Boston University

OpenBU **http://open.bu.edu**

Theses & Dissertations Boston University Theses & Dissertations

2021

On the use of inconsistent normalizers for statistical inference on dependent data

https://hdl.handle.net/2144/43948

Boston University

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

ON THE USE OF INCONSISTENT NORMALIZERS FOR STATISTICAL INFERENCE ON DEPENDENT DATA

by

QIAO PAN

B.S., Nanyang Technological University, 2016

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2021

© 2021 by
QIAO PAN
All rights reserved

Approved by

First Reader

Ting Zhang, PhD
Assistant Professor of Mathematics and Statistics

Second Reader

Ashis Gangopadhyay, PhD
Associate Professor of Mathematics and Statistics

Third Reader

Mark Kon, PhD
Professor of Mathematics and Statistics

Acknowledgments

Throughout my PhD studies at Boston University, I have received a great deal of support and assistance.

First, I would like to thank my advisor, Prof. Ting Zhang, whose expertise was invaluable in formulating the research questions and methodology. His insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I would like to extend my sincere thanks to my collaborators, Liliya Lavitas and Hangcheng Lai, for their help and support in my research. I would also like to show gratitude to my committee, including Prof. Ashis Gangopadhyay, Prof. Mark Kon and Prof. Daniel Sussman. I am grateful for their very valuable comments on this thesis. I'd further like to thank Prof. Yves Atchade, Prof. Solesne Bourguin, Prof. Luis Carvalho, Prof. Hyonho Chun, Prof. Uri Eden, Prof. Ashis Gangopadhyay, Prof. Mark Kon, Prof. Daniel Sussman and Prof. Ting Zhang for their valuable lectures equipped me with critical statistical thinking.

I would also like to thank my colleagues and friends at Boston University. Their kind help and support that have made my study and life at Boston a wonderful time. Finally, I would like to express my gratitude to my parents. Without their understanding and encouragement in the past few years, it would be impossible for me to complete my study.

ON THE USE OF INCONSISTENT NORMALIZERS FOR STATISTICAL INFERENCE ON DEPENDENT DATA

QIAO PAN

Boston University, Graduate School of Arts and Sciences, 2021

Major Professor: Ting Zhang, PhD
Assistant Professor of Mathematics and Statistics

ABSTRACT

Statistical inference, such as confidence interval construction, change point detection and nonparametric regression estimation, has been widely explored in many fields including climate science, economics, finance, industrial engineering and many others. The inference has been well developed in the literature under independent settings, while dependent data, especially time series data, is not uncommon to be observed in these areas. Self-normalization is then proposed to analyze statistical inference for time series data. This thesis first explores asymptotic behavior of optimal weighting in generalized self-normalization, then proposes self-normalized simultaneous confidence regions for high-dimensional time series, and lastly explores unsupervised self-normalized break test for correlation matrix.

The basic idea of self-normalization is that it uses an inconsistent variance estimator as studentizer. The original self-normalizer only considered forward estimators and recently it is generalized to involve both forward and backward estimators with deterministic weights. In the first project, we propose a data-driven weight that corresponds to confidence intervals with minimal lengths and study the asymptotic behavior of such a data-driven weight choice. An interesting dichotomy is found

between linear and nonlinear quantities.

In the second project, we would like to overcome the dimension limitation of self-normalization and propose a different perspective to make statistical inference of general quantities of high-dimensional time series. Taking the advantage of data with sparse signals, we develop an asymptotic theory on the maximal modulus of self-normalized statistics. We further establish a thresholded self-normalization method to produce simultaneous confidence regions. The method is able to detect uncommon signals among NASDAQ100 in 2016-2019 in terms of mean and median log returns.

In the last project, we move on to unsupervised test for correlation matrix breaks. We develop a self-normalized test tailored to detect correlation matrix breaks. This method is unsupervised and directly compares the estimated correlation before and after the hypothesized change point. We apply the test to the stock log returns of 10 companies and volatility indexes of 5 options on individual equities to show its power of detecting correlation matrix breaks.

Contents

List of Tables

List of Figures

List of Abbreviations

$B(\cdot)$	.....	Standard Brownian motion
CDF	.....	Cumulative distribution function
H_0	.....	The null hypothesis
H_a	.....	The alternative hypothesis
$N(\mu, \sigma^2)$	.....	Normal distribution with mean μ and variance σ^2
$\mathbb{R}$	.....	Real numbers
$\mathbb{Z}$	.....	Integers
$\xrightarrow{D}$	.....	Convergence in distribution
$\overset{D}{=}$	.....	Equality in distribution

Chapter 1

Introduction

1.1 Self-Normalization under Independent Settings

Self-normalization has a very long history in analyzing statistical inference that can be traced back to the studies of Student (1908). Student's t-statistic is the first prototypical example of self-normalization, which contains closed form expression for asymptotic variance. Let $(X_i)_{i=1}^n$ be normal random variables sampled from $N(\mu, \sigma^2)$, where (μ, σ^2) are unknown. Denote the sample mean by $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and the sample variance by $S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. To test $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$, the t-statistic is constructed as

$$t_n(\mu_0) = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

Under H_0, $t_n(\mu_0) \overset{D}{=} t_{n-1}$, where t_{n-1} represents t distribution with n-1 degrees of freedom. This Gosset's t-statistic has been further generalized to studentized statistic of general quantities. For any quantity $\theta = \mathrm{T}(\mathrm{F}) \in \mathbb{R}$, where F is the marginal CDF of random sample $(X_i)_{i=1}^n$ and T is a functional. A natural estimator of θ is $\hat{\theta}_n = \mathrm{T}(F_n)$, where F_n is the empirical CDF. Under certain assumptions, we may be able to derive $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \sigma_F^2)$, where σ_F^2 is the asymptotic variance. With this limiting distribution given, we are able to construct confidence intervals for θ based on the studentized quantity

$$G_n(\theta) = \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\hat{\sigma}_F},$$

where $\hat{\sigma}_F^2$ is a consistent estimator of σ_F^2. Therefore, deriving a closed form expression for the asymptotic variance is crucial and makes studentization an important procedure in conducting statistical inference.

1.2 Self-Normalization under Time Series Settings

Similarly, in time series analysis, if we are interested in constructing a confidence interval for the mean, we also need to find a consistent estimator for the asymptotic variance. Suppose we have a strictly stationary time series $(X_t)_{t=1}^n$, and we would like to construct a confidence interval for its mean $\mu = E(X_t)$. Under suitable moment and weak dependence conditions,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, s^2),$$

where $s^2 = \sum_{k=-\infty}^{\infty} \gamma(k)$ with $\gamma(k) = cov(X_0, X_k)$. In spectral analysis $s^2 = 2\pi f(0)$, where f$(\cdot)$ is the spectral density function of X_t. Estimating the density function would be hard and the performance depends on the choice of bandwidth. In the testing context, the optimal bandwidth can be derived but it depends on the derivatives of the spectral density. Estimating the derivatives would be even more difficult than estimating the density function in terms of nonparametric convergence rate, see Sun et al. (2008).

Self-normalization is proposed to avoid choosing the bandwidth parameter. The basic idea of self-normalization in time series analysis is that it uses an inconsistent variance estimator as the studentizer. The resulting test statistic is still (asymptotically) pivotal and its limiting null distribution and critical values can be approximated by Monte Carlo simulations. Therefore, it requires no tuning parameters for some problems or fewer tuning parameters compared to existing inference procedures.

Self-normalization is first proposed by Lobato (2001) helping construct statistical inferences for the mean in time series. Assuming the functional central limit theorem (FCLT) holds, that is,

$$\sqrt{n}(\bar{X}_{1,n} - \mu) \Rightarrow sB(1),$$

where $\Rightarrow$ denotes the weak convergence, $B(\cdot)$ is standard Brownian motion and $s^2 = 2\pi f(0)$. Then by the continuous mapping theorem,

$$U_n = \frac{1}{n^2}\sum_{k=1}^{n}(S_k - k\bar{X}_{1,n})^2 \xrightarrow{D} s^2 \int_0^1 \{B(t) - tB(1)\}^2 dt$$

and

$$\frac{n(\bar{X}_{1,n} - \mu)^2}{U_n} \xrightarrow{D} Q \coloneqq \frac{B(1)^2}{\int_0^1 \{B(t) - tB(1)\}^2 dt}, \tag{1.1}$$

where $S_k = \sum_{i=1}^{k} X_i$. Unlike the t-statistic, U_n is not consistent for s^2 but is proportional to s^2. Since the nuisance parameter s is proportional to both the numerator and the denominator in (1.1), it gets canceled out and the distribution of Q becomes pivotal. Similar to the t distribution, the numerator $B(1)^2$ and the denominator $\int_0^1 B(t) - tB(1)^2 dt$ are independent. Lobato (2001) shows that the distribution of Q has heavier tails than that of χ_1^2 using Monte Carlo simulations.

Shao (2010) extended self-normalization to construct confidence intervals (regions) for a large class of quantities that are encountered in time series analysis. Let $\theta = T(F)$, where T is a functional and F is the marginal CDF of a stationary univariate time series X_t. The estimator of θ is calculated as $\hat{\theta}_{1,n} = T(\hat{F}_{1,n})$. Here we denote $\hat{\theta}_{i,j}$ as an estimator of θ based on the subsample $(X_i, ..., X_j)$, $i \leq j$. Replace S_k/k in the mean self-normalizer by the recursive estimator $T(\hat{F}_{1,k})$, then

$$\sqrt{n}(\hat{\theta}_{1,n} - \theta) \Rightarrow \phi B(1),$$

$$V_n = \frac{1}{n^2}\sum_{t=1}^{n} t^2(\hat{\theta}_{1,t} - \hat{\theta}_{1,n})^2 \xrightarrow{D} \phi^2 \int_0^1 \{B(t) - tB(1)\}^2 dt,$$

and

$$\frac{n(\hat{\theta}_{1,n} - \theta)^2}{V_n} \xrightarrow{D} Q$$

With the developed theory, we are able to construct confidence intervals for θ without user-chosen parameters and the confidence intervals have asymptotically correct coverage. Shao (2010) also describes its connection to the fixed b approach and examine the finite sample performance.

The self-normalizer V_n is not unique and this specific form of V_n was influenced by Lobato (2001), since it corresponds to Lobato's mapping and reduces to Lobato's self-normalizer in the mean case. However, it only uses forward recursive estimator, which can result in an unbalanced use of data. Some modification were introduced by Shao (2015) to overcome this problem. For example, use all recursive subsample estimates in the formation of self-normalizer (W_{2n}) or take the average of the conventional self-normalizers for both the original process and its reversed counterpart (W_{3n}), i.e.,

$$W_{2n} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}(j-i+1)^2(\hat{\theta}_{i,j} - \hat{\theta}_{1,n})^2,$$

and

$$W_{3n} = \frac{1}{n^2}\left\{\sum_{t=1}^{n} t^2(\hat{\theta}_{1,t} - \hat{\theta}_{1,n})^2 + \sum_{t=1}^{n} t^2(\hat{\theta}_{n-t+1,n} - \hat{\theta}_{1,n})^2\right\}/2,$$

Both W_{2n} and W_{3n} are invariant to the direction of application, so we say they are T-symmetric. However, W_{2n} is computationally expensive and does not unify with the conventional approach in the mean case. W_{3n} can be reduced to the conventional self-normalizer in the mean case, while resulting in wider confidence intervals. More explanation and simulation comparison can be found in Shao (2015). To develop an effective self-normalizer, Lavitas and Zhang (2018) further proposed a T-symmetric

generalization to the conventional self-normalizer with theoretical implication of deterministic weight choices, including zero, one-half, and one. Their generalization is T-symmetric and can be reduced to the conventional self-normalizer in the mean case. Moreover, this generalization possibly leads to narrower confidence intervals when compared with the conventional self-normalizer.

1.3 Self-Normalization for Statistical Inference

Besides the intense discussion of self-normalization properties, its applications in constructing statistical inference in various fields have also been widely investigated. For example, Shao and Zhang (2010) proposed a framework for change point detection based on self-normalization. Let $(X_t)_{t=1}^n$ be time series observations and $\theta_t = T(F_t^m)$, where F_t^m is the marginal distribution of $Y_t = (X_t, ..., X_{t+m-1})'$ and T is a functional. The interest is to test for a change point in θ_t, that is,

$$H_0 : \theta_1 = \theta_2 = ... = \theta_n$$

against the alternative hypothesis

$$H_1 : \theta_1 = ... = \theta_k \neq \theta_{k+1} = ... = \theta_n,$$

where the location of the change point $1 \leq k \leq n-1$ is unknown. Shao and Zhang (2010) proposed a self-normalization based test statistics as, for $k = 1, ..., n-1$,

$$G_n \coloneqq \sup_{k=1,...,n-1} T_n(k)' V_n^{-1}(k) T_n(k). \tag{1.2}$$

where

$$T_n(k) = \frac{k}{\sqrt{n}}(\hat{\theta}_{1,k} - \hat{\theta}_{1,n}),$$

and

$$V_n(k) = \frac{1}{n^2}\Big\{ \quad \sum_{t=1}^{k} \quad t^2(\hat{\theta}_{1,t} - \hat{\theta}_{1,k})(\hat{\theta}_{1,t} - \hat{\theta}_{1,k})' \\ + \quad \sum_{k+1}^{n} \quad (n-t+1)^2(\hat{\theta}_{t,n} - \hat{\theta}_{k+1,n})(\hat{\theta}_{t,n} - \hat{\theta}_{k+1,n})'\Big\}.$$

The limiting distribution of G_n and simulated critical values are given in Shao and Zhang (2010). The test doesn't involve any user-chosen parameter or nuisance parameter estimation and is able to test for various quantities with encouraging finite sample performance. However, this test is initially designed against one single change point and it becomes computationally expensive if two or more change points exists. Moreover, the number of change points is needed before conducting the test and the test may suffer severe power loss if the number of change points is misspecified. Zhang and Lavitas (2018) further developed a new unsupervised self-normalized change-point test in terms of requiring no prior knowledge of the number of total change points. In addition, they proposed a contrast-based test to handle quantities other than the mean and the test is more powerful in detecting change points for robust quantities like the median. Self-normalization has also been well explored in many other areas, such as nonparametric time series regression, long memory time series, locally stationary time series, functional time series, and spatial data and spatial-temporal data. Many applications are mentioned in Shao (2015).

Among all these works, self-normalization has presented encouraging finite sample performance. However, most of these works focus on univariate or low-dimensional time series data. Recently, a few works start to investigate applications in high-dimensional settings. Wang and Shao (2020) considered hypothesis testing for high-dimensional time series via self-normalization. They demonstrated that directly applying self-normalization to high-dimensional objects can fail easily due to singulari-

ties or substantial size distortions. To handle the high dimensionality, they considered summarizing the high-dimensional mean by a U-statistic following the work of Chen and Qin (2010), and then applying the conventional self-normalization technique to the one-dimensional U-statistic. Their test mainly targets dense and weak alternatives, whereas the self-normalized test for high dimensional time series for strong and sparse alternatives remains as an interesting topic and needs further investigation.

1.4 Tests for Correlation Matrix Breaks

Recently, there is a growing interest in analyzing correlation matrix, especially for high dimensional data. Correlation coefficient is one of the most often used measure of dependence between random variables. With multivariate data given, changes in correlation matrix reveal the behavior of a system structure and may be more powerful than those of individual variables in describing system characteristics when the variables don't react in an isolated way. Therefore, correlation matrix break testing has a wide application in many fields, such as psychology, economics and EEG analysis, see Mauss et al. (2005); Krishnan et al. (2009); Cabrieto et al. (2017).

Wied et al. (2012) first proposed a CUSUM-type correlation constancy test and then Wied and Galeano (2013) designed a sequential monitoring procedure for testing the constancy of the correlation coefficient of a sequence of random variables. However, the test only considers bivariate correlations, whereas, in many cases, the constancy of the whole correlation matrix is of interest. Wied (2017) extended the methodology to higher dimensions. Let $X_t = (X_{1,t}, X_{2,t}, ..., X_{P,t})_{t=1}^{T}$ be a sequence of P-variate random vectors with correlation matrix $R_t = (\rho_t^{ij})_{1 \leq i,j \leq P}$, where

$$\rho_t^{ij} = \frac{Cov(X_{i,t}, X_{j,t})}{\sqrt{Var(X_{i,t})Var(X_{j,t})}}.$$

We are interested in testing the correlation matrix constancy R_t during the observation period. The null hypothesis is

$$H_0 : \rho_1^{ij} = ... = \rho_T^{ij} \text{ for all i,j,}$$

against the alternative hypothesis

$$H_1 : \rho_1^{ij} = ... = \rho_{t_0}^{ij} \neq \rho_{t_0+1}^{ij} = ... = \rho_T^{ij} \text{ for some i,j,}$$

for some correlation break time $t_0 \in \{1, ..., T-1\}$. The designed test in (Wied, 2017) is

$$DW_T = \sup_{z\in[0,1]} \frac{\tau_T(z)}{\sqrt{T}} \left\| \hat{E}^{-1/2}(vecho(\hat{R}_{1,\tau_T(z)}) - vecho(\hat{R}_{1,T})) \right\|_1 \tag{1.3}$$

where $\tau_T(z) = [2 + z(T-2)], z \in [0,1]$ and
$vecho(\hat{R}_{l,k}) = (\hat{\rho}_{l,k}^{12}, \hat{\rho}_{l,k}^{13}, ..., \hat{\rho}_{l,k}^{1P}, \hat{\rho}_{l,k}^{23}, ..., \hat{\rho}_{l,k}^{2P}, ..., \hat{\rho}_{l,k}^{(P-1)P})'$, $\hat{\rho}_{l,k}^{ij}$ is the sample correlation estimated from the paired sample $\{(X_{it}, X_{jt}), t = l, l+1, ..., k\}$. $\hat{E}$ is a consistent estimator of the $\frac{P(P-1)}{2} \times \frac{P(P-1)}{2}$ long-run covariance matrix E of $\sqrt{T}vecho(\hat{R}_{1,T})$. The test DW_T has the asymptotic null distribution $\sup_{0\leq s\leq 1} \left\| B_0^{\frac{P(P-1)}{2}}(s) \right\|_1$, where $B_0^N(\cdot)$ is an N-dimensional standard Brownian bridge. The asymptotic critical values of the test DW_T depend on P. The test was extended to cases that $P > 2$. However, when P is not small relative to T, DW_T is unstable. For moderate P, DW_T tends to be under-sized and has low power. For large P, it is over-sized due to (near) singularity of variance matrix estimator. The details and simulation results are discussed in Posch et al. (2019); Demetrescu and Wied (2019); Choi and Shin (2020a). To overcome the high dimension limitation, Choi and Shin (2020a) has tried to test correlation breaks based on self-normalization. Compared with the previous work, this test is free of singular problem and has less size distortion in high dimensional settings under multiple dependence structure. However, their test was designed for a single break

alternative and it becomes less powerful when testing against multiple breaks.

1.5 Contributions and Organization of the Thesis

In this chapter, we introduced the concept of self-normalization and provided an overview of its development and application. We can see that even though self-normalization has been investigated under different scenarios, there are still remaining problems. The conventional self-normalizer is not T-symmetric. Limited effort has been made to apply self-normalization to hypothesis testing in high-dimensional settings. A little attention has been paid to test correlation matrix break using self-normalization.

The goal of this thesis is to further explore properties of generalized self-normalization and develop statistical theory and methods to construct inference in multiple high dimensional time series settings. The contribution lies in several aspects as listed below, together with the outline of the thesis. In Chapter 2, we explore asymptotic behavior of optimal weighting in generalized self-normalization for time series. Then we applied self-normalization method to construct simultaneous confidence regions for high-dimensional time series in Chapter 3. Chapter 4 describes an unsupervised break test for correlation matrix in time series based on self-normalization. Lastly, Chapter 5 summarizes the work of this thesis with potential future research directions.

Chapter 2

Asymptotic Behavior of Optimal Weighting in Generalized Self-Normalization for Time Series

2.1 Introduction

In statistics, inference procedures of interested quantities often involve additional nuisance parameters that are not of direct interest but need to be appropriately handled to yield valid results. For example, when one is interested in the mean of independent and identically distributed observations, then the marginal variance appears in the distribution of the sample mean as a nuisance parameter. A common approach used in the literature is to plug in a consistent estimator of the nuisance parameter, such as using the sample variance in the inference of the mean. However, in the time series setting, the nuisance parameter can have a complicated form owing to the dependence, and its estimation itself can be a quite nontrivial problem. For instance, when making inference about the mean in the dependent case, the nuisance parameter that appears in the asymptotic distribution of the sample mean is no longer simply the marginal variance but the sum of autocovariances of all orders, which may require sophisticated procedures for its consistent estimation; see for example Flegal and Jones (2010), Liu and Wu (2010), Politis (2011), Xiao and Wu (2011), Paparoditis and Politis (2012), Zhang (2018) and references therein for recent developments in this direction. When the quantity of interest is beyond the mean, the task of consistently estimating the

associated nuisance parameter can be even more involved. In addition, such plug-in methods have been found to have certain unsatisfactory aspects in their finite-sample performance, mainly due to the difficulty of obtaining efficient and robust nuisance parameter estimates under dependence; see for example the discussion in Kiefer et al. (2000) and Shao (2010).

Self-normalization has emerged as a convenient and flexible alternative for inference of time series, which does not require direct estimation of the nuisance parameter yet still leads to valid statistical inference procedures. The idea was explored by Kiefer et al. (2000) in the linear regression setting with ordinary least squares estimators and by Lobato (2001) for inference of autocorrelations. Shao (2010) extended their results and considered self-normalization techniques for quantities of a more general form, namely quantities that can be expressed as functionals of the cumulative distribution function; see also discussions therein for its connection with the fixed b asymptotic scheme (Kiefer et al., 2000; Kiefer and Vogelsang, 2005). The self-normalization scheme of Shao (2010) uses a sequence of recursive estimators to pivotalize the asymptotic distribution of the statistic of interest, and has been adopted in various statistical inference problems regarding dependent data; see for example Shao and Zhang (2010), Zhou and Shao (2013), Kim et al. (2015b), Bai et al. (2016) and references therein. However, as observed in Shao (2015), this commonly used self-normalization scheme only uses recursive estimators of one direction and as a result may exhibit certain degrees of asymmetry resulting in an unbalanced use of the data. To address this, Shao (2015) considered an alternative that uses recursive estimators from all possible blocks to construct the self-normalizer. However, such an approach requires substantially more computation and its performance in the mean case can be noticeably worse than that of the conventional self-normalizer; see the discussion and simulation results in Shao (2015). Given these concerns, under the

influence of Professor Michael Stein, Shao (2015) considered another approach to involve recursive estimators from both directions, which is to compute the conventional self-normalizer for both the original process and its reversed counterpart and take the average. Although such a self-normalization scheme maintains the same order of computation as that of the conventional self-normalizer, it has been found to have the tendency of yielding wider confidence intervals as observed in the simulation study of Shao (2015).

Recently, Lavitas and Zhang (2018) considered a more sophisticated approach in using recursive estimators from both directions to construct self-normalizers for time series. Their approach exploits the mathematical property of the cumulative sum (CUSUM) process to obtain a decomposition into forward and backward components. Compared with the simple averaging scheme of Shao (2015), the self-normalization scheme of Lavitas and Zhang (2018) is theoretically guaranteed to lead to confidence intervals that are narrower, or at least of the same length, while maintaining the same order of computation. The self-normalization scheme of Zhang and Lavitas (2018) leads to the definition of a generalized class of self-normalizers, which involves a weight choice where the choice of one corresponds to the conventional self-normalizer of Shao (2010). Zhang and Lavitas (2018) studied the theoretical implication of simple deterministic weight choices, including zero, one-half and one. In this chapter, we consider a data-driven weight choice that is optimal in the sense of minimizing the length of the associated confidence interval. We study the asymptotic behavior of such a data-driven weight choice, and find an interesting dichotomy between linear and nonlinear quantities. In particular, for linear quantities such as the mean, we found that the optimal weight obtained by minimizing the length of the associated confidence interval will always have a degenerated distribution, making the generalized self-normalizer equivalent to the conventional self-normalizer of Shao (2010). This means that the

self-normalizer in its original form (Shao, 2010) is already optimal for linear quantities such as the mean. On the other hand, if one is interested in a nonlinear quantity such as the variance, then the optimal weight will generally distribute over an uncountable set, but in the limit it will converge to a discrete distribution. In this case, the conventional self-normalizer of Shao (2010) no longer represents the optimal choice.

This chapter is based on Zhang et al. (2019). In Section 2.2, we introduce the data-driven weight choice that minimizes the length of the associated confidence interval in self-normalized inference procedures. We study its asymptotic behavior for the mean and variance cases in Section 2.3.1. A mathematical framework for handling more general quantities is provided in Section 2.3.2, and the associated asymptotic theory is established. Section 2.4 concludes this chapter. The theoretical proofs, implication and simulation for confidence interval construction can be found in Appendix A.

2.2 Weight Choice in Generalized Self-Normalization

Suppose we observe $X_1, \ldots, X_n$ from a stationary time series with mean μ, then the sample average $\bar{X}_{1,n} = n^{-1}\sum_{i=1}^{n} X_i$ provides a consistent estimator of μ. Under certain regularity conditions (Hannan, 1979; Herrndorf, 1984; Wu, 2005), one can obtain the central limit theorem

$$\sqrt{n}(\bar{X}_{1,n} - \mu) \xrightarrow{D} N(0, g), \tag{2.1}$$

where g is the long-run variance defined as the sum of autocovariances of all orders. Although one can construct confidence intervals for μ based on (2.1), it requires consistent estimation of the long-run variance g which can itself be a nontrivial problem. As an alternative, the self-normalization approach uses a function of recursive estimators to pivotalize the asymptotic distribution to make it free of the nuisance long-run

variance; see also discussions in Kiefer et al. (2000), Lobato (2001) and Shao (2010) for advantages of using self-normalization. Let $S_{1,k} = \sum_{i=1}^{k} X_i$ be the partial sum process, then the self-normalizer introduced in Section 1.2 takes the form

$$U_n = \frac{1}{n^2} \sum_{k=1}^{n} (S_{1,k} - k\bar{X}_{1,n})^2. \tag{2.2}$$

Let $\Rightarrow$ denotes the weak convergence (van der Vaart and Wellner, 1996), the self-normalization approach relies on the following assumption to derive the asymptotic distribution.

(IP) There exists a $s > 0$ such that

$$n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} (X_i - \mu) \Rightarrow sB(t), \quad t \in [0, 1],$$

where $B(\cdot)$ represents a standard Brownian motion.

Assumption (IP) is generally referred to as the invariance principle in the literature, and has been studied under various short-range dependence conditions where it has been found that $s^2 = g$, the long-run variance; see for example Hannan (1979), Herrndorf (1984), Wu (2007), Berkes et al. (2014) and references therein. By assumption (IP), one can show that the self-normalized statistic satisfies

$$\frac{n(\bar{X}_{1,n} - \mu)^2}{U_n} \xrightarrow{D} \frac{s^2 B(1)^2}{\int_0^1 \{sB(t) - tsB(1)\}^2 dt} = \frac{B(1)^2}{\int_0^1 \{B(t) - tB(1)\}^2 dt}, \tag{2.3}$$

where the nuisance parameter s gets cancelled out and does not show up in the asymptotic distribution. Therefore, one can use (2.3) to construct confidence intervals of μ, which is free of the nuisance parameter s due to self-normalization.

In the seminal work of Shao (2010), the above self-normalization scheme was generalized to handle quantities other than the mean. Following their framework, suppose

the quantity of interest θ can be expressed as a functional of the distribution function, namely $\theta = T(F)$, where T is a functional and F is the cumulative distribution function of the random vector Y_i, which can be either the vector $(X_i, \ldots, X_{i+d-1})'$ itself or its transform. Here d is a fixed but arbitrary integer and $'$ denotes the transpose. Let $\hat{F}_{i,j}$ be the empirical distribution function of $Y_i, \ldots, Y_j$, $1 \leq i \leq j \leq N = n - d + 1$, then a natural estimator of $\theta = T(F)$ is given by $\hat{\theta}_{1,N} = T(\hat{F}_{1,N})$. Note that in the self-normalizer (2.2) for the mean case, the partial sum satisfies $S_{1,k} = k\bar{X}_{1,k}$, where $\bar{X}_{1,k}$ can be viewed as a recursive estimator of the mean. Therefore, a natural strategy as proposed by Shao (2010) and advocated in its subsequent works is to replace $\bar{X}_{1,k}$ in (2.2) by the recursive estimator $\hat{\theta}_{1,k} = T(\hat{F}_{1,k})$, $k = 1, \ldots, N$, which gives rise to the self-normalizer

$$V_N = N^{-2} \sum_{k=1}^{N} k^2 \{T(\hat{F}_{1,k}) - T(\hat{F}_{1,N})\}^2. \tag{2.4}$$

Assuming an invariance principle on the process $\{T(\hat{F}_{1,\lfloor Nt \rfloor})\}_{t \in [0,1]}$ similar to that in condition (IP), one can show that

$$\frac{N\{T(\hat{F}_{1,N}) - \theta\}^2}{V_N} \xrightarrow{D} \frac{B(1)^2}{\int_0^1 \{B(t) - tB(1)\}^2 dt},$$

which can then be used to construct confidence intervals for θ. The aforementioned generalization due to Shao (2010) gradually became a standard in self-normalized inference and has been adopted into various settings. However, as noted by Shao (2015), the self-normalizer in (2.4) only uses forward recursive estimators, which can result in an unbalanced use of the data. To address this, Shao (2015) proposed to calculate the self-normalizer (2.4) for the original data $Y_1, \ldots, Y_N$ and its reversed counterpart $Y_N, \ldots, Y_1$ and take the average. This simple averaging scheme of Shao (2015) leads to a self-normalizer that uses both forward and backward recursive es-

timators, but it has been found to have the tendency of producing wider confidence intervals when compared with the conventional self-normalizer of Shao (2010). Recently, Zhang and Lavitas (2018) considered a more sophisticated approach in using forward and backward recursive estimators to construct self-normlizers, which leads to a class of generalized self-normalizers having the form

$$\Lambda_N(w) = N^{-2} \sum_{k=1}^{N} \{wk(\hat{\theta}_{1,k} - \hat{\theta}_{1,N}) - (1-w)(N-k)(\hat{\theta}_{k+1,N} - \hat{\theta}_{1,N})\}^2, \qquad (2.5)$$

where $w \in [0,1]$ is a weight choice, and $\hat{\theta}_{1,k} = T(\hat{F}_{1,k})$ and $\hat{\theta}_{k+1,N} = T(\hat{F}_{k+1,N})$ are the forward and backward recursive estimators respectively. Note that the self-normalizer (2.5) incorporates the forward and backward recursive estimators in a nontrivial manner, and such a mechanism is due to the mathematical property of the CUSUM process; see the discussion in Zhang and Lavitas (2018). The generalized self-normalizer (2.5) includes the conventional self-normalizer (2.4) as a special case by letting $w = 1$. It is also theoretically guaranteed to yield confidence intervals that are narrower, or at least of the same length, when compared with the simple averaging scheme of Shao (2015); see Theorem 1 of Lavitas and Zhang (2018). The aforementioned paper studied the theoretical implication of the generalized self-normalizer (2.5) with a set of deterministic weight choices, such as zero, one-half and one, and we shall here consider a data-driven weight choice that minimizes the length of the associated confidence interval. By Theorem 2 of Lavitas and Zhang (2018), the asymptotic distribution of (2.5) remains the same for any $w \in [0,1]$, and as a result the length of the resulting confidence interval depends on w only through the self-normalizer (2.5). Therefore, by the monotonic relationship, it is equivalent to selecting the weight w by minimizing the self-normalizer (2.5), which motivates us to

consider the data-driven weight choice

$$\hat{w}_N = \underset{w\in[0,1]}{\operatorname{argmin}}\, \Lambda_N(w).$$

If the set of minimizers contains more than one element, then we define $\hat{w}_N$ as the supremum of that set. We study the asymptotic behavior of such a data-driven weight choice in Section 2.3.

2.3 Asymptotic Behavior of Optimal Weighting

2.3.1 Special Cases: Mean and Variance

We shall first consider the case where the quantity of interest is the mean. In this case, we can write

$$\theta = T(F) = \int x dF(x), \tag{2.6}$$

where F is the cumulative distribution function of $Y_i = X_i$ and $N = n$. Therefore, by replacing F with its empirical counterpart we obtain the forward and backward recursive estimators

$$\hat{\theta}_{1,k} = T(\hat{F}_{1,k}) = k^{-1}\sum_{i=1}^{k} X_i, \quad \hat{\theta}_{k+1,N} = T(\hat{F}_{k+1,N}) = (N-k)^{-1}\sum_{i=k+1}^{N} X_i.$$

Note that in this case $k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N}) = -(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})$, we have $\Lambda_N(w) \equiv \Lambda_N(1)$ and thus $\mathrm{pr}(\hat{w}_N = 1) = 1$. This means that, in the mean case, the conventional self-normalizer of Shao (2010) in its original form is already optimal in the sense that it produces the narrowest confidence interval among the class of generalized self-normalizers (Lavitas and Zhang, 2018).

We shall in the following consider the case of the variance, which is a nonlinear

quantity. In this case, we can write

$$\theta = T(F) = \int x^2 dF(x) - \left\{ \int x dF(x) \right\}^2, \tag{2.7}$$

where F is the cumulative distribution function of $Y_i = X_i$ and $N = n$. By replacing F with its empirical counterpart, we obtain the forward and backward recursive estimators

$$\hat{\theta}_{1,k} = T(\hat{F}_{1,k}) = k^{-1} \sum_{i=1}^{k} X_i^2 - \left(k^{-1} \sum_{i=1}^{k} X_i \right)^2 \tag{2.8}$$

and

$$\hat{\theta}_{k+1,N} = T(\hat{F}_{k+1,N}) = (N-k)^{-1} \sum_{i=k+1}^{N} X_i^2 - \left\{ (N-k)^{-1} \sum_{i=k+1}^{N} X_i \right\}^2, \tag{2.9}$$

which relate to the method of moments estimators. Let $S_N^\circ(t) = \{S_{N,1}^\circ(t), S_{N,2}^\circ(t)\}'$ where

$$S_{N,1}^\circ(t) = \sum_{i=1}^{\lfloor Nt \rfloor} (X_i - \mu), \quad S_{N,2}^\circ(t) = \sum_{i=1}^{\lfloor Nt \rfloor} \{X_i^2 - E(X_i^2)\},$$

are the centered partial sum process of X_i and X_i^2 respectively. We shall here make the following assumption.

(IP_v) There exists a positive definite matrix Σ such that

$$N^{-1/2} S_N^\circ(t) \Rightarrow \Sigma B(t), \quad t \in [0,1],$$

where $B(\cdot)$ is a standard two-dimensional Brownian motion.

Assumption (IP_v) requires an invariance principle of the two-dimensional process $(X_i, X_i^2)'$, which can be viewed as a multivariate extension of condition (IP) in Section 2.2 and can be verified under similar short-range dependence conditions; see for example Wu and Zhou (2011) and references therein. The following theorem states that, in case of the variance, the optimal weight choice that corresponds to confidence

intervals with minimal lengths as described in Section 2.2 converges in distribution to a discrete random variable that has a symmetric distribution around one-half.

Theorem 2.3.1. *Assume that* $T(F) = \int x^2 dF(x) - \left\{\int x dF(x)\right\}^2$ *as in the variance case and condition* (IP_v) *holds. Then*

$$\hat{w}_N \xrightarrow{D} Z,$$

where Z *is a Bernoulli random variable whose probability mass function is given by*

$$\mathrm{pr}(Z = 0) = \mathrm{pr}(Z = 1) = 1/2.$$

By Theorem 2.3.1, the data-driven weight choice $\hat{w}_N$ as described in Section 2.2 converges in distribution to a Bernoulli random variable, indicating that the conventional self-normalizer of Shao (2010) in this case no longer represents the optimal choice that minimizes the length of the associated confidence interval in self-normalized inference of time series. This, along with the observation in the mean case, suggests an interesting dichotomy that the asymptotic behavior of the data-driven weight choice $\hat{w}_N$ defined in Section 2.2 can be different depending on the quantity of interest. We shall in the following consider the case with a general functional T and investigate the cause of the dichotomous asymptotic behavior of $\hat{w}_N$.

2.3.2 The General Case: A Framework and Its Asymptotic Theory

In this section, we consider the general case where the functional T is not necessarily of the special forms as considered in Section 2.3.1. For this, we need a framework to study the asymptotic behavior of forward and backward recursive estimators

$$\hat{\theta}_{1,k} = T(\hat{F}_{1,k}), \quad \hat{\theta}_{k+1,N} = T(\hat{F}_{k+1,N}).$$

To achieve this, Shao (2010) in his seminal work proposed to approximate recursive estimators by partial sums of a sequence of influence functions, which was then used

to derive the asymptotic distribution of their self-normalized statistic. However, their approximation scheme cannot be used to fully recover the dichotomous asymptotic behavior as needed here. This is mainly because the approximation scheme of Shao (2010) relies on the partial sum process of influence functions, which has a linear form and is thus always attracted to the first regime of the dichotomy as described in the mean case. To recover the second regime as described in Theorem 2.3.1 and to understand the cause of the dichotomy, we propose to consider the von Mises expansion (von Mises, 1947; Fernholz, 2001) of the general functional T using Gâteaux derivatives, namely

$$T(\hat{F}_{i,j})-T(F) =\int \varphi_1(x)d(\hat{F}_{i,j}-F)(x)+\frac{1}{2}\iint \varphi_2(x,y)d(\hat{F}_{i,j}-F)(x)d(\hat{F}_{i,j}-F)(y)+R_{i,j}, \tag{2.10}$$

where

$$\begin{aligned} \varphi_1(x) &= \left.\frac{\partial T\{(1-\epsilon)F+\epsilon\delta_x\}}{\partial\epsilon}\right|_{\epsilon=0}; \\ \varphi_2(x,y) &= \left.\frac{\partial^2 T\{(1-\epsilon_1-\epsilon_2)F+\epsilon_1\delta_x+\epsilon_2\delta_y\}}{\partial\epsilon_2\partial\epsilon_1}\right|_{\epsilon_1=0,\epsilon_2=0} \end{aligned}$$

are the von Mises kernels with δ_x denoting the Dirac measure at x, and $R_{i,j}$ is the remainder term. Let

$$\psi_1(x) = \varphi_1(x) - \int \varphi_1(x)dF(x)$$

and

$$\psi_2(x) = \varphi_2(x) - \int \varphi_2(x,y)dF(x) - \int \varphi_2(x,y)dF(y) + \iint \varphi_2(x,y)dF(x)dF(y)$$

be the centered von Mises kernels, then by (2.10) we obtain the von Mises expansions

$$\hat{\theta}_{1,k} = \theta + \frac{1}{k}\sum_{i=1}^{k}\psi_1(Y_i) + \frac{1}{2k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\psi_2(Y_i,Y_j) + R_{1,k} \tag{2.11}$$

and

$$\hat{\theta}_{k+1,N} = \theta + \frac{1}{N-k}\sum_{i=k+1}^{N}\psi_1(Y_i) + \frac{1}{2(N-k)^2}\sum_{i=k+1}^{N}\sum_{j=k+1}^{N}\psi_2(Y_i,Y_j) + R_{k+1,N} \quad (2.12)$$

for forward and backward recursive estimators respectively.

Motivated by the special cases considered in Section 2.3.1, we say that the parameter $\theta = T(F)$ is probabilistically linear if the associated functional T is linear. Therefore, the mean parameter (2.6) is probabilistically linear while the variance parameter (2.7) is not. The following theorem states that, for probabilistically linear parameters, the data-driven weight choice $\hat{w}_N$ as described in Section 2.2 has a distribution that puts a unit mass at one thus belongs to the first regime of the dichotomy as described in the mean case.

Theorem 2.3.2. *If $\theta = T(F)$ is probabilistically linear, then (i) the von Mises expansion of recursive estimators admits the form*

$$\hat{\theta}_{1,k} = \theta + \frac{1}{k}\sum_{i=1}^{k}\varphi_1(Y_i), \quad \hat{\theta}_{k+1,N} = \theta + \frac{1}{N-k}\sum_{i=k+1}^{N}\varphi_1(Y_i);$$

and (ii) $\Lambda_N(w) \equiv \Lambda_N(1)$ and thus $\mathrm{pr}(\hat{w}_N = 1) = 1$.

By Theorem 2.3.2, von Mises expansions of recursive estimators of probabilistically linear quantities admit a linear form, which makes the data-driven weight choice $\hat{w}_N$ belong to the first regime of the dichotomy as described in the mean case. We shall in the following consider the case where the functional T is nonlinear. For this, we make the following assumptions.

(IP_g) There exists a nondegenerate stochastic process $W(t) = \{W_1(t), W_2(t), W_3(t)\}'$

such that

$$\left\{ \begin{array}{c} N^{-1/2} \sum_{i=1}^{\lfloor Nt \rfloor} \psi_1(Y_i) \\ N^{-1} \sum_{i=1}^{\lfloor Nt \rfloor} \sum_{j=1}^{\lfloor Nt \rfloor} \psi_2(Y_i, Y_j) \\ N^{-1} \sum_{i=\lfloor Nt \rfloor+1}^{N} \sum_{j=\lfloor Nt \rfloor+1}^{N} \psi_2(Y_i, Y_j) \end{array} \right\} \Rightarrow W(t), \quad t \in [0,1].$$

(N) The remainder terms satisfy

$$\sum_{k=1}^{N} \{k^2 R_{1,k}^2 + (N-k)^2 R_{k+1,N}^2 + N^2 R_{1,N}^2\} = o_p(N).$$

Compared with the commonly adopted invariance principle on the partial sum process as in conditions (IP) and (IP$_v$), Assumption (IP$_g$) in addition requires an invariance principle regarding the von Mises differentiable statistical function, namely the double integral of $\psi_2(x, y)$ with respect to the empirical distribution functions which relates to the U-statistic. Invariance principles of U-statistics and von Mises differentiable statistical functions have been established and studied by Miller and Sen (1972) and Dehling et al. (1984) for independent observations and more recently by Kanagawa and Yoshihara (1994) and Sharipov (2003) for dependent observations; see also references therein. Since $W(t)$ is nondegenerate, it does not overlap with the class of probabilistically linear parameters as considered in Theorem 2.3.2. Assumption (N) basically requires that the remainder term in the von Mises expansion is negligible when compared with the leading terms in the expansion. Note that the data-driven weight choice $\hat{w}_N$ defined in Section 2.2 is restricted to be searched within the unit interval, and we shall here consider in addition its unrestricted version

$$\tilde{w}_N = \operatorname*{argmin}_{w \in \mathbb{R}} \Lambda_N(w).$$

The following theorem provides the connection between $\hat{w}_N$ and $\tilde{w}_N$, and establishes their asymptotic distributions for cases where the functional T is nonlinear.

Theorem 2.3.3. *Assume that $\theta = T(F)$ is probabilistically nonlinear and conditions (IP_g) and (N) hold. If*

$$\limsup_{t \to 0^+} \frac{E\{W_2(t)^2\} + E\{W_3(1-t)^2\}}{t^2} < \infty, \tag{2.13}$$

then (i)

$$\hat{w}_N = \max\{\min(\tilde{w}_N, 1), 0\},$$

namely $\hat{w}_N$ can be obtained from $\tilde{w}_N$ through truncation; (ii)

$$N^{-1/2}\tilde{w}_N \xrightarrow{D} \frac{2\int_0^1 \{tW_1(1) - W_1(t)\}\{t^{-1}W_2(t) + (1-t)^{-1}W_3(t) - W_2(1)\}dt}{\int_0^1 \{t^{-1}W_2(t) + (1-t)^{-1}W_3(t) - W_2(1)\}^2 dt};$$

and (iii)

$$\hat{w}_N \xrightarrow{D} Z^\star, \tag{2.14}$$

where $Z^\star$ is a Bernoulli random variable whose probability mass function is given by

$$\begin{aligned} \mathrm{pr}(Z^\star = 0) &= 1 - \mathrm{pr}(Z^\star = 1) \\ &= \mathrm{pr}\left[\int_0^1 \{tW_1(1) - W_1(t)\}\{t^{-1}W_2(t) + (1-t)^{-1}W_3(t) - W_2(1)\}dt \le 0\right]. \end{aligned}$$

Theorem 2.3.3 states that, for probabilistically nonlinear parameters, the corresponding optimal weight choice $\hat{w}_N$ as described in Section 2.2 converges in distribution to a Bernoulli random variable, suggesting that the conventional self-normalizer of Shao (2010) no longer represents the optimal choice in this case. In addition, it reveals a deeper result, namely the distribution of $\hat{w}_N$ can be viewed as a truncated version of that of an unrestricted optimizer $\tilde{w}_N$ which diverges at a rate of $N^{1/2}$. Therefore, $\hat{w}_N$ typically distributes over the unit interval in finite-sample problems but in the limit it converges to a discrete distribution with a finite support. Inspired by the special cases considered in Section 2.3.1, Theorems 2.3.2 and 2.3.3 confirms

the dichotomous asymptotic behavior of $\hat{w}_N$, and provides a criterion to distinguish the two regimes of the dichotomy, that is, whether the asymptotic distribution of $\hat{w}_N$ will be attracted to the first or second regime of the dichotomy depends on whether the parameter of interest is probabilistically linear or not. We shall in the following consider an application to smooth function models (Bhattacharya and Ghosh, 1978; Hall, 1992; Shao, 2010) to illustrate the implications of the conditions in Theorems 2.3.2 and 2.3.3.

In smooth function models, parameters are modeled as differentiable functions of vector means, namely $\theta = g\{E(Y_i)\}$ where $g: \mathbb{R}^d \to \mathbb{R}$ is a differentiable function and $Y_i \in \mathbb{R}^d$ is a random vector which can be either the vector $(X_i, \ldots, X_{i+d-1})'$ itself or its transform $\{h_1(X_i, \ldots, X_{i+d-1}), \ldots, h_d(X_i, \ldots, X_{i+d-1})\}'$. As commented by Polansky (2008), many standard problems in statistical inference can be studied in terms of a smooth function model; see also the discussion in Hall (1992). Shao (2010) adopted the smooth function model with $Y_i = (X_i, \ldots, X_{i+d-1})'$ to study the practical meaning of their technical conditions for self-normalized inference of time series, and commented that such a class of models is sufficiently wide to include many statistics of practical interest. We shall here consider the situation where Y_i can be of the form $\{h_1(X_i, \ldots, X_{i+d-1}), \ldots, h_d(X_i, \ldots, X_{i+d-1})\}'$. Let $\nu = E(Y_i)$, then the von Mises kernels have the form

$$\varphi_1(x) = \left. \frac{\partial g\left[\int z d\{(1-\epsilon)F(z) + \epsilon\delta_x(z)\}\right]}{\partial\epsilon}\right|_{\epsilon=0} = \nabla g(\nu)'(x-\nu)$$

and

$$\begin{aligned}\varphi_2(x,y) &= \left. \frac{\partial^2 g\left[\int z d\{(1-\epsilon_1-\epsilon_2)F(z) + \epsilon_1\delta_x(z) + \epsilon_2\delta_y(z)\}\right]}{\partial\epsilon_2\partial\epsilon_1}\right|_{\epsilon_1=0,\epsilon_2=0} \\ &= (y-\nu)'H_g(\nu)(x-\nu),\end{aligned}$$

where ∇g and H_g are the gradient and Hessian of g respectively. Therefore,

$$\psi_1(x) = \varphi_1(x) - \int \varphi_1(x) dF(x) = \nabla g(\nu)'(x-\nu),$$

and

$$\begin{aligned} \psi_2(x) &= \varphi_2(x) - \int \varphi_2(x,y) dF(x) - \int \varphi_2(x,y) dF(y) + \iint \varphi_2(x,y) dF(x) dF(y) \\ &= (y-\nu)' H_g(\nu)(x-\nu). \end{aligned}$$

Assume the conventional invariance principle on (Y_i), namely

(IP_y) there exists a positive definite matrix $\Sigma^\star$ such that

$$N^{-1/2} \sum_{i=1}^{\lfloor Nt \rfloor} (Y_i - \nu) \Rightarrow \Sigma^\star B^\star(t), \quad t \in [0,1],$$

where $B^\star(\cdot)$ is a standard d-dimensional Brownian motion,

then

$$\left\{ \begin{array}{c} N^{-1/2} \sum_{i=1}^{\lfloor Nt \rfloor} \psi_1(Y_i) \\ N^{-1} \sum_{i=1}^{\lfloor Nt \rfloor} \sum_{j=1}^{\lfloor Nt \rfloor} \psi_2(Y_i, Y_j) \\ N^{-1} \sum_{i=\lfloor Nt \rfloor+1}^{N} \sum_{j=\lfloor Nt \rfloor+1}^{N} \psi_2(Y_i, Y_j) \end{array} \right\} = \left[\begin{array}{l} \nabla g(\nu)' \left\{ N^{-1/2} \sum_{i=1}^{\lfloor Nt \rfloor} (Y_i - \nu) \right\} \\ N^{-1} \sum_{i=1}^{\lfloor Nt \rfloor} (Y_i - \nu)' H_g(\nu) \sum_{j=1}^{\lfloor Nt \rfloor} (Y_j - \nu) \\ N^{-1} \sum_{i=\lfloor Nt \rfloor+1}^{N} (Y_i - \nu)' H_g(\nu) \sum_{j=\lfloor Nt \rfloor+1}^{N} (Y_j - \nu) \end{array} \right],$$

and thus condition (IP_g) holds with

$$W(t) = \left\{ \begin{array}{c} W_1(t) \\ W_2(t) \\ W_3(t) \end{array} \right\} = \left[\begin{array}{l} \nabla g(\nu)' \Sigma^\star B^\star(t) \\ B^\star(t)' \Sigma^{\star\prime} H_g(\nu) \Sigma^\star B^\star(t) \\ \{B^\star(1) - B^\star(t)\}' \Sigma^{\star\prime} H_g(\nu) \Sigma^\star \{B^\star(1) - B^\star(t)\} \end{array} \right]. \quad (2.15)$$

Let $\iota = (\iota_1, \ldots, \iota_d)'$ be an element in $(\mathbb{Z}^+ \cup \{0\})^d$, and denote $\iota! = \iota_1! \cdots \iota_d!$ and

$|\iota| = \iota_1 + \cdots + \iota_d$. For a vector $\tau = (\tau_1, \ldots, \tau_d) \in \mathbb{R}^d$, we write $\tau^\iota = \tau_1^{\iota_1} \cdots \tau_d^{\iota_d}$ and

$$\partial^\iota g(\tau) = \frac{\partial^{|\iota|} g(\tau)}{\partial^{\iota_1}\tau_1 \cdots \partial^{\iota_d}\tau_d}.$$

Let $\bar{Y}_{1,k} = k^{-1}\sum_{i=1}^{k} Y_i$, then the remainder term in this case can be written as

$$R_{1,k} = \frac{1}{\iota!}\sum_{|\iota|=3} \partial^\iota g\{\nu + c_k(\bar{Y}_{1,k} - \nu)\}(\bar{Y}_{1,k} - \nu)^\iota$$

for some $c_k \in [0,1]$. Assume that the derivative $\partial^\iota g$ is bounded as in Shao (2010), then by condition (IP_y) and Theorem 1 of Wu (2007), one can obtain that $\max_{1\le k\le N}|k(\bar{Y}_{1,k} - \nu)| = O_p(N^{1/2})$, and thus $\max_{1\le k\le N}|k^3 R_{1,k}| = O_p(N^{3/2})$. Let $M_{1,N} = \lfloor N^\rho \rfloor$ for some $\rho \in (3/4, 1)$, then $M_{1,N}^{-4}\max_{1\le k\le N}|k^6 R_{1,k}^2| = O_p(N^{3-4\rho}) = o_p(1)$, and thus

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N} k^2 R_{1,k}^2 &= \frac{1}{N}\sum_{k=1}^{M_{1,N}} k^2 R_{1,k}^2 + \frac{N - M_{1,N}}{N}\frac{\max_{1\le k\le N}|k^3 R_{1,k}|^2}{M_{1,N}^4} \\ &= \frac{M_{1,N}}{N}\left(\frac{1}{M_{1,N}}\sum_{k=1}^{M_{1,N}} k^2 R_{1,k}^2\right) + o_p(1).\end{aligned}$$

Define recursively that $M_{l,N} = \lfloor M_{l-1,N}^\rho \rfloor$ for $2 \le l \le L = \lfloor -\log 4/\log\rho \rfloor + 1$, then by iterating the above argument we have

$$\frac{1}{N}\sum_{k=1}^{N} k^2 R_{1,k}^2 = \frac{M_{L,N}}{N}\left(\frac{1}{M_{L,N}}\sum_{k=1}^{M_{L,N}} k^2 R_{1,k}^2\right) + o_p(1) = O_p\left(\frac{M_{L,N}}{N}\cdot\frac{M_{L,N}^3}{1^4}\right) + o_p(1).$$

Since $M_{L,N}^4 = O(N^{4\rho^L}) = o(N)$, we have $\sum_{k=1}^{N} k^2 R_{1,k}^2 = o_p(N)$, and condition (N) follows by a similar argument for $\sum_{k=1}^{N}(N-k)^2 R_{k+1,N}^2$. We shall now consider (2.13) in the statement of Theorem 2.3.3. By (2.15) and properties of the Brownian motion, $W_3(1-t)$ has the same distribution as $W_2(t)$, and $W_2(t)/t$ has the same distribution

as $W_2(1)$. Therefore, we have

$$\limsup_{t\to 0^+} \frac{E\{W_2(t)^2\} + E\{W_3(1-t)^2\}}{t^2} = 2\limsup_{t\to 0^+} \frac{E\{W_2(t)^2\}}{t^2} = 2\limsup_{t\to 0^+} E\{W_2(1)^2\},$$

which is finite and thus (2.13) in Theorem 2.3.3 is satisfied.

2.4 Conclusion

This chapter studies the asymptotic behavior of a data-driven weight choice that minimizes the length of the associated confidence interval. Our results reveal an interesting dichotomy, namely the asymptotic distribution can be different depending on whether the parameter of interest is probabilistically linear or nonlinear. Note that the approximation scheme used by Shao (2010) in their proof is not enough for the current problem, as their approximation relies on the partial sum process of influence functions, which has a linear form and is thus always attracted to the first regime of the dichotomy. To fully recover the dichotomous asymptotic behavior and to understand its cause, we consider the von Mises expansion (von Mises, 1947; Fernholz, 2001) with Gâteaux derivatives. We also consider an application to the class of smooth function models to illustrate the implications of our technical conditions.

Chapter 3

Self-Normalized Simultaneous Confidence Region Construction for High-Dimensional Time Series

3.1 Introduction

Statistical inference of high-dimensional data has been an active area of research. This is mainly due to the fact that statistical methods developed in the low-dimensional setting can fail frequently in high-dimensional settings due to various issues related to the phenomenon called the curse of dimensionality. To accommodate the high dimensionality and provide valid statistical inference procedures, Bai and Saranadasa (1996), Srivastava and Du (2008) and Srivastava et al. (2013) considered the mean case and proposed to modify the popular low-dimensional test of Hotelling (1931) by removing the off-diagonal components in the sample covariance matrix. Motivated by the work of Bai and Saranadasa (1996), Chen and Qin (2010) proposed to trim additional cross terms to improve the power of the associated test. Cai et al. (2014) considered an approach that uses the maximum of componentwise t-statistics, which is suitable for detecting sparse and strong signals; see also Wang et al. (2019). On the other hand, Gregory et al. (2015) considered a sum of squared approach on componentwise t-statistics, which can be more powerful in detecting dense signals; see also Xu et al. (2016), Chen et al. (2019) and references therein for additional contributions. However, the aforementioned results mainly focused on making inference

about the mean of independent and identically distributed samples, and thus may not be suitable for analyzing high-dimensional time series.

High-dimensional time series have emerged frequently in many fields, and examples include large panel data in economics, functional magnetic resonance imaging (fMRI) data in neuroscience, and high-resolution spatio-temporal data in climate science. A crucial task in addressing statistical inference problems such as hypothesis testing and confidence interval construction for time series data is to appropriately accommodate the effect of temporal dependence, as ignoring it can lead to distorted p-values and erroneous conclusions; see for example the demonstration in Zhang and Wu (2011). Although there is a huge literature on developing rigorous inference procedures for time series data, existing methods mostly focused on the finite- or low-dimensional setting and their validity in the high-dimensional setting remains largely unknown. Recently, Zhang and Wu (2017) and Zhang and Cheng (2018) extended the results of Chernozhukov et al. (2013) and considered Gaussian approximation for sample means of high-dimensional time series. Chen and Wu (2019) considered a sum of squared test for trend functions of high-dimensional linear processes; see also Degras et al. (2012) and Zhang (2013) for related contributions in this direction. However, the aforementioned results mostly require knowledge on the long-run covariance matrix, whose direct estimation can be quite nontrivial in the high-dimensional setting and often involves additional bandwidth parameters. To alleviate this, Wang and Shao (2020) in their important work considered an alternative that applies the technique of self-normalization (Kiefer et al., 2000; Lobato, 2001; Shao, 2010). They demonstrated that directly applying self-normalization to high-dimensional objects can fail easily due to singularities or substantial size distortions. Instead, they considered the mean case and proposed to first summarize the high-dimensional mean by a U-statistic following the work of Chen and Qin (2010), and then the conventional self-normalization

technique becomes straightforwardly applicable to that one-dimensional U-statistic. An application to the covariance was also given by viewing it as the mean of a transformed time series. To handle the temporal dependence, however, the U-statistic of Wang and Shao (2020) has to be trimmed, making it less ideal for constructing simultaneous confidence regions.

Instead of seeking a proxy U-statistic to which the conventional self-normalization can be directly applied as in Wang and Shao (2020), we propose new approaches that accommodate the high dimensionality by studying its effect on the maximum modulus of self-normalized statistics. This chapter is based on Zhang and Pan (2021). We in Section 3.2 use the parabolic cylinder functions from harmonic analysis and their connections with the Meijer G-functions to explicitly characterize the tail behavior of self-normalized distributions. The results then enable us to develop an asymptotic theory on the maximum modulus of self-normalized statistics. In addition, we propose a thresholded self-normalization method, which is capable of taking advantage of data with sparse signals to yield simultaneous confidence regions with much reduced volumes. Monte Carlo simulations are conducted in Section 3.4 to examine the finite-sample performance of the proposed methods and compare with the recent high-dimensional self-normalization method of Wang and Shao (2020). Section 3.5 contains an application to a stock price data to further illustrate the proposed methods. Section 3.6 concludes this chapter, and technical proofs are deferred to the Appendix B.

3.2 Tail Asymptotics of Self-Normalized Distributions

We first provide a brief review of the conventional self-normalization (Shao, 2010) in the univariate setting. For this, suppose we observe $y_1, \ldots, y_n$ from a stationary time

series, which can be a univariate time series itself or a component from a multivariate or even high-dimensional time series, and we are interested in an unknown quantity θ that can be the mean, variance, quantiles, interquartile range, and other quantities associated with the observed time series; see the discussion in Shao (2010). Let $\hat{\theta}_i$ be the recursive estimator of θ based on $y_1, \ldots, y_i$, then the self-normalization method of Shao (2010) considers the use of

$$T_n = \frac{n(\hat{\theta}_n - \theta)^2}{V_n},$$

where

$$V_n = \frac{1}{n^2} \sum_{i=1}^{n} \{i(\hat{\theta}_i - \hat{\theta}_n)\}^2$$

is the self-normalizer which is closely related to the cumulative sum (CUSUM) process in the mean case; see for example the discussion in Zhang and Lavitas (2018). The above formulation also includes the test statistic of Wang and Shao (2020) by letting $\hat{\theta}_i$ be the recursive U-statistic. Assume that we have the invariance principle

(IP) there exists a $\sigma > 0$ such that

$$\{n^{-1/2}\lfloor nt \rfloor(\hat{\theta}_{\lfloor nt \rfloor} - \theta),\ t \in [0,1]\} \Rightarrow \{\sigma B(t),\ t \in [0,1]\},$$

where $B(\cdot)$ is a standard Brownian motion and $\Rightarrow$ denotes the weak convergence in the sense of Hoffmann-Jørgensen (van der Vaart and Wellner, 1996),

which can be verified using the influence function approach of Shao (2010) or the functional delta method of Volgushev and Shao (2014), then by the continuous mapping theorem one can obtain that

$$\lim_{n\to\infty} \mathrm{pr}(T_n > z) = \mathrm{pr}\{Z(B) > z\}, \tag{3.1}$$

where the functional Z has the form

$$Z(B) = \frac{\{\sigma B(1)\}^2}{\int_0^1 \{\sigma B(t) - t\sigma B(1)\}^2 dt} = \frac{\{B(1)\}^2}{\int_0^1 \{B(t) - tB(1)\}^2 dt}. \tag{3.2}$$

Note that the nuisance parameter σ is cancelled out in (3.2), the self-normalized distribution in (3.1) does not depend on any nuisance parameter and can be immediately used to construct confidence intervals for the unknown quantity θ; see also the review paper by Shao (2015) for more discussions and advantages of using self-normalization.

To construct simultaneous confidence regions for high-dimensional time series, however, the uncertainty measured by (3.1) for each component will have to be incorporated into the inference procedure in a simultaneous manner. For this, we propose to study the effect of an increasing dimension on the maximum modulus of self-normalized statistics. Since they all share the same pivotalized limiting distribution given by (3.2), we aim at determining if it belongs to the maximum domain of attraction of a given type. This, however, requires an analytical characterization of the tail distribution of $Z(B)$, which does not seem to be a trivial task. In particular, the specific form of the functional Z makes $Z(B)$ a squared mixed normal random variable, and the problem of deriving closed form formulae for densities of mixed normals may require a case by case study and involve nontrivial tools in probability; see for example Abadir and Paruolo (1997). In Abadir and Paruolo (1997), the density function of two mixed normals were studied, both coming from the limit of optimal bivariate cointegration estimators. However, the limit $Z(B)$ in (3.2) from self-normalized statistics involves a different mixing variate from that investigated by Abadir and Paruolo (1997), and thus their result is not directly applicable. Let $G_{p,q}^{m,n}$ be the Meijer G-function (Beals and Szmigielski, 2013), the following theorem provides a complete characterization of the density function of $Z(B)$. For two positive

functions, we write $f(z) \sim g(z)$ as $z \to \infty$ if $\lim_{z\to\infty} f(z)/g(z) = 1$.

Theorem 3.2.1. *Let $f_Z(\cdot)$ be the density function of $Z(B)$, then for any $z > 0$, it admits the expansion*

$$f_Z(z) = 2^{1/2}\pi^{-1}z^{-3/4}\sum_{k=0}^{\infty}\binom{-1/2}{k}(-1)^k G_{1,3}^{3,0}\left\{(k+1/4)^2 z\Big|_{1/4,1/2,0}^{-1/4}\right\}.$$

In addition, as $z \to \infty$,

$$f_Z(z) \sim (2\pi z)^{-1/2}\exp\big(-z^{1/2}/2\big).$$

Theorem 3.2.1 provides an analytical characterization, especially for the tail behavior. This confirms the observation in Lobato (2001) that the distribution of (3.2) has a heavier tail than χ_1^2-distribution. The following theorem states that the self-normalized distribution in (3.2) belongs to the Gumbel maximum domain of attraction.

Theorem 3.2.2. *Let $F_Z(\cdot)$ be the cumulative distribution function of $Z(B)$, then for any $z \in \mathbb{R}$,*

$$\lim_{p\to\infty}\{F_Z(c_p z + d_p)\}^p = \exp\{-\exp(-z)\},$$

where

$$c_p = 8\log p - 4\log\pi + 12\log 2, \quad d_p = (2\log p - \log\pi + 3\log 2)^2.$$

We shall in the following use the results developed in Theorems 3.2.1 and 3.2.2 to devise self-normalized inference procedures for high-dimensional time series.

3.3 Self-Normalized Inference under High Dimensions

Suppose we observe $Y_i = (y_{1,i}, \ldots, y_{p,i})'$, $i = 1, \ldots, n$, from a high-dimensional stationary time series, and we are interested in making simultaneous inference about $\theta_1, \ldots, \theta_p$, where θ_j is the parameter associated with the j-th component. For exam-

ple, one may consider the special case when $\theta_k = E(y_{k,i})$, $k = 1, \ldots, p$, by testing the null hypothesis of no signal, namely $\theta_1 = \cdots = \theta_p = 0$. Wang and Shao (2020) in their important work proposed an extension of self-normalization to high-dimensional time series for inference of the mean. However, their method relies on the availability of a suitable U-statistic When the quantity of interest cannot be written in the form of a mean, then it can be a nontrivial task to identify a generally suitable U-statistic for such purpose, and thus the method of Wang and Shao (2020) may not be versatile in handling general quantities of high-dimensional time series in a unified manner. In Section 3.3.1, we propose new approaches of using self-normalization for inference of high-dimensional time series that are universally applicable to different target quantities of the observed time series. Section 3.3.2 provides their theoretical justifications.

3.3.1 Proposed Methodology

For θ_k associated with the k-th component, following the notation in Section 3.2, let $\hat{\theta}_{k,i}$ be the recursive estimator of θ_k based on $y_{k,1}, \ldots, y_{k,i}$, then we can form the self-normalized distance

$$T_{k,n} = \frac{n(\hat{\theta}_{k,n} - \theta_k)^2}{V_{k,n}}, \quad V_{k,n} = \frac{1}{n^2} \sum_{i=1}^{n} \{i(\hat{\theta}_{k,i} - \hat{\theta}_{k,n})\}^2 \tag{3.3}$$

for each $k = 1, \ldots, p$. Because of the self-normalization, the distances $T_{k,n}$, $k = 1, \ldots, p$, are all properly normalized to share the same pivotalized limiting distribution given by (3.2) regardless of what θ_k represents. This allows one to have the flexibility of considering different quantities for different components, for example quantiles of different levels. To construct simultaneous confidence regions for $\theta_1, \ldots, \theta_p$ in the

high-dimensional setting, we first consider the maximum modulus

$$\Delta_{p,n} = \max_{1 \le k \le p} T_{k,n}. \tag{3.4}$$

By the theoretical justification in Section 3.3.2, $\Delta_{p,n}$ converges in distribution to a Gumbel random variable after proper centering and scaling, which can then be used to guide the construction of simultaneous confidence regions for $\theta_1, \ldots, \theta_p$. Given that the convergence to extreme value distributions is well known to be slow (Hall, 1979; Kim et al., 2015a; Zhang, 2016), instead of using the asymptotic theory to determine the critical value, we consider the following simulation-assisted maximum self-normalization (SAMSN) procedure.

(i) Given observations $Y_1, \ldots, Y_n$, compute the parameter estimator $\hat{\theta}_{k,n}$ and the associated self-normalizer $V_{k,n}$ for each $k = 1, \ldots, p$.

(ii) Generate independent standard normal random variables $y^\circ_{k,i}$ for $k = 1, \ldots, p$ and $i = 1, \ldots, n$, and calculate theoretical values of the associated $\theta^\circ_1, \ldots, \theta^\circ_p$.

(iii) Compute the self-normalized distances $T^\circ_{k,n}$, $k = 1, \ldots, p$, and their maximum modulus $\Delta^\circ_{p,n} = \max_{1 \le k \le p} T^\circ_{k,n}$.

(iv) Repeat (ii) and (iii) above to obtain the $(1-\alpha)$-th quantile of $\Delta^\circ_{p,n}$, denoted by $q_{1-\alpha}(\Delta^\circ_{p,n})$.

(v) Then we can construct the asymptotic $(1-\alpha)$-th simultaneous confidence region for $\theta_1, \ldots, \theta_p$ as

$$\mathcal{R}^{\text{SAMSN}}_{p,n,1-\alpha} = \{(\theta_1, \ldots, \theta_p) : \ |\hat{\theta}_{k,n} - \theta_k| \le n^{-1/2} q_{1-\alpha}(\Delta^\circ_{p,n})^{1/2} V^{1/2}_{k,n}, \ k = 1, \ldots, p\}. \tag{3.5}$$

Since the distribution of $\Delta_{p,n}$ is approximated by that from Gaussian observations, it relates to the Gaussian approximation of Zhang and Wu (2017) and Zhang and Cheng (2018), but more in the spirit of Wu and Zhao (2007) and Zhang and Wu (2012) to improve the finite-sample performance. Note that the Gumbel convergence will remain the same under weak cross-sectional dependence among different components; see for example Xiao and Wu (2013) and Cai et al. (2014). In addition, for the temporal dependence, due to the self-normalization, it does not require direct estimation of the long-run variance and can be directly applied to quantities beyond the mean. By the results in Section 3.3.2, the constructed simultaneous confidence region for $\theta_1, \ldots, \theta_p$ has projected length $O_p(n^{-1/2} \log p)$ in each of its individual dimensions. Therefore, the confidence region projected onto each of its axes will shrink to the true parameter value if the dimension p increases at a slower rate than $\exp\big(n^{1/2}\big)$, and the same will hold in terms of the total volume due to the geometry of a hypercube. Note that it can also be applied to test the null hypothesis when $\theta_1, \ldots, \theta_p$ equal to some prespecified values $\theta_1^0, \ldots, \theta_p^0$. Without loss of generality we can assume that $\theta_k^0 = 0$, $k = 1, \ldots, p$, and consider testing the null hypothesis

$$H_0 : \ \theta_1 = \cdots = \theta_p = 0; \tag{3.6}$$

The proposed test is able to detect any local alternative where there exists one component whose signal is stronger than $O(n^{-1/2} \log p)$.

As observed by Cai et al. (2014), sparse and strong alternatives prevail in a range of applications, where noticeably strong signals exist only in a small amount of components out of a high-dimensional vector. Although the maximum modulus (3.4) can be used to guide the construction of simultaneous confidence regions, it does not take direct advantage of such a phenomenon. We in the following consider using it as

a regularization tool and propose a new thresholded self-normalization method that can lead to confidence regions with much reduced volumes for data with sparse and strong signals. Let $\mathbb{1}_{\{\cdot\}}$ be the indicator function and $V_{k,n}$ be the self-normalizer in (3.3), we propose to consider the thresholded recursive estimator

$$\tilde{\theta}_{\cdot,i} = (\hat{\theta}_{1,i}\mathbb{1}_{\{V_{1,n}^{-1}\hat{\theta}_{1,n}^2 > \lambda_{p,n}\}}, \ldots, \hat{\theta}_{p,i}\mathbb{1}_{\{V_{p,n}^{-1}\hat{\theta}_{p,n}^2 > \lambda_{p,n}\}})', \tag{3.7}$$

where self-normalized statistics are used for the thresholding, and construct the associated thresholded self-normalizer

$$\Lambda_{p,n} = \frac{1}{n^2}\sum_{i=1}^{n} i^2(\tilde{\theta}_{\cdot,i} - \tilde{\theta}_{\cdot,n})(\tilde{\theta}_{\cdot,i} - \tilde{\theta}_{\cdot,n})'.$$

Write $\theta = (\theta_1, \ldots, \theta_p)'$ and let $\Lambda_{p,n}^+$ be the Moore-Penrose inverse of $\Lambda_{p,n}$, we then consider the thresholded self-normalized distance

$$\Omega_{p,n} = n(\tilde{\theta}_{\cdot,n} - \theta)'\Lambda_{p,n}^+(\tilde{\theta}_{\cdot,n} - \theta), \tag{3.8}$$

and construct sparse simultaneous confidence regions using the following simulation-assisted thresholded self-normalization (SATSN) procedure.

(i) Given observations $Y_1, \ldots, Y_n$ and a threshold $\lambda_{p,n}$, compute the thresholded recursive estimators $\tilde{\theta}_{\cdot,i}$, $i = 1, \ldots, n$, and the associated thresholded self-normalizer $\Lambda_{p,n}$.

(ii) Calculate $\hat{d} = \sum_{k=1}^{p} \mathbb{1}_{\{V_{k,n}^{-1}\hat{\theta}_{k,n}^2 > \lambda_{p,n}\}}$, and generate independent standard normal random variables $y_{k,i}^\circ$ for $k = 1, \ldots, \hat{d}$ and $i = 1, \ldots, n$.

(iii) Compute the non-thresholded self-normalized distance $\Omega_{\hat{d},n}^\circ$ for the generated data.

(iv) Repeat (ii) and (iii) above to obtain the $(1-\alpha)$-th quantile of $\Omega_{\hat{d},n}^\circ$, denoted by

$q_{1-\alpha}(\Omega^{\circ}_{\hat{d},n})$.

(v) Then we can construct the asymptotic $(1-\alpha)$-th simultaneous confidence region for $\theta_1, \ldots, \theta_p$ as

$$\mathcal{R}^{\mathrm{SATSN}}_{p,n,1-\alpha} = \{(\theta_1, \ldots, \theta_p) : \Omega_{p,n} \le q_{1-\alpha}(\Omega^{\circ}_{\hat{d},n})\} \odot (\mathbb{1}_{\{V^{-1}_{1,n}\hat{\theta}^2_{1,n} > \lambda_{p,n}\}}, \ldots, \mathbb{1}_{\{V^{-1}_{p,n}\hat{\theta}^2_{p,n} > \lambda_{p,n}\}}), \tag{3.9}$$

where $\odot$ represents the Hadamard product.

Compared with the SAMSN method (3.5) that relies directly on the maximum modulus of self-normalized statistics, the SATSN method (3.9) takes direct advantage of the signal sparsity to produce simultaneous confidence regions with much reduced volumes that are sparse by themselves.

3.3.2 Theoretical Justification

We shall in this section provide theoretical justifications for the proposed methods. Let $\bar{F}_Z(z) = 1 - F_Z(z)$ be the complementary cumulative distribution function of $Z(B)$, and for $1 \le k_1, \ldots, k_m \le p$ write

$$\bar{F}_{T,k_1,\ldots,k_m,n}(z) = \mathrm{pr}(T_{k_1,n} > z, \ldots, T_{k_m,n} > z), \quad z > 0,$$

then $\bar{F}_{T,k,n}(z) = \mathrm{pr}(T_{k,n} > z)$ represents the marginal complementary cumulative distribution function. We first provide a result concerning the stochastic bound of the maximal self-normalized deviation (3.4), and we make the following assumption.

(A1) For any $z > 0$, $\lim_{n\to\infty} \bar{F}_{T,k,n}(z) = 1 - F_Z(z)$ holds for each $k = 1, \ldots, p$.

Condition (A1) is a direct result of (3.1), which has been extensively studied by Shao (2010), Volgushev and Shao (2014), Zhang and Lavitas (2018) and references therein.

Theorem 3.3.1. *Assume (A1). Then there exists a sufficiently large constant $c > 0$ such that*

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{\Delta_{p,n} > c(\log p)^2\} = 0.$$

By Theorem 3.3.1, the maximal self-normalized deviation $\Delta_{p,n}$ given in (3.4) is of order $O_p\{(\log p)^2\}$, which can also be useful in guiding the selection of the threshold value in (3.7). To understand theoretical properties of the simultaneous confidence region (3.5) produced by the SAMSN procedure, we study the asymptotic distribution of $\Delta_{p,n}$, for which we need the following assumption.

(A2) For any $z > 0$,

$$\lim_{p\to\infty}\lim_{n\to\infty}\left|\frac{\sum_{1\le k_1<\cdots<k_m\le p}\bar{F}_{T,k_1,\ldots,k_m,n}(c_pz+d_p)}{\binom{p}{m}\{1-F_Z(c_pz+d_p)\}^m}-1\right| = 0$$

holds for any $m \ge 1$, where c_p and d_p are sequences defined in Theorem 3.2.2.

Condition (A2) can be viewed as a multivariate generalization of condition (A1). In particular, one can show that condition (A2) will be satisfied for $m = 1$ if condition (A1) holds. For the case when $m > 1$ is considered, condition (A2) is expected to be satisfied when the cross-sectional dependence among different components has a sparse or diagonal structure. Note that condition (A2) typically does not lead to any additional restriction on the temporal dependence part, as that is expected to be handled by the self-normalization. Intuitively, under a sparse or diagonal cross-sectional dependence structure, out of the total $\binom{p}{m}$ summands, most of the m-index combinations $k_1, \ldots, k_m$ will exhibit very weak or no dependence, in which case $\bar{F}_{T,k_1,\ldots,k_m,n}(c_pz+d_p)$ will be very close to $\{1-F_Z(c_pz+d_p)\}^m$, making the sum $\sum_{1\le k_1<\cdots<k_m\le p}\bar{F}_{T,k_1,\ldots,k_m,n}(c_pz+d_p)$ close to $\binom{p}{m}\{1-F_Z(c_pz+d_p)\}^m$. For example, in the diagonal case, one can show that condition (A2) automatically reduces to condition (A1). The following theorem states that the maximal self-normalized deviation

$\Delta_{p,n}$ in (3.4), after proper centering and scaling, follows an asymptotic Gumbel distribution, and that the associated SAMSN procedure proposed in Section 3.3.1 produces simultaneous confidence regions with the desired coverage probability.

Theorem 3.3.2. *Assume (A2). Then for any $z > 0$,*

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{\Delta_{p,n} \le c_p z + d_p\} = \exp\{-\exp(-z)\}.$$

If in addition (A1) holds for independent Gaussian data, then for any $\alpha \in (0,1)$,

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{(\theta_1,\ldots,\theta_p) \in \mathcal{R}^{\mathrm{SAMSN}}_{p,n,1-\alpha}\} = 1-\alpha.$$

Note that in practice $q_{1-\alpha}(\Delta^{\circ}_{p,n})$ has to be replaced by its empirical counterpart. The difference, however, can be made arbitrarily small as we can generate as many independent Gaussian data in the simulation-assisted step as we want. As a result, we assume for simplicity that $q_{1-\alpha}(\Delta^{\circ}_{p,n})$ is available in our theoretical analyses, and alternatively it can be handled by either involving an arbitrarily small constant or another limit when the number of simulated replications goes to infinity. Since $c_p = O(\log p)$ and $d_p = O\{(\log p)^2\}$, the hypercube $\mathcal{R}^{\mathrm{SAMSN}}_{p,n,1-\alpha}$ constructed as the simultaneous confidence region for $\theta_1,\ldots,\theta_p$ has projected length $O_p(n^{-1/2}\log p)$ in each of its individual dimensions. We shall now study asymptotic properties of the thresholded self-normalized deviation $\Omega_{p,n}$ in (3.8) and its associated SATSN procedure that are designed particularly for data with sparse signals. For this, we let $\mathcal{S}_0 = \{k : \ \theta_k = 0\}$ and $\mathcal{S}_1 = \{k : \ \theta_k \neq 0\}$ be subsets of $\{1,\ldots,p\}$ that represent components with zero and nonzero signals respectively, and we use d to denote the cardinality of $\mathcal{S}_1$, namely the total number of nonzero signals. Under signal sparsity, d is expected to be either finite or growing at a very slow rate, so that the conventional self-normalization of Shao (2010) can still be applicable to this sparse subset. We shall in the following summarize this into a mathematical condition. For this, let $\hat{\theta}_{\mathcal{S}_1,i}$

be the recursive estimator on $\mathcal{S}_1$, then one can form the self-normalized distance

$$T_{\mathcal{S}_1,n} = n(\hat{\theta}_{\mathcal{S}_1,n} - \theta_{\mathcal{S}_1})' V_{\mathcal{S}_1,n}^{-1} (\hat{\theta}_{\mathcal{S}_1,n} - \theta_{\mathcal{S}_1}), \quad V_{\mathcal{S}_1,n} = \frac{1}{n^2} \sum_{i=1}^{n} i^2 (\hat{\theta}_{\mathcal{S}_1,i} - \hat{\theta}_{\mathcal{S}_1,n})(\hat{\theta}_{\mathcal{S}_1,i} - \hat{\theta}_{\mathcal{S}_1,n})'.$$

Let $B_d(\cdot)$ be a standard d-dimensional Brownian motion, then the self-normalization method of Shao (2010) and Shao (2015) approximates the distribution of $T_{\mathcal{S}_1,n}$ by that of

$$Z_d(B_d) = B_d(1)' \left[\int_0^1 \{B_d(t) - tB_d(1)\}\{B_d(t) - tB_d(1)\}' dt \right]^{-1} B_d(1).$$

(A3) The number of nonzero components d grows with p slowly enough such that

$$\lim_{p\to\infty} \lim_{n\to\infty} |\mathrm{pr}(T_{\mathcal{S}_1,n} \le z) - \mathrm{pr}\{Z_d(B_d) \le z\}| = 0$$

holds for any $z > 0$.

Let $q_{1-\alpha}\{Z_d(B_d)\}$ be the $(1-\alpha)$-th quantile of $Z_d(B_d)$, the following theorem states that the distribution of the thresholded self-normalized distance $\Omega_{p,n}$ in (3.8) can be asymptotically approximated by that of $Z_d(B_d)$, and that the associated SATSN procedure proposed in Section 3.3.1 produces sparse simultaneous confidence regions with the desired coverage probability.

Theorem 3.3.3. *Assume (A3) and that the threshold satisfies $\lambda_{p,n} = c^\star n^{-1}(\log p)^2$ for some $c^\star > 4$. If the invariance principle (IP) holds marginally for each component with possibly different $\sigma_1, \ldots, \sigma_p$ that are bounded away from zero and infinity and the minimal nonzero signal satisfies*

$$\lim_{p\to\infty} \lim_{n\to\infty} \frac{\min_{k\in\mathcal{S}_1} |\theta_k|}{(n^{-1/2}\log p)(\log d)^{1/2}} = +\infty,$$

then for any $z > 0$,

$$\lim_{p\to\infty} \lim_{n\to\infty} \mathrm{pr}[\Omega_{p,n} \le q_{1-\alpha}\{Z_d(B_d)\}] = 1 - \alpha.$$

If in addition (A3) holds for d-dimensional independent Gaussian data, then for any $\alpha \in (0,1)$,

$$\lim_{p\to\infty}\lim_{n\to\infty} \mathrm{pr}\{(\theta_1,\ldots,\theta_p) \in \mathcal{R}^{\mathrm{SATSN}}_{p,n,1-\alpha}\} = 1-\alpha.$$

Compared with the SAMSN confidence region $\mathcal{R}^{\mathrm{SATSN}}_{p,n,1-\alpha}$ from (3.5), the SATSN procedure takes advantage of the sparse signal structure and produces confidence regions with reduced volumes that only reside in a low-dimensional subspace. We shall here provide a discussion about how to choose the threshold value $\lambda_{p,n}$ in practice. For this, a simple approach is to use Theorem 3.3.3 and choose $\lambda_{p,n} = c^\star n^{-1}(\log p)^2$ for some constant $c^\star > 4$. Alternatively, since the threshold $\lambda_{p,n}$ is mainly used to screen out components with zero signals, as a rule of thumb, one can use Theorem 3.3.2 and choose $\lambda_{p,n}$ as the 99% cut-off value for testing the null hypothesis of no signal. It can be seen from our numerical results in Sections 3.4.1 and 3.4.2 that such a rule of thumb choice seems to perform reasonably well.

3.4 Simulation Results

3.4.1 Simulation Results: The Mean Case

We shall here carry out Monte Carlo simulations to examine the finite-sample performance of the proposed SAMSN and SATSN methods for constructing simultaneous confidence regions for high-dimensional time series. We first consider the mean case, where we can make a comparison with the recently developed self-normalized method of Wang and Shao (2020) based on trimmed U-statistics. We also consider the case of the median in Section 3.4.2, where the method of Wang and Shao (2020) is not directly applicable. Let $(\epsilon_{k,i})_{k,i}$ be an array of independent standard normal random variables, and $(e_{k,i})_i$ be an autoregressive process satisfying the recursion $e_{k,i} = \rho e_{k,i-1} + \epsilon_{k,i}$ for each $k = 1,\ldots,p$. Let $n = 500$ and $e_{\cdot,i} = (e_{1,i},\ldots,e_{p,i})'$, we consider constructing

simultaneous confidence regions for the mean of

$$Y_i = \theta + \Sigma^{1/2} e_{\cdot,i}, \quad i = 1, \ldots, n, \tag{3.10}$$

where $\theta = (3, 5, -4, 5, 0, \ldots, 0)$ represents the true value and Σ takes the block diagonal form $\Sigma = \text{diag}\{B_m, \ldots, B_m, B_{p-m\lfloor p/m \rfloor}\}$ with $B_m = (0.5^{|i-j|})_{1 \le i,j \le m}$. Note that the block diagonal form with $m = 2$ gives a tri-diagonal matrix; see for example the setting in Wang and Shao (2020). For each realization of the process (3.10), we apply the proposed SAMSN and SATSN methods to construct simultaneous confidence regions for its mean, and compare them with the self-normalized U-statistic method of Wang and Shao (2020). Note that the method of Wang and Shao (2020) requires the selection of a trimming parameter to handle the temporal dependence, and we shall here follow their setting and consider choosing the trimming parameter as 5, 10, 20 and 30, and denote the associated methods by WS20_5, WS20_{10}, WS20_{20} and WS20_{30} respectively. For $p \in \{50, 100, 200\}$, $\rho \in \{0.3, 0.6, -0.3, -0.6\}$ and $m \in \{1, 2, 5\}$, the results are summarized in Tables 3.1 and 3.2 based on 1000 realizations for each configuration, from which we can observe the followings.

First, the empirical coverage probabilities of the proposed SAMSN and SATSN methods are reasonably close to their nominal levels, namely 90% in Table 3.1 and 95% in Table 3.2. As a comparison, the performance of the self-normalized U-statistic method of Wang and Shao (2020) can be affected by the choice of trimming in a nonnegligible way, especially when $\rho = -0.6$ for which noticeable size distortions can be identified for WS20_5. Second, we report in parentheses the average projected length (APL) of the constructed simultaneous confidence region on each of the coordinates for the proposed SAMSN and SATSN methods, from which we can see that the SATSN method is capable of taking direct advantage of the signal sparsity to produce

p	m	ρ	$\text{SAMSN}_{(\text{APL})}$	$\text{SATSN}_{(\text{APL})}$	WS20_5	WS20_{10}	WS20_{20}	WS20_{30}
50	1	0.3	$0.885_{(0.637)}$	$0.889_{(0.034)}$	0.883	0.886	0.891	0.889
		0.6	$0.886_{(1.099)}$	$0.863_{(0.058)}$	0.877	0.896	0.884	0.884
		-0.3	$0.905_{(0.345)}$	$0.910_{(0.019)}$	0.884	0.882	0.877	0.883
		-0.6	$0.918_{(0.282)}$	$0.912_{(0.016)}$	0.706	0.914	0.903	0.905
	2	0.3	$0.907_{(0.636)}$	$0.896_{(0.030)}$	0.892	0.890	0.890	0.889
		0.6	$0.895_{(1.103)}$	$0.887_{(0.050)}$	0.857	0.880	0.872	0.870
		-0.3	$0.898_{(0.344)}$	$0.905_{(0.016)}$	0.894	0.897	0.895	0.890
		-0.6	$0.918_{(0.280)}$	$0.926_{(0.014)}$	0.738	0.901	0.894	0.888
	5	0.3	$0.891_{(0.636)}$	$0.896_{(0.028)}$	0.888	0.889	0.873	0.876
		0.6	$0.898_{(1.104)}$	$0.877_{(0.048)}$	0.899	0.896	0.902	0.889
		-0.3	$0.909_{(0.344)}$	$0.894_{(0.016)}$	0.888	0.898	0.891	0.884
		-0.6	$0.920_{(0.282)}$	$0.919_{(0.013)}$	0.716	0.914	0.896	0.895
100	1	0.3	$0.896_{(0.698)}$	$0.894_{(0.017)}$	0.884	0.885	0.888	0.879
		0.6	$0.874_{(1.212)}$	$0.857_{(0.030)}$	0.846	0.885	0.896	0.885
		-0.3	$0.905_{(0.379)}$	$0.886_{(0.009)}$	0.895	0.914	0.904	0.899
		-0.6	$0.915_{(0.311)}$	$0.909_{(0.008)}$	0.641	0.901	0.904	0.895
	2	0.3	$0.906_{(0.699)}$	$0.891_{(0.015)}$	0.900	0.901	0.901	0.887
		0.6	$0.865_{(1.213)}$	$0.875_{(0.025)}$	0.866	0.892	0.881	0.877
		-0.3	$0.915_{(0.380)}$	$0.903_{(0.008)}$	0.899	0.902	0.898	0.902
		-0.6	$0.905_{(0.310)}$	$0.932_{(0.007)}$	0.656	0.910	0.898	0.890
	5	0.3	$0.884_{(0.699)}$	$0.897_{(0.014)}$	0.905	0.905	0.888	0.895
		0.6	$0.885_{(1.212)}$	$0.873_{(0.024)}$	0.869	0.893	0.893	0.888
		-0.3	$0.912_{(0.379)}$	$0.908_{(0.008)}$	0.889	0.889	0.885	0.887
		-0.6	$0.935_{(0.310)}$	$0.925_{(0.006)}$	0.668	0.896	0.901	0.900
200	1	0.3	$0.877_{(0.766)}$	$0.869_{(0.009)}$	0.900	0.900	0.891	0.896
		0.6	$0.882_{(1.334)}$	$0.870_{(0.015)}$	0.796	0.878	0.882	0.872
		-0.3	$0.915_{(0.416)}$	$0.899_{(0.005)}$	0.896	0.894	0.893	0.892
		-0.6	$0.927_{(0.340)}$	$0.909_{(0.004)}$	0.607	0.896	0.904	0.901
	2	0.3	$0.899_{(0.767)}$	$0.900_{(0.007)}$	0.891	0.894	0.883	0.885
		0.6	$0.872_{(1.333)}$	$0.850_{(0.013)}$	0.835	0.896	0.893	0.898
		-0.3	$0.914_{(0.416)}$	$0.904_{(0.004)}$	0.903	0.906	0.907	0.897
		-0.6	$0.936_{(0.340)}$	$0.907_{(0.003)}$	0.598	0.897	0.881	0.896
	5	0.3	$0.893_{(0.765)}$	$0.903_{(0.007)}$	0.901	0.900	0.891	0.877
		0.6	$0.886_{(1.333)}$	$0.894_{(0.012)}$	0.863	0.905	0.896	0.889
		-0.3	$0.906_{(0.417)}$	$0.885_{(0.004)}$	0.888	0.889	0.890	0.889
		-0.6	$0.921_{(0.341)}$	$0.913_{(0.003)}$	0.637	0.911	0.906	0.904

Table 3.1: Empirical coverage probabilities for the 90% simultaneous confidence regions of the high-dimensional mean constructed by the proposed SAMSN and SATSN methods (with average projected lengths in parentheses), and the self-normalized U-statistic method of Wang and Shao (2020) with different trimming parameters.

p	m	ρ	$\text{SAMSN}_{(\text{APL})}$	$\text{SATSN}_{(\text{APL})}$	WS20_5	WS20_{10}	WS20_{20}	WS20_{30}
50	1	0.3	$0.949_{(0.710)}$	$0.946_{(0.040)}$	0.949	0.951	0.951	0.942
		0.6	$0.948_{(1.225)}$	$0.918_{(0.067)}$	0.929	0.944	0.936	0.932
		-0.3	$0.949_{(0.385)}$	$0.951_{(0.022)}$	0.941	0.943	0.934	0.941
		-0.6	$0.959_{(0.315)}$	$0.958_{(0.018)}$	0.835	0.964	0.965	0.960
	2	0.3	$0.951_{(0.709)}$	$0.943_{(0.035)}$	0.949	0.947	0.948	0.953
		0.6	$0.938_{(1.230)}$	$0.935_{(0.058)}$	0.936	0.939	0.935	0.938
		-0.3	$0.954_{(0.384)}$	$0.952_{(0.019)}$	0.941	0.941	0.945	0.950
		-0.6	$0.969_{(0.313)}$	$0.963_{(0.016)}$	0.860	0.953	0.951	0.939
	5	0.3	$0.949_{(0.709)}$	$0.948_{(0.033)}$	0.945	0.945	0.933	0.931
		0.6	$0.946_{(1.230)}$	$0.930_{(0.055)}$	0.938	0.942	0.944	0.943
		-0.3	$0.959_{(0.383)}$	$0.939_{(0.018)}$	0.946	0.948	0.944	0.943
		-0.6	$0.961_{(0.314)}$	$0.952_{(0.015)}$	0.845	0.954	0.951	0.952
100	1	0.3	$0.948_{(0.769)}$	$0.944_{(0.020)}$	0.945	0.938	0.937	0.942
		0.6	$0.921_{(1.334)}$	$0.916_{(0.034)}$	0.904	0.939	0.945	0.946
		-0.3	$0.958_{(0.417)}$	$0.945_{(0.011)}$	0.955	0.947	0.951	0.947
		-0.6	$0.958_{(0.342)}$	$0.956_{(0.009)}$	0.787	0.956	0.954	0.951
	2	0.3	$0.958_{(0.769)}$	$0.945_{(0.017)}$	0.955	0.951	0.953	0.947
		0.6	$0.933_{(1.335)}$	$0.931_{(0.029)}$	0.933	0.942	0.935	0.939
		-0.3	$0.955_{(0.418)}$	$0.949_{(0.010)}$	0.952	0.947	0.945	0.944
		-0.6	$0.957_{(0.341)}$	$0.964_{(0.008)}$	0.800	0.954	0.946	0.950
	5	0.3	$0.947_{(0.769)}$	$0.946_{(0.016)}$	0.942	0.946	0.943	0.941
		0.6	$0.938_{(1.335)}$	$0.932_{(0.028)}$	0.927	0.948	0.953	0.945
		-0.3	$0.962_{(0.417)}$	$0.947_{(0.009)}$	0.944	0.949	0.940	0.944
		-0.6	$0.962_{(0.341)}$	$0.968_{(0.007)}$	0.802	0.945	0.952	0.935
200	1	0.3	$0.928_{(0.834)}$	$0.926_{(0.010)}$	0.947	0.942	0.945	0.949
		0.6	$0.948_{(1.451)}$	$0.932_{(0.017)}$	0.882	0.945	0.940	0.941
		-0.3	$0.955_{(0.453)}$	$0.941_{(0.005)}$	0.953	0.947	0.949	0.947
		-0.6	$0.962_{(0.370)}$	$0.953_{(0.005)}$	0.806	0.946	0.950	0.950
	2	0.3	$0.939_{(0.835)}$	$0.945_{(0.008)}$	0.944	0.945	0.945	0.944
		0.6	$0.923_{(1.451)}$	$0.917_{(0.015)}$	0.901	0.953	0.952	0.946
		-0.3	$0.958_{(0.453)}$	$0.945_{(0.005)}$	0.945	0.948	0.945	0.946
		-0.6	$0.975_{(0.371)}$	$0.955_{(0.004)}$	0.779	0.946	0.936	0.954
	5	0.3	$0.941_{(0.833)}$	$0.943_{(0.008)}$	0.945	0.943	0.939	0.934
		0.6	$0.932_{(1.450)}$	$0.937_{(0.014)}$	0.920	0.950	0.952	0.947
		-0.3	$0.952_{(0.453)}$	$0.938_{(0.004)}$	0.941	0.948	0.942	0.945
		-0.6	$0.963_{(0.371)}$	$0.961_{(0.004)}$	0.811	0.955	0.955	0.958

Table 3.2: Empirical coverage probabilities for the 95% simultaneous confidence regions of the high-dimensional mean constructed by the proposed SAMSN and SATSN methods (with average projected lengths in parentheses), and the self-normalized U-statistic method of Wang and Shao (2020) with different trimming parameters.

a much smaller simultaneous confidence region as its APL is significantly smaller than that of SAMSN. In contrast, due to the use of a trimmed U-statistic, the shape of the simultaneous confidence region produced by the method of Wang and Shao (2020) can be very difficult to understand, and as a result it can be quite a nontrivial challenge to find its average projected length. Third, the simultaneous confidence regions constructed by the proposed SAMSN and SATSN methods center at $\bar{Y}_{1,n} = n^{-1}\sum_{i=1}^{n} Y_i$ and its thresholded counterpart, while the center of the simultaneous confidence region induced by the method of Wang and Shao (2020) can be difficult to identify. Compared with the self-normalized U-statistic method of Wang and Shao (2020), the proposed SAMSN and SATSN methods seem to be more desirable for the purpose of constructing simultaneous confidence regions for high-dimensional time series.

3.4.2 Simulation Results: The Median Case

We shall in this section apply the proposed SAMSN and SATSN methods to construct simultaneous confidence regions for the median of the process in (3.10), for which the self-normalized U-statistic method of Wang and Shao (2020) is not directly applicable. The results are summarized in Table 3.3, from which we can observe the followings. First, the empirical coverage probabilities of the proposed SAMSN and SATSN methods for the median case, though being slightly worse that those in Tables 3.1 and 3.2 for the mean case, are still reasonably close to their 90% and 95% nominal levels. Second, when compared with the SAMSN method, the SATSN method continues to produce much smaller simultaneous confidence regions in the median case due to its capability of taking direct advantage of the signal sparsity. Third, the average projected lengths in Table 3.3 for the median are generally longer than those in Tables 3.1 and 3.2 for the mean, indicating that simultaneous confidence regions

constructed for the median can be bigger than those constructed for the mean at the same nominal level.

			$\text{SAMSN}_{(\text{APL})}$		$\text{SATSN}_{(\text{APL})}$	
p	m	ρ	90%	95%	90%	95%
50	1	0.3	$0.904_{(0.894)}$	$0.962_{(1.014)}$	$0.904_{(0.042)}$	$0.938_{(0.050)}$
		0.6	$0.921_{(1.457)}$	$0.964_{(1.653)}$	$0.902_{(0.068)}$	$0.943_{(0.080)}$
		-0.3	$0.877_{(0.606)}$	$0.943_{(0.688)}$	$0.898_{(0.030)}$	$0.945_{(0.035)}$
		-0.6	$0.859_{(0.600)}$	$0.921_{(0.681)}$	$0.883_{(0.030)}$	$0.932_{(0.035)}$
	2	0.3	$0.922_{(0.894)}$	$0.961_{(1.015)}$	$0.896_{(0.039)}$	$0.945_{(0.046)}$
		0.6	$0.914_{(1.462)}$	$0.964_{(1.659)}$	$0.909_{(0.062)}$	$0.953_{(0.073)}$
		-0.3	$0.882_{(0.606)}$	$0.935_{(0.688)}$	$0.887_{(0.029)}$	$0.932_{(0.034)}$
		-0.6	$0.865_{(0.600)}$	$0.917_{(0.681)}$	$0.881_{(0.029)}$	$0.927_{(0.034)}$
	5	0.3	$0.916_{(0.891)}$	$0.953_{(1.011)}$	$0.870_{(0.038)}$	$0.941_{(0.045)}$
		0.6	$0.919_{(1.468)}$	$0.955_{(1.666)}$	$0.891_{(0.060)}$	$0.941_{(0.071)}$
		-0.3	$0.875_{(0.604)}$	$0.926_{(0.685)}$	$0.890_{(0.028)}$	$0.940_{(0.033)}$
		-0.6	$0.851_{(0.599)}$	$0.905_{(0.679)}$	$0.887_{(0.029)}$	$0.931_{(0.034)}$
100	1	0.3	$0.906_{(1.030)}$	$0.957_{(1.170)}$	$0.903_{(0.021)}$	$0.949_{(0.025)}$
		0.6	$0.936_{(1.682)}$	$0.965_{(1.911)}$	$0.879_{(0.034)}$	$0.937_{(0.040)}$
		-0.3	$0.888_{(0.700)}$	$0.942_{(0.795)}$	$0.888_{(0.015)}$	$0.937_{(0.017)}$
		-0.6	$0.866_{(0.692)}$	$0.923_{(0.786)}$	$0.882_{(0.015)}$	$0.937_{(0.017)}$
	2	0.3	$0.933_{(1.028)}$	$0.965_{(1.167)}$	$0.890_{(0.020)}$	$0.945_{(0.023)}$
		0.6	$0.954_{(1.689)}$	$0.978_{(1.918)}$	$0.886_{(0.031)}$	$0.941_{(0.037)}$
		-0.3	$0.877_{(0.700)}$	$0.932_{(0.795)}$	$0.897_{(0.014)}$	$0.940_{(0.017)}$
		-0.6	$0.847_{(0.693)}$	$0.922_{(0.787)}$	$0.865_{(0.014)}$	$0.922_{(0.017)}$
	5	0.3	$0.930_{(1.031)}$	$0.967_{(1.171)}$	$0.893_{(0.019)}$	$0.944_{(0.023)}$
		0.6	$0.935_{(1.686)}$	$0.971_{(1.915)}$	$0.875_{(0.030)}$	$0.934_{(0.035)}$
		-0.3	$0.895_{(0.701)}$	$0.936_{(0.796)}$	$0.894_{(0.014)}$	$0.942_{(0.017)}$
		-0.6	$0.870_{(0.691)}$	$0.929_{(0.784)}$	$0.884_{(0.014)}$	$0.938_{(0.017)}$
200	1	0.3	$0.929_{(1.157)}$	$0.965_{(1.309)}$	$0.899_{(0.011)}$	$0.948_{(0.012)}$
		0.6	$0.932_{(1.891)}$	$0.966_{(2.140)}$	$0.866_{(0.017)}$	$0.916_{(0.020)}$
		-0.3	$0.862_{(0.784)}$	$0.934_{(0.887)}$	$0.894_{(0.007)}$	$0.943_{(0.009)}$
		-0.6	$0.818_{(0.777)}$	$0.896_{(0.879)}$	$0.872_{(0.007)}$	$0.936_{(0.009)}$
	2	0.3	$0.922_{(1.159)}$	$0.965_{(1.312)}$	$0.896_{(0.010)}$	$0.947_{(0.012)}$
		0.6	$0.934_{(1.890)}$	$0.973_{(2.138)}$	$0.881_{(0.015)}$	$0.940_{(0.018)}$
		-0.3	$0.883_{(0.784)}$	$0.938_{(0.887)}$	$0.884_{(0.007)}$	$0.930_{(0.008)}$
		-0.6	$0.847_{(0.775)}$	$0.916_{(0.877)}$	$0.880_{(0.007)}$	$0.925_{(0.009)}$
	5	0.3	$0.932_{(1.158)}$	$0.973_{(1.310)}$	$0.879_{(0.010)}$	$0.942_{(0.011)}$
		0.6	$0.956_{(1.892)}$	$0.978_{(2.140)}$	$0.888_{(0.015)}$	$0.936_{(0.018)}$
		-0.3	$0.890_{(0.785)}$	$0.938_{(0.888)}$	$0.881_{(0.007)}$	$0.933_{(0.008)}$
		-0.6	$0.847_{(0.777)}$	$0.916_{(0.879)}$	$0.879_{(0.007)}$	$0.924_{(0.009)}$

Table 3.3: Empirical coverage probabilities for the 90% and 95% simultaneous confidence regions of the high-dimensional median constructed by the proposed SAMSN and SATSN methods (with average projected lengths in parentheses).

3.5 Application to a Stock Price Data

We shall here apply the developed results to a financial data that contains daily adjusted close prices for a set of NASDAQ-100 companies during the period from 2016 to 2019. The set of companies are chosen based on the NASDAQ-100 list in April 2020, and we removed a few with incomplete data (for example, ZOOM started trading in 2019). A complete list can be found in Table 3.4, and we consider constructing simultaneous confidence regions for the mean and median of their daily rates of return.

ATVI	ADBE	AMD	ALXN	ALGN	GOOGL	GOOG	AMZN	AMGN	ADI
ANSS	AAPL	AMAT	ASML	ADSK	ADP	BIDU	BIIB	BMRN	BKNG
AVGO	CDNS	CDW	CERN	CHTR	CHKP	CTAS	CSCO	CTXS	CTSH
CMCSA	CPRT	CSGP	COST	CSX	DXCM	DLTR	EBAY	EA	EXC
EXPE	FB	FAST	FISV	GILD	IDXX	ILMN	INCY	INTC	INTU
ISRG	JD	KLAC	KHC	LRCX	LBTYA	LBTYK	LULU	MAR	MXIM
MELI	MCHP	MU	MSFT	MDLZ	MNST	NTAP	NTES	NFLX	NVDA
NXPI	ORLY	PCAR	PAYX	PYPL	PEP	QCOM	REGN	ROST	SGEN
SIRI	SWKS	SPLK	SBUX	SNPS	TMUS	TTWO	TSLA	TXN	TCOM
ULTA	UAL	VRSN	VRSK	VRTX	WBA	WDAY	WDC	XEL	XLNX

Table 3.4: The complete list of stock tickers used in the data analysis in Section 3.5.

In our analysis, n = 1003 and p = 100. For the mean rate of return, both the SAMSN and SATSN methods reject the null hypothesis of zero return at 1% significance level, which is in line with the test of Wang and Shao (2020) if the trimming parameter in their test is chosen to be 20. However, the test of Wang and Shao (2020) can be affected by the trimming, and its use of a trimmed U-statistic to summarize the high-dimensional mean makes it difficult to provide any additional information about

which stock has a nonzero mean beyond rejecting the null hypothesis. In contrast, the proposed SAMSN and SATSN methods are able to produce simultaneous confidence regions for the high-dimensional mean, based on which practitioners can further obtain individual level information such as which stock has a nonzero mean that caused the rejection of the null. In particular, simultaneous confidence regions constructed by both the SAMSN and SATSN methods suggest that it is the stock CTAS that has a nonzero mean rate of return, which corroborates with the assumption of signal sparsity. The 99% SAMSN simultaneous confidence region is in the shape of a hypercube and its projection onto the CTAS dimension is 0.00124 ± 0.00122. On the other hand, the 99% SATSN simultaneous confidence region in this case is a sparse ellipsoid which occupies 0.00124 ± 0.00069 on the dimension of CTAS. Note that the SATSN method produces a narrower interval than the SAMSN method.

Stock	SAMSN	SATSN
ISRG	0.00188 ± 0.00139	0.00188 ± 0.00107
KLAC	0.00206 ± 0.00197	0.00206 ± 0.00152
NVDA	0.00247 ± 0.00207	0.00247 ± 0.00160

Table 3.5: Projected 99% simultaneous confidence regions produced by the SAMSN and SATSN methods for the median rate on each of the stock dimensions.

We then consider the median case, for which the proposed SAMSN and SATSN methods identify three stocks ISRG, KLAC and NVDA with positive median rates of return, meaning that they have more days of gaining than losing. The 99% simultaneous confidence regions produced by the SAMSN and SATSN methods are summarized in Table 3.5 in terms of their projections onto the dimensions ISRG, KLAC and NVDA respectively. Note that the simultaneous confidence region produced by the SATSN method in this case is in the shape of a low-dimensional ellipsoid,

and we report the maximal area of its projection onto each of the ISRG, KLAC and NVDA axes. The method of Wang and Shao (2020) is no longer applicable in this case as it was developed only for the mean case, and therefore we do not include it for comparison. It can be seen from our empirical examples that the proposed SAMSN and SATSN methods not only complement the method of Wang and Shao (2020) by being able to produce simultaneous confidence regions that provide practitioners with valuable information beyond simply rejecting the null hypothesis of no signal, they are also readily applicable to quantities beyond the mean which can be a very useful feature in practice.

3.6 Conclusion

The chapter aims at proposing new approaches of using self-normalization to make statistical inference of general quantities of high-dimensional time series. Instead of seeking a proxy U-statistic on which the conventional self-normalization can be applied as in Wang and Shao (2020), we propose to study the effect of an increasing dimension on the maximum modulus of self-normalized statistics. We analyze the tail behaviour of the self-normalized distribution with parabolic cylinder functions and then develop an asymptotic theory on the maximum modulus of self-normalized statistics. In addition, we propose a thresholded self-normalization method, which can yield simultaneous confidence regions with much reduced volumes. Our methods seem to perform reasonably well in our simulation study and application. Compared with existing results on self-normalized inference, this work is not only the first to consider self-normalized simultaneous confidence region in the high-dimensional setting, but also directly applicable to quantities beyond the mean with a unified inference protocol that can be very convenient to use in practice.

Chapter 4

Unsupervised Self-Normalized Break Test for Correlation Matrix

4.1 Introduction

Correlation coefficient is a statistical method, and it is widely applied to many areas, such as psychology, economics and EEG analysis in trend analysis and classification analysis. This is because the correlation coefficient of a system characterizes its system structure and it becomes a favorable metric when variables of a system don't react in isolated ways. For example, in emotion psychology, researchers believe that to cope with environmental opportunities or threats quickly and efficiently, one need to enable physiological, experiential, and behavioral reactions synchronously (Mauss et al., 2005). Economists observe that during financial crisis, the correlation structure of financial returns changes informatively, see, e.g., Krishnan et al. (2009). Therefore, correlations among financial returns are widely used in risk management. In an EEG report, correlation coefficient changes can also be used to detect the presence of epileptiform discharges in different electrodes (Zhou et al., 2020). Hence, correlation matrix break testing has attracts much attention with a wide application. Many other examples are mentioned in (Cabrieto et al., 2017).

To reveal the change pattern of correlation structure, many methods are developed in testing correlation constancy. Wied et al. (2012) proposed a CUSUM-type correlation coefficient break test and then Wied and Galeano (2013) applied the test

to detect structural changes on global minimum-variance portfolios. The test was further extended to detect multiple breaks in the correlation structure of random variables in (Galeano and Wied, 2014). This multiple-breaks-test was widely applied in financial areas. For example, Adams and Glück (2015) applied the test to identify the impact of financialization in commodity markets and Berens et al. (2015) used it to forecast value-at-risk of financial portfolio. However, the tests only consider bivariate correlations condition, which doesn't account for high dimensional time series data.

Recently, Wied (2017) constructed a new test with bootstrap variance matrix estimator to detect change in correlation matrix. As mentioned in Chapter 1, the test becomes unstable and has over-size issue when the number of variables (P) is relatively large to the time series dimension (T) of data. Moreover, when P is moderate to T, it tends to be under-sized and have low power. To overcome these weaknesses, Choi and Shin (2020a) integrated the idea of self-normalization in Lobato (2001) and Shao and Zhang (2010) and proposed a self-normalization break test for correlation matrix, CS_T. The test CS_T doesn't involve estimating covariance matrix of the sample correlations, and therefore it doesn't have issues such as instability or singularity. Moreover, Choi and Shin (2020a) showed that CS_T has better size performance than DW_T (Eq 1.3) proposed by Wied (2017) under various simulation settings.

However, CS_T was mainly designed for a single break alternative and it becomes less powerful when testing against multiple breaks. They proposed a method of scanning subsamples to detect multiple breaks, whereas the performance highly depends on how many subsamples they choose to split the data. Zhang and Lavitas (2018) proposed a contrast-based approach in generalizing self-normalized statistics tailored for change-point testing. In this chapter, we follow Zhang and Lavitas (2018) scheme and propose an unsupervised correlation matrix break test based on self-normalization.

This chapter is based on Pan et al. (2021). The method can test against multiple breaks without knowing the number of breaks as the prior information.

This chapter is organized as follows: Section 4.2 reviews the idea of the self-normalized correlation matrix break test and the extended scanning test proposed by Choi and Shin (2020a) and then introduces our unsupervised correlation matrix break test. Section 4.2.1 describes the framework of the test procedure and the theoretical properties are given in Section 4.2.2. Moreover, we introduce an approximation scheme in Section 4.2.3 so that the computational cost of testing procedure could be largely reduced. The simulation performance is illustrated in Section 4.3 and then we apply the method with two real data sets in Section 4.4 and Section 4.5. Lastly, Section 4.6 concludes this chapter.

4.2 Self-Normalized Test for Correlation Matrix Breaks

To illustrate the idea, let's assume we have a sequence of P-variate random vectors, $X_t = (X_{1,t}, X_{2,t}, ..., X_{P,t})_{t=1}^{T}$ with correlation matrix $R_t = (\rho_t^{ij})_{1 \leq i,j \leq P}$, where

$$\rho_t^{ij} = \frac{Cov(X_{i,t}, X_{j,t})}{\sqrt{Var(X_{i,t})Var(X_{j,t})}}.$$

We are interested in testing the correlation matrix constancy during the observation period, so the null hypothesis is

$$H_0 : \rho_1^{ij} = ... = \rho_T^{ij} \text{ for all i,j,}$$

against the alternative hypothesis

$$H_1 : \rho_1^{ij} = ... = \rho_{t_0}^{ij} \neq \rho_{t_0+1}^{ij} = ... = \rho_T^{ij} \text{ for some i,j,}$$

for some correlation break time $t_0 \in \{1, ..., T-1\}$. Choi and Shin (2020a) proposed a self-normalized correlation matrix break test extended from Shao and Zhang (2010). The test is

$$CS_T = \sup_{z\in[0,1]} \frac{z^2(1-z)^2\{\sum_{1\le i<j\le P}(\hat{\rho}^{ij}_{1,[Tz]} - \hat{\rho}^{ij}_{[Tz]+1,T})\}^2}{\frac{1}{T}\sum_{t=1}^{[Tz]}\{\sum_{1\le i<j\le P}\frac{t}{T}(\hat{\rho}^{ij}_{1,t} - \hat{\rho}^{ij}_{1,[Tz]})\}^2 + \frac{1}{T}\sum_{t=[Tz]+1}^{[T]}\{\sum_{1\le i<j\le P}(1-\frac{t}{T})(\hat{\rho}^{ij}_{t,T} - \hat{\rho}^{ij}_{[Tz]+1,T})\}^2}$$

where $\hat{\rho}^{ij}_{l,k}$ is the sample correlation estimated from the the paired sample $\{(X_{it}, X_{jt}), t = l, l+1, ..., k\}$. The asymptotic null distribution of CS_T is established by Choi and Shin (2020a).

Theorem 4.2.1 (Choi and Shin (2020a), Theorem 3.3)**.** *With certain assumptions, under H_0, with fixed P and $T \to \infty$,*

$$CS_T \xrightarrow{D} \sup_{z\in[0,1]} \frac{[B_0(z)]^2}{\int_0^z [B_0(s) - \frac{s}{z}B_0(z)]^2 ds + \int_z^1 [B_0(1) - B_0(s) - \frac{1-s}{1-z}\{B_0(1) - B_(z)\}]^2 ds} \tag{4.1}$$

CS_T has the same limiting distribution as G_n (Eq 1.2) mentioned in Chapter 1.3. The limiting distribution is free of P and typical cut-off values are $q_{0.01} = 68.6$, $q_{0.05} = 40.1$, $q_{0.1} = 29.6$. CS_T is exempted from estimating the covariance matrix and has capability of handling high dimensional scenarios. However, it is mainly designed against single correlation matrix break and suffers sever power loss in multiple breaks situations. This can be shown in our simulation results in Section 4.3. We would like to apply the framework from Zhang and Lavitas (2018) and propose an unsupervised self-normalized test against multiple breaks.

4.2.1 Proposed Methodology

Our test is tailored to correlation matrix break test against multiple change points as the alternative. We consider the problem of testing the null hypothesis

$$H_0 : \rho_1^{ij} = ... = \rho_T^{ij} \text{ for all i,j,}$$

versus the alternative that there exist $M \geq 1$ breaks points $k_0 = 1 < k_1 < ... < k_M < T = k_{M+1}$ such that

$$\rho_{t_0}^{ij} \neq \rho_{t_0+1}^{ij},\ t_0 \in \{k_1, ..., k_M\}$$

and $\rho_{t_0}^{ij} = \rho_{t_0+1}^{ij}$ otherwise. Choi and Shin (2020a) mentioned that their method could also be extended for multiple break alternatives using the spirit of scan method. They split the sample size T into L subsamples as $A = \{t : t \in (\frac{(l-1)T}{L}, \frac{lT}{L}], l = 1, ..., L\}$ and $(L-1)$ subsamples as $B = \{t : t \in (\frac{T}{2L} + \frac{(l-1)T}{L}, \frac{T}{2L} + \frac{lT}{L}], l = 1, ..., L-1\}$. The set of samples corresponding to $A \cup B$ is similar to a set of moving window samples. Then, the subsample extension of CS_T is $CS_T^S = \max\{\max_{1\leq l\leq L} CS_{lT}^A, \max_{1\leq l\leq L-1} CS_{lT}^B\}$. The test CS_T^S rejects H_0 at level α if $CS_T^S > q_{L\frac{\alpha}{2}}$, where $q_{L\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$-quantile of $\max_{1\leq l\leq L} CS_{(l)}$ and $CS_{(l)}, l = 1, ..., L$ are iid from the limiting distribution in Eq (4.1). Bonferroni-type adjusted p-values of $CS_T^S = q$, $CS_{lT}^A = q$ or $CS_{lT}^B = q$ can be approximated by $2pr\{\max_{1\leq l\leq L} CS_{(l)} > q\}$ by Choi and Shin (2020b). Compared with CS_T, this test largely improve the power when testing against multiple breaks. However, the performance highly depends on the choice of L.

We would like to propose a test that is powerful against both single and multiple breaks alternatives and is free of nuisance parameter. Let $\hat{\theta}_{j_1,j_2} = \sum_{1\leq i<j\leq p} \hat{\rho}_{j_1,j_2}^{ij}$. Then, calculate the forward recursive statistics as

$$D_{T,f}(j_1, j_2, j_3) = \frac{(j_2 - j_1 + 1)(j_3 - j_2)}{(j_3 - j_1 + 1)^{3/2}}(\hat{\theta}_{j_1,j_2} - \hat{\theta}_{j_2+1,j_3}),$$

$$L_{T,f}(j_1, j_2, j_3) = \sum_{i=j_1}^{j_2} \frac{(i - j_1 + 1)^2 (j_2 - i)^2}{(j_3 - j_1 + 1)^2 (j_2 - j_1 + 1)^2} (\hat{\theta}_{j_1,i} - \hat{\theta}_{i+1,j_2})^2,$$

$$R_{T,f}(j_1, j_2, j_3) = \sum_{i=j_2}^{j_3} \frac{(i - 1 - j_2)^2 (j_3 - i + 1)^2}{(j_3 - j_1 + 1)^2 (j_3 - j_2)^2} (\hat{\theta}_{i,j_3} - \hat{\theta}_{j_2+1,i-1})^2,$$

and backward recursive statistics as

$$D_{T,b}(j_1, j_2, j_3) = \frac{(j_2 - j_1)(j_3 - j_2 + 1)}{(j_3 - j_1 + 1)^{3/2}} (\hat{\theta}_{j_2,j_3} - \hat{\theta}_{j_1,j_2-1}),$$

$$L_{T,b}(j_1, j_2, j_3) = \sum_{i=j_1}^{j_2-1} \frac{(i - j_1 + 1)^2 (j_2 - 1 - i)^2}{(j_3 - j_1 + 1)^2 (j_2 - j_1)^2} (\hat{\theta}_{j_1,i} - \hat{\theta}_{i+1,j_2-1})^2,$$

$$R_{T,b}(j_1, j_2, j_3) = \sum_{i=j_2}^{j_3} \frac{(i - j_2)^2 (j_3 - i + 1)^2}{(j_3 - j_1 + 1)^2 (j_3 - j_2 + 1)^2} (\hat{\theta}_{i,j_3} - \hat{\theta}_{j_2,i-1})^2.$$

Let

$$\Xi_{T,f}(j_1, j_2, j_3) = L_{T,f}(j_1, j_2, j_3) + R_{T,f}(j_1, j_2, j_3),$$

and

$$\Xi_{T,b}(j_1, j_2, j_3) = L_{T,b}(j_1, j_2, j_3) + R_{T,b}(j_1, j_2, j_3).$$

Then, the test statistic is given by

$$\begin{aligned} U_T = & \max_{(l_1,l_2) \in \Omega_T(\epsilon)} D_{T,f}(1, l_1, l_2)' \Xi_{T,f}(1, l_1, l_2)^{-1} D_{T,f}(1, l_1, l_2) \\ & + \max_{(m_1,m_2) \in \Omega_T(\epsilon)} D_{T,b}(m_1, m_2, T)' \Xi_{T,b}(m_1, m_2, T)^{-1} D_{T,b}(m_1, m_2, T), \end{aligned}$$

where $\Omega(\epsilon) = \{(t_1, t_2) : \epsilon \leq t_1 < t_2 \leq 1 - \epsilon, t_2 - t_1 \geq \epsilon\}$ and $\Omega_T(\epsilon) = \{(\lfloor T t_1 \rfloor, \lfloor T t_2 \rfloor) : (t_1, t_2) \in \Omega(\epsilon)\}$.

4.2.2 Theoretical Properties

Unlike the scanning test by Choi and Shin (2020a), which detects change points by scanning samples with a moving window according to a prespecified window width,

this test is unsupervised and detects change points by recursive scanning. The following theorem provides the asymptotic property of U_T under the null and alternative hypothesis.

Theorem 4.2.2 (Zhang and Lavitas (2018), Theorem 3)**.** *Under certain conditions, (i) under the null hypothesis, we have*

$$U_T \xrightarrow{D} U(B) \coloneqq \sup_{(r_1,r_2)\in\Omega(\epsilon)} D(B,0,r_1,r_2)'\Xi(B,0,r_1,r_2)^{-1}D(B,0,r_1,r_2) + \sup_{(s_1,s_2)\in\Omega(\epsilon)} D(B,s_1,s_2,1)'\Xi(B,s_1,s_2,1)^{-1}D(B,s_1,s_2,1)$$

where

$$D(B,t_1,t_2,t_3) = \frac{1}{\sqrt{t_3-t_1}}\left[B(t_2)-B(t_1)-\frac{t_2-t_1}{t_3-t_1}\{B(t_3)-B(t_1)\}\right],$$

$$\Xi(B,t_1,t_2,t_3) = \frac{1}{(t_3-t_1)^2}\left(\int_{t_1}^{t_2}\left[B(s)-B(t_1)-\frac{s-t_1}{t_2-t_1}\{B(t_2)-B(t_1)\}\right]^2 ds + \int_{t_2}^{t_3}\left[B(t_3)-B(s)-\frac{t_3-s}{t_3-t_2}\{B(t_3)-B(t_2)\}\right]^2 ds\right);$$

and (ii) under the alternative with $\min_{0\le i\le M}|k_{i+1}-k_i|/T>\epsilon$ *and* $\theta_{k_i+1}-\theta_{k_i}=T^{-1/2}C_i$ *for some* $C_i\neq 0, i=1,...,M$*, we have* $U_T\to\infty$ *in probability if* $\min_{1\le i\le M}|C_i|\to\infty$*.*

Hence, we reject H_0 if $U_T > q_\alpha$, where q_α is the α-quantile of $U(B)$. Typical simulated critical values of $U(B)$ can be found in Zhang and Lavitas (2018) as $q_{0.1} = 142.237$, $q_{0.05} = 171$ and $q_{0.01} = 246.437$.

4.2.3 An Approximation Scheme

The test statistic U_T involves a maximum of $O(T^2)$ terms. It is independent of the total number of change points, so the testing procedure and its computational cost remains the same regardless the numbers of change points. To further reduce the computational cost, Zhang and Lavitas (2018) implemented a grid approximation

scheme, with which the maximum only needs to be searched over $O(T)$ partitions. Let $G_\epsilon = \{(1+k\epsilon)/2, k \in \mathbb{Z}\} \cap [0,1]$ be a grid on the unit interval, which is symmetric around the middle point. Let $G_{\epsilon,T,f} = \{(\lfloor Tt_1 \rfloor, \lfloor Tt_2 \rfloor) : (t_1, t_2) \in ([0,1] \times G_\epsilon) \cap \Omega(\epsilon)\}$ and $G_{\epsilon,T,b} = \{(\lfloor Tt_1 \rfloor, \lfloor Tt_2 \rfloor) : (t_1, t_2) \in (G_\epsilon \times [0,1]) \cap \Omega(\epsilon)\}$. The test statistic is

$$U_T^\diamond = \max_{(l_1,l_2)\in G_{\epsilon,T,f}} D_{T,f}(1,l_1,l_2)' \Xi_{T,f}(1,l_1,l_2)^{-1} D_{T,f}(1,l_1,l_2)$$
$$+ \max_{(m_1,m_2)\in G_{\epsilon,T,b}} D_{T,b}(m_1,m_2,T)' \Xi_{T,b}(m_1,m_2,T)^{-1} D_{T,b}(m_1,m_2,T).$$

The asymptotic distribution of $U_T^\diamond$ is given in the following theorem along with its consistency.

Theorem 4.2.3 (Zhang and Lavitas (2018), Proposition 2)**.** *With certain assumptions, under the null hypothesis, we have*

$$U_T^\diamond \xrightarrow{D} U^\diamond(B)$$

where

$$U^\diamond(B) = \sup_{(r_1,r_2)\in([0,1]\times G_\epsilon)\cap\Omega(\epsilon)} D(B,0,r_1,r_2)' \Xi(B,0,r_1,r_2)^{-1} D(B,0,r_1,r_2)$$
$$+ \sup_{(s_1,s_2)\in(G_\epsilon\times[0,1])\cap\Omega(\epsilon)} D(B,s_1,s_2,1)' \Xi(B,s_1,s_2,1)^{-1} D(B,s_1,s_2,1).$$

Under the alternative with $\min_{0\le i\le M} |k_{i+1} - k_i|/T > 3\epsilon/2$ *and* $\theta_{k_i+1} - \theta_{k_i} = T^{-1/2} C_i$ *for some* $C_i \neq 0, i = 1, ..., M$*, we have* $U_T^\diamond \to \infty$ *in probability if* $\min_{1\le i\le M} |C_i| \to \infty$.

Hence, we reject H_0 if $U_T^\diamond > q_\alpha$, where q_α is the α-quantile of $U^\diamond(B)$. The cut-off values can be obtained by simulating the distribution of $U^\diamond(B)$. Typical cut-off values are $q_{0.1} = 118.379$, $q_{0.05} = 144.723$ and $q_{0.01} = 208.462$. Zhang and Lavitas (2018) showed with simulation results that the above grid approximation scheme gives reasonably good performance in terms of both size and power.

Remark **(Trimming parameter ϵ)** As mentioned before, we don't estimate or tune any nuisance parameter in the process. However, the procedure involves a

trimming parameter ϵ. This trimming parameter accounts for the limiting distribution and its approximation, which is different from the smoothing parameter in nonparametric estimation. This effect was discussed by Zhou and Shao (2013) and they suggested a rule of thumb choice of $\epsilon = 0.1$.

4.3 Simulation Results

In this section, we would like to compare the finite performance of our method U_T with CS_T and CS_T^S proposed in Choi and Shin (2020a). In their paper, they claimed that their method has stabler sizes than the previous test in Wied (2017) as P gets larger under multiple dependent structure and has reasonable power against single break while no power for up-down (or down-up) type double breaks. Here we use the same setting as Choi and Shin (2020a) and the results show that we obtained significant increase in power with a little increase in size distortion.

The data was generated from a vector autoregression of order 1,

$$X_t = \phi X_{t-1} + a_t. \tag{4.2}$$

where $X_t = (X_{1t}, ..., X_{Pt})'$ and the error $a_t = (a_{1t}, ..., a_{Pt})'$ has zero mean and covariance matrix $\Sigma_t = \Xi_t R_t \Xi_t$ with $\Xi_t = diag(\sigma_{1t}, ..., \sigma_{Pt})$ and correlation matrix R_t. Serial independence and dependence are considered by $\phi \in \{0, 0.8\}$. We consider homoscedastic and conditionally heteroscedastic errors a_t given by

$$a_t = \Sigma_t^{1/2} e_t, \sigma_{it}^2 = (1 - \alpha_1 - \beta_1) + \alpha_1 a_{i,t-1}^2 + \beta_1 \sigma_{i,t-1}^2, i = 1, ..., P,$$

$(\alpha_1, \beta_1) \in \{(0, 0), (0.1, 0.89)\}$, where $e_t = (e_{1t}, ..., e_{Pt})'$ is a vector of independent standard normal errors. Size comparison is made for the correlation matrix $R_t =$

$R_0 = \begin{pmatrix} 1 & \rho_0 & \cdots & \rho_0 \\ & 1 & \cdots & \rho_0 \\ & & \ddots & \\ & & & 1 \end{pmatrix}$, $\rho_0 \in \{0, 0.5\}$. We compare the empirical size, empirical power and empirical size-adjusted power based on 1000 repetitions with $(P, T) \in \{50, 200\} * \{50, 200\}$ among our test U_T, Choi and Shin (2020a) test CS_T and Choi and Shin (2020a) scan test $CS_T^S(L)$ with multiple choices of window width L.

The results of sizes comparison with homoscedastic and conditionally heteroscedastic errors are given in Table 4.1 and Table 4.2 respectively. CS_T is designed specifically against single break and has the best performance in sizes comparison in terms of closest empirical sizes to 5% with both homoscedastic and conditionally heteroscedastic errors. The performance of CS_T^S largely depends on the choice of L. When $L = 2$ and $L = 3$, the performance are acceptable, while $L = 5$ and $L = 8$, the empirical size may be far away from the test level and the results generally present a large size distortion. The problem is that there is no clear guidance of choosing L. Our test U_T seems to deliver more reasonable performance under various settings than CS_T^S. It has size distortion compared to CS_T, but we will next show that CS_T can suffer serious power loss against multiple breaks.

Homoscedasticity										
ϕ	ρ	P	T	U_T	CS_T	CS_T^S (L=2)	CS_T^S (L=3)	CS_T^S (L=4)	CS_T^S (L=5)	CS_T^S (L=8)
0	0	50	50	0.091	0.052	0.077	0.169	0.265	0.451	0.954
0	0.5	50	200	0.070	0.060	0.057	0.100	0.113	0.117	0.174

Table 4.1: Empirical sizes of the 5% level tests with homoscedastic errors constructed by (i) the proposed unsupervised test (U_T); (ii) the self-normalized test in (Choi and Shin, 2020a) (CS_T); and (iii) the scanning self-normalized test in (Choi and Shin, 2020a) with different values of window width ($CS_T^S(L)$).

Conditional Heteroscedasticity										
ϕ	ρ	P	T	U_T	CS_T	CS_T^S (L=2)	CS_T^S (L=3)	CS_T^S (L=4)	CS_T^S (L=5)	CS_T^S (L=8)
0	0.5	50	200	0.067	0.059	0.053	0.076	0.121	0.124	0.186
0.8	0.5	50	50	0.176	0.065	0.077	0.147	0.267	0.412	0.975

Table 4.2: Empirical sizes of the 5% level test with conditional heteroscedastic errors constructed by (i) the proposed unsupervised test (U_T); (ii) the self-normalized test in (Choi and Shin, 2020a) (CS_T); and (iii) the scanning self-normalized test in (Choi and Shin, 2020a) with different values of window width ($CS_T^S(L)$).

We next compare powers of the tests CS_T, CS_T^S and U_T. The data is generated from model (4.2) with the correlation matrix $R_t = R_{01}I(t \leq t_0) + R_{02}I(t > t_0)$ having a correlation change at $t_0 \in \{T/10, T/4\}$ and the values R_{01} and R_{02} with equal correlation $\rho_0 = 0$ and 0.2, respectively. We further investigate the power under double breaks. For double breaks, we consider $R_t = R_{01}I(1 \leq t \leq \frac{T}{3}) + R_{02}I(\frac{T}{3} < t \leq \frac{2T}{3}) + R_{03}I(\frac{2T}{3} < t \leq T)$. The values of ρ_0 in R_{01}, R_{02}, R_{03} are 0.2, 0.7, 0.2 respectively. $(P, T) \in \{4, 10\} * \{200, 2000\}$. Table 4.3 gives the result of empirical powers with single break and Table 4.4 presents the result with double breaks.

Single break ($\rho : 0 \to 0.2$)									
P	T	Break	U_T	CS_T	CS_T^S (L=2)	CS_T^S (L=3)	CS_T^S (L=4)	CS_T^S (L=5)	CS_T^S (L=8)
4	200	T/10	0.131	0.067	0.089	0.130	0.132	0.143	0.176
10	200	T/4	0.811	0.653	0.676	0.504	0.360	0.319	0.237

Table 4.3: Empirical powers of the 5% level tests against single breaks constructed by (i) the proposed unsupervised test (U_T); (ii) the self-normalized test in (Choi and Shin, 2020a) (CS_T); and (iii) the scanning self-normalized test in (Choi and Shin, 2020a) with different values of window width ($CS_T^S(L)$).

Double breaks ($\rho : 0.2 \to 0.7 \to 0.2$)								
P	T	U_T	CS_T	CS_T^S (L=2)	CS_T^S (L=3)	CS_T^S (L=4)	CS_T^S (L=5)	CS_T^S (L=8)
4	200	0.986	0.002	0.900	0.877	0.763	0.602	0.418
10	200	1	0	0.995	0.985	0.920	0.824	0.561

Table 4.4: Empirical powers of the 5% level tests against double breaks constructed by (i) the proposed unsupervised test (U_T); (ii) the self-normalized test in (Choi and Shin, 2020a) (CS_T); and (iii) the scanning self-normalized test in (Choi and Shin, 2020a) with different values of window width ($CS_T^S(L)$).

From Table 4.3 and Table 4.4, we can see that CS_T presents some power under single breaks, but the power drop to 0 against double breaks. It indeed has good size performance under no break setting, but in real cases of no prior knowledge of number of breaks, the test result of CS_T may be misleading. Similar situations happen in CS_T^S results. When there is no break, CS_T^S could give sizes as good as 0.057 or as bad as 0.975, while with double breaks, CS_T^S could give powers as good as 0.995 or as bad as 0.418. No suggestion about choosing L will result in unreliable conclusions in real application. We will show that opposite conclusions could be drawn based on different choices of L in Section 4.4 and Section 4.5. In contrast, our method gives a steady power under multiple settings. When we have a break at $T/10$ with only 200 time points, it is pretty hard to detect the break, whereas our test still has comparable power. Other than that, U_T gives empirical power over 0.8 for different settings. This is a significant improvement from CS_T especially for testing double breaks. As mentioned before, U_T gives better power with a cost of size distortion, so we may compare the performance of empirical size-adjusted power (Table 4.5 and Table 4.6).

Table 4.5 and Table 4.6 show that U_T has the highest empirical size-adjusted

power under various settings compared with CS_T and CS_T^S with different values of L, so it is able to provide more reliable results in real applications when we have no information of data structure and number of correlation matrix breaks.

Single break ($\rho : 0 \to 0.2$)									
P	T	Break	U_T	CS_T	CS_T^S (L=2)	CS_T^S (L=3)	CS_T^S (L=4)	CS_T^S (L=5)	CS_T^S (L=8)
4	200	T/10	0.127	0.077	0.081	0.084	0.077	0.100	0.089
10	200	T/4	0.793	0.627	0.709	0.461	0.272	0.229	0.139

Table 4.5: Empirical size-adjusted powers of the 5% level tests against single breaks constructed by (i) the proposed unsupervised test (U_T); (ii) the self-normalized test in (Choi and Shin, 2020a) (CS_T); and (iii) the scanning self-normalized test in (Choi and Shin, 2020a) with different values of window width ($CS_T^S(L)$).

Double breaks ($\rho : 0.2 \to 0.7 \to 0.2$)								
P	T	U_T	CS_T	CS_T^S (L=2)	CS_T^S (L=3)	CS_T^S (L=4)	CS_T^S (L=5)	CS_T^S (L=8)
4	200	0.967	0	0.908	0.809	0.634	0.395	0.234
10	200	0.998	0	0.996	0.972	0.873	0.756	0.353

Table 4.6: Empirical size-adjusted powers of the 5% level tests against double breaks constructed by (i) the proposed unsupervised test (U_T); (ii) the self-normalized test in (Choi and Shin, 2020a) (CS_T); and (iii) the scanning self-normalized test in (Choi and Shin, 2020a) with different values of window width ($CS_T^S(L)$).

4.4 Application to a Stock Price Data

We would like to show the performance of our test with a real data set of the stock log returns of the top 10 companies of the NASDAQ 100 index on August 1, 2019. The company names are listed in Table 4.7. The period of stock returns considered

here is 06/01/2010 - 06/30/2019 ($T = 2286$), see Figure 4·1. The stock prices can be downloaded from Yahoo Finance (`https://finance.yahoo.com/`). The data presents weak serial dependence and strong conditional heteroscedasticity (Choi and Shin, 2020a).

Microsoft Corp (MSFT)	Apple Inc(AAPL)	Amazon.com Inc (AMZN)
Alphabet Inc (GOOG)	Cisco Systems Inc (CSCO)	Intel Corp (INTC)
Comcast Corp (CMCSA)	PepsiCo Inc (PEP)	Adobe Inc (ADBE)
Netflix Inc(NFLX)		

Table 4.7: The list of company names (and tickers) in the stock price data set.

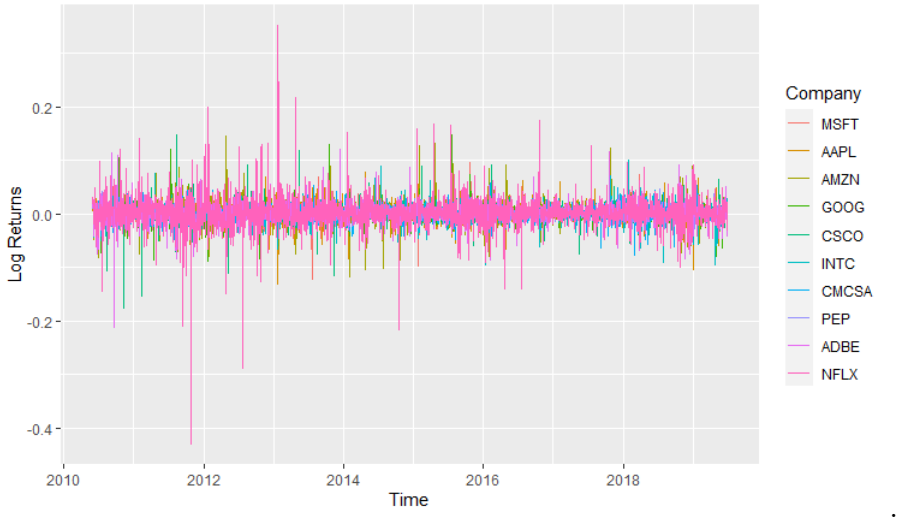

Figure 4·1: Log returns of stock prices for the top 10 companies of NASDAQ 100 index on August 1, 2019 during the period of 06/01/2010 - 06/30/2019.

We apply all the three tests, U_T, CS_T and CS_T^S, to the data. The test statistics and cut-off values at 5% significance level are shown in Table 4.8. CS_T concludes no correlation matrix break exists for stock returns. Since CS_T is designed for single break alternative, the non-rejection conclusion may be a consequence of more than one

break. The scanning test results are uninformative. CS_T^S with $L = 4$ says no break exists, while CS_T^S with $L = 5$ gets the opposite conclusion claiming the existence of correlation matrix break. This shows that their results highly depends on the choice of L. Our method reports a test statistic that is slightly larger than the cut-off value, so we conclude that correlation matrix breaks exist, which agrees with $CS_T^S(L = 5)$.

	U_T	CS_T	CS_T^S(L=4)	CS_T^S(L=5)
Test statistic	171.196	18.915	41.913	106.996
Cut-off value	171	40.1	76.6	84.6

Table 4.8: Test statistics and the corresponding cut-off values at 5% significance level for the log return of stock price data.

The data spans a long time of 9 years of 109 months involving European debt crisis (June/2012) and Greek debt crisis (June/2015). European events induce correlation matrix change for American stock returns, so it is likely that more than one break exist. Here, we would like to use time lagged cross correlations plots to show the correlation patterns between two stocks, for example Amazon and Cisco. Time lagged cross correlations plot is a great way to visualize the fine-grained dynamic interaction between two signals such as the leader-follower relationship and how they shift over time. Figure 4·2 splits the time series into 20 even chunks and computes the cross correlation in each window. The figure shows that in the seventh or eighth window (row), blue mainly occupies the row. This suggests Cisco mainly leads the interaction. However, Amazon leads the interaction in some other windows, like the thirteenth and sixteenth row. Figure 4·3 computes the time lagged cross correlation continuously resulting in a smoother plot. The plot shows that the interaction is first leaded by Cisco. Then Amazon and Cisco lead the interaction alternatively. In the end, Amazon mainly leads the interaction.

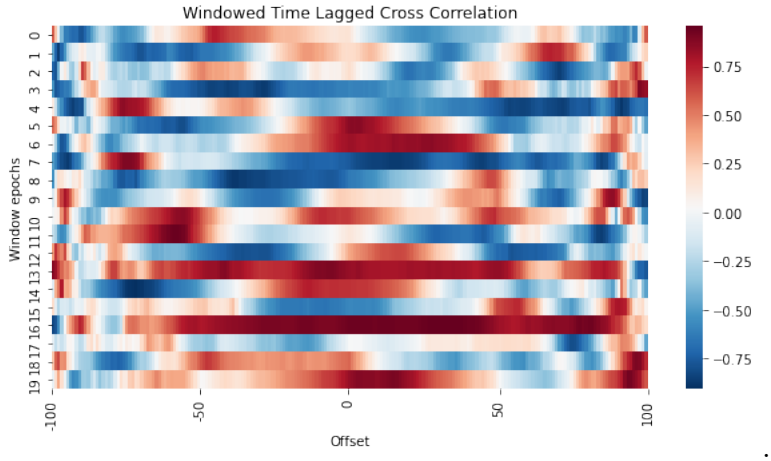

Figure 4·2: Windowed time lagged cross correlation between log returns of Amazon and Cisco.

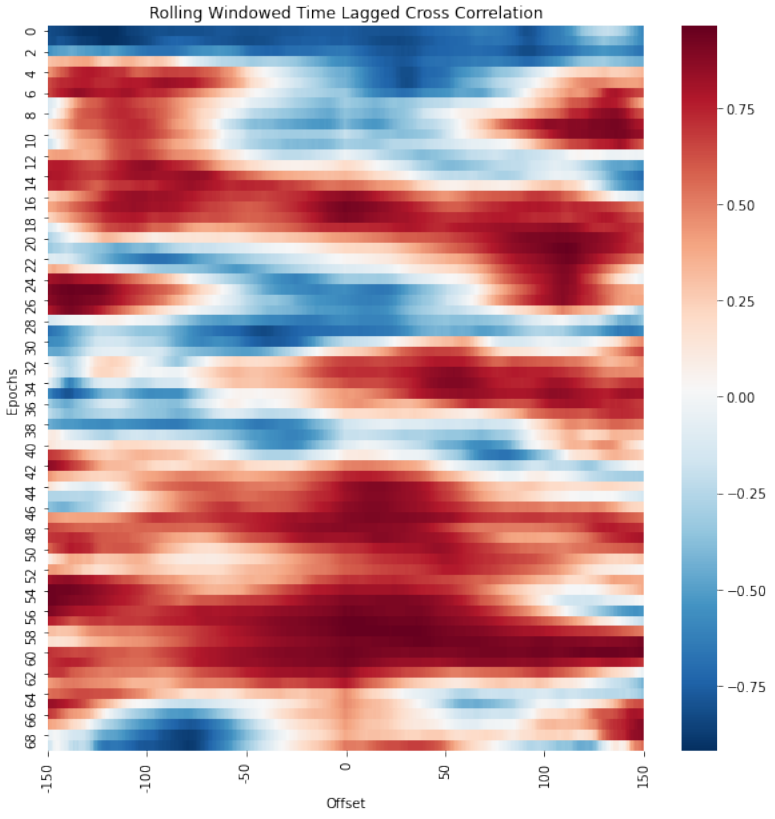

Figure 4·3: Rolling windowed time lagged cross correlation between log returns of Amazon and Cisco.

In summary, the time lagged cross correlation plots show that the correlation structure between a pair of stocks indeed changes during the selected period and the correlation structure among all ten stocks is also likely to change during the European and Greek debt crisis. Our test is able to detect the break, whereas CS_T couldn't and CS_T^S can discover the break only with valid choice of L.

4.5 Application to a Option Data

We would like to show another real data application, which contains five volatility indexes for options on individual equities. The company names are listed in Table 4.9. The period of volatility indexes considered here is 06/01/2010 - 06/30/2019 ($T = 2286$), see Figure 4·4. The volatility indexes can be downloaded from Chicago Board Options Exchange (CBOE) website (`http://www.cboe.com/products/vix-index-volatility/volatility-indexes`). The data presents strong serial dependence and strong conditional heteroscedasticity (Choi and Shin, 2020a).

Company name:	Amazon	Apple	Goldman Sachs	Google	IBM
Company ticker:	VXAZN	VXAPL	VXGS	VXGOG	VXIBM

Table 4.9: The list of company names and tickers in the option volatility index data set.

Again, we apply all the three tests and Table 4.10 presents the test statistics and corresponding cut-off values at 5% significance level. CS_T concludes no correlation matrix break exists for volatility index data. $CS_T^S(L = 4)$ points out the existence of breaks, while $CS_T^S(L = 5)$ has a test statistic that is slightly smaller than the cut-off value and concludes that correlation matrix break doesn't exist. Our method reports a test statistic that is much larger than the cut-off value and concludes that correlation matrix breaks exist. This result agrees with $CS_T^S(L = 4)$.

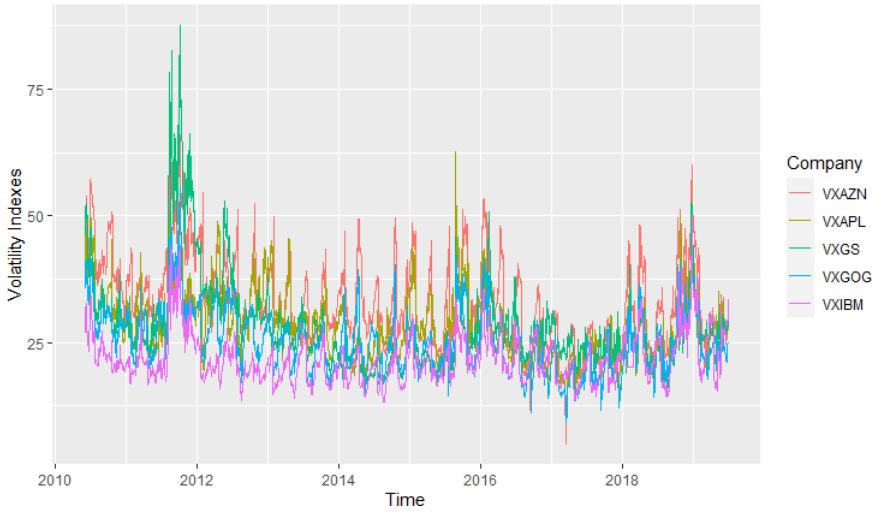

Figure 4·4: Volatility index of 5 options during the period of 06/01/2010 - 06/30/2019.

	U_T	CS_T	CS_T^S(L=4)	CS_T^S(L=5)
Test statistic	304.944	3.403	199.855	84.314
Cut-off value	171	40.1	76.6	84.6

Table 4.10: Test statistics and the corresponding cut-off values at 5% significance level for the volatility index data.

As we explained in the previous application, the selected period involves European and Greek debt, so it is likely to have more than one break. From the windowed and rolling windowed time lagged cross correlation plots between Amazon and Google (Figure 4·5 and Figure 4·6), we can see that the alternation frequency of leading interaction changes with time. The leadership changes more frequently between ten to thirty-two windows and forty-two to fifty-eight windows.

Therefore, it is very likely that correlation structure of volatility index among the five options changed during the period. Our method is able to detect the break while CS_T cannot. Without a valid choice of L, the conclusion of CS_T^S could be unreliable. In real cases, when we have no prior information about the number of breaks, it may

be hard to draw a conclusion based on the results of CS_T and CS_T^S.

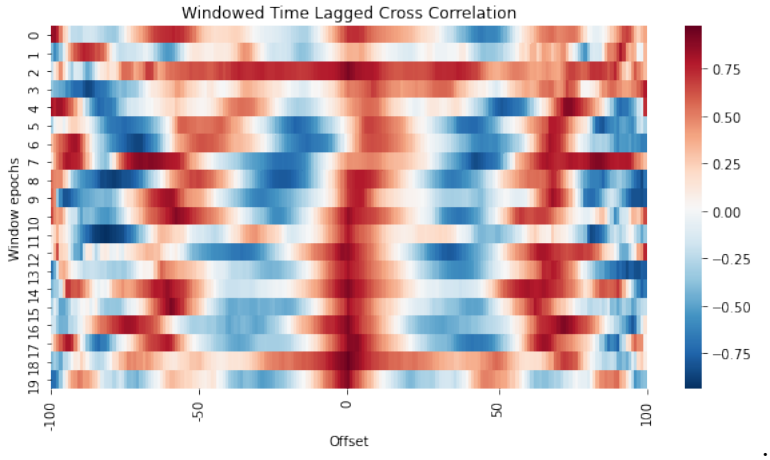

Figure 4·5: Windowed time lagged cross correlation between volatility index of Amazon and Google.

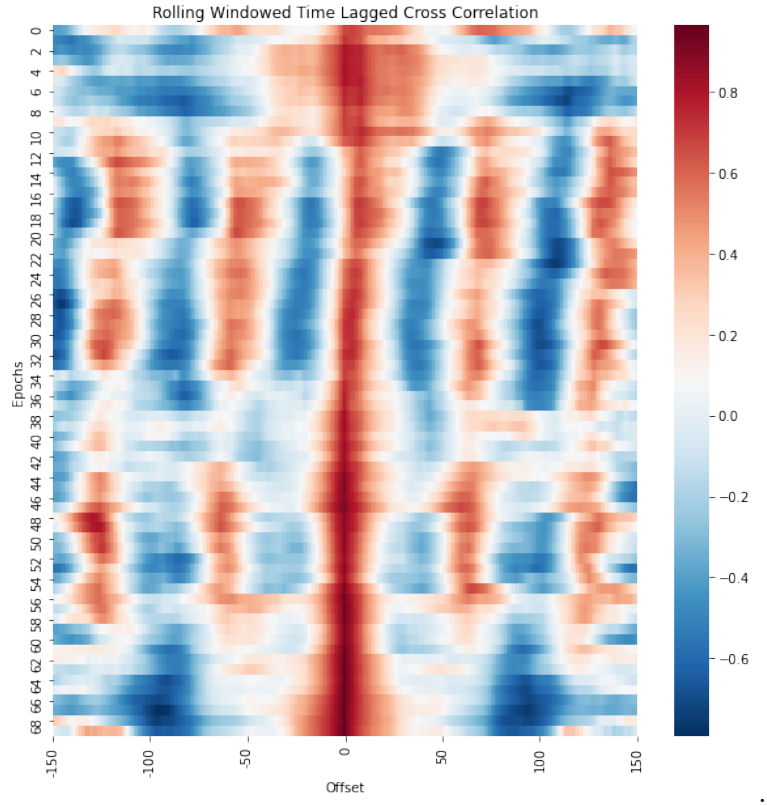

Figure 4·6: Rolling windowed time lagged cross correlation between volatility index of Amazon and Google.

4.6 Conclusion

In this chapter, we proposed a contrast-based unsupervised correlation matrix break test based on self-normalization. This method can test against multiple breaks without knowing the number of breaks as the prior information. The simulation shows the method is more powerful than the existing methods especially against multiple breaks. The application shows that the method is able to detect the correlation matrix break in stock log returns and volatility indexes and outperformed Choi and Shin (2020a) since their conclusion is sensitive to the choice of L.

Chapter 5

Conclusions

5.1 Summary of the Thesis

The contributions of this thesis to self-normalization properties and self-normalized statistical inference come in several aspects. We start from analyzing a generalization of traditional self-normalization. As mentioned in the introduction, the original self-normalizer uses only recursive estimators of forward direction. In Chapter 2, we would like to consider both the forward and backward estimators and propose a data-driven weight to balance the forward and backward estimators. Then we move on to self-normalized statistical inference involving constructing simultaneous confidence regions for high dimensional time series and hypothesis test for correlation matrix breaks. Self-normalization is originally developed for univariate or low-dimensional multivariate time series, and its direct application to high-dimensional time series can fail easily due to singularities or substantial size distortions. In Chapter 3, we design a self-normalized statistic with maximal modulus and construct simultaneous confidence regions for high-dimensional time series with much reduced volumes. Chapter 4 describes a new contrast-based unsupervised test for correlation matrix break with self-normalization.

In Chapter 2, we focus on a data-driven weight that corresponds to confidence intervals with minimal lengths. We study the asymptotic behavior of such a data-driven weight choice, and find an interesting dichotomy between linear and nonlinear quan-

tities. We apply the von Mises expansion (von Mises, 1947) with Gâteaux derivatives to explain the dichotomous asymptotic behavior. Another interesting phenomenon is that, for nonlinear quantities, the data-driven weight typically distributes over an uncountable set in finite-sample problems but in the limit, it converges to a discrete distribution with a finite support. The phenomenon are built on technical conditions that can be verified with smooth function models. We then compare the average confidence intervals length and empirical coverage probability constructed for different quantities using our method and literature methods in simulations. Our method indeed provide the minimal interval lengths with the cost of larger size distortion.

Chapter 3 aims at proposing new approaches of using self-normalization to make statistical inference of general quantities of high-dimensional time series. To achieve this goal, we first use the parabolic cylinder functions to characterize the tail behavior of self-normalized distributions, and then develop an asymptotic theory on the maximal modulus of self-normalized statistics. In addition, we incorporate regularization technique into the self-normalization paradigm and further propose a thresholded self-normalization method, which is capable of taking advantage of data with sparse signals to produce simultaneous confidence regions with much reduced volumes. Monte Carlo simulations are conducted to examine the finite-sample performance of the proposed methods and compare with existing self-normalized methods. A real data application is also provided to illustrate the developed results.

In Chapter 4, we propose an unsupervised self-normalization test for correlation matrix constancy. The test statistic is contrast-based and directly compares the correlation before and after the hypothesized change point. This unsupervised test doesn't require prespecifying the number of change points and outperforms the existing tests when more than one change points exist. The framework is extended from the unsupervised self-normalized change-point test by Zhang and Lavitas (2018). They

established the asymptotic distribution and theoretical consistency of the test. Here, we compare and discuss the performance under multiple simulation settings and with two real data applications.

5.2 Potential Future Research Topics

In Chapter 3, we introduced that Wang and Shao (2020) designed a test against dense alternatives, while our test targets at sparse alternatives. In some cases, we may not have the prior knowledge on whether the alternative is sparse or not. It may be interesting to develop an adaptive test in the high-dimensional time series setting.

In Chapter 4, we borrow the theoretical results from (Zhang and Lavitas, 2018). It would be great to establish theoretical conditions and provides the asymptotic distribution and theoretical consistency tailored for correlation matrix break tests. The test procedure involves a trimming parameter ϵ. The performance may be further improved if we come up with a data-driven value for ϵ. As mentioned in Chapter 4, we extended the correlation matrix break tests framework from testing against one single break to multiple breaks. However, since we test breaks in terms of $\sum_{1\leq i<j\leq P}\rho^{ij}$, the test has no power against cancelling breaks. We may further consider vectorized test statistics to detect cancelling breaks.

We shall leave these topics for future research.

Appendix A

Appendix A: Theoretical Justification, Implication and Simulation for Chapter 2

Theoretical Justification

Proof. (Theorem 2.3.1) Let $\bar{X}_{i,j} = (j-i+1)^{-1}\sum_{k=i}^{j} X_k$ be the sample average of $X_i, \ldots, X_j$, $1 \le i \le j \le N$, then

$$\bar{X}_{i,j}^2 = (\bar{X}_{i,j} - \mu)^2 + 2\mu(\bar{X}_{i,j} - \mu) + \mu^2,$$

and thus by (2.8) and (2.9) we have

$$\begin{aligned} k(\hat{\theta}_{1,k} - \hat{\theta}_{1,N}) &= \sum_{i=1}^{k} X_i^2 - \frac{k}{N}\sum_{i=1}^{N} X_i^2 - k(\bar{X}_{1,k}^2 - \bar{X}_{1,N}^2) \\ &= S_{N,2}^{\circ}(k/N) - \frac{k}{N}S_{N,2}^{\circ}(1) - \left[\frac{\{S_{N,1}^{\circ}(k/N)\}^2}{k} - \frac{k}{N}\cdot\frac{\{S_{N,1}^{\circ}(1)\}^2}{N}\right] \\ &\quad -2\mu\left\{S_{N,1}^{\circ}(k/N) - \frac{k}{N}S_{N,1}^{\circ}(1)\right\} \end{aligned}$$

and

$$\begin{aligned} (N-k)(\hat{\theta}_{k+1,N} - \hat{\theta}_{1,N}) &= \sum_{i=k+1}^{N} X_i^2 - \frac{N-k}{N}\sum_{i=1}^{N} X_i^2 - (N-k)(\bar{X}_{k+1,N}^2 - \bar{X}_{1,N}^2) \\ &= S_{N,2}^{\circ}(1) - S_{N,2}^{\circ}(k/N) - \frac{N-k}{N}S_{N,2}^{\circ}(1) \\ &\quad - \left[\frac{\{S_{N,1}^{\circ}(1) - S_{N,1}^{\circ}(k/N)\}^2}{N-k} - \frac{N-k}{N}\cdot\frac{\{S_{N,1}^{\circ}(1)\}^2}{N}\right] \\ &\quad -2\mu\left\{S_{N,1}^{\circ}(1) - S_{N,1}^{\circ}(k/N) - \frac{N-k}{N}S_{N,1}^{\circ}(1)\right\} \end{aligned}$$

respectively. By adding the aforementioned two equations, we obtain that

$$k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})+(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N}) = \frac{\{S^{\circ}_{N,1}(1)\}^2}{N}-\frac{\{S^{\circ}_{N,1}(k/N)\}^2}{k} -\frac{\{S^{\circ}_{N,1}(1)-S^{\circ}_{N,1}(k/N)\}^2}{N-k}.$$

Let $B_{\Sigma}(t)=\{B_{\Sigma,1}(t),B_{\Sigma,2}(t)\}'$ be a centered Brownian motion with the same covariance structure as $\Sigma B(t)$ and write

$$\Psi(B_{\Sigma,1},t)=B_{\Sigma,1}(1)^2-\frac{B_{\Sigma,1}(t)^2}{t}-\frac{\{B_{\Sigma,1}(1)-B_{\Sigma,1}(t)\}^2}{1-t}, \tag{A.1}$$

then by assumption (IP_v) we have the convergence of the cumulative distribution function

$$\begin{aligned} & \mathrm{pr}\left[\sum_{k=1}^{N}k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})\{k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})+(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\}\le 0\right] \\ = & \mathrm{pr}\left[\frac{1}{N^{3/2}}\sum_{k=1}^{N}k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})\{k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})+(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\}\le 0\right] \\ \to & \mathrm{pr}\left(\int_0^1[B_{\Sigma,2}(t)-tB_{\Sigma,2}(1)-2\mu\{B_{\Sigma,1}(t)-tB_{\Sigma,1}(1)\}]\Psi(B_{\Sigma,1},t)dt\le 0\right). \end{aligned}$$

Similarly, note that $B_{\Sigma,2}(1)-B_{\Sigma,2}(t)-(1-t)B_{\Sigma,2}(1)=tB_{\Sigma,2}(1)-B_{\Sigma,2}(t)$ and $B_{\Sigma,1}(1)-B_{\Sigma,1}(t)-(1-t)B_{\Sigma,1}(1)=tB_{\Sigma,1}(1)-B_{\Sigma,1}(t)$, we have

$$\begin{aligned} & \mathrm{pr}\left[\sum_{k=1}^{N}(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\{k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})+(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\}\le 0\right] \\ = & \mathrm{pr}\left[\frac{1}{N^{\frac{3}{2}}}\sum_{k=1}^{N}(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\{k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})+(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\}\le 0\right] \\ \to & \mathrm{pr}\left(\int_0^1[tB_{\Sigma,2}(1)-B_{\Sigma,2}(t)-2\mu\{tB_{\Sigma,1}(1)-B_{\Sigma,1}(t)\}]\Psi(B_{\Sigma,1},t)dt\le 0\right). \end{aligned}$$

Since $B_\Sigma(t)$ has the same distribution as $-B_\Sigma(t)$, by (A.1) we have

$$\begin{aligned} & \text{pr}\left(\int_0^1 [B_{\Sigma,2}(t) - tB_{\Sigma,2}(1) - 2\mu\{B_{\Sigma,1}(t) - tB_{\Sigma,1}(1)\}]\Psi(B_{\Sigma,1},t)dt \le 0\right) \\ = & \text{pr}\left(\int_0^1 [-\{B_{\Sigma,2}(t) - tB_{\Sigma,2}(1)\} + 2\mu\{B_{\Sigma,1}(t) - tB_{\Sigma,1}(1)\}]\Psi(-B_{\Sigma,1},t)dt \le 0\right) \\ = & \text{pr}\left(\int_0^1 [B_{\Sigma,2}(t) - tB_{\Sigma,2}(1) - 2\mu\{B_{\Sigma,1}(t) - tB_{\Sigma,1}(1)\}]\Psi(B_{\Sigma,1},t)dt \ge 0\right), \end{aligned}$$

and thus

$$\text{pr}\left[\sum_{k=1}^{N} k(\hat\theta_{1,k} - \hat\theta_{1,N})\{k(\hat\theta_{1,k} - \hat\theta_{1,N}) + (N-k)(\hat\theta_{k+1,N} - \hat\theta_{1,N})\} \le 0\right] \to 1/2 \quad \text{(A.2)}$$

and

$$\text{pr}\left[\sum_{k=1}^{N} (N-k)(\hat\theta_{k+1,N} - \hat\theta_{1,N})\{k(\hat\theta_{1,k} - \hat\theta_{1,N}) + (N-k)(\hat\theta_{k+1,N} - \hat\theta_{1,N})\} \le 0\right] \to 1/2. \quad \text{(A.3)}$$

By viewing the generalized self-normalizer (2.5) as a function of w, one can show that the second order partial derivative satisfies

$$\inf_{w\in[0,1]} \frac{\partial^2 \Lambda_N(w)}{\partial w^2} \ge 0,$$

and thus for its minimizer $\hat w_N$ we have

$$\begin{aligned} \text{pr}(\hat w_N = 1) &= \text{pr}\left\{\left.\frac{\partial \Lambda_N(w)}{\partial w}\right|_{w=1} \le 0\right\} \\ &= \text{pr}\left[\sum_{k=1}^{N} k(\hat\theta_{1,k} - \hat\theta_{1,N})\{k(\hat\theta_{1,k} - \hat\theta_{1,N}) + (N-k)(\hat\theta_{k+1,N} - \hat\theta_{1,N})\} \le 0\right], \end{aligned}$$

which converges to 1/2 by (A.2). On the other hand, by a similar argument using (A.3), we can obtain the convergence

$$\text{pr}\left\{\left.\frac{\partial \Lambda_N(w)}{\partial w}\right|_{w=0} \ge 0\right\} \to 1/2.$$

Note that in this case,

$$1-\mathrm{pr}(\hat{w}_N=1)\geq \mathrm{pr}(\hat{w}_N=0)\geq \mathrm{pr}\left\{\left.\frac{\partial\Lambda_N(w)}{\partial w}\right|_{w=0}\geq 0,\ \left.\frac{\partial^2\Lambda_N(w)}{\partial w^2}\right|_{w=0}>0\right\},\ \text{(A.4)}$$

where

$$\mathrm{pr}\left\{\left.\frac{\partial^2\Lambda_N(w)}{\partial w^2}\right|_{w=0}>0\right\}\geq 1-\mathrm{pr}\left[\sum_{k=1}^{N}\{k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})+(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\}^2=0\right]$$
$$\to 1$$

because of the assumed positive definiteness of Σ. Therefore, by (A.4) and the sandwich theorem, we have $\mathrm{pr}(\hat{w}_N=0)\to 1/2$ and the result follows. □

Proof. (Theorem 2.3.2) For probabilistically linear parameters, the functional T is linear and thus

$$T(\hat{F}_{i,j})-T(F)=\frac{1}{j-i+1}\sum_{k=i}^{j}T(\delta_{X_k})-T(F)=\frac{1}{j-i+1}\sum_{k=i}^{j}\{T(\delta_{X_k})-T(F)\}.$$

Since

$$T\{(1-\epsilon)F+\epsilon\delta_x\}=(1-\epsilon)T(F)+\epsilon T(\delta_x),$$

we have

$$\varphi_1(x)=\left.\frac{\partial T\{(1-\epsilon)F+\epsilon\delta_x\}}{\partial\epsilon}\right|_{\epsilon=0}=T(\delta_x)-T(F),$$

and (i) follows. In this case, we have

$$k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})=\sum_{i=1}^{k}\varphi_1(Y_i)-\frac{k}{N}\sum_{i=1}^{N}\varphi_1(Y_i)$$

and

$$(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})=\sum_{i=k+1}^{N}\varphi_1(Y_i)-\frac{N-k}{N}\sum_{i=1}^{N}\varphi_1(Y_i),$$

based on which we can show that

$$\begin{aligned}\Lambda_N(w) &= \frac{1}{N^2}\sum_{k=1}^{N}[k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})-(1-w)\{k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})+(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\}]^2\\ &= \frac{1}{N^2}\sum_{k=1}^{N}\{k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})\}^2 = \Lambda_N(1),\end{aligned}$$

and (ii) follows. □

Proof. (Theorem 2.3.3) By the von Mises expansions (2.11) and (2.12), we have

$$\begin{aligned}k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N}) &= \sum_{i=1}^{k}\psi_1(Y_i)-\frac{k}{N}\sum_{i=1}^{k}\psi_1(Y_i)+k(R_{1,k}-R_{1,N})\\ &\quad+\frac{1}{2}\left\{\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k}\psi(Y_i,Y_j)-\frac{k}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\psi(Y_i,Y_j)\right\}\end{aligned}$$

and

$$\begin{aligned}(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N}) &= \sum_{i=k+1}^{N}\psi_1(Y_i)-\frac{N-k}{N}\sum_{i=k+1}^{N}\psi_1(Y_i)+(N-k)(R_{k+1,N}-R_{1,N})\\ &\quad+\frac{1}{2}\left\{\frac{1}{N-k}\sum_{i=k+1}^{N}\sum_{j=k+1}^{N}\psi(Y_i,Y_j)-\frac{N-k}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\psi(Y_i,Y_j)\right\}.\end{aligned}$$

Let

$$\Upsilon_{i,j}(\psi_1)=\sum_{k=i}^{j}\psi_1(k),\quad \Xi_{i,j}(\psi_2)=\sum_{k=i}^{j}\sum_{l=i}^{j}\psi_2(Y_k,Y_l),$$

then by the linearity of partial sums we have $\Upsilon_{1,k}(\psi_1)+\Upsilon_{k+1,N}(\psi_1)=\Upsilon_{1,N}(\psi_1)$, and thus

$$\begin{aligned}k(\hat{\theta}_{1,k}-\hat{\theta}_{1,N})+(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N}) &= kR_{1,k}+(N-k)R_{k+1,N}-NR_{1,N}\\ &\quad+\frac{1}{2}\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}.\end{aligned}$$

For any given $M \geq 0$, by the invariance principle (IP_g), we have the convergence

$$\begin{aligned} & \mathrm{pr}\left[\frac{1}{N}\sum_{k=1}^{N}\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}^2 > M\right] \\ \to \quad & \mathrm{pr}\left[\int_0^1\left\{\frac{W_2(t)}{t}+\frac{W_3(t)}{1-t}-W_2(1)\right\}^2 dt > M\right], \end{aligned}$$

and thus

$$\frac{1}{N}\sum_{k=1}^{N}\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}^2 = O_p(1).$$

Let $\Delta_{k,N} = k\hat{\theta}_{1,k}+(N-k)\hat{\theta}_{k+1,N}-N\hat{\theta}_{1,N}$, then by condition (N) the remainder terms satisfy

$$\sum_{k=1}^{N}\{kR_{1,k}+(N-k)R_{k+1,N}-NR_{1,N}\}^2 = o_p(N), \tag{A.5}$$

and thus by Hölder's inequality we have

$$\begin{aligned} & \left|\frac{1}{N}\sum_{k=1}^{N}\Delta_{k,N}^2-\frac{1}{4N}\sum_{k=1}^{N}\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}^2\right| \\ \leq \quad & \frac{1}{N}\left[\sum_{k=1}^{N}\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}^2\right]^{1/2}\times o_p(N^{1/2})+o_p(1), \end{aligned} \tag{A.6}$$

which converges to zero in probability. On the other hand, since $(N-k)^2R_{1,N}^2 \leq N^2R_{1,N}^2$, by condition (N) and Hölder's inequality we have

$$\begin{aligned} & \left|\sum_{k=1}^{N}(N-k)(R_{k+1,N}-R_{1,N})\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}\right| \\ \leq \quad & \left[\sum_{k=1}^{N}\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}^2\right]^{1/2}\times o_p(N^{1/2}) = o_p(N), \end{aligned}$$

and

$$\begin{aligned}
&\left|\sum_{k=1}^{N}\{kR_{1,k}+(N-k)R_{k+1,N}-NR_{1,N}\}\left(\Upsilon_{k+1,N}-\frac{N-k}{N}\Upsilon_{1,N}\right)\right.\\
&\qquad\left.+\sum_{k=1}^{N}\{kR_{1,k}+(N-k)R_{k+1,N}-NR_{1,N}\}\left\{\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{(N-k)\Xi_{1,N}(\psi_2)}{N^2}\right\}\right|\\
\le&\left(\left[\sum_{k=1}^{N}\left(\Upsilon_{k+1,N}-\frac{N-k}{N}\Upsilon_{1,N}\right)^2\right]^{\frac{1}{2}}+\left[\sum_{k=1}^{N}\left\{\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{(N-k)\Xi_{1,N}(\psi_2)}{N^2}\right\}^2\right]^{\frac{1}{2}}\right)\\
&\quad\times\left[\sum_{k=1}^{N}\{kR_{1,k}+(N-k)R_{k+1,N}-NR_{1,N}\}^2\right]^{1/2},
\end{aligned}$$

which by the invariance principle (IP_g) and (A.5) is of order $o_p(N^{3/2})$. Therefore, we have

$$\begin{aligned}
&\left|\sum_{k=1}^{N}(N-k)(\hat{\theta}_{k+1,N}-\hat{\theta}_{1,N})\Delta_{k,N}\right.\\
&\qquad\left.-\frac{1}{2}\sum_{k=1}^{N}\left(\Upsilon_{k+1,N}-\frac{N-k}{N}\Upsilon_{1,N}\right)\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}\right|\\
\le&\frac{1}{4}\left|\sum_{k=1}^{N}\left\{\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{(N-k)\Xi_{1,N}(\psi_2)}{N^2}\right\}\left\{\frac{\Xi_{1,k}(\psi_2)}{k}+\frac{\Xi_{k+1,N}(\psi_2)}{N-k}-\frac{\Xi_{1,N}(\psi_2)}{N}\right\}\right|\\
&\quad+\left|\sum_{k=1}^{N}(N-k)(R_{k+1,N}-R_{1,N})\{kR_{1,k}+(N-k)R_{k+1,N}-NR_{1,N}\}\right|+o_p(N^{3/2}).
\end{aligned}$$

Since

$$\begin{aligned}
&\left|\sum_{k=1}^{N}(N-k)(R_{k+1,N}-R_{1,N})\{kR_{1,k}+(N-k)R_{k+1,N}-NR_{1,N}\}\right|\\
\le&\left[\sum_{k=1}^{N}(N-k)^2(R_{k+1,N}-R_{1,N})^2\right]^{1/2}\left[\sum_{k=1}^{N}\{kR_{1,k}+(N-k)R_{k+1,N}-NR_{1,N}\}^2\right]^{1/2},
\end{aligned}$$

which is of order $o_p(N)$ by condition (N), we have

$$\left| \frac{1}{N^{1/2}} \sum_{k=1}^{N} (N-k)(\hat{\theta}_{k+1,N} - \hat{\theta}_{1,N})\Delta_{k,N} - \frac{1}{2} \sum_{k=1}^{N} \left(\frac{\Upsilon_{k+1,N}}{N^{1/2}} - \frac{(N-k)\Upsilon_{1,N}}{N^{3/2}} \right) \right.$$
$$\left. \left\{ \frac{\Xi_{1,k}(\psi_2)}{k} + \frac{\Xi_{k+1,N}(\psi_2)}{N-k} - \frac{\Xi_{1,N}(\psi_2)}{N} \right\} \right| \to 0 \qquad \text{(A.7)}$$

in probability. Note that by the invariance principle (IP_g), one has the convergence

$$\begin{bmatrix} & \frac{1}{4N} \sum_{k=1}^{N} \left\{ \frac{\Xi_{1,k}(\psi_2)}{k} + \frac{\Xi_{k+1,N}(\psi_2)}{N-k} - \frac{\Xi_{1,N}(\psi_2)}{N} \right\}^2 \\ \frac{1}{2N} \sum_{k=1}^{N} \left(\frac{\Upsilon_{k+1,N}}{N^{1/2}} - \frac{(N-k)\Upsilon_{1,N}}{N^{3/2}} \right) \left\{ \frac{\Xi_{1,k}(\psi_2)}{k} + \frac{\Xi_{k+1,N}(\psi_2)}{N-k} - \frac{\Xi_{1,N}(\psi_2)}{N} \right\} & \end{bmatrix}$$
$$\xrightarrow{D} \begin{bmatrix} \frac{1}{4} \int_0^1 \left\{ \frac{W_2(t)}{t} + \frac{W_3(t)}{1-t} - W_2(1) \right\}^2 dt \\ \frac{1}{2} \int_0^1 \{W_1(1) - W_1(t) - (1-t)W_1(1)\} \left\{ \frac{W_2(t)}{t} + \frac{W_3(t)}{1-t} - W_2(1) \right\} dt \end{bmatrix},$$

and therefore by (A.6) and (A.7) we have

$$\begin{bmatrix} & \frac{1}{N} \sum_{k=1}^{N} \{k\hat{\theta}_{1,k} + (N-k)\hat{\theta}_{k+1,N} - N\hat{\theta}_{1,N}\}^2 \\ \frac{1}{N^{3/2}} \sum_{k=1}^{N} (N-k)(\hat{\theta}_{k+1,N} - \hat{\theta}_{1,N}) \{k\hat{\theta}_{1,k} + (N-k)\hat{\theta}_{k+1,N} - N\hat{\theta}_{1,N}\} & \end{bmatrix}$$
$$\xrightarrow{D} \begin{bmatrix} \frac{1}{4} \int_0^1 \left\{ \frac{W_2(t)}{t} + \frac{W_3(t)}{1-t} - W_2(1) \right\}^2 dt \\ \frac{1}{2} \int_0^1 \{W_1(1) - W_1(t) - (1-t)W_1(1)\} \left\{ \frac{W_2(t)}{t} + \frac{W_3(t)}{1-t} - W_2(1) \right\} dt \end{bmatrix}.$$

For the generalized self-normalizer (2.5), by viewing it as a function of w, one can show that its second order partial derivative satisfies

$$\inf_{w \in \mathbb{R}} \frac{\partial^2 \Lambda_N(w)}{\partial w^2} \geq 0,$$

where

$$\mathrm{pr}\left\{\frac{\partial^2 \Lambda_N(w)}{\partial w^2} = 0\right\} \to 0$$

holds for all $w \in \mathbb{R}$. Therefore, unrestricted minimizer $\tilde{w}_N$ of $\Lambda_N(w)$ can be obtained by solving the equation

$$\left.\frac{\partial \Lambda_N(w)}{\partial w}\right|_{w=\tilde{w}_N} = 0,$$

which gives us the explicit expression

$$\tilde{w}_N = \frac{\sum_{k=1}^{N}(N-k)(\hat{\theta}_{k+1,N} - \hat{\theta}_{1,N})\{k(\hat{\theta}_{1,k} - \hat{\theta}_{1,N}) + (N-k)(\hat{\theta}_{k+1,N} - \hat{\theta}_{1,N})\}}{\sum_{k=1}^{N}\{k(\hat{\theta}_{1,k} - \hat{\theta}_{1,N}) + (N-k)(\hat{\theta}_{k+1,N} - \hat{\theta}_{1,N})\}^2},$$

and (ii) follows. If $\tilde{w}_N < 0$, then by the positiveness of the second order partial derivative, the first order partial derivative must be nonnegative uniformly on the unit interval, and thus the minimum within the unit interval will be achieved at $\hat{w}_N = 0$. On the other hand, if $\tilde{w}_N > 1$, then by a similar argument one can show that the minimum within the unit interval in this case will be achieved at $\hat{w}_N = 1$, which entails (i). For (iii), note that $\{W_1(t), W_2(t), W_3(t)\}'$ has the same distribution as $\{-W_1(t), W_2(t), W_3(t)\}'$, we have

$$\begin{aligned}\mathrm{pr}(\hat{w}_N = 0) \quad &= \mathrm{pr}(\tilde{w}_N \le 0) \\ &\to \mathrm{pr}\left[\int_0^1 \{tW_1(1) - W_1(t)\}\{t^{-1}W_2(t) + (1-t)^{-1}W_3(t) - W_2(1)\}dt \le 0\right].\end{aligned}$$

We shall now prove that, with probability tending to one, $\hat{w}_N$ does not take values in the open set $(0,1)$. For this, let $\zeta(W)$ denote the limit of $N^{-1/2}\tilde{w}_N$ as derived in (ii), then for any given $\varepsilon > 0$ there exists an $\varepsilon^\star$ such that

$$\mathrm{pr}\{0 < \zeta(W) < \varepsilon^\star\} < \varepsilon/2.$$

Therefore, by the weak convergence in (ii), there exists an $N^\star > 0$ such that

$$|\mathrm{pr}(0 < N^{-1/2}\tilde{w}_N < \varepsilon^\star) - \mathrm{pr}\{0 < \zeta(W) < \varepsilon^\star\}| < \varepsilon/2 \tag{A.8}$$

holds for all $N \ge N^\star$. Let $N^\circ = \max\{N^\star, (\lfloor 1/\varepsilon^\star \rfloor + 1)^2\}$, then for any $N \ge N^\circ$, we have

$$N^{-1/2} \le (\lfloor 1/\varepsilon^\star \rfloor + 1)^{-1} \le \varepsilon^\star,$$

and thus

$$\begin{aligned} \mathrm{pr}(0 < \hat{w}_N < 1) &= \mathrm{pr}(0 < N^{-1/2}\tilde{w}_N < N^{-1/2}) \\ &\leq \mathrm{pr}(0 < N^{-1/2}\tilde{w}_N < \varepsilon^\star) < \varepsilon \end{aligned}$$

by (A.8), indicating that

$$\lim_{N\to\infty} \mathrm{pr}(0 < \hat{w}_N < 1) = 0.$$

Therefore,

$$\lim_{N\to\infty} \mathrm{pr}(\hat{w}_N = 1) = 1 - \lim_{N\to\infty} \mathrm{pr}(\hat{w}_N = 0) - \lim_{N\to\infty} \mathrm{pr}(0 < \hat{w}_N < 1) = 1 - \lim_{N\to\infty} \mathrm{pr}(\hat{w}_N = 0),$$

and (iii) follows. □

Implication for Confidence Interval Construction

By the discussion in Section 2.2, we consider the data-driven weight choice $\hat{w}_N$ that minimizes the generalized self-normalizer (2.5), which in turn is expected to lead to confidence intervals with the minimal length. To provide a theoretical guarantee for this, we need an asymptotic theory on the self-normalized quantity

$$\frac{N(\hat{\theta}_{1,N} - \theta)^2}{\Lambda_N(\hat{w}_N)}. \tag{A.9}$$

Compared with the asymptotic results in Shao (2010) and Lavitas and Zhang (2018), the key difference in (A.9) is the use of a data-driven weight choice, and it is not quite clear if this will make the asymptotic distribution of (A.9) different from that in the aforementioned papers. Fortunately, by Theorems 2.3.2 and 2.3.3, the distribution of $\hat{w}_N$ will be attracted to a finite set with probability tending to one for both probabilistically linear and nonlinear quantities, which makes it possible to derive the asymptotic distribution for (A.9).

Corollary A.0.1. *Assume conditions of either Theorem 2.3.2 or 2.3.3. Then we*

have

$$\text{pr}\{\Lambda_N(\hat{w}_N) = \Lambda_N(0)\} + \text{pr}\{\Lambda_N(\hat{w}_N) = \Lambda_N(1)\} \to 1. \tag{A.10}$$

If in addition the invariance principle as in Theorem 2 of Lavitas and Zhang (2018) holds, then

$$\frac{N(\hat{\theta}_{1,N} - \theta)^2}{\Lambda_N(\hat{w}_N)} \xrightarrow{D} \frac{B(1)^2}{\int_0^1 \{B(t) - tB(1)\}^2 dt}, \tag{A.11}$$

where $B(\cdot)$ represents a standard Brownian motion.

Proof. By Theorems 2.3.2 and 2.3.3, for both probabilistically linear and nonlinear quantities we have

$$\text{pr}(\hat{w}_N = 0) + \text{pr}(\hat{w}_N = 1) \to 1,$$

and thus (A.10) follows. For (A.11), let $\Lambda_N^\star = \min\{\Lambda_N(0), \Lambda_N(1)\}$, then by the argument above we have

$$\text{pr}\{\Lambda_N(\hat{w}_N) = \Lambda_N^\star\} \to 1.$$

By the invariance principle, following a similar proof as that of Theorem 2 in Lavitas and Zhang (2018), one can obtain a joint convergence result for $\{N^{1/2}(\hat{\theta}_{1,N} - \theta), \Lambda_N(0), \Lambda_N(1)\}'$, which leads to the convergence

$$\frac{N(\hat{\theta}_{1,N} - \theta)^2}{\Lambda_N^\star} \xrightarrow{D} \frac{B(1)^2}{\int_0^1 \{B(t) - tB(1)\}^2 dt}.$$

Since $\Lambda_N(\hat{w}_N)/\Lambda_N^\star = 1$ with probability tending to one, the result (A.11) follows by Slutsky's theorem. □

By Corollary A.0.1, the data-driven self-normalization scheme (A.9) is guaranteed to yield narrower confidence intervals when compared with the conventional self-normalizer $\Lambda_N(1)$ of Shao (2010), denoted by S10 hereafter, and the T-symmetric self-normalizer $\Lambda_N(1/2)$ of Lavitas and Zhang (2018), denoted by LZ18 hereafter.

Monte Carlo Simulation

In addition to the above asymptotic result, we in the following provide a small Monte Carlo study to examine its finite-sample effectiveness in reducing the confidence in-

	Median			*Variance*			*acf1*		
ρ	S10	LZ18	$\Lambda_N(\hat{w}_N)$	S10	LZ18	$\Lambda_N(\hat{w}_N)$	S10	LZ18	$\Lambda_N(\hat{w}_N)$
0.3	0.597	0.580	0.497	0.602	0.593	0.573	0.349	0.338	0.295
0.6	0.960	0.941	0.819	1.100	1.062	1.000	0.320	0.298	0.238
0.8	1.796	1.767	1.545	2.689	2.470	2.242	0.293	0.252	0.186
-0.3	0.405	0.392	0.324	0.609	0.605	0.594	0.335	0.331	0.304
-0.6	0.403	0.387	0.315	1.128	1.124	1.105	0.288	0.283	0.242
-0.8	0.456	0.433	0.342	2.812	2.808	2.766	0.231	0.223	0.169

Table A.1: Average lengths of 95% confidence intervals constructed for the median, marginal variance and first-order autocorrelation (acf1) of the autoregressive process (A.12) using (i) the conventional self-normalizer of Shao (2010), denoted by S10; (ii) the T-symmetric self-normalizer of Lavitas and Zhang (2018), denoted by LZ18; and (iii) the data-driven self-normalizer $\Lambda_N(\hat{w}_N)$. The results are based on 50,000 realizations for each configuration.

terval length. For this, we consider the autoregressive model

$$X_i = \rho X_{i-1} + \epsilon_i, \quad i = 1, \ldots, n, \tag{A.12}$$

where (ϵ_i) is a sequence of independent standard normal random variables. Let $n = 200$ and $\rho \in \{\pm 0.3, \pm 0.6, \pm 0.8\}$ to represent situations with different strengths of dependence. The results are summarized in Table A.1-A.3, from which we can observe the followings.

First, by Table A.1, the data-driven self-normalizer $\Lambda_N(\hat{w}_N)$ can be useful in reducing the finite-sample length of the resulting confidence interval when compared with the conventional S10 self-normalizer of Shao (2010) and the T-symmetric LZ18 self-normalizer of Lavitas and Zhang (2018), which corroborates with our theoretical findings. Second, it can be seen from Table A.2 that the minimal length benefit of $\Lambda_N(\hat{w}_N)$ typically comes at a price of a larger size distortion. To further investigate the tradeoff between the length and coverage probability, we in Table A.3 consider

	Median			*Variance*			*acf1*		
ρ	S10	LZ18	$\Lambda_N(\hat{w}_N)$	S10	LZ18	$\Lambda_N(\hat{w}_N)$	S10	LZ18	$\Lambda_N(\hat{w}_N)$
0.3	0.935	0.941	0.902	0.941	0.937	0.930	0.945	0.942	0.917
0.6	0.933	0.938	0.903	0.922	0.914	0.899	0.949	0.945	0.911
0.8	0.925	0.928	0.894	0.885	0.869	0.844	0.956	0.949	0.906
-0.3	0.931	0.941	0.892	0.941	0.939	0.936	0.942	0.942	0.925
-0.6	0.932	0.942	0.888	0.932	0.931	0.928	0.940	0.941	0.918
-0.8	0.929	0.946	0.880	0.910	0.910	0.907	0.939	0.944	0.905

Table A.2: Empirical coverage probability of 95% confidence intervals constructed for the median, marginal variance and first-order autocorrelation (acf1) of the autoregressive process (A.12) using (i) the conventional self-normalizer of Shao (2010), denoted by S10; (ii) the T-symmetric self-normalizer of Lavitas and Zhang (2018), denoted by LZ18; and (iii) the data-driven self-normalizer $\Lambda_N(\hat{w}_N)$. The results are based on 50,000 realizations for each configuration.

the size-adjusted length defined as $2N^{-1/2}\Lambda_N(w)^{1/2}q_{1-\alpha}^{1/2}$, where $q_{1-\alpha}$ is the empirical $(1-\alpha)$th quantile of $\Lambda_N(w)^{-1}N(\hat{\theta}_{1,N}-\theta)^2$. Note that $\Lambda_N(w)q_{1-\alpha}$ depends on the realization through the self-normalizer $\Lambda_N(w)$, and thus $\Lambda_N(w)q_{1-\alpha}$ is different from the $(1-\alpha)$th quantile of $N(\hat{\theta}_{1,N}-\theta)^2$. From Table A.3, we can see that the data-driven choice $\Lambda_N(\hat{w}_N)$ yields the shortest size-adjusted length for the autocorrelation case. However, the results are quite comparable for the variance case, and the advantage seems to disappear for the median case. Therefore, although the data-driven self-normalizer $\Lambda_N(\hat{w}_N)$ is guaranteed to lead to narrowest confidence intervals among the class of generalized self-normalizers, its superiority in terms of the size-adjusted length is not uniform and depends on the quantity of interest. The problem of identifying the exact class of quantities for which the data-driven choice provides the best finite-sample tradeoff between the length and coverage probability is left as a possible future research topic.

	Median			*Variance*			*acf1*		
ρ	S10	LZ18	$\Lambda_N(\hat{w}_N)$	S10	LZ18	$\Lambda_N(\hat{w}_N)$	S10	LZ18	$\Lambda_N(\hat{w}_N)$
0.3	0.653	0.616	0.641	0.635	0.642	0.645	0.359	0.353	0.347
0.6	1.068	1.018	1.045	1.315	1.322	1.321	0.323	0.307	0.285
0.8	2.097	2.017	2.075	3.980	3.936	3.906	0.280	0.254	0.230
-0.3	0.458	0.416	0.440	0.644	0.647	0.646	0.351	0.347	0.347
-0.6	0.457	0.408	0.436	1.270	1.274	1.275	0.306	0.298	0.285
-0.8	0.527	0.445	0.494	3.631	3.645	3.647	0.247	0.233	0.210

Table A.3: Average size-adjusted lengths of 95% confidence intervals constructed for the median, marginal variance and first-order autocorrelation (acf1) of the autoregressive process (A.12) using (i) the conventional self-normalizer of Shao (2010), denoted by S10; (ii) the T-symmetric self-normalizer of Lavitas and Zhang (2018), denoted by LZ18; and (iii) the data-driven self-normalizer $\Lambda_N(\hat{w}_N)$. The results are based on 50,000 realizations for each configuration.

Appendix B

Appendix B: Theoretical Justification for Chapter 3

Proof. (Theorem 3.2.1) Let U be the functional that satisfies $U(B) = \int_0^1 \{B(t) - tB(1)\}^2 dt$, then by properties of the Brownian motion we have

$$\text{cov}\{B(s) - sB(1), B(t) - tB(1)\} = s \wedge t - st,$$

and thus by Theorem 1 of Tolmatz (2002), the density function of $U(B)$ can be expressed as

$$f_U(u) = u^{-5/4} \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \cdot \Gamma(1/2 - k)} \exp\{-(k + 1/4)^2/u\} D_{3/2}\{(2k + 1/2)u^{-1/2}\},$$

where $\Gamma(v) = \int_0^\infty x^{v-1} \exp(-x) dx$ is the gamma function and $D_a(\cdot)$ is the parabolic cylinder function (Whittaker, 1902) that commonly appears in harmonic analysis. We need a little preparation before being able to provide an explicit formula for the density of $Z(B)$. In particular, by the results on page 251 and page 615 of Prudnikov et al. (1992), we have

$$\begin{aligned}
&\int_0^\infty u^{-3/4} \exp\{-uy^2/2 - (k + 1/4)^2/u\} D_{3/2}\{(2k + 1/2)u^{-1/2}\} du \\
= \;& 2^{3/4}(2\pi)^0 \{2^{5/2}\pi^{1/2}\Gamma(-3/2)\}^0 (y^2/2)^{-1/4} G_{3,1}^{0,3} \left\{ (k + 1/4)^{-2} y^{-2} \Big|_{5/4}^{3/4,1/2,1} \right\} \\
= \;& 2y^{-1/2} G_{3,1}^{0,3} \left\{ (k + 1/4)^{-2} y^{-2} \Big|_{5/4}^{3/4,1/2,1} \right\},
\end{aligned}$$

and thus by a change of variable one can obtain that

$$\begin{aligned}\Xi_k(z) &= \int_0^\infty \left(\frac{u}{2\pi z}\right)^{1/2} u^{-5/4}\exp\{-uz/2-(k+1/4)^2/u\}D_{3/2}\{(2k+1/2)u^{-1/2}\}du \\ &= (2/\pi)^{1/2}z^{-3/4}G_{3,1}^{0,3}\left\{\frac{1}{(k+1/4)^2 z}\Big|_{5/4}^{3/4,1/2,1}\right\}.\end{aligned}$$

On the other hand, by properties of the gamma function, we have $\Gamma(1/2)=\pi^{1/2}$ and

$$\frac{1}{k!\cdot\Gamma(1/2-k)} = \frac{\prod_{l=1}^k(1/2-l)}{k!\cdot\Gamma(1/2)} = \pi^{-1/2}\binom{-1/2}{k}.$$

Note that

$$\mathrm{cov}\{B(t)-tB(1),B(1)\}=0$$

holds for all $t\in[0,1]$, the density function of $Z(B)$ then follows the relation

$$\begin{aligned}f_Z(z) &= \int_0^\infty f_{Z|U}(z\mid u)f_U(u)du \\ &= \sum_{k=0}^\infty \frac{(-1)^k}{k!\cdot\Gamma(1/2-k)}\Xi_k(z) \\ &= 2^{1/2}\pi^{-1}z^{-3/4}\sum_{k=0}^\infty\binom{-1/2}{k}(-1)^k G_{3,1}^{0,3}\left\{\frac{1}{(k+1/4)^2 z}\Big|_{5/4}^{3/4,1/2,1}\right\}.\end{aligned}$$

The first claim follows by using the mathematical equivalency on page 209 of Erdélyi (1953). For the second claim, we shall apply the asymptotic expansion of the parabolic cylinder function as in Abadir and Paruolo (1997), through which one can obtain that, as $|y|\to\infty$,

$$\sum_{k=0}^\infty\binom{-1/2}{k}(-1)^k G_{3,1}^{0,3}\left\{\frac{1}{(k+1/4)^2 y^2}\Big|_{5/4}^{3/4,1/2,1}\right\} \sim 2^{-1}\pi^{1/2}|y|^{1/2}\exp(-|y|/2).$$

Therefore, by letting $z = y^2$ we have

$$\begin{aligned} f_Z(z) &= 2^{1/2}\pi^{-1}z^{-3/4}\sum_{k=0}^{\infty}\binom{-1/2}{k}(-1)^k G_{3,1}^{0,3}\left\{\frac{1}{(k+1/4)^2 z}\Big|_{5/4}^{3/4,1/2,1}\right\} \\ &\sim (2\pi z)^{-1/2}\exp\big(-z^{1/2}/2\big), \end{aligned}$$

and the second claim follows. □

Proof. (Theorem 3.2.2) Let $\varpi(u) = 4u^{1/2}$, $u > 0$, then $\varpi(\cdot)$ is a positive function whose derivative satisfies $d\varpi(u)/du = 2u^{-1/2} \to 0$ as $u \to \infty$. Then for any $z \in \mathbb{R}$,

$$\lim_{u\to\infty}[\{u+\varpi(u)z\}^{1/2} - u^{1/2}] = \lim_{u\to\infty}\frac{\varpi(u)z}{\{u+\varpi(u)z\}^{1/2}+u^{1/2}} = 2z,$$

and thus by Theorem 3.2.1,

$$\begin{aligned} \lim_{u\to\infty}\frac{1-F_Z\{u+\varpi(u)z\}}{1-F_Z(u)} &= \lim_{u\to\infty}\frac{(1+2zu^{-1/2})f_Z\{u+\varpi(u)z\}}{f_Z(u)} \\ &= \lim_{u\to\infty}\frac{\{u+\varpi(u)z\}^{-1/2}\exp\big[-\{u+\varpi(u)z\}^{1/2}/2\big]}{u^{-1/2}\exp(-u^{1/2}/2)} \\ &= \lim_{u\to\infty}\exp\big(-[\{u+\varpi(u)z\}^{1/2}-u^{1/2}]/2\big) = \exp(-z). \end{aligned}$$

Therefore, by Section 3.3.3 of Embrechts et al. (1997), the distribution function $F_Z(\cdot)$ belongs to the Gumbel maximum domain of attraction, and it suffices to show that c_p and d_p can be chosen in the given form. For this, note that $d_p \to \infty$ as $p \to \infty$, we have, by Theorem 3.2.1,

$$\begin{aligned} \lim_{p\to\infty} p\{1-F_Z(d_p)\} &= \lim_{p\to\infty} p^2 f_Z(d_p)\cdot 4p^{-1}(2\log p - \log\pi + 3\log 2) \\ &= \lim_{p\to\infty} 4p(2\pi)^{-1/2}\exp\{-\log p + (\log\pi - 3\log 2)/2\} \\ &= 4(2\pi)^{-1/2}\cdot\pi^{1/2}2^{-3/2} = 1. \end{aligned}$$

Since $c_p = \varpi(d_p)$ holds for all large p, Theorem 3.2.2 follows by the results in Mikosch and Yslas (2019). □

Proof. (Theorem 3.3.1) By Theorem 3.2.1,

$$\begin{aligned}\lim_{p\to\infty} p[1-F_Z\{c(\log p)^2\}] &= \lim_{p\to\infty} p^2 f_Z\{c(\log p)^2\}\cdot(2cp^{-1}\log p)\\ &= \lim_{p\to\infty}(2cp)\cdot(2\pi c)^{-1/2}\exp\{-(c^{1/2}\log p)/2\}\\ &= \lim_{p\to\infty}(2c)^{1/2}\pi^{-1/2}p^{1-c^{1/2}/2},\end{aligned}$$

whose limit is zero if $c>4$. Then by condition (A1) and Boole's inequality, we have

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{\Delta_{p,n}>c(\log p)^2\}\le\lim_{p\to\infty}p[1-F_Z\{c(\log p)^2\}]=0$$

if we choose the constant c to be sufficiently large, and Theorem 3.3.1 follows. □

Proof. (Theorem 3.3.2) We first prove the first claim, and it suffices to show that

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{\Delta_{p,n}>c_pz+d_p\}=1-\exp\{-\exp(-z)\}.$$

For $k=1,\ldots,p$, let $A_{k,p,n}$ be the set that represents the event $\{T_{k,n}>c_pz+d_p\}$, then we can write $\{\Delta_{p,n}>c_pz+d_p\}=\bigcup_{k=1}^{p}A_{k,p,n}$ as their union, and by the Bonferroni inequality we have

$$\mathrm{pr}(\Delta_{p,n}>c_pz+d_p)\ge\sum_{m=1}^{2M}(-1)^{m-1}\sum_{1\le k_1<\cdots<k_m\le p}\mathrm{pr}\left(\bigcap_{l=1}^{m}A_{k_l,p,n}\right),$$

and

$$\mathrm{pr}(\Delta_{p,n}>c_pz+d_p)\le\sum_{m=1}^{2M-1}(-1)^{m-1}\sum_{1\le k_1<\cdots<k_m\le p}\mathrm{pr}\left(\bigcap_{l=1}^{m}A_{k_l,p,n}\right).$$

By condition (A2),

$$\lim_{p\to\infty}\lim_{n\to\infty}\sum_{1\le k_1<\cdots<k_m\le p}\mathrm{pr}\left(\bigcap_{l=1}^{m}A_{k_l,p,n}\right)=\lim_{p\to\infty}\lim_{n\to\infty}\binom{p}{m}\{1-F_Z(c_pz+d_p)\}^m,$$

while by the proof of Theorem 3.2.2,

$$\lim_{p\to\infty}\lim_{n\to\infty}p\{1-F_Z(c_pz+d_p)\}=\lim_{p\to\infty}\lim_{n\to\infty}p\{1-F_Z(d_p)\}\cdot\frac{1-F_Z(c_pz+d_p)}{1-F_Z(d_p)}=\exp(-z).$$

Combining all the above results, we have

$$\sum_{m=1}^{2M} \frac{(-1)^{m-1}}{m!} \exp(-mz) \leq \lim_{p\to\infty} \lim_{n\to\infty} \mathrm{pr}(\Delta_{p,n} > c_p z + d_p) \leq \sum_{m=1}^{2M-1} \frac{(-1)^{m-1}}{m!} \exp(-mz)$$

holds for any fixed $M > 0$. Note that by the Taylor expansion, the series

$$\sum_{m=1}^{M} \frac{(-1)^{m-1}}{m!} \exp(-mz) = 1 - \sum_{m=0}^{M} \frac{(-1)^{m}}{m!} \exp(-mz) \to 1 - \exp\{-\exp(-z)\}$$

as $M \to \infty$, and the first claim follows. Note that for independent Gaussian data, condition (A1) entails condition (A2), and thus by the same argument one can show that $c_p^{-1}(\Delta^\circ_{p,n} - d_p)$ shares the same asymptotic distribution as $c_p^{-1}(\Delta_{p,n} - d_p)$. As a result,

$$\lim_{p\to\infty} \lim_{n\to\infty} \mathrm{pr}\{(\theta_1, \ldots, \theta_p) \in \mathcal{R}^{\mathrm{SAMSN}}_{p,n,1-\alpha}\} = \lim_{p\to\infty} \lim_{n\to\infty} \mathrm{pr}\{\Delta_{p,n} \leq q_{1-\alpha}(\Delta^\circ_{p,n})\} = 1 - \alpha,$$

and the second claim follows. □

Proof. (Theorem 3.3.3) By the proof of Theorem 3.3.1,

$$\lim_{p\to\infty} \lim_{n\to\infty} \mathrm{pr}\{\Delta_{p,n} \leq c(\log p)^2\} = 1$$

holds for any $c > 4$, and therefore

$$\lim_{p\to\infty} \lim_{n\to\infty} \mathrm{pr}\left(\max_{k\in\mathcal{S}_0} V_{k,n}^{-1} \hat{\theta}^2_{k,n} \leq \lambda_{p,n}\right) = 1.$$

On the other hand,

$$\mathrm{pr}\left(\min_{k\in\mathcal{S}_1} V_{k,n}^{-1} \hat{\theta}^2_{k,n} > \lambda_{p,n}\right) \geq \mathrm{pr}(V_{k,n}^{-1}\theta_k^2 > 2\lambda_{p,n},\ 2|\hat{\theta}_{k,n} - \theta_k| \leq 2^{-1}|\theta_k|,\ \forall k \in \mathcal{S}_1),$$

where

$$\mathrm{pr}(|\hat{\theta}_{k,n} - \theta_k| \leq 4^{-1}|\theta_k|,\ \forall k \in \mathcal{S}_1) \geq 1 - \sum_{k\in\mathcal{S}_1} \mathrm{pr}(|\hat{\theta}_{k,n} - \theta_k| > 4^{-1}|\theta_k|).$$

By the invariance principle (IP) and the condition on the minimal nonzero signal, we

have

$$\lim_{p\to\infty}\lim_{n\to\infty}\sum_{k\in\mathcal{S}_1}\mathrm{pr}(|\hat{\theta}_{k,n}-\theta_k|>4^{-1}|\theta_k|)\leq\lim_{p\to\infty}\lim_{n\to\infty}\sum_{k\in\mathcal{S}_1}\mathrm{pr}\left(|\hat{\theta}_{k,n}-\theta_k|>4^{-1}\min_{k\in\mathcal{S}_1}|\theta_k|\right)$$
$$\leq\lim_{p\to\infty}\lim_{n\to\infty}\sum_{k\in\mathcal{S}_1}\mathrm{pr}(n^{1/2}|\hat{\theta}_{k,n}-\theta_k|>c^{\diamond}\log p)=0$$

holds for a sufficiently large choice of $c^{\diamond}>0$, and therefore

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}(2|\hat{\theta}_{k,n}-\theta_k|\leq 2^{-1}|\theta_k|,\ \forall k\in\mathcal{S}_1)=1.$$

Note that

$$\begin{aligned}\mathrm{pr}(V_{k,n}^{-1}\theta_k^2>2\lambda_{p,n},\ \forall k\in\mathcal{S}_1) &\geq \mathrm{pr}\left(\min_{k\in\mathcal{S}_1}\theta_k^2>2\lambda_{p,n}\max_{k\in\mathcal{S}_1}V_{k,n}\right)\\ &\geq 1-\sum_{k\in\mathcal{S}_1}\mathrm{pr}\left\{V_{k,n}\geq(2\lambda_{p,n})^{-1}\min_{k\in\mathcal{S}_1}\theta_k^2\right\},\end{aligned}$$

where

$$\lim_{p\to\infty}\lim_{n\to\infty}\sum_{k\in\mathcal{S}_1}\mathrm{pr}\left\{V_{k,n}\geq(2\lambda_{p,n})^{-1}\min_{k\in\mathcal{S}_1}\theta_k^2\right\}=0$$

when $d<\infty$ is fixed, and by Theorem 11 of Tolmatz (2002) we have

$$\lim_{p\to\infty}\lim_{n\to\infty}\sum_{k\in\mathcal{S}_1}\mathrm{pr}\left\{V_{k,n}\geq(2\lambda_{p,n})^{-1}\min_{k\in\mathcal{S}_1}\theta_k^2\right\}\leq\lim_{p\to\infty}\lim_{n\to\infty}\sum_{k\in\mathcal{S}_1}\mathrm{pr}\left\{V_{k,n}\geq c^{\sharp}\log d\right\}=0$$

holds for a sufficiently large choice of $c^{\sharp}>0$. Combining the above results, we obtain that

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\left(\min_{k\in\mathcal{S}_1}V_{k,n}^{-1}\theta_k^2>2\lambda_{p,n}\right)=1,$$

and

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\left(\min_{k\in\mathcal{S}_1}V_{k,n}^{-1}\hat{\theta}_{k,n}^2>\lambda_{p,n}\right)=1.$$

Therefore, we have $\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}(\Omega_{p,n}=T_{\mathcal{S}_1,n})=1$, and as a result,

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}[\Omega_{p,n}\leq q_{1-\alpha}\{Z_d(B_d)\}]=\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}[T_{\mathcal{S}_1,n}\leq q_{1-\alpha}\{Z_d(B_d)\}]=1-\alpha,$$

which entails the first claim. For the second claim, a similar argument can be used

to show that $\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}(\hat{d}=d)=1$, and thus by condition (A3),

$$\begin{aligned}\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{\Omega_{p,n}\le q_{1-\alpha}(\Omega^{\circ}_{\hat{d},n})\} &= \lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{Z_d(B_d)\le q_{1-\alpha}(\Omega^{\circ}_{\hat{d},n})\}\\ &= \lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{\Omega^{\circ}_{\hat{d},n}\le q_{1-\alpha}(\Omega^{\circ}_{\hat{d},n})\}=1-\alpha.\end{aligned}$$

Since the sparse vector $(\mathbb{1}_{\{V_{1,n}^{-1}\hat{\theta}^2_{1,n}>\lambda_{p,n}\}},\ldots,\mathbb{1}_{\{V_{p,n}^{-1}\hat{\theta}^2_{p,n}>\lambda_{p,n}\}})$ used to form the Hadamard product in (3.9) satisfies

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\left(\max_{1\le k\le p}|\mathbb{1}_{\{V_{k,n}^{-1}\hat{\theta}^2_{k,n}>\lambda_{p,n}\}}-\mathbb{1}_{\{k\in\mathcal{S}_1\}}|=0\right)=1,$$

we have

$$\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{(\theta_1,\ldots,\theta_p)\in\mathcal{R}^{\mathrm{SATSN}}_{p,n,1-\alpha}\}=\lim_{p\to\infty}\lim_{n\to\infty}\mathrm{pr}\{\Omega_{p,n}\le q_{1-\alpha}(\Omega^{\circ}_{\hat{d},n})\}=1-\alpha,$$

and Theorem 3.3.3 follows. □

Bibliography

Abadir, K. M. and Paruolo, P. (1997). Two mixed normal densities from cointegration analysis. *Econometrica*, 65:671–680.

Adams, Z. and Glück, T. (2015). Financialization in commodity markets: A passing trend or the new normal? *Journal of Banking & Finance*, 60:93–111.

Bai, S., Taqqu, M. S., and Zhang, T. (2016). A unified approach to self-normalized block sampling. *Stochastic Processes and their Applications*, 126:2465–2493.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2):311–329.

Beals, R. and Szmigielski, J. (2013). Meijer G-functions: a gentle introduction. *Notices of the American Mathematical Society*, 60(7):866–872.

Berens, T., Weiß, G. N., and Wied, D. (2015). Testing for structural breaks in correlations: Does it improve value-at-risk forecasting? *Journal of Empirical Finance*, 32:135–152.

Berkes, I., Liu, W., and Wu, W. B. (2014). Komlós-major-tusnády approximation under dependence. *The Annals of Probability*, 42(2):794–817.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal edgeworth expansion. *The Annals of Statistics*, 6(2):434–451.

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Grassmann, M., and Ceulemans, E. (2017). Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods. *Behavior Research Methods*, 49(3):988–1005.

Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B*, 76(2):349–372.

Chen, L. and Wu, W. B. (2019). Testing for trends in high-dimensional time series. *Journal of the American Statistical Association*, 114:869–881.

Chen, S. X., Li, J., and Zhong, P.-S. (2019). Two-sample and ANOVA tests for high dimensional means. *The Annals of Statistics*, 47(3):1443–1474.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Choi, J.-E. and Shin, D. W. (2020a). A self-normalization break test for correlation matrix. *Statistical Papers*, pages 1–21.

Choi, J.-E. and Shin, D. W. (2020b). Subsample scan test for multiple breaks based on self-normalization. Working paper.

Degras, D., Xu, Z., Zhang, T., and Wu, W. B. (2012). Testing for parallelism among trends in multiple time series. *IEEE Transactions on Signal Processing*, 60(3):1087–1097.

Dehling, H., Denker, M., and Philipp, W. (1984). Invariance principles for von mises and U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67:139–167.

Demetrescu, M. and Wied, D. (2019). Testing for constant correlation of filtered series under structural change. *The Econometrics Journal*, 22(1):10–33.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

Erdélyi, A. (1953). *Higher Transcendental Functions, Volume 1*. McGraw-Hill, New York.

Fernholz, L. T. (2001). On multivariate higher order von mises expansions. *Metrika*, 53:123–140.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in markov chain monte carlo. *The Annals of Statistics*, 38:1034–1070.

Galeano, P. and Wied, D. (2014). Multiple break detection in the correlation structure of random variables. *Computational Statistics & Data Analysis*, 76:262–282.

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., and Lahiri, S. N. (2015). A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association*, 110(510):837–849.

Hall, P. (1979). On the rate of convergence of normal extremes. *Journal of Applied Probability*, 16(2):433–439.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hannan, E. (1979). The central limit theorem for time series regression. *Stochastic Processes and their Applications*, 9:281–289.

Herrndorf, N. (1984). A functional central limit theorem for weakly dependent sequences of random variables. *The Annals of Probability*, 12(1):141–153.

Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.

Kanagawa, S. and Yoshihara, K. (1994). The almost sure invariance principles of degenerate U-statistics of degree two for stationary random variables. *Stochastic Processes and their Applications*, 49:347–356.

Kiefer, N. M. and Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, 21:1130–1164.

Kiefer, N. M., Vogelsang, T. J., and Bunzel, H. (2000). Simple robust testing of regression hypotheses. *Econometrica*, 68(3):695–714.

Kim, K. H., Zhang, T., and Wu, W. B. (2015a). Parametric specification test for nonlinear autoregressive models. *Econometric Theory*, 31:1078–1101.

Kim, S., Zhao, Z., and Shao, X. (2015b). Nonparametric functional central limit theorem for time series regression with application to self-normalized confidence interval. *Journal of Multivariate Analysis*, 133:277–290.

Krishnan, C., Petkova, R., and Ritchken, P. (2009). Correlation risk. *Journal of Empirical Finance*, 16(3):353–367.

Lavitas, L. and Zhang, T. (2018). A time-symmetric self-normalization approach for inference of time series. *Journal of Time Series Analysis*, 39:748–762.

Liu, W. and Wu, W. B. (2010). Asymptotics of spectral density estimates. *Econometric Theory*, 26:1218–1245.

Lobato, I. N. (2001). Testing that a dependent process is uncorrelated. *Journal of the American Statistical Association*, 96:1066–1076.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., and Gross, J. J. (2005). The tie that binds? coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2):175.

Mikosch, T. and Yslas, J. (2019). Gumbel and Fréchet convergence of the maxima of independent random walks. *arXiv e-prints*, page arXiv:1904.04607.

Miller, R. G. and Sen, P. K. (1972). Weak convergence of U-statistics and von Mises' differentiable statistical functions. *The Annals of Mathematical Statistics*, 43(1):31–41.

Pan, Q., Xu, B., and Zhang, T. (2021). Testing structural breaks in correlation matrix using unsupervised self-normalization. Working paper.

Paparoditis, E. and Politis, D. N. (2012). Nonlinear spectral density estimation: thresholding the correlogram. *Journal of Time Series Analysis*, 33:386–397.

Polansky, A. M. (2008). *Observed Confidence Levels: Theory and Application*. Chapman & Hall/CRC.

Politis, D. N. (2011). Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices. *Econometric Theory*, 27:703–744.

Posch, P. N., Ullmann, D., and Wied, D. (2019). Detecting structural changes in large portfolios. *Empirical Economics*, 56(4):1341–1357.

Prudnikov, A. B., Brychkov, Yu. A., and Marichev, O. I. (1992). *Integrals and Series, Volume 4: Direct Laplace Transforms*. Gordon and Breach Science Publishers, New York.

Shao, X. (2010). A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:343–366.

Shao, X. (2015). Self-normalization for time series: a review of recent developments. *Journal of the American Statistical Association*, 110:1797–1817.

Shao, X. and Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association*, 105:1228–1240.

Sharipov, O. Sh. (2003). Invariance principle for U-statistics and von Mises' functionals of weakly dependent observations. *Theory of Probability & Its Applications*, 47:730–733.

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99:386–402.

Srivastava, M. S., Katayama, S., and Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114:349–358.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Sun, Y., Phillips, P. C., and Jin, S. (2008). Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing. *Econometrica*, 76(1):175–194.

Tolmatz, L. (2002). On the distribution of the square integral of the Brownian bridge. *The Annals of Probability*, 30(1):253–269.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Verlag, New York.

Volgushev, S. and Shao, X. (2014). A general approach to the joint asymptotic analysis of statistics from sub-samples. *Electronic Journal of Statistics*, 8(1):390–431.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348.

Wang, R. and Shao, X. (2020). Hypothesis testing for high-dimensional time series via self-normalization. *The Annals of Statistics*, page to appear.

Wang, W., Lin, N., and Tang, X. (2019). Robust two-sample test of high-dimensional mean vectors under dependence. *Journal of Multivariate Analysis*, 169:312–329.

Whittaker, E. T. (1902). On the functions associated with the parabolic cylinder in harmonic analysis. *Proceedings of the London Mathematical Society*, 35:417–427.

Wied, D. (2017). A nonparametric test for a constant correlation matrix. *Econometric Reviews*, 36(10):1157–1172.

Wied, D. and Galeano, P. (2013). Monitoring correlation change in a sequence of random variables. *Journal of Statistical Planning and Inference*, 143(1):186–196.

Wied, D., Krämer, W., and Dehling, H. (2012). Testing for a change in correlation at an unknown point in time using an extended functional delta method. *Econometric Theory*, pages 570–589.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154.

Wu, W. B. (2007). Strong invariance principles for dependent random variables. *The Annals of Probability*, 35(6):2294–2320.

Wu, W. B. and Zhao, Z. (2007). Inference of trends in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:391–410.

Wu, W. B. and Zhou, Z. (2011). Gaussian approximations for non-stationary multiple time series. *Statistica Sinica*, 21(3):1397–1413.

Xiao, H. and Wu, W. B. (2011). A single-pass algorithm for spectrum estimation with fast convergence. *IEEE Transactions on Information Theory*, 57:4720–4731.

Xiao, H. and Wu, W. B. (2013). Asymptotic theory for maximum deviations of samplecovariance matrix estimates. *Stochastic Processes and their Applications*, 123:2899–2920.

Xu, G., Lin, L., Wei, P., and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika*, 103:609–624.

Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919.

Zhang, T. (2013). Clustering high-dimensional time series based on parallelism. *Journal of the American Statistical Association*, 108(502):577–588.

Zhang, T. (2016). Testing for jumps in the presence of smooth changes in trends of nonstationary time series. *Electronic Journal of Statistics*, 10(1):706–735.

Zhang, T. (2018). A thresholding-based prewhitened long-run variance estimator and its dependence-oracle property. *Statistica Sinica*, 28(1):319–338.

Zhang, T. and Lavitas, L. (2018). Unsupervised self-normalized change-point testing for time series. *Journal of the American Statistical Association*, 113:637–648.

Zhang, T., Lavitas, L., and Pan, Q. (2019). Asymptotic behavior of optimal weighting in generalized self-normalization for time series. *Journal of Time Series Analysis*, 40(5):831–851.

Zhang, T. and Pan, Q. (2021). Self-normalized simultaneous confidence regions for high-dimensional time series. Working paper.

Zhang, T. and Wu, W. B. (2011). Testing parametric assumptions of trends of a nonstationary time series. *Biometrika*, 98:599–614.

Zhang, T. and Wu, W. B. (2012). Inference of time-varying regression models. *The Annals of Statistics*, 40(3):1376–1402.

Zhang, X. and Cheng, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli*, 24:2640–2675.

Zhou, T., Mei, Z., Zhu, X., and Huang, Z. (2020). Synchrony detection of epileptic eeg signals based on attention and pearson's correlation coefficient. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 531–535. IEEE.

Zhou, Z. and Shao, X. (2013). Inference for linear models with dependent errors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:323–343.

CURRICULUM VITAE