigma}^2_{jack}(L_n)+\mathcal{O}_p\left(\ell^{-\frac{1}{2}}\right). \end{aligned}$$

In summary we have the following theorem

Theorem 2.4.1. *If* $\max_j \mathbb{E}\left(R_n^{(j)^2}\right)=\mathcal{O}\left(\frac{1}{n^2}\right)$, *then we have* $n\hat{\sigma}^2_{jack}(T_n)=n\hat{\sigma}^2_{jack}(L_n)+\mathcal{O}_p\left(\ell^{-\frac{1}{2}}\right)$. *Under the assumptions in section 3, we have* $n\hat{\sigma}^2_{jack}$ *converges to* σ^2_{as}, *where*

$$\sigma^2_{as} = \sum_{k=-\infty}^{\infty}\mathbb{E}IF(X_0,F)IF(X_k,F). \tag{2.20}$$

More precise we look at the linearization of T on ρ_n instead of F. Assume that $IF\left(y,\rho_n\right)$

exists, then we can write

$$\begin{aligned} T_n^{(j)} &= T\left(\rho_n\right)+(n-\tilde{L}_j)^{-1}\sum_{i=1}^{n}(1-\mathbb{1}_{([j,j+\tilde{L}_j-1])}(i))IF\left(X_i,\rho_n\right)+S_n^{(j)} \\ &= M_n^{(j)}+S_n^{(j)}. \end{aligned}$$

The distance between $\rho_n^{(j)}$ and ρ_n is much closer than the the distance between $\rho_n^{(j)}$ and F. The expected value of the total variation distance between $\rho_n^{(j)}$ and ρ_n is

$$\begin{aligned} \mathbb{E}\left(d_{TV}\left(\rho_n^{(j)},\rho_n\right)\right) &\leq \sum_{i=1}^{n}\mathbb{E}\left|\left(1-\mathbb{1}_{[j,j+\tilde{L}_j-1]}(i)\right)\frac{1}{n-\tilde{L}_j}-\frac{1}{n}\right| \\ &\leq \frac{\ell}{n}+\frac{\ell}{n-2\ell\log n} \\ &= \mathscr{O}\left(\frac{\ell}{n}\right). \end{aligned} \tag{2.21}$$

From this, it is reasonable to expect $S_n^{(j)}$ to be at most of the order $\mathscr{O}_p\left(\frac{\ell^2}{n^2}\right)$. Then we have the following result

Theorem 2.4.2. *If* $\sum S_n^{(j)^2}=\mathscr{O}_p\left(\ell^4 n^{-3}\right)$ *and* $\hat{\sigma}_{jack}(M_n)=\mathscr{O}_p\left(\frac{1}{n}\right)$. *Then we have*

$$n\hat{\sigma}^2_{jack}(T_n)=n\hat{\sigma}^2_{jack}(M_n)+\mathscr{O}_p\left(\ell^{\frac{3}{2}}n^{-1}\right). \tag{2.22}$$

Proof.

$$\begin{aligned} \left|n\hat{\sigma}^2_{jack}(T_n)-n\hat{\sigma}^2_{jack}(M_n)\right| &= \left|n\hat{\sigma}^2_{jack}(S_n)+2\frac{(n-\ell)^2}{\ell m}\sum_{j=1}^{m}\left(M_n^{(j)}-M_n^{\cdot}\right)\left(S_n^{(j)}-S_n^{\cdot}\right)\right| \\ &\leq n\hat{\sigma}^2_{jack}(S_n)+2n\left(\hat{\sigma}^2_{jack}(S_n)\hat{\sigma}^2_{jack}(M_n)\right)^{\frac{1}{2}} \\ &\leq \frac{(n-\ell)^2}{\ell m}\sum S_n^{(j)^2}+2n\left(\hat{\sigma}^2_{jack}(S_n)\hat{\sigma}^2_{jack}(M_n)\right)^{\frac{1}{2}} \\ &= \mathscr{O}_p\left(\ell^3 n^{-2}\right)+\mathscr{O}_p\left(\ell^{\frac{3}{2}}n^{-1}\right). \end{aligned} \tag{2.23}$$

□

The sufficient conditions for $\sum(S_n^{(j)})^2 = \mathcal{O}_p\left(\ell^4 n^{-3}\right)$ for the different kinds of statistics are given in the lemma 4.1-4.3 of (Kunsch 1989). Then for the difference between $L_n^{(j)}$ and $M_n^{(j)}$, assumption (C) in (Kunsch 1989) describes the order of the difference between $n\hat{\sigma}^2_{jack}(M_n)$ and $n\hat{\sigma}^2_{jack}(L_n)$ that

$$n\hat{\sigma}^2_{jack}(M_n) \quad = \quad n\hat{\sigma}^2_{jack}(L_n) + \mathcal{O}_p\left(n^{-\frac{1}{2}}\right) + \mathcal{O}_p\left(\ell n^{-1}\right). \tag{2.24}$$

Then we have a natural consequence

Theorem 2.4.3. *Under the former conditions, we have*

$$n\hat{\sigma}^2_{jack}(T_n) = n\hat{\sigma}^2_{jack}(L_n) + \mathcal{O}_p(\max(\ell^{\frac{3}{2}} n^{-1}, n^{-\frac{1}{2}})). \tag{2.25}$$

We can see that when $\ell \to \infty$ and $\ell = o\left(n^{\frac{1}{2}}\right)$, the nonlinear part will decay and linear part will dominate all effects. And under conditions on the $IF(X,F)$, the $n\hat{\sigma}^2_{jack}(L_n)$ will converge in probability to the asymptotic variance of T_n.

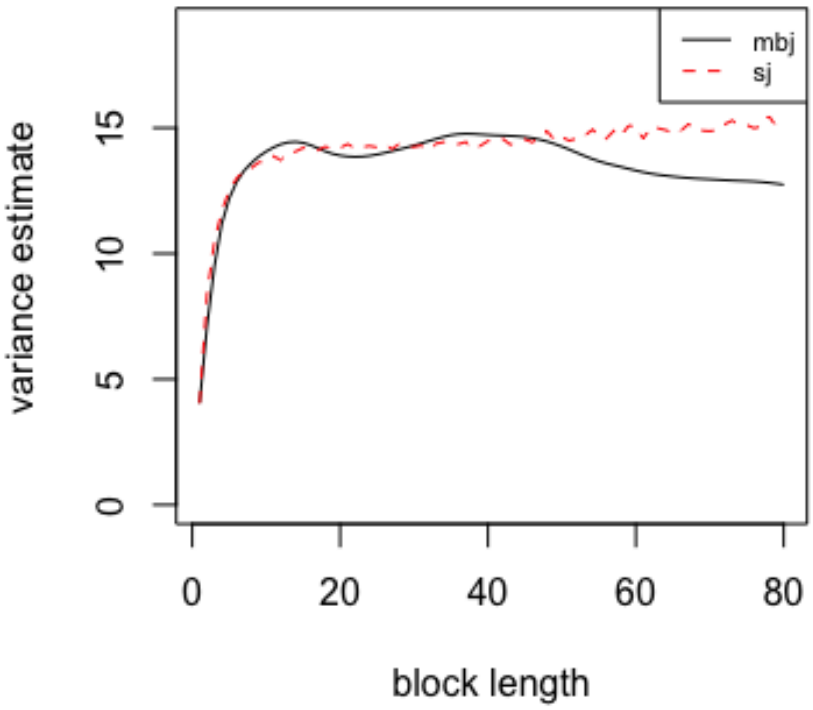
2.5 Simulation

In this section we will use some numerical examples to illustrate what we have developed in the former sections. We will apply our stationary jackknife variance estimator in different scenarios. Also we will compare our method Stationary Jackknife(SJ) with several existing resampling variance estimator such as Moving Block Bootstrap (MBB), Stationary Bootstrap(SB) and Moving Block Jackknife(MBJ).

2.5.1 Two Simulated Data

In this part er simulated two data set under two model that $X_t = Z_t + Z_{t-1} + Z_{t-2} + Z_{t-3}$ and $X_t = Z_t - Z_{t-1} + Z_{t-2} - Z_{t-3} + Z_{t-4}$, where $Z_t \overset{\text{iid}}{\sim} N(0,1)$. We generate N=1000 observations X_t from each model and we want to estimate the quantity $var(\sqrt{N}\bar{X}_t)$ by the stationary jackknife and the moving block jackknife. Moreover we want to investigate their behaviors in a wide range of block length. From the section of optimal block length we know that the optimal block length is of order $N^{\frac{1}{3}}$ and in our case this quantity is $1000^{\frac{1}{3}} = 10$. In this simulation, the block length range $l \in [1,80]$ which contains the range of the optimal block length for both stationary jackknife and moving block jackknife and some extreme block length. In the stationary jackknife the corresponding parameter p in truncated geometric distribution will be $\frac{1}{\ell}$. In the first model we could see that $\sum_{t=1}^{1000} X_t \simeq 4\sum_{t=1}^{1000} Z_t$, So $var(\sqrt{1000}\bar{X}_t) \simeq 16\, var(\sqrt{1000}\bar{Z}_t) = 16$. The figure 1 is the simulation result for this model. From the second model we could see that $\sum_{t=1}^{1000} X_t \simeq \sum_{t=1}^{1000} Z_t$ So we would expect that $var(\sqrt{1000}\bar{X}_t) \simeq var(\sqrt{1000}\bar{Z}_t) = 1$. The figure 2 is the simulation result for the second model. The main difference in these two models is that the autocovariances in model 1 is always positive while in the second model the autocovariance alternate in sign until they become 0 for lag k greater than 4. From the result, we could see (especially from the second graph) that as block length increases, the moving block jackknife estimator is more sensitive to block length and it will give a bad estimator when block length is not suitable. Whereas the stationary jackknife gives the reasonable estimator is a wider range of the block length. This means that when we don't know the exact best block length for the variance estimator, the stationary jackknife estimator may be the better choice for estimation.

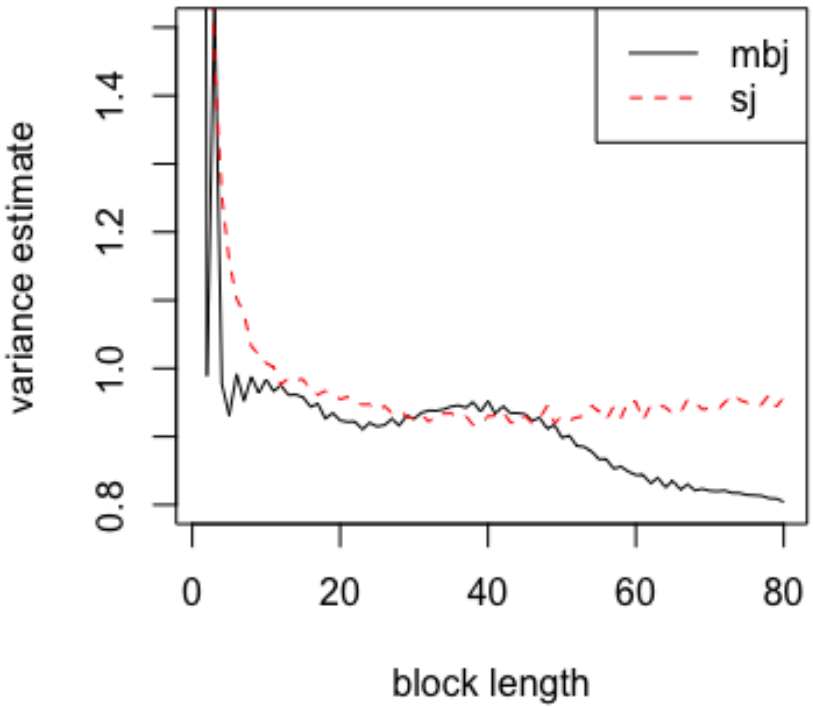
Figure 2.1: The Stationary Jackknife and Block Jackknife variance estimator for model $X_t = Z_t + Z_{t-1} + Z_{t-2} + Z_{t-3}$



2.5.2 Application in time series model

In this section, we are interested in estimating the variance of sample mean of different time series data with the optimal block length order. In each model we will use all four resampling methods to estimate the variance. In each model we will generate N=1000 data as our original data X_t.And the block length is an important variable in all methods. So we will use different values of block length $\ell \in \{N^{\frac{1}{3}}, 2N^{\frac{1}{3}}, 3N^{\frac{1}{3}}\}$ in our simulation. For MBB and SB, the variance estimator is a level-2 estimator, so we need a positive number B as the number of bootstrap replicates to calculate the Monte-Carlo approximation of MBB and SB

Figure 2.2: The Stationary Jackknife and Block Jackknife variance estimator for model $X_t = Z_t - Z_{t-1} + Z_{t-2} - Z_{t-3} + Z_{t-4}$



estimator for this level-2 parameter. We choose B=50 in our situation. Lastly in order to get the estimator of standard error of variance estimator, we do 200 simulations for each model.

AR(1) model

The first example is the AR(1) model $X_t = \phi X_{t-1} + w_t$, where $w_t \overset{\text{iid}}{\sim} N(0,1)$. And we consider the AR parameter ϕ vary in $\{-0.3, 0.5, 0.8\}$. We want to estimate the standardized variance of the mean of the data. $\hat{\sigma}^2 = \text{Var}(\sqrt{N}\bar{X}_N)$ Figure 2.3 shows the results of three different AR(1)

processes. The black solid line in each facet is the true variance of the arithmetic mean. We use the coefficients $\{1,2,3\}$ to represent the corresponding block length $\{N^{\frac{1}{3}}, 2N^{\frac{1}{3}}, 3N^{\frac{1}{3}}\}$. In black interval on each bar represents the confidence interval of the estimation.

ARMA(2,1) model

The second model is the ARMA(2,1) model $X_t - 0.6X_{t-1} + 0.05X_{t-2} = W_t + 0.2W_{t-1}$, where $w_t \overset{\text{iid}}{\sim} N(0,1)$. This is a causal and invertible ARMA processes. Details in (Brockwell and Davis 1990). And this time series model could be expressed as $(1-0.5B)(1-0.1B)X_t = (1+0.2B)W_t$ where B is the backward shift operator. The $var(\sqrt{N}\bar{X}_N)$ is converge to a finite quantity that means σ^2_{as} is well-defined. But the computation of this value is a little bit complex, so we use the standardized standard variance of the mean value of data from the 200 replicate simulation to approximate the true value. Figure 2.4 shows the result under the ARAM model. The red solid line represents the true variance 6.994.

Conclusion

From the comparison of the four methods in estimating the standardized variance of the mean in different time series method, we could see that estimators produced by these four methods could be regarded as the consistency estimator. And usual the block jackknife and stationary jackknife will give us the variance estimator with smaller bias. Also since variance estimator is a level two parameter, so the bootstrap method needs the extra Monte-Carlo approximation which will need more computation compared with the Jackknife method.

2.5.3 Simulation for AR(1) parameter

The AR(1) model $x_t = \phi x_{t-1} + \epsilon_t$, where $\epsilon_t \overset{i.i.d}{\sim} N(0,1)$. We denote $\hat{\phi}$ the least square estimator of ϕ, and this is a nonlinear statistic of $\{x_t\}$. We want to investigate the behavior of

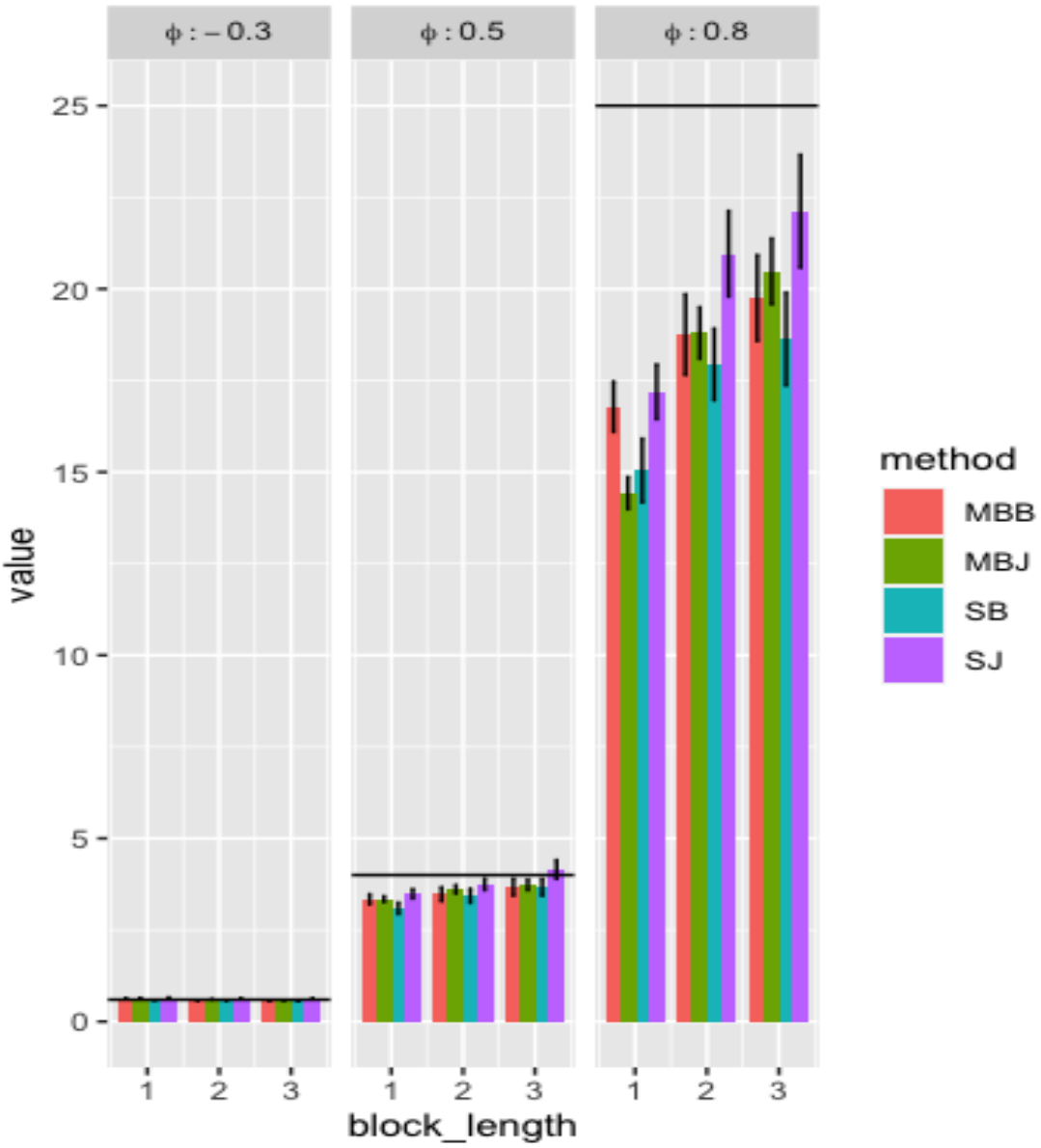


Figure 2.3: Comparison of four resampling methods in the AR processes

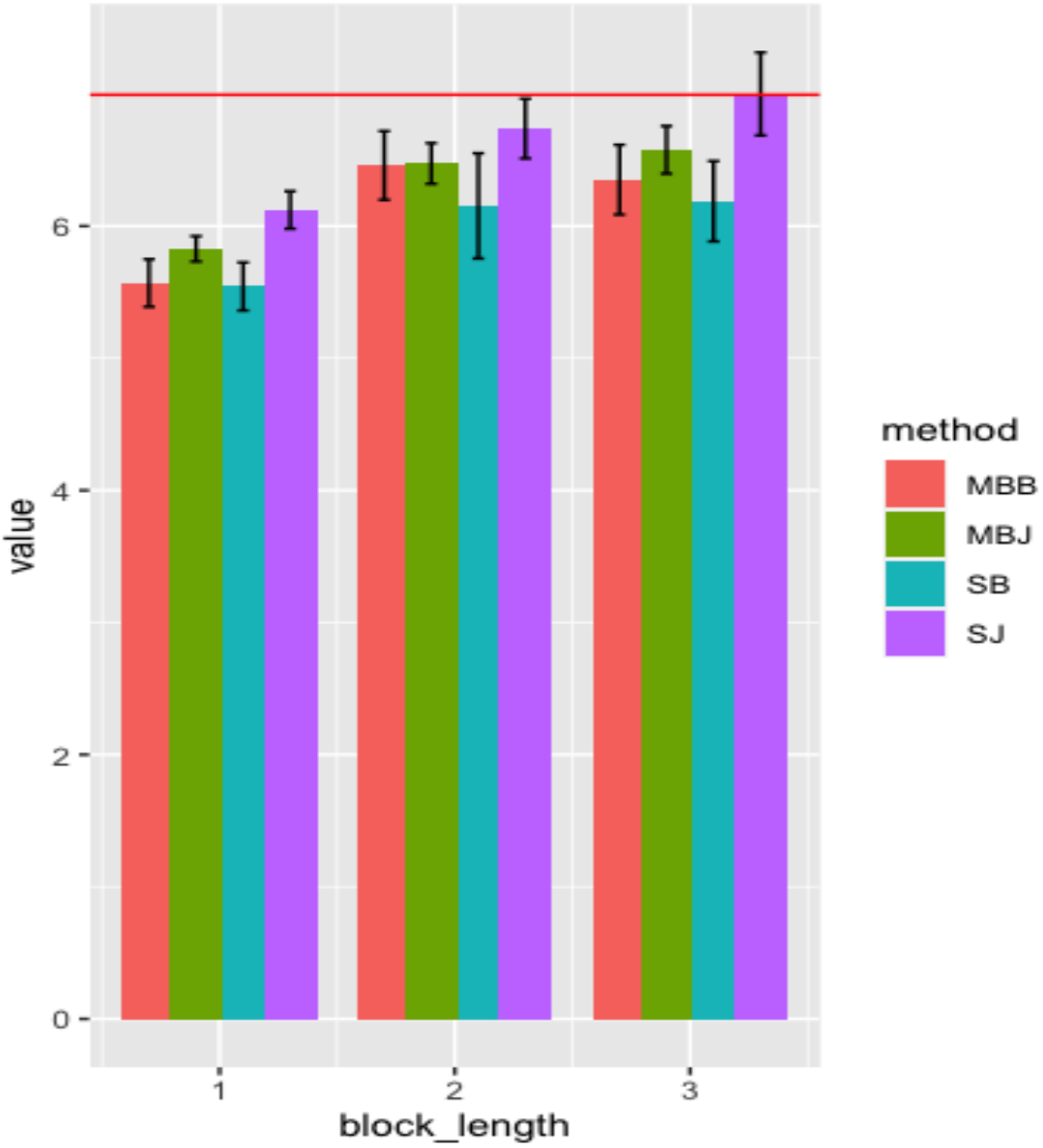


Figure 2.4: Comparison of four resampling methods in the ARMA processes

stationary jackknife estimator on the variance estimation of $\hat{\phi}$. Based on the theoretical result, we have $\sqrt{n}(\hat{\phi}-\phi) \xrightarrow{d} N(0,1-\phi^2)$. In the simulation setting, $\phi = 0.5$, data length is 200 and the block length l varies in $\{1,2,3,4,5\}$. Result will be shown in table 3. From

Table 2.1: the variance estimate by stationary jackknife for AR(1) parameter with different block length

	bl=1	bl=2	bl=3	bl=4	bl=5
SJ	0.65(0.00)	0.72(0.01)	0.74(0.01)	0.75(0.01)	0.75(0.01)
True	0.75	0.75	0.75	0.75	0.75

the result, we can see that the stationary jackknife gives the reasonable estimators to the variance of nonlinear statistics.

2.5.4 Trimmed mean

In this part, we check the stationary jackknife variance estimator's performance on the trimmed mean of an AR process. The trimmed mean is a method of averaging after removing the designated percentage of the smallest and the largest values. In our case, we focus on the 60% trimmed mean. This means that we need to remove the 20% extreme values on both ends of the data beforehand. Formally, for the observations $\{X_i\}_{i=1}^n$ it is defined as

$$\mu_{tm} = \frac{1}{[0.6n]} \sum_{i=[0.2n]}^{[0.8n]} X_{(i)},$$

where $[a]$ is the largest integer not exceed a and $X_{(i)}$ is i-th order statistics of $\{X_i\}$.

In the simulation, the quantity we are interested in is the variance of the scaled mean $Var\left(\sqrt{[0.6n]}\mu_{tm}\right)$. We investigate the stationary jackknife estimator performance for the trimmed mean in different AR(1) models. We choose the AR(1) coefficient α in $[0.2, 0.5, 0.8]$

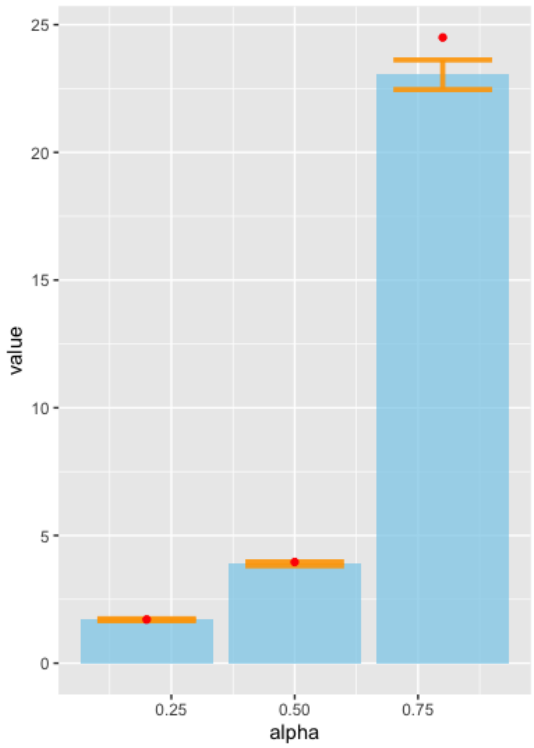


Figure 2.5: The stationary jackknife estimation on the trimmed mean

and in each AR(1) model we generate $n = 600$ samples. In each setting, we repeat the experiment for 200 times to get the expectation and the standard error. For the different strength of the correlation, we choose the different expected block length. The block lengths are 9,18 and 27 for the cases when AR(1) coefficients are 0.2,0.5 and 0.8. The block lengths are approximate to $n^{\frac{1}{3}}$,$2n^{\frac{1}{3}}$ and $3n^{\frac{1}{3}}$. The true variance of the trimmed mean is approximated by the Monte Carlo method. The number of repetition in the Monte Carlo method is 10000. For the convenience of the visualization, we scale both the estimated value and the true value by the sample size. Figure 2.5 shows the result of the stationary jackknife estimation. The three blue bars represent the value of the stationary jackknife estimator, the orange intervals represent the confidence interval for the estimation and the red points are the true variance under three different data generating processes.

From the result, we can see the stationary jackknife variance estimation is pretty close to the true value. Moreover the true variance is inside the confidence interval of the stationary jackknife estimation.

CHAPTER

3

JACKKNIFE LASSO ESTIMATOR

In the high dimensional data situation, Lasso(Tibshirani 1996) was introduced to provide the variable selection and shrinkage estimation on the same time.

3.1 Lasso Problem

Consider the following linear regression model,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \quad i = 1, 2, \cdots, n, \tag{3.1}$$

where y_i is the response and $\mathbf{x}_i = \left(\mathbf{x}_{i,1}, \cdots, \mathbf{x}_{i,p}\right) \in \mathbb{R}^p$ is a p-dimensional covariate vec-

tor, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p)$ is the regression coefficient and $\{\epsilon_1, \epsilon_2 \cdots, \epsilon_n\}$ are independent and identically distributed errors with mean 0 and variance σ^2.

The Lasso estimator is defined as the minimizer of the l_1 penalty least square criterion function,

$$\hat{\boldsymbol{\beta}}_n = \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\mathbf{u}\right)^2 + \lambda_n \sum_{j=1}^{p} \left|\mathbf{u}_j\right|, \tag{3.2}$$

where λ_n is a regularization parameter.

The nature of L_1 penalty leads to a sparse solution, which is favorable in the high dimensional settings. On

An important problem in the context of the Lasso estimator is variance estimation consistency of it. In this paper, we investigate the jackknife-kind method in estimating the variance of the Lasso. In the following sections, we first show the standard jackknife variance estimator fails to provide the consistent estimator. Then we develop a modified version of the delete-d jackknife to estimate the variance of the Lasso and prove its validity.

3.2 Adaptive Lasso

Although Lasso enjoys the sim variable selection and shrinkage estimation, it needs some specific conditions (Zhao and Yu 2006) to ensure the variable selection consistency. It doesn't enjoy the oracle property(Fan and Li 2001).

Definition 3.2.1. *Denote $\mathscr{A} = \{\beta_j \neq 0\}$ and δ is a fitting process, then $\hat{\boldsymbol{\beta}}(\delta)$ is the coefficient estimation by the fitting processes δ. δ is called an oracle processes if it satisfies the following two conditions*

1. *Identifies the right subset model, $\{j : \hat{\boldsymbol{\beta}}_j \neq 0\} = \mathscr{A}$.*

2. *Has the optimal estimation rate,* $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathscr{A}}-\boldsymbol{\beta}\right)\rightarrow N(\mathbf{0},\Sigma)$*, where* Σ *is the covariance matrix when knowing the true nonzero subset.*

Lasso assigns the equal weight on the penalty for each coefficient. The adaptive lasso(Zou 2006) assigns different weights to different coefficients and is a kind of weighted lasso

$$\tilde{\boldsymbol{\beta}}_n = \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min}\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i'\mathbf{u}\right)^2+\lambda_n\sum_{j=1}^{p}\mathbf{w}_j\left|\mathbf{u}_j\right|, \tag{3.3}$$

where the weight vector $\mathbf{w}=\frac{1}{|\hat{\boldsymbol{\beta}}_{ols}|^{\gamma}}$.

3.3 The standard Jackknife

The asymptotic properties of the Lasso estimator has been investigated in (Knight and Fu 2000). Assume the following regularity conditions for the regression design,

$$\mathbf{C}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\rightarrow\mathbf{C}, \tag{3.4}$$

where $\mathbf{C}$ is a non-negative definite matrix and

$$\frac{1}{n}\max_{1\leq i\leq n}\mathbf{x}_i'\mathbf{x}_i\rightarrow 0. \tag{3.5}$$

The limiting distribution of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)$ is shown in the following theorem.

Theorem 3.3.1. *If* $\frac{\lambda_n}{\sqrt{n}}\rightarrow\lambda_0\geq 0$ *and* $\mathbf{C}$ *is nonsingular, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\xrightarrow{d}\underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min}(V(\mathbf{u})), \tag{3.6}$$

where

$$V(\mathbf{u}) \;=\; -2\mathbf{u}'\mathbf{w}+\mathbf{u}'\mathbf{C}\mathbf{u}+\lambda_0\sum_{j=1}^{p}\left[u_j\,\mathrm{sign}\left(\boldsymbol{\beta}_j\right)\mathbb{I}\left(\boldsymbol{\beta}_j\neq 0\right)+\left|u_j\right|\mathbb{I}\left(\boldsymbol{\beta}_j=0\right)\right],$$

and $\mathbf{w}$ *has a* $N\left(\mathbf{0},\sigma^2\mathbf{C}\right)$ *distribution. Moreover we denote* $T_\infty=\arg\min_{\mathbf{u}\in\mathbb{R}^p}V(u)$.

Now we consider the standard jackknife on estimating the variance of the Lasso estimator. The standard jackknife, also called the delete-1 jackknife, estimates the variance of the estimator by the sample variance of the jackknife replications. Each jackknife replication is computed by the whole data except one.

For simplicity, we investigate the consistency of the delete-1 jackknife estimator on the simple linear regression. Model (3.1) is reduced to

$$y_i \;=\; x_i\beta+\epsilon_i. \tag{3.7}$$

Moreover, we assume the covariate x has been centered and standardized which means we have $\frac{1}{n}\sum_{i=1}^n x_i^2=1$, and the variance of error $\sigma=1$. The corresponding Lasso estimator is defined as

$$\hat{\beta}_n \;=\; \arg\min_{u\in\mathbb{R}}\sum_{i=1}^{n}(y_i-x_i\,u)^2+\lambda_n|u|.$$

Then $\sqrt{n}\left(\hat{\beta}_n-\beta\right)$ is the minimizer of the following objective function $V_n(u)$,

$$\begin{aligned}
V_n(u) &= \sum_{i=1}^{n}\left[\left(y_i-x_i\left(\beta+\frac{u}{\sqrt{n}}\right)\right)^2-\epsilon_i^2\right]+\lambda_n\left(\left|\beta+\frac{u}{\sqrt{n}}\right|-\left|\beta\right|\right)\\
&= u^2-\frac{\sum_{i=1}^n x_i\epsilon_i}{\sqrt{n}}u+\lambda_n\left(\mathrm{sign}\left(\beta\right)\frac{u}{\sqrt{n}}\mathbb{I}_{(\beta\neq 0)}+\left|\frac{u}{\sqrt{n}}\right|\mathbb{I}_{(\beta=0)}\right)\\
&= u^2-z_n\,u+\lambda_n\left(\mathrm{sign}\left(\beta\right)\frac{u}{\sqrt{n}}\mathbb{I}_{(\beta\neq 0)}+\left|\frac{u}{\sqrt{n}}\right|\mathbb{I}_{(\beta=0)}\right),
\end{aligned}$$

where $z_n = \frac{\sum_{i=1}^n x_i \epsilon_i}{\sqrt{n}}$.

Denote $\hat{u}_n$ is the minimizer of the $V_n(u)$, then $\hat{u}_n$ satisfies the following condition

$$(\hat{u}_n - z_n) + \frac{\lambda_n}{\sqrt{n}}\left(\text{sign}(\beta)\mathbb{I}_{(\beta\neq 0)} + \hat{k}\mathbb{I}_{(\beta=0)}\right) = 0, \tag{3.8}$$

where $\hat{k} = \frac{\partial |u|}{\partial u}$ and $\lambda_n = \frac{\lambda_n}{2}$. Then

$$\hat{u}_n \quad = \quad \begin{cases} z_n - \frac{\lambda_n}{\sqrt{n}}\,\text{sign}\left(\beta\right), & \beta \neq 0, \\ \text{sign}(z_n)\left(|z_n| - \frac{\lambda_n}{\sqrt{n}}\right)_+, & \beta = 0, \end{cases}$$

where $(x)_+ = \max(x, 0)$.

The j-th jackknife replication for the Lasso estimator is,

$$\hat{\beta}_n^{(j)} \quad = \quad \underset{u\in\mathbb{R}}{\arg\min} \sum_{i=1, i\neq j}^{n} \left(y_i - x_i u\right)^2 + \lambda_n |u|.$$

Similarly $\sqrt{n-1}\left(\hat{\beta}_n^{(j)} - \beta\right)$ is the minimizer of the objective function $V_n^{(j)}(u)$,

$$\begin{aligned} V^{(j)}(u) \quad &= \quad \frac{n - x_j^2}{n-1} u^2 - \frac{\sum_{i=1, i\neq j}^n x_i \epsilon_i}{\sqrt{n-1}} u + \lambda_n \left(\text{sign}\left(\beta\right) \frac{u}{\sqrt{n-1}} \mathbb{I}_{(\beta\neq 0)} \right. \\ &+ \quad \left. \left| \frac{u}{\sqrt{n-1}} \right| \mathbb{I}_{(\beta=0)} \right). \end{aligned} \tag{3.9}$$

Denote $\hat{u}_n^{(j)}$ is the minimizer of the $V_n^{(j)}(u)$, then $\hat{u}_n^{(j)}$ satisfies the condition,

$$\left(\frac{n - x_j^2}{n-1} \hat{u}_n^{(j)} - z_n^{(j)} \right) + \frac{\lambda_n}{\sqrt{n-1}} \left(\text{sign}(\beta)\mathbb{I}_{\left(\beta\neq 0\right)} + \hat{k}\mathbb{I}_{(\beta=0)} \right) = 0, \tag{3.10}$$

where $z_n^{(j)} = \frac{\sum_{i=1,i\neq j}^n x_i\epsilon_i}{\sqrt{n-1}}$. Then

$$\hat{u}_n^{(j)} = \begin{cases} \frac{n-1}{n-x_j^2}\left(z_n^{(j)} - \frac{\lambda_n}{\sqrt{n-1}}\operatorname{sign}(\beta)\right) & \beta \neq 0 \\ \frac{n-1}{n-x_j^2}\operatorname{sign}\left(z_n^{(j)}\right)\left(|z_n^{(j)}| - \frac{\lambda_n}{\sqrt{n-1}}\right)_+ & \beta = 0 \end{cases}.$$

The jackknife estimator for $var(\sqrt{n}(\hat{\beta}-\beta_0))$ is the sample variance of $\{\hat{u}^{(j)} - \hat{u}_n\}$.

Theorem 3.3.2. *In the simple linear regression* (3.7)*, suppose the true coefficient* $\beta = 0$*, the standard jackknife variance estimator is a inconsistent estimator for the variance of lasso estimator under the conditions (1):* $\lim_{n\to\infty}\frac{\lambda_n}{\sqrt{n}} = c$*, (2):* $max_i \frac{x_i^2}{n} \to 0$*, (3):* $\frac{\sum_{i=1}^n x_i^2}{n} = 1$.

Proof. First we prove $\max_j \left|z_n^{(j)} - z_n\right| = o_p\left(\frac{1}{n^\alpha}\right) \quad \forall\, j$ and $0 < \alpha < \frac{1}{2}$. We have

$$\begin{aligned} \left|z_n^{(j)} - z_n\right| &= \left|\frac{\sum_{i=1}^n x_i\epsilon_i - x_j\epsilon_j}{\sqrt{n-1}} - \frac{\sum_{i=1}^n x_i\epsilon_i}{\sqrt{n}}\right| \\ &= \left|\frac{\sum_{i=1}^n x_i\epsilon_i - x_j\epsilon_j}{\sqrt{n-1}} - \frac{\sum_{i=1}^n x_i\epsilon_i}{\sqrt{n-1}} + (\frac{1}{\sqrt{n-1}} - \frac{1}{\sqrt{n}})\sum_{i=1}^n x_i\epsilon_i\right| \\ &\leq \left|\frac{x_j\epsilon_j}{\sqrt{n-1}}\right| + \left|\sum_{i=1}^n x_i\epsilon_i\right| \frac{1}{\sqrt{n(n-1)}\left(\sqrt{n}+\sqrt{n-1}\right)}. \end{aligned}$$

Now we only need to prove $\max_j \left|\frac{x_j\epsilon_j}{\sqrt{n-1}}\right| = o_p\left(\frac{1}{n^\alpha}\right)$ and $\left|\frac{\sum_{i=1}^n x_i\epsilon_i}{\sqrt{n(n-1)}(\sqrt{n}+\sqrt{n-1})}\right| = o_p\left(\frac{1}{n^\alpha}\right)$.

Denote $\epsilon_0 = \max_j\{|\epsilon_1|, |\epsilon_2| \cdots |\epsilon_n|\}$, then for $\forall\, t$ we have the following,

$$\begin{aligned} e^{t\mathbb{E}\epsilon_0} &\leq \mathbb{E}e^{t\epsilon_0} \\ &= \mathbb{E}\max_i e^{t|\epsilon_i|} \\ &\leq \sum_{j=1}^n \mathbb{E}e^{t|\epsilon_j|} \\ &\leq 2ne^{\frac{t^2}{2}}. \end{aligned}$$

Therefore $\mathbb{E}\epsilon_0 \le \sqrt{2\log 2n}$. Notice that for any $\delta > 0$, we have

$$\begin{aligned} P\left(n^{\alpha}\max_j|\frac{x_j\epsilon_j}{\sqrt{n-1}}| > \delta\right) &\le \frac{\max_j\left|x_j\right|\mathbb{E}\max_j\left|\epsilon_j\right|n^{\alpha-\frac{1}{2}}}{\delta} \\ &\le \frac{\max_j\left|x_j\right|\sqrt{2\log 2n}}{\delta n^{\frac{1-2\alpha}{2}}} \\ &\to 0. \end{aligned}$$

Similarly we have for any $\delta > 0$,

$$\begin{aligned} P\left(n^{\alpha}\frac{\left|\sum_{i=1}^n x_i\epsilon_i\right|}{\sqrt{n(n-1)}\left(\sqrt{n}+\sqrt{n-1}\right)} > \delta\right) &\le \frac{var\left(\sum_{i=1}^n x_i\epsilon_i\right)}{n^{3-2\alpha}\delta^2} \\ &= \frac{\sum_{i=1}^n x_i^2}{n}\frac{1}{n^{2-2\alpha}\delta^2} \\ &\to 0. \end{aligned}$$

Thus $\max_j\left|z_n^{(j)} - z_n\right| = o_p\left(\frac{1}{n^{\alpha}}\right)$, where $0 < \alpha < \frac{1}{2}$.

Next we consider $\left|\hat{u}_n^{(j)} - \hat{u}_n\right|$, we have,

$$\begin{aligned} \left|\hat{u}_n^{(j)} - \hat{u}_n\right| &= \left|\frac{n-1}{n-x_j^2}\operatorname{sign}\left(z_n^{(j)}\right)\left(\left|z_n^{(j)}\right| - \frac{\lambda_n}{\sqrt{n-1}}\right)_+ - \operatorname{sign}(z_n)\left(|z_n| - \frac{\lambda_n}{\sqrt{n}}\right)_+\right| \\ &\le \left|\frac{n-1}{n-x_j^2}\operatorname{sign}\left(z_n^{(j)}\right)\left(|z_n^{(j)}| - \frac{\lambda_n}{\sqrt{n-1}}\right)_+ - \operatorname{sign}\left(z_n^{(j)}\right)\left(|z_n^{(j)}| - \frac{\lambda_n}{\sqrt{n-1}}\right)_+\right| \\ &+ \left|\operatorname{sign}\left(z_n^{(j)}\right)\left(\left|z_n^{(j)}\right| - \frac{\lambda_n}{\sqrt{n-1}}\right)_+ - \operatorname{sign}\left(z_n^{(j)}\right)\left(|z_n^{(j)}| - \frac{\lambda_n}{\sqrt{n}}\right)_+\right| \\ &+ \left|\operatorname{sign}\left(z_n^{(j)}\right)\left(\left|z_n^{(j)}\right| - \frac{\lambda_n}{\sqrt{n}}\right)_+ - \operatorname{sign}(z_n)\left(|z_n| - \frac{\lambda_n}{\sqrt{n}}\right)_+\right| \\ &= A_j + B_j + C_j. \end{aligned}$$

For A_j we have,

$$\begin{aligned}
\max_j A_j &\leq \max_j \left| \frac{x_j^2-1}{n-x_j^2} \left(\left| z_n^{(j)} \right| - \frac{\lambda_n}{\sqrt{n-1}} \right) \right| \\
&\leq \max_j \left| \frac{x_j^2-1}{n-x_j^2} \right| \max_j \left| \left| z_n^{(j)} \right| - \frac{\lambda_n}{\sqrt{n-1}} \right| \\
&= \mathcal{O}\left(\frac{1}{n}\right) \mathcal{O}_p(1) \\
&= \mathcal{O}_p\left(\frac{1}{n}\right).
\end{aligned}$$

Moreover we have $\max_j B_j \leq \frac{\lambda_n}{(\sqrt{n}+\sqrt{n-1})\sqrt{n(n-1)}}$, so that $\max_j B_j = \mathcal{O}_p(\frac{1}{n})$.

Finally we aim at proving $\max_j C_j = o_p(\frac{1}{n^\alpha})$ where $0 < \alpha < \frac{1}{2}$. For any $\delta > 0$, denote $P_{1,j} = P\left(n^\alpha C_j > \delta\right)$, then we have,

$$\begin{aligned}
P_{1,j} &= P\left(n^\alpha \left| \operatorname{sign}\left(z_n^{(j)}\right) \left(\left| z_n^{(j)} \right| - \frac{\lambda_n}{\sqrt{n}} \right)_+ - \operatorname{sign}(z_n) \left(|z_n| - \frac{\lambda_n}{\sqrt{n}} \right)_+ \right| > \delta, \operatorname{sign}\left(z_n^{(j)}\right) = \operatorname{sign}(z_n) \right) \\
&+ P\left(n^\alpha \left| \operatorname{sign}\left(z_n^{(j)}\right) \left(\left| z_n^{(j)} \right| - \frac{\lambda_n}{\sqrt{n}} \right)_+ - \operatorname{sign}(z_n) \left(|z_n| - \frac{\lambda_n}{\sqrt{n}} \right)_+ \right| > \delta, \operatorname{sign}\left(z_n^{(j)}\right) \neq \operatorname{sign}(z_n) \right) \\
&\leq P\left(n^\alpha \left| z_n^{(j)} - z_n \right| > \delta \right) + P\left(\left| z_n - z_n^{(j)} \right| > \frac{\lambda_n}{\sqrt{n}} \right) \\
&\to 0.
\end{aligned}$$

The last inequality is due to the fact that if $\operatorname{sign}(z_n) \neq \operatorname{sign}\left(z_n^{(j)}\right)$ then we must have $\left| z_n - z_n^{(j)} \right| \geq \frac{\lambda_n}{\sqrt{n}}$, otherwise both $|z_n|$ and $\left| z_n^{(j)} \right|$ will drop in $\left[0, \frac{\lambda_n}{\sqrt{n}}\right]$. Moreover we already known $\max_j \left| z_n^{(j)} - z_n \right| = o_p\left(\frac{1}{n^\alpha}\right)$, so we have,

$$\begin{aligned}
\max_j P_{1,j} &\leq P\left(n^\alpha \max_j \left| z_n - z_n^{(j)} \right| > \delta \right) + P\left(\max_j \left| z_n - z_n^{(j)} \right| > \frac{\lambda_0}{\sqrt{n}} \right) \\
&\to 0
\end{aligned}$$

This means $\max_j C_j = o_p(\frac{1}{n^\alpha})$, then we successfully prove the result

$$\hat{\sigma}^2_{jack} = \frac{\sum_{i=1}^n (\hat{u}_n^{(j)} - \hat{u}_n^{(\cdot)})^2}{n} \leq \max_j \left| \hat{u}_n^{(j)} - \hat{u}_n \right|^2 = o_p(1),$$

where $\hat{u}_n^{(\cdot)} = \frac{\sum_{j=1}^n \hat{u}_n^{(j)}}{n}$. □

3.4 Delete-d Jackknife Estimator

In this section, we introduce a new method to provide the consistent estimation for the variance of the Lasso estimator. As we have shown in the last section, the standard jackknife variance estimator is not a reliable estimator for the Lasso estimator. The reason is that the Lasso estimator is not smooth enough especially when the true regression parameter is zero. This gives us the motivation to consider the delete-d jackknife variance estimator, which is effective in dealing with the non smooth statistics. Unlike the standard jackknife, we randomly delete d observations when computing the replications of the delete-d jackknife. We now propose a modified version of the delete-d jackknife variance estimator for the Lasso estimator.

For the Lasso estimator $\hat{\boldsymbol{\beta}}_n$ in (3.2), we define the j-th delete-d jackknife replication of it as,

$$\hat{\boldsymbol{\beta}}_{n,d}^{(j)} = \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^r \left(y_i^{(j)} - \mathbf{x}_i^{(j)'} \mathbf{u} \right)^2 + \sqrt{\frac{rd}{n^2}} \lambda_n \sum_{j=1}^p \left| \mathbf{u}_j \right|, \quad 1 \leq j \leq N, \tag{3.11}$$

where $r = n - d$, $N = \binom{n}{d}$ and $\left\{ \left(\mathbf{x}_i^{(j)}, y_i^{(j)} \right) \right\}_{i=1}^r$ is the remaining data in the j-th replication.

Then we define the **Modified Delete-d Jackknife** variance estimator $\hat{\mathbf{V}}^2_{n,d}$ as,

$$\hat{\mathbf{V}}^2_{n,d} = \frac{r}{dN}\sum_{j=1}^{N}\left(\hat{\boldsymbol{\beta}}^{(j)}_{n,d}-\hat{\boldsymbol{\beta}}^{(\cdot)}_{n,d}\right)\left(\hat{\boldsymbol{\beta}}^{(j)}_{n,d}-\hat{\boldsymbol{\beta}}^{(\cdot)}_{n,d}\right)', \tag{3.12}$$

where $\hat{\boldsymbol{\beta}}^{(\cdot)}_{n,d} = \frac{\sum_{j=1}^{N}\hat{\boldsymbol{\beta}}^{(j)}_{n,d}}{N}$.

3.5 Jackknifing the Lasso estimator

In this section, we shows the consistency of the modified delete-d jackknife estimator. We focus on two quantities $T_n = \sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right)$ and $T^\star_n = \sqrt{\frac{nr}{d}}\left(\hat{\boldsymbol{\beta}}^\star_{n,d} - \boldsymbol{\beta}\right)$ where $\hat{\boldsymbol{\beta}}^\star_{n,d}$ is a random piece sampled from the jackknife replications $\left\{\hat{\boldsymbol{\beta}}^{(j)}_{n,d}\right\}$. Let $G_n(\cdot)$ denote the distribution of T_n and $\tilde{G}_n(\cdot)$ denote the conditional distribution of $T^\star_n$ given the observations. More specifically for any $B \in \mathbb{B}(\mathbb{R}^p)$, $\tilde{G}_n(B) = P_\star\left(T^\star_n \in B\right)$ where $P_\star$ is the conditional probability based on the observations.

3.5.1 Consistency of the Distributional Approximation

The first result shows that the modified delete-d jackknife gives a valid approximation to the distribution T_n.

Theorem 3.5.1. *Suppose that the following assumptions hold:*

(C.1) $n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \to \mathbf{C}$, *where* $\mathbf{C}$ *is positive definite.*

(C.2) $\sup_i \|\mathbf{x}_i\|^4 = \mathcal{O}(1)$.

(C.3) $\lambda_n n^{-\frac{1}{2}} \to \lambda_0 \geq 0$.

(C.4) The errors $\{\epsilon_i\}_{i=1}^n$ *are independent and identically distributed with* $\mathbb{E}\epsilon_i = 0$ *and* $\mathrm{Var}(\epsilon_i) = \sigma^2 \in (0,\infty)$.

(C.5) $d/n \to \alpha \in (0,1)$.

Then with probability 1

$$\varrho\left(\tilde{G}_n(\cdot), G_n(\cdot)\right) \to 0 \quad a.s. \quad n \to \infty,$$

where $\varrho(\cdot,\cdot)$ *denotes the Prohorov metric on the set of all probability measure on* $(\mathbb{R}^p, \mathbb{B}(\mathbb{R}^p))$.

Proof. We notice that $T_n^\star$ is the minimizer of the following objective function,

$$V_n^\star(\mathbf{u}) = \sum_{i=1}^{r}\left(y_i^\star - \mathbf{x}_i^{\star\prime}\left(\boldsymbol{\beta} + \sqrt{\frac{d}{nr}}\mathbf{u}\right)\right)^2 + \sqrt{\frac{rd}{n^2}}\lambda_n \sum_{j=1}^{p}\left|\boldsymbol{\beta}_j + \sqrt{\frac{d}{nr}}u_j\right|, \tag{3.13}$$

where $\{(\mathbf{x}_i^\star, y_i^\star)\}$ are r observations sampled without replacement from the observations.

It is equivalent to investigate a centered version $V_n^\star(\mathbf{u}) = V_n^\star(\mathbf{u}) - V_n^\star(\mathbf{0})$. Then we have,

$$\begin{aligned} V_n^\star(\mathbf{u}) &= \frac{d}{n}\mathbf{u}'\frac{\sum_{i=1}^r \mathbf{x}_i^\star \mathbf{x}_i^{\star\prime}}{r}\mathbf{u} - 2\mathbf{u}'\frac{\sum_{i=1}^r \epsilon_i^\star \mathbf{x}_i^\star}{\sqrt{\frac{rn}{d}}} + \sqrt{\frac{rd}{n^2}}\lambda_n \sum_{j=1}^{p}\left(\left|\beta_j + \frac{u_j}{\sqrt{\frac{rn}{d}}}\right| - |\beta_j|\right) \\ &= \frac{d}{n}\mathbf{u}'\frac{\sum_{i=1}^r \mathbf{x}_i^\star \mathbf{x}_i^{\star\prime}}{r}\mathbf{u} - 2\mathbf{u}'\frac{\sum_{i=1}^r \epsilon_i^\star \mathbf{x}_i^\star}{\sqrt{\frac{rn}{d}}} + \sqrt{\frac{rd}{n^2}}\frac{\lambda_n}{\sqrt{n}}\left(\sum_{j=1}^{p_0}\operatorname{sign}(\beta_j)\frac{u_j}{\sqrt{\frac{r}{d}}}\right. \\ &+ \left.\sum_{j=p_0+1}^{p}\frac{|u_j|}{\sqrt{\frac{r}{d}}}\right). \end{aligned}$$

Denote the limit of $V_n^\star(\mathbf{u})$ as $V^\star(\mathbf{u})$, now we prove $V^\star(\mathbf{u})$ is equivalent to the limiting estimating equation of T_n, that is,

$$V^\star(\mathbf{u}) = \alpha\mathbf{u}'\mathbf{C}\mathbf{u} - 2\alpha\mathbf{u}'\mathbf{w}_1 + \lambda_0\alpha\left(\sum_{j=1}^{p_0}\operatorname{sign}(\beta_j)u_j + \sum_{j=p_0+1}^{p}|u_j|\right),$$

where $\mathbf{w}_1$ is a random vector with the normal distribution $N(\mathbf{0}, \sigma^2\mathbf{C})$.

First we have

$$\begin{aligned}\frac{\sum_{i=1}^{r}\mathbf{x}_i^{\star}\mathbf{x}_i^{\star\prime}}{r} &\rightarrow \mathbb{E}_{\star}\mathbf{x}_i^{\star}\mathbf{x}_i^{\star\prime}\\ &= \frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'}{n}\\ &= \mathbf{C}_n.\end{aligned}$$

According to the condition, $\mathbf{C}_n \rightarrow \mathbf{C}$ as $n \rightarrow \infty$.

Then for any nonzero vector $\mathbf{a} \in \mathbb{R}^p$, we define $z_i = \mathbf{a}'\mathbf{x}_i\epsilon_i$ and $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z}_n)^2$.

We have

$$\begin{aligned}\mathbb{E}\left(\frac{\sum_{i=1}^{n}z_i^2}{n}\right) &= \mathbf{a}'\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'}{n}\mathbf{a}\mathbb{E}\left(\epsilon_i^2\right)\\ &= \mathbf{a}'\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'}{n}\mathbf{a}\sigma^2\\ &\rightarrow \mathbf{a}'\mathbf{C}\mathbf{a}\sigma^2, \qquad (3.14)\end{aligned}$$

and

$$\begin{aligned}Var\left(\frac{\sum_{i=1}^{n}z_i^2}{n}\right) &= \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathbf{a}'\mathbf{x}_i|^4}{n}Var\left(\epsilon_i^2\right)\\ &\leq 2\frac{1}{n}\|\mathbf{a}\|^4\sup_i\|\mathbf{x}_i\|^4\\ &\rightarrow 0. \qquad (3.15)\end{aligned}$$

Based on (3.14) and (3.15), we have $\frac{\sum_{i=1}^{n}z_i^2}{n} \xrightarrow{p} \mathbf{a}'\mathbf{C}\mathbf{a}\sigma^2$. Similarly, we can prove that $\bar{z}_n \xrightarrow{p} 0$. Therefore we have

$$\hat{\sigma}_n^2 \xrightarrow{p} \mathbf{a}'\mathbf{C}\mathbf{a}\sigma^2. \qquad (3.16)$$

Now we denote $J(t) = P_\star\left(\left(\frac{nr}{d}\right)^{\frac{1}{2}} \frac{\left(\bar{z}_r^\star - \bar{z}_n\right)}{\hat{\sigma}_n} \leq t\right)$, where $\bar{z}_r^\star = \sum_{i=1}^r \mathbf{a}'\mathbf{x}_i^\star \epsilon_i^\star / r$. According to the Berry-Esseen bound for simple random sample without replacement(Höglund 1978), we have

$$\sup_t |J(t) - \Phi(t)| \leq \frac{C}{(rd/n)^{\frac{1}{2}}} \frac{n^{-1}\sum_{i=1}^n |z_i - \bar{z}_n|^3}{\hat{\sigma}_n^3}, \tag{3.17}$$

where $\Phi(t)$ is the cdf of the standard normal distribution.

Moreover we have

$$\begin{aligned}
\frac{\sum_{i=1}^n |z_i - \bar{z}_n|^3}{n^{\frac{3}{2}}} &= \frac{1}{n^{\frac{1}{2}}} \frac{\sum_{i=1}^n |z_i - \bar{z}_n|^3}{n} \\
&\leq 8n^{-\frac{1}{2}} \frac{\sum_{i=1}^n |z_i|^3}{n} \\
&\leq 8\|\mathbf{a}\|^3 \sup_i \|\mathbf{x}_i\|^3 \frac{1}{n^{\frac{3}{2}}} \sum_{i=1}^n |\epsilon_i|^3 \\
&\xrightarrow{as} 0.
\end{aligned}$$

The last convergence is due to the Marcinkiewicz-Zygmund Strong Law of Large Number(Athreya and Lahiri 2006) on $\left\{\epsilon_i^3\right\}$. Then the right hand side of (3.17) goes to 0 as n goes to infinity.

When we combine (3.16) and (3.17) together, we have

$$\mathscr{L}_\star\left(\left(\frac{nr}{d}\right)^{\frac{1}{2}} \frac{\sum_{i=1}^r \mathbf{x}_i^\star \epsilon_i^\star}{r}\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2 \mathbf{C}\right), \tag{3.18}$$

where $\mathscr{L}_\star(\cdot)$ is the conditional distribution on the observations. This implies that

$$V_n^\star(\mathbf{u}) \to V^\star(\mathbf{u}).$$

Following the arguments as in (Knight and Fu 2000), one can establish the weak convergence

$$\mathscr{L}_\star\left(V_n^\star(\cdot)\right) \xrightarrow{d} \mathscr{L}\left(V(\cdot)\right),$$

on the space of all functions on $\mathbb{R}^p$ that are uniformly bounded on compact subsets of $\mathbb{R}^p$. This in turn implies

$$\varrho\left(\tilde{G}_n(\cdot), G_n(\cdot)\right) \to 0 \quad as \quad n \to \infty.$$

The theorem is proved. □

3.5.2 Bias and variance consistency

Theorem 3.5.2. *Under the assumptions of the Theorem 3, we have*

$$\mathbb{E}_\star\left(T_n^\star\right) \to \mathbb{E}(T_\infty),$$

and

$$\left(Var_\star\left(T_n^\star\right)\right)_{p\times p} \to \left(Var\left(T_\infty\right)\right)_{p\times p},$$

with probability 1.

Proof. The idea is to prove $\{\|T_n^\star\|^2\}$ is uniformly integrable with probability 1. The proof can be finished step by step by following the proof in (Chatterjee and Lahiri 2011). We omit the details here. □

3.6 Modified Bootstrap Estimator

For the later comparison in section 3.7, we introduce the modified Bootstrap method(Chatterjee and Lahiri 2011) which provides a consistent estimation for the Lasso variance estimation. (Chatterjee and Lahiri 2010) has shown the residual Bootstrap estimator will converge to a random measure. The modified Bootstrap method is based on the residual bootstrap and its key idea is to force the components of the Lasso estimator $\hat{\boldsymbol{\beta}}$ to exact 0 when it closes to 0. Specifically, define the threshold $\{a_n\}$ that satisfies

$$a_n + \left(n^{-\frac{1}{2}} \log n\right) a_n^{-1} \quad \rightarrow \quad 0.$$

Then define the modified lasso estimator $\tilde{\boldsymbol{\beta}}_n = \left(\tilde{\boldsymbol{\beta}}_{n,1}, \cdots \tilde{\boldsymbol{\beta}}_{n,p}\right)$ as

$$\tilde{\boldsymbol{\beta}}_{n,j} = \hat{\boldsymbol{\beta}}_{n,j} \mathbb{I}\left(\hat{\boldsymbol{\beta}}_{n,j} \geq a_n\right).$$

The modified bootstrap is based on the residual $\{\tilde{r}_i\}$ where $\tilde{r}_i = y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}$. Each time, we select the random samples $\{e_i^\star\}_{i=1}^n$ from the centered residuals $\{\tilde{r}_i - \bar{\tilde{r}}_i\}$. Then compute the responses $\{y_i^\star\}$ based on the relationship $y_i^\star = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} + e_i^\star$.

3.7 Numerical results

In this section, we implement a simulations to evaluate the performance of the variance estimation of the lasso estimator by the jackknife method and do a comparison with the modified bootstrap method.

3.7.1 Basic settings

In the simulation we assume the true coefficient is $[1,3,0]$, the sample size $n = 250$. For the lasso estimator, the regularization parameter λ_n satisfies $\frac{\lambda_n}{\sqrt{n}} \to 0.5$. For the jackknife method, since it is impossible to get all the combination when the deleting number d is large, we use the random sample to avoid this bottleneck. For the modified bootstrap estimator, the threshold is chosen by minimizing the mean square error of the variance estimation. For the true value, we use the Monte Carlo method to approximate it, the number of the Monte Carlo replication is 1000.

3.7.2 Result

For the variance comparison, we care about the some specific position of the covariance matrix of $\boldsymbol{\beta}$. We detect the variance in the positions (1,1),(1,3),(3,3) which represent the variance of the nonzero coefficient, covariance between the zero and nonzero coefficient and the variance of the zero coefficient. The result is shown in Table 3.1.

Table 3.1: The variance estimation for the lasso estimation based on $\beta = (1,3,0)$

Position	Jackknife	Modified Bootstrap	True value
(1,1)	1.019	0.990	1.025
(1,3)	-0.004	0.008	-0.011
(3,3)	0.766	0.654	0.734

From the result, we can see that the jackknife method provides a reasonable variance estimation for the covaraince matrix of the lasso estimator. It also correctly estimate the sign of the interested position of the covariance matrix.

CHAPTER

4

RESAMPLING THE MARKOWITZ PORTFOLIO OPTIMIZATION

4.1 Markowitz Problem

Markowitz porfolio selection is a cornerstone in the modern economics(Markowitz 1952). Suppose $\{\mathbf{y}_t\}_{t=1}^n$ are the returns for the stocks, where $\mathbf{y}_t \in \mathbb{R}^{p\times 1}$. Denote $\Sigma_n, \boldsymbol{\mu}_n$ as the covariance matrix and the mean of $\mathbf{y}_t$. The Markowitz problem aims at finding the best Portfolio allocation(Jobson and Korkie 1980) $\mathbf{w}_n$ to minimize the risk of the portfolio given the ex-

pected return is μ_0. This can be formatted as following

$$\begin{aligned} \arg\min_{\mathbf{w}} \quad & \mathbf{w}^T \Sigma_n \mathbf{w}, \\ \text{s.t} \quad & \mathbf{1}^T \mathbf{w} = 1, \\ & \boldsymbol{\mu}_n^T \mathbf{w} = \mu_0. \end{aligned}$$

The explicit formula for $\mathbf{w}_n$ is

$$\mathbf{w}_n = \frac{C_n - \mu_0 A_n}{A_n C_n - B_n^2} \Sigma_n^{-1} \mathbf{1} + \frac{\mu_0 A_n - B_n}{A_n C_n - B_n^2} \Sigma_n^{-1} \boldsymbol{\mu}_n,$$

where $A_n = \mathbf{1}^T \Sigma_n^{-1} \mathbf{1}$, $B_n = \mathbf{1}^T \Sigma_n^{-1} \boldsymbol{\mu}_n$ and $C_n = \boldsymbol{\mu}_n^T \Sigma_n^{-1} \boldsymbol{\mu}_n$. Then the optimal risk is

$$R = \frac{A_n \mu_0^2 - 2 B_n \mu_0 + C_n}{A_n C_n - B_n^2}. \tag{4.1}$$

Moreover, if μ_0 is not fixed, we can get the global minimal risk

$$\begin{aligned} R_g &= (A_n)^{-1}, \\ &= \left(\mathbf{1}^T \Sigma_n^{-1} \mathbf{1} \right)^{-1}. \end{aligned} \tag{4.2}$$

4.2 Factor model

Factor models(Bai 2003)(Fan et al. 2008) have been widely used both theoretically and empirically in economics and finance. Suppose the return $\{\mathbf{y}_t\}$ has the factor structure

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{Q}\mathbf{f}_t + \boldsymbol{\epsilon}_t, \tag{4.3}$$

where $\boldsymbol{\mu}$ is the mean vector of the processes $\{y_t\}$, $\mathbf{Q} \in \mathbb{R}^{p\times K}$ is an unknown fixed factor loading matrix, $\{\mathbf{f}_t\}_{t=1}^n$ is the K dimensional latent factor processes with $\mathbf{f}_t \sim N(\mathbf{0}, \Sigma_{\mathbf{f}})$, $\{\boldsymbol{\epsilon}_t\}$ is a vector white noise processes with $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \Sigma_0)$. After we have the estimators $\hat{\mathbf{Q}}$ and $\hat{\mathbf{f}}$ for $\mathbf{Q}$ and $\mathbf{f}$, the corresponding estimator for $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}} \;=\; \bar{\mathbf{y}} - \hat{\mathbf{Q}}\bar{\hat{\mathbf{f}}}.$$

We can use the centered version of $\mathbf{y}_i = \mathbf{y}_i - \bar{\mathbf{y}}$ and $\mathbf{f}_i = \mathbf{f}_i - \bar{\mathbf{f}}$ to eliminate $\mathbf{f}$. In the following part we treat $\mathbf{y}$ and $\mathbf{f}$ as the centered version for simplicity.

Denote $\mathbf{Y} = \left(\mathbf{y}_1, \mathbf{y}_2 \cdots \mathbf{y}_n\right)$, $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2 \cdots \mathbf{f}_n)$ and $\mathbf{E} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \cdots \boldsymbol{\epsilon}_n)$. The model (4.3) can be expressed as

$$\mathbf{Y} = \mathbf{Q}\mathbf{F} + \mathbf{E}. \tag{4.4}$$

$\mathbf{Q}$ cannot be uniquely determined by the model (4.4). The reason is that for any estimator $\left(\hat{\mathbf{Q}}, \hat{\mathbf{F}}\right)$, $\left(\hat{\mathbf{Q}}\mathbf{T}, \mathbf{T}^{-1}\hat{\mathbf{F}}\right)$ is also the suitable estimator for the model (4.4) if $\mathbf{T}$ is an invertible matrix, but the linear space of columns of $\mathbf{Q}$ can be uniquely determined by (4.4). By adapting the suitable transformation, we can assume the normalization of $\mathbf{Q}$ satisfying $\frac{1}{p}\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$.

4.3 Estimation procedure

Suppose that we already knew the number of factors in the factor model. Some existing works(Bai and Ng 2002) filled the gap between the whether we have the information about the number of factors. The columns of the estimated factor loading $\hat{\mathbf{Q}}$ are $\sqrt{p}$ times the

eigenvectors corresponding to the first K eigenvalue of $\frac{\mathbf{YY}^T}{np}$, so we have

$$\frac{\mathbf{YY}^T}{np}\hat{\mathbf{Q}} = \hat{\mathbf{Q}}\Lambda, \tag{4.5}$$

where $\Lambda = diag(\lambda_1, \lambda_2 \cdots \lambda_K)$ with $\lambda_1, \cdots \lambda_k$ are the first K eigenvalues of $\frac{1}{np}\mathbf{YY}^T$. From 4.5, we have

$$\begin{aligned}\hat{\mathbf{Q}} &= \frac{\mathbf{YY}^T}{np}\hat{\mathbf{Q}}\Lambda^{-1} \\ &= \frac{\mathbf{QFF}^T\mathbf{Q}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np} + \frac{\mathbf{QFE}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np} + \frac{\mathbf{EF}^T\mathbf{Q}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np} + \frac{\mathbf{EE}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np}.\end{aligned}$$

If we denote $\mathbf{R} = \frac{\mathbf{FF}^T\mathbf{Q}^T\hat{\mathbf{Q}}\Lambda}{np}$, we have

$$\hat{\mathbf{Q}} - \mathbf{QR} = \frac{\mathbf{QFE}^T\hat{\mathbf{Q}}\Lambda^{-1} + \mathbf{EF}^T\mathbf{Q}^T\hat{\mathbf{Q}}\Lambda^{-1} + \mathbf{EE}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np} \tag{4.6}$$

After estimating the factor loading $\mathbf{Q}$, we further estimate the common factors and the error terms.

$$\begin{aligned}\hat{\mathbf{F}} &= \left(\hat{\mathbf{Q}}^T\hat{\mathbf{Q}}\right)^{-1}\hat{\mathbf{Q}}^T\mathbf{Y} = \frac{1}{p}\hat{\mathbf{Q}}^T\mathbf{Y}. \\ \hat{\mathbf{E}} &= \mathbf{Y} - \hat{\mathbf{Q}}\hat{\mathbf{F}}.\end{aligned}$$

Now we have

$$\begin{aligned}
\mathbf{R}\hat{\mathbf{F}} &= \frac{\mathbf{R}\hat{\mathbf{Q}}^T\mathbf{Y}}{p} \\
&= \frac{\mathbf{R}\hat{\mathbf{Q}}^T(\mathbf{Q}\mathbf{F}+\mathbf{E})}{p} \\
&= \frac{\mathbf{R}\hat{\mathbf{Q}}^T\hat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{F}+\mathbf{R}\hat{\mathbf{Q}}^T\left(\mathbf{Q}\mathbf{R}-\hat{\mathbf{Q}}\right)\mathbf{R}^{-1}\mathbf{F}+\mathbf{R}\hat{\mathbf{Q}}^T\mathbf{E}}{p} \\
&= \mathbf{F}+\frac{\mathbf{R}\hat{\mathbf{Q}}^T\left(\mathbf{Q}\mathbf{R}-\hat{\mathbf{Q}}\right)\mathbf{R}^{-1}\mathbf{F}+\mathbf{R}\hat{\mathbf{Q}}^T\mathbf{E}}{p}.
\end{aligned}$$

Under the model (4.4), the covariance matrix of $\mathbf{y}_t$ can be expressed as

$$\Sigma_{\mathbf{y}} = \mathbf{Q}\Sigma_{\mathbf{f}}\mathbf{Q}^T+\Sigma_0.$$

The corresponding estimation for the covariance matrix Σ is

$$\hat{\Sigma}_{\mathbf{y}} = \hat{\mathbf{Q}}\hat{\Sigma}_{\mathbf{f}}\hat{\mathbf{Q}}^T+\hat{\Sigma}_0, \tag{4.7}$$

where $\hat{\Sigma}_{\mathbf{f}}=\frac{\hat{\mathbf{F}}\hat{\mathbf{F}}^T}{n}$ and $\hat{\Sigma}_0=\mathbf{I}_p\circ\frac{\mathbf{E}\mathbf{E}^T}{n}$. Here $\mathbf{A}\circ\mathbf{B}$ is the Hadamand product of the same size matrices $\mathbf{A}$ and $\mathbf{B}$.

4.4 Estimation consistency

The optimal risk R and the global optimal risk R_g consist of quadratic term of $\Sigma_{\mathbf{y}}^{-1}$ and $\hat{\Sigma}_{\mathbf{y}}^{-1}$. We are interested in the quantity $\|\Sigma_{\mathbf{y}}^{-1}-\hat{\Sigma}_{\mathbf{y}}^{-1}\|$ where $\|\mathbf{A}\|$ represents the spectral norm of $\mathbf{A}$, that is $\|\mathbf{A}\|=\lambda_{max}\left(\mathbf{A}\mathbf{A}^T\right)^{\frac{1}{2}}$.

Before going to the theorem, we first introduce some assumptions

1. Assumption 1: K is fixed.

2. Assumption 2: The factor loading matrix $\mathbf{Q} \in \mathbb{R}^{p\times K}$ is a deterministic matrix with $\max_{ij}\left|\mathbf{Q}_{ij}\right| \leq c_1$.

3. Assumption 3: The factor processes $\{\mathbf{f}_t\}$ are i.i.d with the covariance matrix Σ_f. The eigenvalues of Σ_f are bounded that $d_1 \leq \lambda_{min}\left(\Sigma_f\right) \leq \lambda_{max}\left(\Sigma_f\right) \leq d_2$.

4. Assumption 4: The factor processes $\{\mathbf{f}_t\}$ are independent with the error term $\{\boldsymbol{\epsilon}_t\}$. Moreover, the eigenvalues of the covariance matrix of $\boldsymbol{\epsilon}_t$ is bounded.

Then we introduce the theorem for the consistency of the estimation of the inverse of the covariance matrix

Theorem 4.4.1. *According to the Assumptions 1 - 4, the convergence rate for the estimated inverse of the covariance matrix $\hat{\Sigma}_{\mathbf{y}}$ in 4.7 under the factor model structure settings 4.3 is expressed as the following*

$$\|\hat{\Sigma}_{\mathbf{y}}^{-1} - \Sigma_{\mathbf{y}}^{-1}\| \quad = \quad \mathcal{O}_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}}\right). \tag{4.8}$$

Moreover, in the high dimensional setting that the dimension p is in the same order of n, $p = \mathcal{O}(n)$, we have

$$\|\hat{\Sigma}_{\mathbf{y}}^{-1} - \Sigma_{\mathbf{y}}^{-1}\| \quad = \quad \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right). \tag{4.9}$$

Proof. According to the Sherman-Morrison-Woodbury formula, we have

$$\hat{\Sigma}_{\mathbf{y}}^{-1} \quad = \quad \hat{\Sigma}_0^{-1} - \hat{\Sigma}_0^{-1}\hat{\mathbf{Q}}\left[\hat{\Sigma}_{\mathbf{f}}^{-1} + \hat{\mathbf{Q}}^T\hat{\Sigma}_0^{-1}\hat{\mathbf{Q}}\right]^{-1}\hat{\mathbf{Q}}^T\hat{\Sigma}_0^{-1}. \tag{4.10}$$

Then we have

$$\begin{aligned}
\|\hat{\Sigma}_{\mathbf{y}}^{-1}-\Sigma_{\mathbf{y}}^{-1}\| &\leq \|\hat{\Sigma}_0^{-1}-\Sigma_0^{-1}\|+\|\left(\hat{\Sigma}_0^{-1}-\Sigma_0^{-1}\right)\hat{\mathbf{Q}}\left[\hat{\Sigma}_{\mathbf{f}}^{-1}+\hat{\mathbf{Q}}'\hat{\Sigma}_0^{-1}\hat{\mathbf{Q}}\right]^{-1}\hat{\mathbf{Q}}'\hat{\Sigma}_0^{-1}\| \\
&+ \|\Sigma_0^{-1}\hat{\mathbf{Q}}\left[\hat{\Sigma}_{\mathbf{f}}^{-1}+\hat{\mathbf{Q}}'\hat{\Sigma}_0^{-1}\hat{\mathbf{Q}}\right]^{-1}\hat{\mathbf{Q}}'\left(\hat{\Sigma}_0^{-1}-\Sigma_0^{-1}\right)\| \\
&+ \|\Sigma_0^{-1}\left(\hat{\mathbf{Q}}-\mathbf{Q}\mathbf{R}\right)\left[\hat{\Sigma}_{\mathbf{f}}^{-1}+\hat{\mathbf{Q}}'\hat{\Sigma}_0^{-1}\hat{\mathbf{Q}}\right]^{-1}\hat{\mathbf{Q}}'\Sigma_0^{-1}\| \\
&+ \|\Sigma_0^{-1}\mathbf{Q}\mathbf{R}\left[\hat{\Sigma}_{\mathbf{f}}^{-1}+\hat{\mathbf{Q}}'\hat{\Sigma}_0^{-1}\hat{\mathbf{Q}}\right]^{-1}\left(\hat{\mathbf{Q}}-\mathbf{Q}\mathbf{R}\right)'\Sigma_0^{-1}\| \\
&+ \|\Sigma_0^{-1}\mathbf{Q}\gamma\left[\left(\hat{\Sigma}_{\mathbf{f}}^{-1}+\hat{\mathbf{Q}}'\hat{\Sigma}_0^{-1}\hat{\mathbf{Q}}\right)^{-1}-\mathbf{R}^{-1}\left(\Sigma_{\mathbf{f}}^{-1}+\mathbf{Q}'\Sigma_0^{-1}\mathbf{Q}\right)^{-1}\left(\mathbf{R}'\right)^{-1}\right]\left(\mathbf{R}'\right)\mathbf{Q}'\Sigma_0^{-1}\|, \\
&= \mathscr{K}_1+\mathscr{K}_2+\mathscr{K}_3+\mathscr{K}_4+\mathscr{K}_5+\mathscr{K}_6.
\end{aligned}$$

First we look at $\mathscr{K}_1$, we have

$$\begin{aligned}
\|\hat{\Sigma}_0-\Sigma_0\| &= \max_{i\leq p}\left|\left(\hat{\Sigma}_0\right)_{i,i}-(\Sigma)_{i,i}\right| \\
&= \max_{i\leq p}|\hat{\sigma}_i-\sigma_i|.
\end{aligned}$$

Moreover

$$\hat{\sigma}_i-\sigma_i = \frac{1}{n}\sum_{j=1}^{n}\hat{\mathbf{E}}_{ij}^2-\mathbf{E}_{ij}^2=\frac{1}{n}\sum_{j=1}^{n}\left(\left(\mathbf{E}_{ij}-\hat{\mathbf{E}}_{ij}\right)^2+2\mathbf{E}_{ij}\left(\hat{\mathbf{E}}_{ij}-\mathbf{E}_{ij}\right)\right).$$

For $\frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{E}_{ij}-\hat{\mathbf{E}}_{ij}\right)^2$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{E}_{ij}-\hat{\mathbf{E}}_{ij}\right)^2 &= \frac{e_i^\top\left(\mathbf{E}-\hat{\mathbf{E}}\right)\left(\mathbf{E}-\hat{\mathbf{E}}\right)^\top e_i}{n} \\
&= \frac{e_i^\top\left(\hat{\mathbf{Q}}\hat{\mathbf{F}}-\mathbf{Q}\mathbf{F}\right)\left(\hat{\mathbf{Q}}\hat{\mathbf{F}}-\mathbf{Q}\mathbf{F}\right)^\top e_i}{n} \\
&= \frac{e_i^\top\left(\hat{\mathbf{Q}}-\mathbf{Q}\mathbf{R}\right)\hat{\mathbf{F}}\hat{\mathbf{F}}^\top\left(\hat{\mathbf{Q}}-\mathbf{Q}\mathbf{R}\right)^\top e_i}{n} + \frac{e_i^\top\mathbf{Q}\left(\mathbf{R}\hat{\mathbf{F}}-\mathbf{F}\right)\left(\mathbf{R}\hat{\mathbf{F}}-\mathbf{F}\right)^\top\mathbf{Q}^\top e_i}{n} \\
&= \frac{e_i^\top\left(\hat{\mathbf{Q}}-\mathbf{Q}\mathbf{R}\right)\hat{\mathbf{F}}\hat{\mathbf{F}}^\top\left(\hat{\mathbf{Q}}-\mathbf{Q}\mathbf{R}\right)^\top e_i}{n} + \frac{e_i^\top\mathbf{Q}\mathbf{R}\hat{\mathbf{Q}}^\top\mathbf{E}\mathbf{E}^\top\hat{\mathbf{Q}}\mathbf{R}^\top\mathbf{Q}^\top e_i}{np^2} \\
&+ \frac{e_i^\top\mathbf{Q}\mathbf{R}\hat{\mathbf{Q}}\left(\mathbf{Q}\mathbf{R}-\hat{\mathbf{Q}}\right)\mathbf{R}^{-1}\mathbf{F}\mathbf{F}^\top\left(\mathbf{R}^{-1}\right)^\top\left(\mathbf{Q}\mathbf{R}-\hat{\mathbf{Q}}\right)^\top\hat{\mathbf{Q}}^\top\mathbf{R}^\top\mathbf{Q}^\top e_i}{np^2} \\
&= L_{1i}+L_{2i}+L_{3i}.
\end{aligned}$$

For L_{1i}, we have $\|L_{1i}\| = \frac{1}{n}\|e_i^T\left(\hat{\mathbf{Q}}-\mathbf{Q}\gamma\right)\hat{F}\|^2$. Since we have

$$\begin{aligned}
\|\frac{1}{n}\hat{F}\hat{F}^T\| &\leq \frac{1}{p}\|\hat{\mathbf{Q}}\hat{\mathbf{Q}}^T\|\|\frac{\mathbf{Y}\mathbf{Y}^T}{np}\| \\
&= \mathcal{O}_p(1).
\end{aligned}$$

So

$$\begin{aligned}
\max_i\|L_{1i}\| &= \max_i\|e_i^T\frac{\mathbf{Q}\mathbf{F}\mathbf{E}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np}+e_i^T\frac{\mathbf{E}\mathbf{F}^T\mathbf{Q}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np}+e_i\frac{\mathbf{E}\mathbf{E}^T\hat{\mathbf{Q}}^{-1}}{np}\|^2 \\
&\leq 3\max_i\left(\|e_i^T\frac{\mathbf{Q}\mathbf{F}\mathbf{E}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np}\|^2+\|e_i^T\frac{\mathbf{E}\mathbf{F}^T\mathbf{Q}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np}\|^2+\|e_i^T\frac{\mathbf{E}\mathbf{E}^T\hat{\mathbf{Q}}\Lambda^{-1}}{np}\|^2\right) \\
&= \mathcal{O}_p\left(\max\left(\frac{1}{n},\frac{1}{p}\right)\right)+\mathcal{O}_p\left(\frac{1}{n}\right)+\mathcal{O}_p\left(\frac{1}{p}\right) \\
&= \mathcal{O}_p\left(\frac{1}{p}+\frac{1}{n}\right).
\end{aligned}$$

Similar we have

$$\max_i \|L_{2i}\| = \mathcal{O}_p\left(\max\left(\frac{1}{n},\frac{1}{p}\right)\right).$$

and

$$\max_i \|L_{3i}\| = \mathcal{O}_p\left(\max\left(\frac{1}{n},\frac{1}{p}\right)\right).$$

From all things above, we have

$$\max_i \frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{E}_{ij}-\hat{\mathbf{E}}_{ij}\right)^2 = \mathcal{O}_p\left(\frac{1}{p}+\frac{1}{n}\right). \tag{4.11}$$

Then

$$\begin{aligned}\|\hat{\Sigma}_0-\Sigma_0\| &= \max_{i\le p}|\hat{\sigma}_i-\sigma_i| \\ &= \mathcal{O}_p\left(\sqrt{\frac{1}{p}+\frac{1}{n}}\right)\end{aligned}$$

Since the eigenvalues of $\hat{\Sigma}_0$ and Σ are bounded, we have

$$\|\hat{\Sigma}_0^{-1}-\Sigma_0^{-1}\| = \mathcal{O}_p\left(\sqrt{\frac{1}{p}+\frac{1}{n}}\right).$$

Now we look at $\mathscr{K}_2$ and $\mathscr{K}_3$. By using the fact that

$$\begin{aligned}\|\hat{\Sigma}_{\mathbf{f}}^{-1}+\hat{\mathbf{Q}}\hat{\Sigma}_0^{-1}\hat{\mathbf{Q}}^T\| &\ge \lambda_{min}\left(\hat{\mathbf{Q}}\hat{\mathbf{Q}}^T\right) \\ &= \mathcal{O}_p(p),\end{aligned}$$

we have $\|\mathscr{K}_2\|$ and $\|\mathscr{K}_3\|$ are both $\mathscr{O}_p\left(\sqrt{\frac{1}{p}+\frac{1}{n}}\right)$. It is trival that

$$\|\mathscr{K}_4\|, \|\mathscr{K}_5\| \quad = \quad \mathscr{O}_p\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{p}}\right).$$

Lastly $\|\mathscr{K}_6\| = \mathscr{O}_p\left(\sqrt{\frac{1}{p}+\frac{1}{n}}\right)$.

In total, we have

$$\|\hat{\Sigma}_{\mathbf{y}} - \Sigma_{\mathbf{y}}\| \quad = \quad \mathscr{O}_p\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{p}}\right).$$

□

4.5 Numerical results

4.5.1 basic settings

The true model is a three factor model,

$$\mathbf{y}_t \quad = \quad \mathbf{q}_1 f_{1t} + \mathbf{q}_2 f_{2t} + \mathbf{q}_3 f_{3t} + \mathbf{e}_t,$$

where $\mathbf{Q} = \left(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3\right)$ is an orthogonal matrix, $\mathbf{f}_t = \left(f_1, f_2, f_3\right)'$ are independent Gaussian distribution with the mean vector

$$\boldsymbol{\mu}_f \quad = \quad \begin{bmatrix} 0.118 \\ 0.065 \\ 0.104 \end{bmatrix},$$

and the covariance matrix

$$Cov(\mathbf{f}) = \begin{bmatrix} 31.275 & -0.875 & -5.1 \\ -0.875 & 7.875 & -0.05 \\ -5.1 & -0.05 & 4.825 \end{bmatrix}.$$

and

$$\mathbf{e}_t \overset{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I}_{20}).$$

We consider the convergence properties of Σ_y, Σ_y^{-1}, R and $\mathbf{w}_n$. For the matrices Σ_y and Σ_y^{-1}, we use the mean square error to represent the estimation error. For two matrices $\mathbf{A}$ and $\mathbf{B} \in \mathbb{R}^{p\times p}$ with the same dimension, the mean square error of $\mathbf{A}$ and $\mathbf{B}$, $MSE(\mathbf{A},\mathbf{B})$ is defined as

$$MSE(\mathbf{A},\mathbf{B}) = \frac{1}{p^2}\mathrm{Tr}\Big((\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^{\intercal}\Big).$$

For the vector optimal weight $\mathbf{w}_n$ and the scalar optimal risk R, we use the mean absolute error to measure the convergence performance.

In the simulation, we fix the dimension of $\mathbf{y}_t$ as 20. The sample size varies in the following set $[20, 50, 100, 200, 500]$. For each setting, we repeat the simulation 500 times to record the mean value and the variance of the metric. The results are shown in below tables. Table 4.1 shows the convergence performance of estimating the covariance matrix Σ_y. The mean absolute value of the elements of the true covariance matrix Σ_y is 1.318. Table 4.2 shows the result for the precision matrix $\overset{\circ}{}{}_y^{-1}$.The mean absolute value of the elements of the true precision matrix Σ_y^{-1} is 0.094. Table 4.3 shows the convergence performance of the optimal risk R. The true optimal risk is 0.065. The convergence result of the optimal weight is in

Table 4.1: The convergence performance of Σ_y

Sample size(n)	20	50	100	200	500
Mean(MSE)	0.875	0.433	0.294	0.230	0.187
Var(MSE)	0.334	0.052	0.015	0.006	0.003

Table 4.2: The convergence performance of Σ_y^{-1}

Sample size(n)	20	50	100	200	500
Mean(MSE)	2.210e-2	6.277e-3	3.768e-3	2.885e-3	2.454e-3
Var(MSE)	1.131e-4	3.600e-6	7.063e-7	3.216e-7	2.707e-7

Table 4.4. The maximum and the minimum element in the optimal weight $\mathbf{w}_n$ are 0.107 and 0.005 respectively.

4.6 Confidence interval of the optimal risk

In this section, we provide a resampling based method to construct the confidence interval for the optimal risk of the Markowitz Problem.

4.6.1 numerical result

In this simulation, we apply the same factor structure as described in 4.5. The only difference is that we let the dimension of the returns $\mathbf{y}_t$ and the number of the observations n both go to ∞. To be more specifically, we set the relationship $p = \frac{n}{10}$ and let n varies in

Table 4.3: The convergence performance of risk R

Sample size(n)	20	50	100	200	500
Mean(MAE)	0.017	0.009	0.007	0.006	0.005
Var(MAE)	5.182e-5	2.771e-5	1.869e-5	1.634e-5	1.298e-5

Table 4.4: The convergence performance of the optimal weight $\mathbf{w}_n$

Sample size(n)	20	50	100	200	500
Mean(MAE)	0.020	0.015	0.012	0.011	0.010
Var(MAE)	1.723e-5	8.479e-6	7.117e-6	6.196e-6	6.577e-6

Algorithm 1 Resampling method for the confidence interval of the optimal risk

1: Based on $\mathbf{Y}=\left(\mathbf{y}_1,\cdots,\mathbf{y}_n\right)$ and predetermined number of factor K, estimate the $\hat{\boldsymbol{\mu}}$, $\hat{\mathbf{Q}}$ and $\hat{\mathbf{f}}_t$.
2: Compute the residuals $\hat{\epsilon}_t$ from $\hat{\epsilon}_t=\mathbf{y}_t-\hat{\boldsymbol{\mu}}-\hat{\mathbf{Q}}\hat{\mathbf{f}}_t$.
3: For each element $\hat{\epsilon}_{t,i}$ of $\hat{\epsilon}_t$, we generate a standard normal random number v and let $\tilde{\epsilon}_{t,i}=v*\hat{\epsilon}_{t,i}$.
4: For $l=1,2\cdots B$, we compute the optimal risk $\hat{R}^{(l)}$ based on $\{y_t^\star\}$, where $y_t^\star=\hat{\boldsymbol{\mu}}+\hat{\mathbf{Q}}\hat{\mathbf{f}}_t+\mathbf{e}_t^\star$ and $\{\mathbf{e}_t^\star\}$ is resampled from the centered residuals $\{\tilde{\epsilon}_t\}$.
5: Construct the $1-\alpha$ confidence interval by $\left[\hat{R}^{(l)}\left(\frac{\alpha}{2}\right),\hat{R}^{(l)}\left(\frac{1-\alpha}{2}\right)\right]$

$[400,800,1600,3200]$. For each time we compute the 90% confidence interval of the optimal risk, we choose the Bootstrap replication $B=100$ to get the corresponding sample quantile. For each given combination of $\left(n,p\right)$ we repeat the simulation for 100 times to record the coverage probability and the average length of the confidence interval. Fig 4.1 and Fig 4.2 show the coverage probability and the average length of the Bootstrap confidence interval.

We can see that as we increase the sample size to 1600, the coverage probability is approximate to the expected 90%. The length of the confidence interval is decreasing as the sample size increases. This means that the confidence interval is precision in the high dimensional case.

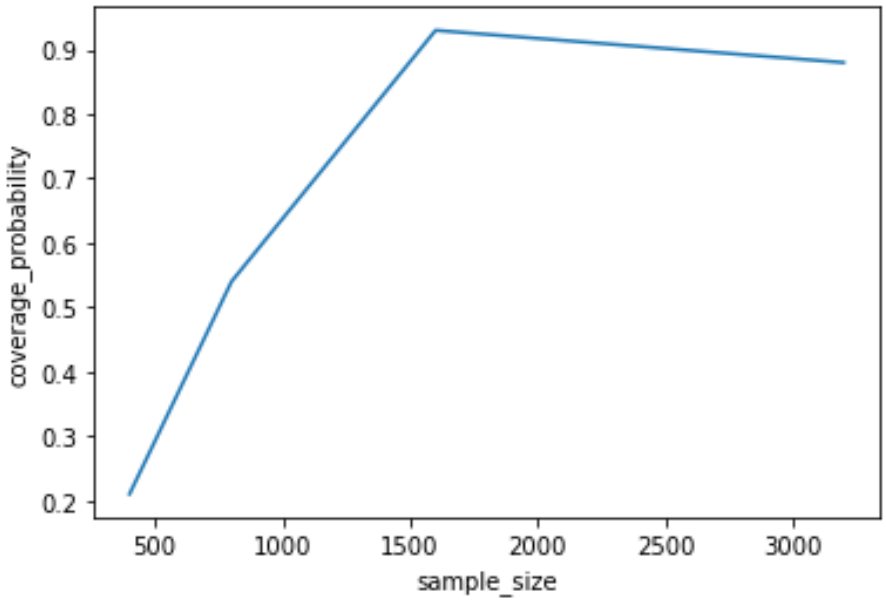


Figure 4.1: The coverage probability of the 90% confidence interval for the optimal risk

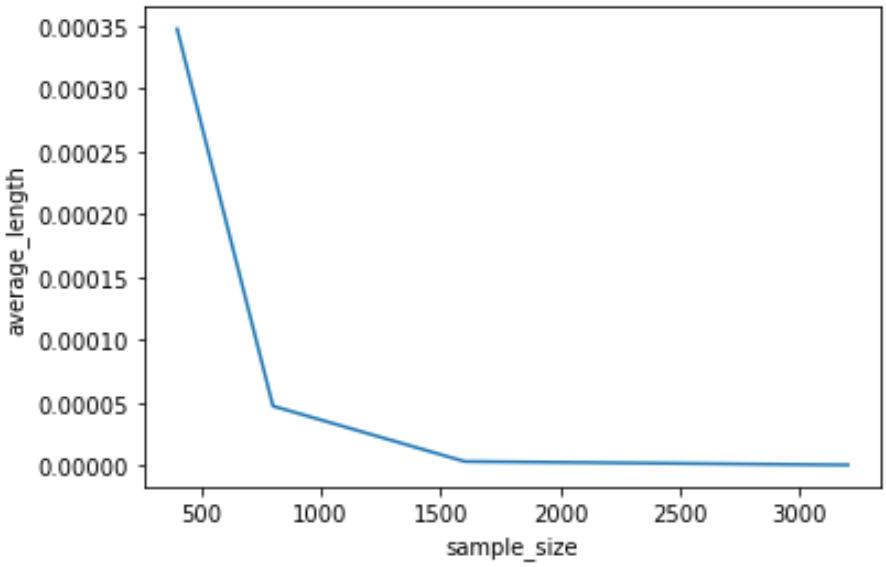


Figure 4.2: The average length of the 90% confidence interval for the optimal risk

REFERENCES

Athreya, K. and Lahiri, S. (2006). *Measure Theory and Probability Theory*. Springer.

Bai, J. (2003). Inferential theory for factor model of large dimensions. *Econometrica*, 71:135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate fator models. *Econometrica*, 70:191–221.

Brockwell, P. J. and Davis, R. A. (1990). *Time Series: Theory and Methods(Second Edition)*. Springer.

Chatterjee, A. and Lahiri, S. (2010). Asymptotic Properties of the Residual Bootstrap for Lasso Estimators. *Preedings of the American Mathemati- cal Society*, 138:608–611.

Chatterjee, A. and Lahiri, S. (2011). Bootstrapping Lasso Estimators. *Journal of the American Statistical Association*, 106(494):608–625.

Doukhan, P. (1994). *Mixing: Properties and Examples*, volume 85. Springer-Verlag, New York.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics*, 147:186–197.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96:1348–1360.

Höglund, T. (1978). Sampling from a Finite Population. A Remainder Term Estimate. *Scandinavian Journal of Statistics*, 5(1):69–71.

Jobson, J. and Korkie, B. (1980). Estimation for Markowitz efficient profolios. *Journal of the American Statistical Association*, 75:544–554.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-Type Estimators. *The Annals of Statistics*, 28:1356–1378.

Kunsch, H. R. (1989). The Jackknife and the Bootstrap for General Stationary Observation. *The Annals of Statistics*, 17.

Markowitz, H. (1952). Portfolio selection. *J.Finance*, 7:77–91.

Miller, R. G. (1974). The jackknife- a review. *Biometrika*, 61.

Politis, D. N. and Romano, J. P. (1989). The Stationary Bootstrap. *Journal of the American Statistical Association*, 89(428).

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*, 58:267–288.

Tukey, J. (1958). Bias and confidence is not quite large samples(abstract). *Ann. Math. Statist*, 29(614).

Zhao, p. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research(JMLR)*, 101:1418–1429.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429.