

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Towards a Theoretical Foundation of the Model-free Bootstrap for Regression and Time Series

Permalink

<https://escholarship.org/uc/item/9h17p76f>

Author

Wang, Yiren

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Towards a Theoretical Foundation of the Model-free Bootstrap for Regression and Time Series

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

Yiren Wang

Committee in charge:

Professor Dimitris N. Politis, Chair
Professor Ery Arias-Castro
Professor Loki Natarajan
Professor Danna Zhang
Professor Wenxin Zhou

2021

Copyright
Yiren Wang, 2021
All rights reserved.

The dissertation of Yiren Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To the sleepless nights.

EPIGRAPH

And what, Socrates, is the food of the soul?

Surely, I said, knowledge is the food of the soul.

—Plato

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
List of Tables	x
Acknowledgements	xi
Vita	xiii
Abstract of the Dissertation	xiv
Chapter 1 Model-free Bootstrap and Conformal Prediction in Regression: Conditionality, Conjecture Testing, and Pertinent Prediction Intervals	1
1.1 Abstract	1
1.2 Introduction	2
1.3 Model-free vs. model-based regression	5
1.3.1 Model-based regression	5
1.3.2 Model-free regression	9
1.3.3 On choosing conditional CDF and quantile estimator	13
1.4 Preliminary asymptotic analysis	15
1.5 The conditionality problem	18
1.6 Further analysis of Model-free bootstrap	23
1.6.1 Asymptotic pertinence and better coverage guarantees	23
1.6.2 Required technical assumptions	25
1.6.3 Results for MFB pertinence and Claim 1.6.1	28
1.7 Conjecture testing	29
1.7.1 Prediction vs. estimation	29
1.7.2 Conjecture testing: definition and discussion	30
1.8 Numerical experiments	33
1.8.1 Synthetic data	33
1.8.2 VaR prediction for intraday stock returns data	36
1.9 Additional proofs	41
1.9.1 Asymptotic analysis for CP	41
1.9.2 Asymptotic analysis for MFB	43
1.9.3 Proof of Lemma 1.6.1	45

	1.9.4	Proofs of Theorems 1.6.2, 1.6.3 and 1.6.4	46
	1.10	Acknowledgement	47
Chapter 2		Model-free Bootstrap for a General Class of Stationary Time Series . . .	49
	2.1	Abstract	49
	2.2	Introduction	50
	2.3	Model-free bootstrap algorithm	53
	2.3.1	The MFB algorithm	53
	2.3.2	MFB for confidence intervals	55
	2.3.3	MFB for prediction intervals	56
	2.3.4	Appropriate estimators for $\hat{F}$, $\tilde{\Phi}^{-1}$ and $\hat{\Sigma}_n$	59
	2.3.5	Connection with previous work on MFB for time series . . .	61
	2.4	Assumptions and preliminary results	62
	2.4.1	Acceptable forms of $\{Y_t\}$	62
	2.4.2	Short range dependence (SRD) assumptions on $\{Y_t\}$	64
	2.5	Model-free bootstrap validity for estimation	69
	2.5.1	Bootstrap validity for the mean	69
	2.5.2	Bootstrap validity for smooth functions of linear statistics .	74
	2.5.3	Approximately linear statistics and sample quantiles	77
	2.5.4	Bootstrap validity for kernel smoothed spectral density . . .	78
	2.6	Model-free bootstrap for prediction	80
	2.7	Numerical results	82
	2.7.1	Performance for bootstrap confidence intervals	83
	2.7.2	Performance of bootstrap prediction intervals	87
	2.8	Additional proofs	88
	2.8.1	Preliminary results under m -approximation assumptions . .	88
	2.8.2	Proof of relative lemmas and theorems	90
	2.9	Acknowledgement	105
Chapter 3		Model-free Bootstrap Prediction Regions for Multivariate Time Series . .	106
	3.1	Abstract	106
	3.2	Introduction	107
	3.3	Description of algorithm	107
	3.3.1	Models of interest	108
	3.3.2	Connection to copula-based models	109
	3.3.3	The model-free bootstrap algorithm	110
	3.3.4	Generating joint prediction band for univariate time series .	117
	3.4	Theoretical results	119
	3.5	Numerical results	122
	3.5.1	Synthetic data experiment	122
	3.5.2	Joint prediction bands for heavy-tailed financial returns . . .	124
	3.6	Acknowledgement	130

Chapter 4	Bootstrap Methods for Quantile Autoregressive Time Series	132
4.1	Abstract	132
4.2	Introduction	133
4.3	Bootstrap algorithms	136
4.3.1	Model-based forward bootstrap	136
4.3.2	Model-free bootstrap	139
4.4	Simulations	142
4.4.1	Asymptotic performance of bootstrap for inference	142
4.4.2	Coverage performance of prediction intervals	144
4.5	Acknowledgement	145
Appendix A	MFB under misspecification	146
Bibliography		149

LIST OF FIGURES

Figure 1.1:	Asymptotic results for prediction intervals. (a): results based on the nonparametric CDF estimator; (b): results based on the quantile regression based CDF estimator.	35
Figure 1.2:	Histogram of coverage probabilities at $N = 50$. Left: results based on the nonparametric CDF estimator; right: results based on the quantile regression based CDF estimator.	37
Figure 3.1:	CVR plots with $n = 100$	124
Figure 3.2:	CVR plots with $n = 200$	125
Figure 3.3:	CVR plots with $n = 300$	126
Figure 3.4:	CVR plots with $n = 400$	127
Figure 3.5:	CVR plots with $n = 500$	128
Figure 3.6:	ECVR plots for AAPL.	129
Figure 3.7:	ECVR plots for GME	130
Figure 3.8:	ECVR plots for AMZN.	131
Figure 3.9:	ECVR plots for TSLA.	131
Figure 4.1:	Approximation performance of the bootstrap. ($n = 1000$)	143
Figure 4.2:	Approximation performance of the bootstrap. ($n = 5000$)	143
Figure 4.3:	Approximation performance of the bootstrap. ($n = 10000$)	144

LIST OF TABLES

Table 1.1:	Comparison between conformal prediction and Model-free bootstrap.	29
Table 1.2:	Analogies between parameter estimation and prediction.	30
Table 1.3:	Empirical acceptance rate results for GME.	40
Table 1.4:	Empirical acceptance rate results for AMC.	41
Table 2.1:	Empirical CVR for the mean parameter across 3 models.	85
Table 2.2:	Empirical CVR for lag 1 autocovariance for the first 2 models.	86
Table 2.3:	Empirical CVR for lag 2 autocovariance for model 3.	87
Table 2.4:	Empirical CVR for prediction intervals across the three models.	89
Table 4.1:	Empirical coverage performance of 1-step ahead PIs between different algorithms.	145

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor Professor Dimitris N. Politis. I am truly inspired by his wisdom, professionalism, profound knowledge in statistics and enthusiasm for research. This dissertation would not have been possible without his direction and suggestions, as well as the countless discussions and editing work. I am deeply grateful for all the guidance and support received from him, and in addition, his patience and encouragement during the beginning stage of my research.

I also thank Professors Ery Arias-Castro, Loki Natarajan, Danna Zhang and Wenxin Zhou for taking their time to serve in my doctoral committee and providing valuable feedbacks. Special thanks to Professor Loki Natarajan, for offering me many opportunities in both teaching and research. In addition, to Professor Ruth Williams, for her introduction to financial mathematics and stochastic systems, from which I learned a lot. To all other professors I took classes from at UCSD for sharing with me their knowledge. Lastly, to UCSD, for giving me an unforgettable experience in the past five years.

Many people have provided me with valuable help during my graduate study to whom I am thankful: to Anne, Ashley, Nishant, Porter, Yingjia, etc, for their assistance on numerous matters. In particular, to Srinjoy, for the many discussions about intriguing problems and help as a friend; to my awesome friends Tong and Haochen, for their care and support, as well as all the fun conversations we had regardless of us being thousands of miles apart.

Finally, to my parents, who thrived on providing me the best despite all the hardships. I am forever indebted to them, for giving me life, raising me up, and love.

Chapter 1, in full, is a reprint of the materials in the paper “Model-free Bootstrap and Conformal Prediction in Regression: Conditionality, Conjecture Testing, and Pertinent Prediction Intervals”, Yiren Wang, Dimitris N. Politis. It has been submitted for consideration of publication in a major statistics journal. The dissertation author is the primary investigator and author of this material.

Chapter 2, in full, is a reprint of the materials in the paper “Model-free Bootstrap for a General Class of Stationary Time Series”, Yiren Wang, Dimitris N. Politis. It has been accepted for publication in Bernoulli. The dissertation author is the primary investigator and author of this material.

Chapter 3, in full, is a version of the paper “Model-free Bootstrap Prediction Regions for Multivariate Time Series”, Yiren Wang, Dimitris N. Politis. It is currently being prepared for publication in a major statistics journal. The dissertation author is the primary investigator and author of this material.

Chapter 4, in full, is a version of the paper “Bootstrap Methods for Quantile Autoregressive Time Series”, Yiren Wang, Dimitris N. Politis. It is currently being prepared for publication in a major statistics journal. The dissertation author is the primary investigator and author of this material.

The following grants/award partially supported the writing of this dissertation which I would like to acknowledge: National Science Foundation Grants DMS 16-13026 and DMS 19-14556; the Richard Libby graduate research award.

VITA

2016	B. S. in Mathematics, Renmin University of China
2016-2021	M. S. in Statistics, University of California San Diego
2017-2021	Graduate Teaching Assistant, University of California San Diego
2019-2021	Graduate Research Assistant, University of California San Diego
2021	Ph. D. in Mathematics, University of California San Diego

PUBLICATIONS

Yiren Wang, Dimitris N. Politis, “Model-free Bootstrap for a General Class of Stationary Time Series”, accepted by *Bernoulli*, 2021.

Yiren Wang, Dimitris N. Politis, “Model-free Bootstrap and Conformal Prediction in Regression: Conditionality, Conjecture Testing, and Pertinent Prediction Intervals”, submitted.

ABSTRACT OF THE DISSERTATION

Towards a Theoretical Foundation of the Model-free Bootstrap for Regression and Time Series

by

Yiren Wang

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2021

Professor Dimitris N. Politis, Chair

The model-free bootstrap (MFB), first introduced in Politis [2013] followed by the monograph of Politis [2015], and further investigated in a series of papers (cf. Pan and Politis [2016b], Chen and Politis [2019], Das and Politis [2021], etc), is a recent advent in the bootstrap literature. The principle of MFB is to (invertibly) transform the original data to a space of i.i.d. variables, wherein the standard i.i.d. bootstrap is performed for the variables, and then the inverse transform is used to obtain bootstrap samples in the original data space. Because of the wide selection of applicable transforms, the MFB framework can be easily extended for complex data scenarios – such as regression and time series. The term "model-free" relates to the fact that the

transformations are estimated without model assumptions, i.e., nonparametrically.

The main purpose of this dissertation is to build a theoretical foundation for the MFB under different setups. Specifically, in chapter 1, we analyze the MFB under model-free regression setup, and compare it with other methods of interest focusing on conditional coverage of prediction intervals. The concept of pertinent prediction intervals is extended to this setup, and we propose the notion of conjecture testing for predictive inference. In chapter 2, we establish bootstrap validity of various statistics for a general class of univariate time series under very basic dependence assumptions, using the recently developed m -approximation technique. In chapter 3, the MFB algorithm of chapter 2 is further extended to handle multivariate time series, wherein we specifically focus on the generation of prediction regions. We also link the MFB under multivariate context with the well-known copula models. Finally in chapter 4, we propose a new MFB algorithm for quantile autoregressive processes based on the MFB framework for general Markov process by Pan and Politis [2016b], using an augmented quantile regression estimator.

Chapter 1

Model-free Bootstrap and Conformal Prediction in Regression: Conditionality, Conjecture Testing, and Pertinent Prediction Intervals

1.1 Abstract

Predictive inference under a general regression setting is gaining more interest in the big-data era. In terms of going beyond point prediction to develop prediction intervals, two main threads of development are conformal prediction and Model-free prediction. Recently, Chernozhukov et al. [2019] proposed a new conformal prediction approach exploiting the same uniformization procedure as in the Model-free Bootstrap of Politis [2015]. Hence, it is of interest to compare and further investigate the performance of the two methods. In the paper at hand, we contrast the two approaches via theoretical analysis and numerical experiments with a focus

on conditional coverage of prediction intervals. We discuss suitable scenarios for applying each algorithm, underscore the importance of conditional vs. unconditional coverage, and show that, under mild conditions, the Model-free bootstrap yields prediction intervals with guaranteed better conditional coverage compared to quantile estimation. We also extend the concept of ‘pertinence’ of prediction intervals in Politis [2015] to the nonparametric regression setting, and give concrete examples where its importance emerges under finite sample scenarios. Finally, we define the new notion of ‘conjecture testing’ that is the analog of hypothesis testing as applied to the prediction problem; we also devise a modified conformal score to allow conformal prediction to handle one-sided ‘conjecture tests’, and compare to the Model-free bootstrap.

Keywords: Asymptotic validity, confidence intervals, significance, nonparametric regression, predictive inference, undercoverage.

1.2 Introduction

Regression is a ubiquitous tool in statistical literature and practice. To discuss it, let $X \in \mathbb{R}^d$ denote the random regressor/covariate, and $Y \in \mathbb{R}$ the (univariate) response associated with X . A general form of a regression setup is the following: the relationship between the response Y and covariate X is described via their (assumed continuous) joint cumulative distribution function (CDF) F , i.e.,

$$(X, Y) \sim F \tag{1.2.1}$$

with the observations $\{(X_i, Y_i)\}_{i=1}^n$ arising from (1.2.1) are assumed independent, identically distributed (i.i.d.). This general regression setup has allowed researchers to study valid statistical procedures for inference and prediction problems.

The focus in this paper is predictive inference, specifically, the aim to generate a valid prediction interval (PI) for the response Y_f associated with a future covariate x_f of interest; the PI should have a predetermined coverage level/size. Formally, we call $\mathcal{C}_{1-\alpha}(X_f)$ an ‘oracle’ PI of

size $1 - \alpha$ at a new covariate $X_f = x_f$, if under (1.2.1),

$$P(Y_f \in \mathcal{C}_{1-\alpha}(X_f) | X_f = x_f) = 1 - \alpha.$$

In practice, the oracle PI is unknown; it has to be estimated through a statistical procedure, i.e., given data $\{(X_i, Y_i)\}_{i=1}^n$ the statistician produces an estimate $\hat{\mathcal{C}}_{1-\alpha}(X_f)$ such that $P(Y_f \in \hat{\mathcal{C}}_{1-\alpha}(X_f) | X_f = x_f) \approx 1 - \alpha$. This estimated PI is called *conditionally valid* by some authors, see e.g., Lei and Wasserman [2014]; we will further discuss the issue of conditionality in prediction intervals in Section 1.5.

Under the general regression setup of (1.2.1), there are at least three avenues in constructing prediction intervals, namely: naive quantile estimation, conformal prediction, and the Model-free Bootstrap (MFB); they are all discussed in detail in Section 1.3.2.

To give a short preview: quantile estimation(QE) amounts to estimating the conditional distribution of Y given X ; the term ‘naive’ is due to the fact that using the quantiles of the estimated conditional distribution of Y_f given $X_f = x_f$ to create a PI for Y_f sweeps under the carpet the variability of estimation which invariably results in undercoverage, i.e., the size of the PI will be less than then nominal $1 - \alpha$ in finite samples.

First proposed in Vovk et al. [2005], conformal prediction (CP, for short) is an ingenious technique for generating prediction intervals for i.i.d. or just exchangeable data $Z_1 \dots, Z_n$; its basic idea is to gauge how well a new data point Z_{n+1} “conforms” to the observed sample $\{Z_i\}_{i=1}^n$ via a self-chosen “conformity score”. A tutorial on conformal prediction can be found in Shafer and Vovk [2008].

There has been a recent surge in extending the conformal prediction idea to non-i.i.d. statistical setups. Notably, Lei et al. [2013](see also Lei et al. [2018]) extended this idea to the regression setup (1.2.1), where the newly proposed conformal prediction algorithm constructs $\hat{\mathcal{C}}_{1-\alpha}$ that satisfies the following no-undercoverage guarantee:

$$\mathbb{P}(Y_f \in \hat{C}_{1-\alpha}(X_f)) \geq 1 - \alpha. \quad (1.2.2)$$

The PI is marginally valid as the probability is evaluated unconditionally, based on the randomness of $\{(X_i, Y_i)\}_{i=1}^n \cup (X_f, Y_f)$; its coverage guarantee relies on the exchangeability of the $n + 1$ data pairs (X_i, Y_i) as implied by (1.2.1). However, the PI generated by the algorithm does not guarantee conditional validity as shown in Lei and Wasserman [2014] where a new algorithm was proposed to construct a PI with asymptotic "local validity". In Romano et al. [2019], the authors proposed a conformalized quantile regression approach which offers further improvement for heteroscedastic data. A recent endeavor to extend regression based conformal prediction to guarantee conditional validity is due to Chernozhukov et al. [2019]; see Section 1.3.2 for details.

The fact that the bootstrap –by generating a large number of artificial samples– can also capture the variability of estimated quantities inherent in prediction makes it an ideal candidate for predictive inference. Recall that the original bootstrap of Efron [1979] is applicable to i.i.d. data that are then sampled with replacement to produce the bootstrap resample. Politis [2015] proposed the “*Model-free prediction principle*” which amounts to using an invertible transformation to map a complex, non-i.i.d. dataset, to a dataset consisting of i.i.d. variables. The motivation is that prediction is easier –even trivial– with i.i.d. data; one can carry out the prediction in the i.i.d. domain and then map back via the inverse transformation. Furthermore, the “*Model-free bootstrap*” amounts to performing the standard bootstrap in the i.i.d. domain and then map back, yielding bootstrap PIs in the complex data domain. In a general regression setting, Politis [2013] suggested the use of the probability integral transform (PIT) –based on locally estimated conditional distributions– in order to transform the response data $Y_1, \dots, Y_n$ to a sample of n i.i.d. variables.

Our main focus of the paper is to understand the commonalities and differences between quantile estimation, conformal prediction and the Model-free Bootstrap (MFB). In Section 1.3, we contrast model-free to model-based regression, and discuss several different PI constructions.

In Section 1.4, we carry out a preliminary asymptotic analysis for the coverage probability under the different methods. Then, in Section 1.5 we discuss how PI validity could/should be assessed, i.e., conditionally or not, and prove a key lemma that reconciles conditional with unconditional validity in a big data setting. In Section 1.6, we discuss further differences of the three algorithms in terms of their ranges of applicability, and show theoretically that MFB intervals achieve better coverage compared to QE. The implicit reason is that MFB can produce *pertinent* PIs, a notion introduced by Politis [2015] in the setting of model-based, nonparametric regression; we extend this notion to model-free regression. Finally, in Section 1.7 we introduce the new concept of *conjecture testing* that is the analog of hypothesis testing as applied to the prediction problem; we also devise a modified conformal score to allow conformal prediction to handle one-sided ‘conjecture tests’, and compare to the Model-free bootstrap. Our theoretical findings are backed by simulation studies in Section 1.8.

1.3 Model-free vs. model-based regression

1.3.1 Model-based regression

The study of interval estimation is almost a century old, with its origins dating back to Neyman and Jeffreys [1937]. The extension to prediction intervals(PI) was soon to follow; see Patel [1989] and Geisser [1993] for a historical introduction. The traditional –although now outdated– approach was to construct $\hat{C}_{1-\alpha}$ based on the normal distribution assumption. To describe it, consider first the rather trivial case where the Y_i do not depend on the X_i at all.

Case I: I.i.d. responses Assume for a moment that the responses $Y_1, \dots, Y_n$ are i.i.d.; to conform with the classical tradition, let us assume that the common distribution is $N(\mu, \sigma^2)$. Then the interval

$$C_{1-\alpha}(x_f) = (\mu + z_{\alpha/2}\sigma, \mu - z_{\alpha/2}\sigma) \quad (1.3.1)$$

is an oracle PI of size $1 - \alpha$. Here, $z_{\alpha/2}$ is the lower $\alpha/2$ quantile of the standard normal distribution.

Since both parameters of the normal distribution are unknown, (1.3.1) has to be estimated by replacing μ and σ by their unbiased estimators $\hat{\mu}$ and $\hat{\sigma}$, i.e., sample mean and sample variance of data $\{Y_i\}_{i=1}^n$. However, in order to achieve the exact size $1 - \alpha$, we also need to adjust the quantile to be from a t_{n-1} distribution instead of the normal distribution, forming the estimated PI as

$$\hat{C}_{1-\alpha}(x_f) = \left(\hat{\mu} + t_{n-1, \alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n}}, \hat{\mu} - t_{n-1, \alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n}} \right) \quad (1.3.2)$$

where $t_{n-1, \alpha/2}$ is the lower $\alpha/2$ quantile of the t distribution with $n - 1$ degrees of freedom. The above is based on the fact that

$$\frac{Y_f - \hat{\mu}}{\hat{\sigma} \sqrt{1 + \frac{1}{n}}} \sim t_{n-1}$$

instead of $\mathcal{N}(0, 1)$.

If we did not account for the randomness of $\{Y_i\}_{i=1}^n$, that precipitates the randomness of $\hat{\mu}$ and $\hat{\sigma}$, then we would simply use (1.3.1) plugging in $\hat{\mu}$ and $\hat{\sigma}$ for μ and σ respectively. Since $z_{\alpha/2} > t_{n-1, \alpha/2}$, this would lead to *undercoverage*, as it does not take into account the variance related to the estimation procedure. The estimation variance is also present in the extra $1/n$ term inside the square root; although $1/n$ vanishes asymptotically, ignoring it in finite samples further induces undercoverage.

The above simple example provides a framework towards understanding the contribution of this paper, i.e., that it is imperative to take into account the variance related to model-fitting and estimation of all features that are to be plugged-in the oracle PI. We can now attempt some further generalizations.

Case II: Model-based linear regression Assume the data are generated by the model

$$Y_i = \beta'X_i + Z_i \text{ for } i = 1, \dots, n \quad (1.3.3)$$

where $\beta \in \mathbb{R}^d$ is the (unknown) parameter vector, and the errors Z_i are i.i.d. $N(0, \sigma^2)$.

Remark 1.3.1. We may now assume that the regressors X_i are fixed, i.e., non-random, or that they are random but independent of the errors; in the latter case, the following discussion will be understood in terms of the conditional probability given $X_i = x_i$ for $i = 1, \dots, n$.

Letting $\hat{\beta}$ and $\hat{\sigma}^2$ denote the Least Squares (LS) estimator of β and σ^2 , the analog of (1.3.2) is

$$\hat{\mathcal{C}}_{1-\alpha}(x_f) = \left(\hat{\beta}'x_f + t_{n-d, \alpha/2} \hat{\sigma} \sqrt{1 + x_f'(X'X)^{-1}x_f}, \hat{\beta}'x_f - t_{n-d, \alpha/2} \hat{\sigma} \sqrt{1 + x_f'(X'X)^{-1}x_f} \right) \quad (1.3.4)$$

where X is the $n \times d$ matrix having X_i' as its i th row.

Case III: Model-based nonparametric regression Assume now the data are generated by the model

$$Y_i = \mu_{X_i} + Z_i \text{ for } i = 1, \dots, n \quad (1.3.5)$$

where μ_x is an unknown (but assumed smooth) function of $x \in \mathbb{R}^d$, and the errors Z_i are i.i.d. $N(0, \sigma^2)$; the caveat of Remark 1.3.1 still applies.

The quantity μ_x can be estimated by any smoothing method, e.g., kernel smoothing, local polynomials, etc. resulting in the estimator m_x . There is no exact distribution (like the t distribution) to use here, so we must resort to asymptotic normality of m_x . The analog of (1.3.4) is the PI

$$\hat{\mathcal{C}}_{1-\alpha}(x_f) = (m_{x_f} - z_{\alpha/2} \hat{\sigma} V_{x_f}, m_{x_f} + z_{\alpha/2} \hat{\sigma} V_{x_f}) \quad (1.3.6)$$

that only has *approximate/asymptotic* coverage $1 - \alpha$. Here, $\hat{\sigma}$ is the natural estimate of σ , and V_{x_f} is a feature of the smoothing method; see eq. (3.17) of Politis [2015] for a concrete example based on kernel smoothing.

Model (1.3.5) can be extended to allow for heteroscedasticity, i.e., letting

$$Y_i = \mu_{X_i} + \sigma_{X_i} \varepsilon_i \text{ for } i = 1, \dots, n \quad (1.3.7)$$

where μ_x and σ_x are unknown (but assumed smooth) functions of $x \in \mathbb{R}^d$, and the errors ε_i are i.i.d. $N(0, 1)$. An analogous PI to (1.3.6) ensues, having first estimated σ_x via nonparametric smoothing. Nevertheless, this PI, as well as the previously mentioned PIs, i.e., (1.3.2), (1.3.4) and (1.3.6), are only valid if/when the regression errors are *exactly* normal, an assumption that is rarely justifiable nowadays.

Furthermore, it is of interest to go beyond the assumption of a regression equation driven by i.i.d. errors since even the most general/flexible such equation, namely (1.3.7), is unnecessarily restrictive. For example, (1.3.7) implies that the skewness and kurtosis of the response Y_i is constant, i.e., does not depend on X_i . In that respect, (1.3.7) can not be assumed to hold even for `cps71`, a dataset available within the `np` package of R that has served as a workhorse for nonparametric regression for decades; see Politis [2013] for details.

Note that in the random regressor case, a model equation such as (1.3.7), or any one of the more restrictive homogeneous and/or parametric versions such as (1.3.3) or (1.3.5), all imply (1.2.1) –under the caveat of Remark 1.3.1– but not vice versa. Therefore, assumption (1.2.1) is less restrictive even than general nonparametric regression as defined by eq. (1.3.7). Working under assumption (1.2.1) has been termed *Model-free regression* by Politis [2015] in the sense that it is devoid of a model equation such as (1.3.3), (1.3.5) or (1.3.7); this is the subject of the following subsection.

1.3.2 Model-free regression

We now revert to the “Model-free” setup of eq. (1.2.1) which is re-stated for convenience here:

$$\{(X_i, Y_i)\} \text{ for } i = 1, \dots, n \text{ are i.i.d. from } F. \quad (1.3.8)$$

Note that the above implies a random design; it is also possible to conduct Model-free regression with a fixed design as discussed in Politis [2015] but we limit the present discussion to setup (1.3.8).

Naive quantile estimation. The quantile estimation (QE) approach is straightforward; an oracle PI would be

$$(F_{y|x_f}^{-1}(\alpha/2), F_{y|x_f}^{-1}(1 - \alpha/2)). \quad (1.3.9)$$

Since the conditional quantile $F_{y|x}^{-1}(p) = \inf\{y \in \mathbb{R} : F(y|x) \geq p\}$ is unknown, we may estimate it by the appropriate quantile of the CDF estimator, i.e.,

$$\hat{F}_{y|x}^{-1}(p) = \inf\{y \in \mathbb{R} : \hat{F}(y|x) \geq p\}.$$

There is a wide range of selections when it comes to estimating the conditional CDF estimation, which we will mention in section 1.3.3. Having settled on a choice of estimator, we could plug it in (1.3.9) resulting in the PI

$$(\hat{F}_{y|x_f}^{-1}(\alpha/2), \hat{F}_{y|x_f}^{-1}(1 - \alpha/2)). \quad (1.3.10)$$

The above plug-in approach is termed ‘naive’ because it does not account for the estimation error of the quantiles; hence, the PI (1.3.10) is typically characterized by undercoverage in finite samples. In the next section, asymptotic analysis shows both the bias and variance of this error

will affect coverage performance.

Model-free bootstrap. Consider an arbitrary point predictor $\hat{Y}_f$ for the response to be observed at future point x_f ; mathematically, $\hat{Y}_f$ is a function of the observed data $\{(X_i, Y_i)\}$ and x_f , i.e., $\hat{Y}_f = g(\{(X_i, Y_i)\}_{i=1}^n, x_f)$; $\hat{Y}_f$ can also be regarded as some functional of the estimated predictive distribution, denoted by $\mathcal{J}(\hat{F}_{y|x_f})$. In addition, let $\mathcal{J}(F_{y|x_f})$ be the true point predictor that $\hat{Y}_f$ is estimating. The prediction error $R_f = Y_f - \hat{Y}_f$ is termed the "predictive root" in bootstrap literature; through resampling R_f , we can approximate the quantiles of the distribution of R_f , which then can be used to calculate the bootstrap PI centered at $\hat{Y}_f$.

The original bootstrap of Efron [1979] is applicable to i.i.d. data that are then sampled with replacement to produce the bootstrap resample. Politis [2015] proposed the "*Model-free prediction principle*" which amounts to using an invertible transformation to map a complex, non-i.i.d. dataset, to a dataset consisting of i.i.d. variables. The motivation is that prediction is easier –even trivial– with i.i.d. data; one can carry out the prediction in the i.i.d. domain and then map back via the inverse transformation. Furthermore, the "*Model-free bootstrap*" (MFB) amounts to performing the standard bootstrap in the i.i.d. domain and then map back, yielding bootstrap PIs in the complex data domain.

In the model-free regression setting, Politis [2013] suggested the use of the probability integral transform (PIT) –based on locally estimated conditional distributions– in order to transform the response data $Y_1, \dots, Y_n$ to a sample of n i.i.d. variables. To elaborate, define the ranks $U_i = F(Y_i|X_i)$ that are i.i.d. $Unif(0, 1)$ due to the PIT, assuming $F_{y|x}$ is continuous in y for all x ; hence, this transformation performs a *uniformization* of the responses. Meanwhile, we can recover the data Y_i by the identity $Y_i = F_{y|X_i}^{-1}(U_i)$, i.e., by utilizing the inverse PIT. Asymptotic validity of the MFB for bootstrap confidence –as opposed to prediction– intervals was recently shown in Wang and Politis [2021a].

Remark 1.3.2. Note that by letting $U_i = F(Y_i|X_i = x_i)$ where x_i was the realization of X_i , and finding that the distribution of U_i conditionally on $X_i = x_i$ is $Unif(0, 1)$, i.e., not depending

on the value of x_i , implies that (a) the unconditional distribution of U_i is also $Unif(0, 1)$, and (b) that U_i is independent of X_i . Hence, the collection of conditional CDFs $\{F_{y|x} : x \in \mathbb{R}^d\}$ connects two equivalent probability spaces: one is the space of $\{(X_i, Y_i)\}_{i=1}^n$, the other is the space of $\{(X_i, U_i)\}_{i=1}^n$, where X_i and U_i are two mutually independent sequences. In the latter, the regression/prediction problem has been trivialized, bringing it to the i.i.d. setting of Case I of Section 1.3.1.

Here, we propose the following model-free bootstrap algorithm for predictive inference under the distributional setup in (1.3.8). The original MFB algorithm of Politis [2015] treated the design points X_i as fixed during the resampling; the algorithm below is similar albeit the X_i 's are also resampled.

Algorithm 1.3.1. MFB prediction interval at future point x_f —random regressor case:

1. Use the given data $\{(X_i, Y_i)\}_{i=1}^n$ to estimate the conditional CDF $\hat{F}_n(\cdot | X_i)$ and its inverse.
2. Calculate estimated ranks $\hat{U}_i = \hat{F}_n(Y_i | X_i)$.
3. Choose a type of predictor $\hat{Y}_f = g(\{(X_i, Y_i)\}_{i=1}^n, x_f)$, i.e., choose the function g , that will be the center of the prediction interval. The predictive root is defined by

$$R_f = Y_f - \hat{Y}_f.$$

4. Create $X_1^*, \dots, X_n^*$ by resampling the regressors $\{X_i\}_{i=1}^n$, i.e., by randomly sampling with replacement from the set $\{X_1, \dots, X_n\}$.
5. Create $U_1^*, \dots, U_n^*$ by resampling the $\{\hat{U}_i\}_{i=1}^n$. Let $Y_i^* = \hat{F}^{-1}(U_i^* | X_i^*)$. Let Y_f^* be a sample from distribution $\hat{F}(y | X = x_f)$, and $\hat{Y}_f^* = g(\{(X_i^*, Y_i^*)\}_{i=1}^n, x_f)$. The bootstrap predictive root is

$$R_f^* = Y_f^* - \hat{Y}_f^*.$$

6. The above step is repeated B times to create B replicates of R_f^* . Denote by q_α^* the α -quantile of the empirical distribution of the B replicates of R_f^* . The bootstrap prediction interval is then defined as

$$(\hat{Y}_f + q_{\alpha/2}^*, \hat{Y}_f + q_{1-\alpha/2}^*). \quad (1.3.11)$$

For comparison, we state below the original algorithm model-free bootstrap algorithm, i.e., Algorithm 4.4.1 in Politis [2015]

Algorithm 1.3.2. MFB prediction interval at future point x_f —fixed regressor case:

The algorithm is identical to Algorithm 1.3.1 except that we replace Step 4 by:

4'. Let $X_i^* = X_i$.

Remark 1.3.3. Politis [2015] also proposed some different variations to the basic Algorithm 1.3.2. For example, the *Limit* model-free bootstrap omits step (2) that calculates estimated ranks. Instead, in step 3 the U_i^* s are directly sampled from a $Unif(0, 1)$ distribution, thus reducing computational cost and simplifying the analysis for proofs. For predictive inference, the author also recommends using the *predictive* model-free approach, which estimates the t^{th} rank $\hat{U}_t^{(-t)} = \hat{F}^{(-t)}(Y_t|X_t)$ where $\hat{F}^{(-t)}$ is estimated through the delete- t dataset $\{(X_i, Y_i)\}_{i=1}^n / (X_t, Y_t)$.

Distributional conformal prediction. As mentioned in the Introduction, a recent endeavor to extend conformal prediction to guarantee conditional validity in regression is distributional conformal prediction (CP). This was proposed by Chernozhukov et al. [2019] who used the aforementioned uniformization transformation –based on a locally estimated PIT– to calculate the conformity score as follows.

For a candidate $y \in \mathbb{R}$ for which we want to test whether it belongs to the PI, $\hat{F}_{n+1}(\cdot|x)$ is an estimator for the conditional distribution function $F(\cdot|x)$, based on $n+1$ data $\{(X_i, Y_i)\}_{i=1}^n \cup (X_f, y)$. Define the sample conditional ranks $\hat{U}_i'(X_f, y) = \hat{F}_{n+1}(Y_i|X_i)$ for $i = 1, \dots, n$ and $\hat{U}_{n+1}'(X_f, y) =$

$\hat{F}_{n+1}(y|X_f)$. The conformity scores are defined as $\hat{V}_i(y) = |\hat{U}_i'(X_f, y) - 1/2|$, and the p -value for the null hypothesis that y conforms to the data is

$$\hat{p}(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}(\hat{V}_i(y) \geq \hat{V}_{n+1}(y)).$$

The $1 - \alpha$ prediction interval is defined as the range of y where $\hat{p}(y) > \alpha$, i.e.

$$\hat{\mathcal{C}}_{1-\alpha}(X_f) = \{y : \hat{p}(y) > \alpha\}.$$

With certain assumptions, Chernozhukov et al. [2019] showed that this procedure will guarantee not only unconditional validity, i.e., a finite-sample guarantee of unconditional coverage not less than $1 - \alpha$, but also asymptotic *conditional* validity in the following sense:

$$\mathbb{P}(Y_f \in \hat{\mathcal{C}}_{1-\alpha}(X_f) | X_f = x_f) \rightarrow 1 - \alpha$$

as $n \rightarrow \infty$. Interestingly, this is what bootstrap prediction intervals, such as (1.3.11), typically satisfy. Hence, it is of interest to further compare the CP intervals with the MFB ones.

1.3.3 On choosing conditional CDF and quantile estimator

The conditional CDF and its quantiles plays an important role in all three above mentioned methods. The two quantities are directly related in that we can construct a valid quantile estimator from a CDF estimator, and vice versa. For example, by inverting the CDF estimator, we get the quantile estimator; to construct the CDF estimator from estimated quantiles, the following relationship is convenient:

$$F_{y|x}(u) = \int_0^1 \mathbb{1}(F_{y|x}^{-1}(\tau) \leq u) d\tau. \quad (1.3.12)$$

As one of the fundamental problem in statistics, there exists a plethora of estimators with

guarantees of consistency. Considering the volume of different types of distributions, methods that do not rely on a particular parametric model for the conditional distribution are preferred in practice. We now list two such methods which will be used for our comparisons of the three PI methods.

1. **Nonparametric CDF estimation.** This method constructs the estimator via nonparametric kernel smoothing as follows:

$$\hat{F}(y|x) = \frac{\frac{1}{n} \sum_{i=1}^n W_h(X_i, x) K(\frac{y-Y_i}{h_0})}{\bar{W}_h(x)}, \quad (1.3.13)$$

with $W_h(X_i, x) = \frac{1}{h^d} \prod_{s=1}^d w(\frac{X_{i,s}-x_s}{h})$ and $\bar{W}_h(x) = \frac{1}{n} \sum_{i=1}^n W_h(X_i, x)$; $w(\cdot)$ is a univariate, symmetric density function with bounded support, and K is a proper CDF; h and h_0 are bandwidths that converge to 0, whose optimal rates depend on asymptotic analysis. Under finite sample setting, we choose h and h_0 according to existing empirical rule or through cross validation, see Ch. 6, Li and Racine [2006]. Nonparametric kernel estimation works well when the dimension d of the covariates is low, e.g, $d \leq 3$. When dealing with high-dimensional covariates, using the formulation in Equation (1.3.13) may lead to poor performance as a result of curse of dimensionality. Instead of using a product kernel, Ch.4, Politis [2015] demonstrated a simplifying approach by exploiting the univariate kernel combined with a certain form of distance function $d_\theta(\cdot, \cdot)$ in $\mathbb{R}_d$, and $W_h(X_i, x) = \frac{1}{h} w(d_\theta(\cdot, \cdot)/h)$. This helps when the dimension of X_i is high, or even with functional-type covariates whose dimension is infinite.

Equation (1.3.13) is suitable for estimation when x is not a boundary point in the range of covariates. Near the boundary, local linear –as opposed to local averaging– estimation is preferable; see Das and Politis [2020] for details.

2. **Regression quantiles.** The regression quantile approach can alleviate the curse of dimensionality when estimating conditional quantiles with high-dimensional covariates, and

has attracted attention both in statistics and econometrics. It seeks to estimate $F_{y|x}^{-1}(\tau)$ by searching in the linear span of the covariates $\{X'\beta, \beta \in \mathbb{R}^d\}$ such that the expected quantile loss $\mathbb{E}\rho_\tau(Y - X'\beta)$ is minimized. Here, $\rho_\tau(r) = r(\tau - \mathbb{1}(r < 0))$ is the check function. The rationale behind using regression quantiles is that the global minimizer $q(x)$ to $\mathbb{E}\rho_\tau(Y - q(x))$ is exactly the τ -quantile $F_{y|x}^{-1}(\tau)$, and we restrict our search for $q(x)$ in the linear span of x . β is estimated by the minimizer of the sample quantile loss:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n \rho_\tau(Y_i - X_i'\beta), \quad (1.3.14)$$

which is often calculated via numerical algorithms. The optimization problem in eq. (1.3.14) can be further adjusted to estimate $\hat{\beta}$ under high-dimensional sparse scenarios, see Chap.15 of Koenker et al. [2017].

1.4 Preliminary asymptotic analysis

The methods introduced in Section 1.3 share some common features. Besides the fact that all of them use the conditional CDF function as an indispensable tool, the distributional conformal prediction and Model-free bootstrap both utilize the PIT to get estimated ranks $\hat{U}_t$ as a middle step towards further inference. However, there are also important differences that distinguish them apart, and may result in different performance. In this section, we conduct asymptotic analysis for the three methods, with a focus on the coverage probability unconditional on the data $(\mathbf{X}_n, \mathbf{Y}_n) = \{(X_i, Y_i)\}_{i=1}^n$, and conditioning on the future covariate, i.e.

$$\mathbb{P}_{x_f}(Y_f \in \hat{\mathcal{C}}_{1-\alpha}(X_f)) := \mathbb{P}(Y_f \in \hat{\mathcal{C}}_{1-\alpha}(X_f) | X_f = x_f). \quad (1.4.1)$$

As a first step, we perform analysis for the QE method to examine the factors that can affect coverage. This, along with analysis for the other two methods in the appendix, facilitates a

conclusion on the commonalities of factors that contribute to coverage bias, which we later put forth as a claim below.

Naive quantile estimation. In the example of QE, the estimated conditional quantile $\hat{F}_{y|x}^{-1}(p)$ is used to construct the prediction interval. Rewriting the conditional coverage probability $\mathbb{P}_{x_f}(Y_{n+1} \in \hat{C}_{1-\alpha}(X_{n+1}))$ under this setup, we have:

$$\mathbb{P}_{x_f}\left(Y_{n+1} \in \left(\hat{F}_{y|x_f}^{-1}(\alpha/2), \hat{F}_{y|x_f}^{-1}(1 - \alpha/2)\right)\right). \quad (1.4.2)$$

Denoting the quantile estimation errors $A_n = \hat{F}_{y|x_f}^{-1}(\alpha/2) - F_{y|x_f}^{-1}(\alpha/2)$ and $B_n = \hat{F}_{y|x_f}^{-1}(1 - \alpha/2) - F_{y|x_f}^{-1}(1 - \alpha/2)$, the above probability equals to

$$\mathbb{E}_{B_n} F_{y|x_f}\left(F_{y|x_f}^{-1}(1 - \alpha/2) + B_n\right) - \mathbb{E}_{A_n} F_{y|x_f}\left(F_{y|x_f}^{-1}(\alpha/2) + A_n\right) \quad (1.4.3)$$

By a second order Taylor expansion, expression(1.4.3) equals to

$$\begin{aligned} & (1 - \alpha) + f_{y|x_f}\left(F_{y|x_f}^{-1}(1 - \alpha/2)\right) \mathbb{E}(B_n) - f_{y|x_f}\left(F_{y|x_f}^{-1}(\alpha/2)\right) \mathbb{E}(A_n) \\ & + \frac{1}{2} f'_{y|x_f}\left(F_{y|x_f}^{-1}(1 - \alpha/2)\right) \mathbb{E}(B_n^2) - \frac{1}{2} f'_{y|x_f}\left(F_{y|x_f}^{-1}(\alpha/2)\right) \mathbb{E}(A_n^2) + \text{higher-order terms.} \end{aligned} \quad (1.4.4)$$

With the decomposition $\mathbb{E}(A_n^2) = \text{Var}(A_n) + \mathbb{E}^2(A_n)$, Equation(1.4.4) can also be written in the form of a bias-variance expansion.

Different choices of CDF estimators affect the rate of convergence for coverage bias. In the case of the nonparametric CDF estimator, we have the rate for for the bias and variance as follows (also see Theorem 6.4, Li and Racine [2006]):

Lemma 1.4.1. *When $\hat{F}_{y|x_f}$ is the nonparametric kernel estimator, both A_n and B_n converge to 0 in probability at rate of $\mathcal{O}_p(h^2 + 1/\sqrt{nh})$, with $\mathbb{E}(A_n) = \mathcal{O}(h^2)$ and $\text{Var}(A_n) = \mathcal{O}(1/nh)$.*

This lemma can be used to indicate the rate of the optimal bandwidth h such that the asymp-

otic rate of convergence of the nonparametric kernel estimator $\hat{F}_{y|x_f}$ is optimal; as usual, with nonnegative kernels the optimal rate for h is of order $n^{-1/5}$.

In the case of quantile regression, we can achieve asymptotic efficiency, i.e., a $\sqrt{n}$ -convergence rate for the estimated quantiles, *provided the implied linear regression model for the quantiles is correct*; when the model is false, however, the bias can be non-negligible and further leads to incorrect coverage.

The takeaway of the above calculations is that the first two moments regarding the model estimation error, A_n and B_n , contribute the most to the bias in conditional coverage probability. Consequently, we do not have a no-undercoverage guarantee in finite samples for this simple method. As it turns out, asymptotic analysis for the other two methods also agrees with this conclusion, which we state as a claim below.

Claim 1.4.1. *Let e_{total} be the total estimation error regarding the procedures of QE, CP or MFB. Then the first two factors contributing to the coverage bias are the first and second moments of e_{total} , i.e.,*

$$\text{Coverage bias} = \mathcal{O}(\mathbb{E}_n e_{total}) + \mathcal{O}(\text{Var}_n(e_{total})).$$

Granted, e_{total} takes on different forms for the three methods. For example, since QE and CP are essentially based on the estimated quantiles, e_{total} will be the sum of the quantile estimation errors. However, in MFB the total error also includes the error of predictor estimation, and e_{total} of MFB has extra terms comparing with to QE/CP. Since all errors must converge to 0 in order for each method to be asymptotically valid, validity for MFB requires more assumptions, as we later show in Theorem 1.6.2. We argue that extra terms in e_{total} do not necessarily mean more undercoverage – the sign of coefficients for the extra error terms controls this matter.

The derivation of asymptotic results for CP and MFB are presented in the Appendix section. Note that asymptotic theory will play a less important role in a finite-sample setting, and the performance between them can vary as well. Hence, empirical experiments and simulation

are typically conducted to compare predictive performance in finite samples.

1.5 The conditionality problem

One exciting result about conformal prediction which inspired many recent explorations in this area is the marginal/unconditional validity property as given in eq. (1.2.2). The basis of (1.2.2) relies on both $(\mathbf{X}_n, \mathbf{Y}_n)$ and (X_f, Y_f) being random and exchangeable, or more relevantly, i.i.d. in the model-free regression setup (1.3.8); in addition, the coverage probability should be unconditional with respect to the data. Discussions about appropriateness of assumptions are mostly focused on the data itself, such as the i.i.d. assumption for the data $(\mathbf{X}_n, \mathbf{Y}_n)$ which is standard in the regression literature; or equi-distribution assumption for the future pair (X_f, Y_f) , which led to recent advances in conformal prediction that proposed solutions when there is a distributional shift for X_f , see e.g, Tibshirani et al. [2019].

What is equally important but less focused on, is the appropriate level of conditionality with which we should evaluate the PI's coverage probability. More specifically, at which level should we integrate randomness of data into the probability evaluation. This so-called conditionality problem has been a crucial concept in statistical inference going beyond prediction intervals. Ever since the proposal of the conditionality principle of Cox [1958] and its connection to the likelihood principle (see Birnbaum [1962]), statisticians have debated on the correctness of these principles and their implications in statistical inference. For a review, see Robins and Wasserman [2000], which also points out that the issue of conditionality does not involve mathematical rigor; instead, arguments regarding its virtue are purely a statistical concern.

In resonance with conditionality for parameter inference problems, new findings are recently discovered regarding the impact of conditioning for prediction problems under a regression setup; see e.g., Rosset and Tibshirani [2020] and Bates et al. [2021], where new tools are also developed for better predictive inference under conditioning. In this section, we study the evalu-

ation of coverage probabilities under different levels of conditioning, and derive relationships between them. We connect different conditioning setups with real world scenarios, and suggest the appropriate level of conditioning under different circumstances. The conditionality problem also leads to further analysis of the MFB algorithm in Section 1.6.

As a first step, we introduce three probability measures with different levels of conditioning. For simplicity, they are denoted as $\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3$ respectively:

$$\begin{aligned}\mathbb{P}_1(\cdot) &= \mathbb{P}_{x_f}(\cdot) = \mathbb{P}(\cdot | X_f = x_f), \\ \mathbb{P}_2(\cdot) &= \mathbb{P}(\cdot | X_f = x_f; \mathbf{X}_n), \\ \mathbb{P}_3(\cdot) &= \mathbb{P}(\cdot | X_f = x_f; (\mathbf{X}_n, \mathbf{Y}_n)).\end{aligned}\tag{1.5.1}$$

We have the following tower property regarding the hierarchical relationship between $\mathbb{P}_k : k \in \{1, 2, 3\}$.

Lemma 1.5.1. *(Tower property) Let $A \in \sigma(\mathbf{X}_n, \mathbf{Y}_n, X_f, Y_f)$ be an arbitrary measurable event; here $\sigma(\cdot)$ denotes σ -algebra. Then, we have that $\mathbb{P}_2(A) = \mathbb{E}_{\mathbf{Y}_n | \mathbf{X}_n} \mathbb{P}_3(A)$ and $\mathbb{P}_1(A) = \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n} \mathbb{P}_3(A)$.*

Remark 1.5.1. Bootstrap calculations (including the MFB) are typically conducted conditionally on the data. The so-called ‘bootstrap world’ is governed by probability measure $\mathbb{P}_3$, and bootstrap validity is typically proven in probability (or almost surely) with respect to the ‘real world’ probability $\mathbb{P}$. Note that showing validity (or asymptotic validity) under $\mathbb{P}_3$ is stronger, implying validity (or asymptotic validity) under $\mathbb{P}_2$ or $\mathbb{P}_1$. To see why, assume $\mathbb{P}_3(A) \rightarrow 1 - \alpha$ in probability; then the expectation of $\mathbb{P}_3(A)$ will also tend to $1 - \alpha$ by the bounded convergence theorem, i.e., $\mathbb{P}_1(A) \rightarrow 1 - \alpha$ and $\mathbb{P}_2(A) \rightarrow 1 - \alpha$. Hence, working under a $\mathbb{P}_3$ performance measure which is more stringent may be generally preferable. Remark 1.5.2 below offers a reconciliation of the different viewpoints in a big data setting.

Consider the following data scenarios.

A. Scenarios with respect to data. The data $\{(X_i, Y_i)\}_{i=1}^n$ belong to one of the following

conditions:

1. Both X_i and Y_i are random, with i.i.d. $(X_i, Y_i) \sim F$.
2. $(X_1, \dots, X_n)$ are fixed covariates; $\{Y_i\}_{i=1}^n$ are considered random with distribution $Y_i|X_i \sim F_{y|X_i}$.
3. The data $\{(X_i, Y_i)\}_{i=1}^n = (\mathbf{X}_n, \mathbf{Y}_n)$ are observations that were drawn from the joint distribution F but are now treated as fixed.

The above three scenarios are directly coupled with the above probability measures, i.e., scenario (A.i) corresponds to $\mathbb{P}_i$, for any $i = 1, 2, 3$.

Remark 1.5.2. [Big data vs. limited data.] By definition of $\mathbb{P}_3$ in eq. (1.5.1), the instance-specific $\mathbb{P}_3$ is a random probability measure which inherits its randomness from the observations $(\mathbf{X}_n, \mathbf{Y}_n)$; from lemma 1.5.1 it follows that $\mathbb{P}_1$ and $\mathbb{P}_2$ are expectations of $\mathbb{P}_3$. By the Strong Law of Large Numbers, $\mathbb{P}_3(A)$ will converge to $\mathbb{P}_1(A)$ for any $A \in \sigma(\mathbf{X}_n, \mathbf{Y}_n, X_f, Y_f)$ as the sample size $n \rightarrow \infty$. Therefore, under a **big data** regime, $\mathbb{P}_1$ and $\mathbb{P}_2$ are not very different from $\mathbb{P}_3$, and are equally suitable for coverage evaluation. As setups (A.1) and (A.2) are coupled with $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively, we see that treating the data as random has validity under the big data regime. On the other hand, when n is finite, $\mathbb{P}_3$ can be very different from the other two probability measures; this phenomenon was recently put forth in Bates et al. [2021]. Contrary to the big data regime, the finite- n case is a **limited data** regime where the data should be regarded as fixed observations, i.e., setup (A.3); this suggests the use of $\mathbb{P}_3$ as being more appropriate for coverage evaluation with limited data.

To make more concrete the above big data vs. limited data dichotomy, we provide three leading examples to illustrate scenarios (A.1), (A.2) and (A.3) respectively.

1. **Randomized experiments.** Commonly known as A/B testing, randomized experiments is a prevalent tool for treatment effect estimation. Given a large pool of data $\{X_i\}_{i=1}^n$, it

involves random selection and assignment of candidates into control/treatment groups. Formally put, $\mathcal{S}_1, \mathcal{S}_2 \subset \{1, \dots, n\}$ are disjoint random index sets with $\max(|\mathcal{S}_1|, |\mathcal{S}_2|) \ll n$, and $\{X_{s_1}\}_{s_1 \in \mathcal{S}_1}, \{X_{s_2}\}_{s_2 \in \mathcal{S}_2}$ are the covariates of treatment/control groups, with responses $\{Y_{s_1}\}_{s_1 \in \mathcal{S}_1}, \{Y_{s_2}\}_{s_2 \in \mathcal{S}_2}$ are collected for these covariates. In this case, the data pair (X, Y) in each group should be regarded as random, and $\mathbb{P}_1$ should be the correct measure for conditional coverage evaluation.

2. **Regression for categorical data.** Given continuous covariates X and response Y , binning (categorization) is a common practice that transforms X into a handful of categories in order to eliminate unnecessary randomness in the covariates and reduce complexity of data. While Y can still be considered random, binning is a procedure that interpolates between random design and fixed design. It is therefore reasonable to consider $\mathbb{P}_2$ when the binning procedure produces fixed covariates.
3. **Few-shot learning and modeling of rare events.** The phrase "few-shot learning" is adapted from the machine learning community which describes scenarios where training data (X, Y) is extremely limited in quantity; this is naturally a limited data scenario. In the case of rare events, X and Y are observations that are extreme values compared to the population distribution. In this case, the coverage probability conditioning on these observations are clearly distinct from one without conditioning. Therefore, $\mathbb{P}_3$ should be the correct probability measure for coverage.

Remark 1.5.3. [Conditionality under prediction.] The conditionality principle was originally proposed in a parameter estimation context; in one of its many versions it states the following: Let θ be a parameter of interest, then inference with respect to θ , such as evaluating coverage of a CI, should be conditioned on all ancillary statistics, and also relevant subsets; see Robins and Wasserman [2000] for details.

Under a prediction context, the different levels of conditioning also represent whether

we think the data is ancillary and/or relevant subsets with respect to PI coverage. In this regard, $\mathbb{P}_2$ and $\mathbb{P}_3$ can be better measures of coverage comparing to $\mathbb{P}_1$ under certain circumstances. In the above examples, example 2 is related to viewing the covariates as ancillary statistics; while example 3 is related to viewing the entire observations as relevant subsets.

Together with scenarios (A.1), (A.2) and (A.3), we also introduce a second set of scenarios about the future covariate X_f .

B. Distribution assumption for the future covariate.

1. X_f is random with distribution F_X .
2. X_f is a fixed future design point in $\mathbb{R}^d$.

(B.1) assumes that the future covariate is randomly chosen; therefore the marginal coverage $\mathbb{P}(Y_f \in \mathcal{C}(X_f))$ is actually a weighted average of conditional coverages: $\mathbb{E}\mathbb{P}_{x_f}(Y_f \in \mathcal{C}(X_f))$. Being able to achieve a $1 - \alpha$ marginal coverage does not imply an $1 - \alpha$ conditional coverage. On the other hand, an $1 - \alpha$ conditional coverage will guarantee $1 - \alpha$ marginal coverage. To this aspect, validity under (B.2) is stronger than that under (B.1).

In recent years, CP has gained increasing interest in the machine learning community in that it is able to guarantee no undercoverage for a finite number of data under exchangeability; this is a consequence of (A.1) and (B.1)—see Lei et al. [2018]. However, CP does not guarantee conditional validity. The CP variant of Chernozhukov et al. [2019] offers improvement to Lei et al. [2018] in that it achieves asymptotic conditional validity under (A.1) and (B.2). Yet *asymptotically* this method is on par with the performance of naive QE when considering conditional coverage, as demonstrated in section 1.4 and appendix 1.9.1. Thus, CP may be better suited for big data scenarios, where data and future covariate can be regarded random.

It has been known that there is no algorithm that has finite sample no-undercoverage guarantees for conditional validity; see Lemma 1 of Lei and Wasserman [2014]. Instead, correctness of a prediction algorithm is demonstrated via asymptotic conditional validity. As we

will show rigorously in the following section, the MFB approach delivers asymptotic coverage validity under any combination of conditions in scenario A and B. Granted, this alone does not justify using MFB over QE and CP as both of them are also asymptotically conditionally valid. Therefore, besides asymptotic validity, we will introduce two extra properties of MFB prediction intervals, namely *pertinence* and *improved coverage* over the other two methods.

1.6 Further analysis of Model-free bootstrap

Recall the two MFB algorithms from Section 1.3.2. Algorithm 1.3.1 is referred to as the random regressor scheme, while Algorithm 1.3.2 is the fixed regressor scheme—which can also be considered when inference is to be made conditionally on the design covariates $\mathbf{X}_n$. Algorithm 1.3.1 takes advantage of the distributional setup of the covariates, and conforms to the setup where the study of interest is the coverage probability unconditional on the data $\mathbf{X}_n, \mathbf{Y}_n$. Since both $\mathbb{P}_2$ and $\mathbb{P}_3$ assume that the design covariates $\mathbf{X}_n$ is fixed, Algorithm 1.3.2 is tailored for scenarios (A.2) and (A.3), while Algorithm 1.3.1 is more suitable for (A.1). Asymptotically, the two algorithms are equivalent.

1.6.1 Asymptotic pertinence and better coverage guarantees

In terms of interval prediction algorithms, while asymptotic validity guarantees correctness, it alone does not justify good finite sample coverage performance. Section 3.6 of Politis [2015] made a clear illustration of this point and defined the notion of “pertinence” of prediction intervals. Formally speaking, a “pertinent” prediction interval should be able to capture both the variability of future response Y_f correctly, as well as the variance due to estimation. Although the latter is usually asymptotically negligible, being able to capture it will improve finite sample performance. As far as we know, to capture the variance due to estimation (without resorting to simplifying assumptions such as normality) requires some form of bootstrap.

Politis [2015] formulated the notion of “pertinent” prediction intervals in a model-based regression setting such as (1.3.7). Here, we extend the definition of asymptotic pertinence to the model-free regression setup of (1.3.8).

Definition 1.6.1. Asymptotic pertinence of a model-free prediction interval. A prediction interval generated by the bootstrap predictive root R_f^* (as defined in Algorithms 1.3.1 and 1.3.2) is asymptotically *pertinent* if it satisfies all of the following:

- (1). Both R_f and R_f^* can be decomposed into the following representations:

$$R_f = \varepsilon_f + e_f,$$

$$R_f^* = \varepsilon_f^* + e_f^*$$

Here, $\varepsilon_f = Y_f - \mathcal{J}(F_{y|x_f})$ is the non-degenerate variable concerning the distribution of future response Y_f , and e_f is model estimation error which converges to 0 in probability; ε_f^* and e_f^* are their bootstrap analogues.

- (2). $\sup_x |\mathbb{P}^*(\varepsilon_f^* \leq x) - \mathbb{P}(\varepsilon_f \leq x)| \xrightarrow{P} 0$ as $n \rightarrow \infty$.
- (3). There is a diverging sequence of positive numbers a_n such that $a_n e_f$ and $a_n e_f^*$ converge to nondegenerate distributions, and $\sup_x |\mathbb{P}^*(a_n e_f^* \leq x) - \mathbb{P}(a_n e_f \leq x)| \xrightarrow{P} 0$ as $n \rightarrow \infty$.
- (4). ε_f is independent of e_f in the real world, i.e., in unconditional probability $\mathbb{P}_1$; similarly, ε_f^* is independent of e_f^* in the bootstrap world, i.e., in conditional probability $\mathbb{P}_3$.

In section 1.6.2, we give closed forms for the decompositions of part (1) of Definition 1.6.1, and show that the MFB algorithm produces asymptotic pertinent prediction intervals, a unique advantage compared to both QE and CP, which do not satisfy the decomposition in Definition 1.6.1. The effect of pertinence will also be shown in section 1.8, where we run parallel comparison for experiments with large vs. small sample size n , and show that asymptotic validity and pertinence each plays a more important role in the two separate cases.

In addition, we now state a general claim that will be substantiated in what follows.

Claim 1.6.1. *Under mild conditions, the MFB has better coverage probability compared with QE regardless of the type of CDF/quantile estimator one chooses, and regardless of which of the three probability measures $\mathbb{P}_k$ one chooses to evaluate coverage.*

To elaborate, we will show that there exist an integer N and $\alpha_0 \in (0, 1)$, such that for all $\alpha < \alpha_0$ and $n \geq N$, we have

$$\mathbb{P}_k \left(Y_f \in \hat{\mathcal{C}}_{1-\alpha}^{MF}(X_f) \right) > \mathbb{P}_k \left(Y_f \in \hat{\mathcal{C}}_{1-\alpha}^{QE}(X_f) \right) \text{ for any } k \in \{1, 2, 3\}. \quad (1.6.1)$$

Note that for the case of $k = 3$, Equation (1.6.1) holds in probability.

Remark 1.6.1. In Section 1.4, we showed that QE and CP have almost identical asymptotic expressions in terms of conditional coverage probability, see Equation (1.9.2) in Appendix 1.9.1. Thus for large n , we may replace $\mathbb{P}_k \left(Y_f \in \hat{\mathcal{C}}_{1-\alpha}^{QE}(X_f) \right)$ with $\mathbb{P}_k \left(Y_f \in \hat{\mathcal{C}}_{1-\alpha}^{CP}(X_f) \right)$ in Equation (1.6.1), showing that MFB will also have better coverage than CP for large enough n .

1.6.2 Required technical assumptions

The idea of proving the property of asymptotic pertinence and also Equation (1.6.1) goes as follows. We first consider the setup of (A.3), which is both the natural real-world scenario as well as the correct setup in the bootstrap world. We show that under (A.3) the decomposition of Definition 1.6.1 holds for MFB, in contrast with QE and CP. Next, to prove Equation (1.6.1), we make mild assumption on the estimator $\hat{F}_{y|x_f}$, under which the convolution of $\hat{F}_{y|x_f}$ with a (asymptotic negligible) Gaussian kernel will lead to more heavy-tailedness of the distribution. In the last step, lemma 1.5.1 is invoked to extend our result under (A.3) to (A.1) and (A.2).

Assumption 1.6.1. *We make the following general assumptions:*

(C1). The sequence of conditional CDF estimators $\hat{F}_{y|x}$ (indexed by n) is consistent to a continuous CDF $F_{y|x}$ in the sup norm:

$$\sup_u |\hat{F}_{y|x}(u) - F_{y|x}(u)| \xrightarrow{P} 0.$$

(C2). The MFB is asymptotically valid for the inference problem with respect to $\hat{Y}_f$ in the following sense, as $n \rightarrow \infty$:

1. The predictor $\hat{Y}_f = g(\{(X_i, Y_i)\}_{i=1}^n, x_f)$ and $\hat{Y}_f^* = g(\{(X_i^*, Y_i^*)\}_{i=1}^n, x_f)$ in the bootstrap world satisfy a central limit theorem: $\exists \sigma_\infty^2 \in \mathbb{R}_+, \tau_n \rightarrow \infty$ such that

$$\begin{aligned} \tau_n (\hat{Y}_f - \mathbb{E} \hat{Y}_f) &\xrightarrow{D} \mathcal{N}(0, \sigma_\infty^2); \\ \tau_n (\hat{Y}_f^* - \mathbb{E}^* \hat{Y}_f^*) &\xrightarrow{D^*} \mathcal{N}(0, \sigma_\infty^2), \text{ in probability.} \end{aligned}$$

Where $\mathbb{E} \hat{Y}_f = \mathcal{J}(F_{y|x_f}) + o_p(1/\tau_n)$ and $\mathcal{J}$ is a functional that maps the CDF function to $\mathbb{R}$; similarly, $\mathbb{E}^* \hat{Y}_f^* = \mathcal{J}(\hat{F}_{y|x_f}) + o_{p^*}(1/\tau_n)$.

2. $\mathcal{J}(\hat{F}_{y|x_f}) \xrightarrow{P} \mathcal{J}(F_{y|x_f})$.

(C3). For large enough n , $\hat{F}_{y|x_f}$ belongs to the following class of CDFs:

$$\mathcal{F} = \{F : \exists \sigma_0, u_0 \in \mathbb{R}_+, \forall \sigma < \sigma_0, u > u_0, \bar{F}(u) < \overline{F \star \phi_\sigma}(u); \bar{F}(-u) > \overline{F \star \phi_\sigma}(-u)\}.$$

Here, $\bar{F}(u) = 1 - F(u)$ is the tail distribution function and $\star$ is the convolution operator; ϕ_σ is the density function of the normal distribution $\mathcal{N}(0, \sigma^2)$.

Remark 1.6.2. In practice, two types of predictors are typically considered: the mean which minimizes the L^2 loss between future response Y_f and the predictor $\hat{Y}_f$, and the median that minimizes the L^1 loss. For the median case, under assumption (C1) and Polya's theorem, item 2 in

(C2) is guaranteed and no further assumptions are necessary; while in the case of the mean, on top of (C1), we also need uniform integrability of the sequence $\hat{F}_{y|x_f}$ to ensure $\mathcal{J}(\hat{F}_{y|x_f}) \xrightarrow{P} \mathcal{J}(F_{y|x_f})$.

The above conditions can be easily achieved in general. (C1) and (C2) are classic assumptions regarding validity of statistical procedures. Plus, we make no assumption on the rate of convergence for either the conditional CDF estimator nor the predictor at which we center prediction intervals. (C3) is a less intuitive assumption: it implies that the distribution of $Y_{x_f} + \sigma Z$ has heavier tails than that of Y_{x_f} when $Y_{x_f} \sim \hat{F}_{y|x_f}$ and $Z \sim \mathcal{N}(0, 1)$ independent of Y_{x_f} . However, we will show there is in fact a large class of distributions that falls into this class. We start with the following lemma:

Lemma 1.6.1. *Let $f(u)$ be a density function defined on $\mathbb{R}$ where (i) $\exists u_0 > 0$ such that $f(u)$ is third-order differentiable and $|f^{(3)}(u)|$ is bounded for $|u| > u_0$. (ii) $f(u)$ is convex for $|u| > u_0$, i.e. $f''(u) > 0$. Then the CDF of f satisfies (C3).*

Remark 1.6.3. (a). Condition (ii) in lemma 1.6.1 is satisfied by a large class of distributions. Since (C3) is a property of the tail distribution function, it is natural to consider distributions that are in the class of regularly varying functions $\mathbf{RV}(\rho)$ at infinity, where the tail CDF $\bar{F}(u)$ (as well as $\bar{F}(-u)$) satisfies

$$\frac{\bar{F}(cu)}{\bar{F}(u)} \xrightarrow{u \rightarrow \infty} c^{-\rho}, \rho > 0;$$

so that the tail CDF behaves like a power function $u^{-\rho}$ at infinity (multiplied by a slowly varying function $L(u)$). If L, L' and L'' behave "nicely", such that

1. L is positive and third-order differentiable near infinity;
2. L, L' and L'' all belong to the following class of functions

$$\mathcal{H} := \left\{ H : \lim_{|u| \rightarrow \infty} \frac{H'(u)}{H(u)/u} = 0 \right\}$$

then $\exists C > 0$, $f''(u) \sim Cu^{-(\rho+3)}L(u) > 0$. The class $\mathcal{H}$ is a subset of slowly varying functions, see Resnick [2007].

(b). Another interpretation related to the convolution problem in (C3) is the following. Given $\hat{F}_{y|x_f}$ that is consistent to $F_{y|x_f}$ in the sup norm, then we have that $\hat{F}_{y|x_f} \star \phi_\sigma$ is a C^∞ CDF that approximates $\hat{F}_{y|x_f}$ in the limit when $\sigma \rightarrow 0$. In addition, we have the following result:

$$\sup_x |\hat{F}_{y|x_f} \star \phi_\sigma(x) - F_{y|x_f} \star \phi_\sigma(x)| \leq \sup_x |\hat{F}_{y|x_f}(x) - F_{y|x_f}(x)|, \quad (1.6.2)$$

where the equality holds only when $|\hat{F}_{y|x_f}(x) - F_{y|x_f}(x)|$ equals to the sup norm at every x . Then the implication using the predictive root for bootstrap is two-fold: First of all, the root distribution is more smooth than $\hat{F}_{y|x_f}$ itself, and this can be beneficial to the bootstrap procedure, see Silverman and Young [1987]. Secondly, the convoluted CDF has better approximation performance than the plain CDF $\hat{F}_{y|x_f}$, this means bootstrapping the predictive root requires less data than only bootstrapping the future response in order to achieve the same performance. This agrees with our numerical findings in Section 1.8.

1.6.3 Results for MFB pertinence and Claim 1.6.1

The main theoretical results for our paper are summarized in the following theorems:

Theorem 1.6.2. (Asymptotic validity) *Assume (C1) with $x = x_f$; then the prediction intervals $\hat{C}(x_f)$ of QE is asymptotically valid. Further assume (C1) for all x , then CP is also asymptotically valid.*

$$\mathbb{P}_k \left(Y_f \in \hat{C}_{1-\alpha}(X_f) \right) \rightarrow 1 - \alpha \text{ as } n \rightarrow \infty \text{ for } k \in \{1, 2, 3\}. \quad (1.6.3)$$

(The convergence is in probability sense for $k = 2, 3$.) If we also assume $\hat{Y}_f^* - \hat{Y}_f \xrightarrow{P} 0$, then Equation (1.6.3) holds for the MFB as well.

Theorem 1.6.3. (Asymptotic Pertinence of MFB) *Assume (C1) and (C2); then, the predictive*

root R_f and R_f^* in the Model-free bootstrap is asymptotically pertinent in the sense of Definition 1.6.1.

Theorem 1.6.4. (Better performance guarantee) Assume (C1), (C2) and (C3); then the Model-free bootstrap has higher coverage comparing to the quantile estimation method in the sense of Equation (1.6.1).

The last theorem is in effect verifying our Claim 1.6.1. Proofs of the three theorems are found in Appendix 1.9.4.

We summarize our findings in Table 1.1.

Table 1.1: Comparison between conformal prediction and Model-free bootstrap.

	Conformal prediction	Model-free bootstrap
Applicable scenarios	Big data Random design Random future covariate	Big data and limited data Random/ fixed design Random/ fixed future covariate
Validity	Marginal validity Asymptotic conditional validity	Asymptotic marginal validity Asymptotic conditional validity
Advantage	Finite-sample marginal coverage guarantee	Pertinence of prediction intervals Guaranteed better coverage than QE.

1.7 Conjecture testing

1.7.1 Prediction vs. estimation

Consider the two general classes of statistical inference: parameter estimation and prediction. While prediction intervals are the analogs of confidence intervals in estimation, there is no clearly defined analog of hypothesis testing as applied to prediction. We aim to fill this gap here by defining the new notion of *conjecture* testing. Table 1.2 illustrates these analogies.

Conjecture testing is different from hypothesis testing in the following ways. First of all, conjecture testing attempts to answer a question regarding a future response Y_f which is

Table 1.2: Analogies between parameter estimation and prediction.

Inference type	Parameter estimation	Prediction
Point	Point estimation	Point prediction
Interval	Confidence interval	Prediction interval
Test	Hypothesis test	Conjecture test

an unobserved random variable; by contrast, (frequentist) hypothesis testing aims to answer a question regarding an unknown fixed (nonrandom) parameter θ . Secondly, the rejection region of a hypothesis test is derived from the distribution of the target statistic under the null hypothesis; by contrast, the probability measure by which a conjecture gets tested is one of the three mentioned, i.e., $\mathbb{P}_k$ (for some k), which remains unchanged when the null changes. In addition, contrary to hypothesis testing which often relies on central limit theorems to prove consistency of a particular test, validity of conjecture testing is directly associated with validity of prediction intervals, which essentially requires consistent estimation of the predictive distribution $F_{y|x_f}$. In this regard, as sample size grows, the acceptance region of hypothesis testing will shrink towards a single point θ , while the acceptance region of conjecture testing will not degenerate.

1.7.2 Conjecture testing: definition and discussion

Statistical hypothesis testing at a certain significance level α is a fundamental tool in statistical inference. Consider a parameter of interest θ whose value is estimated by a statistic $\hat{\theta}_n$. After defining the null hypothesis H_0 and the alternative H_1 , a frequentist hypothesis test can be conducted by figuring out the threshold of the test statistic that ensures $P_{H_0}(\text{reject } H_0) \leq \alpha$. Alternatively, the test can be conducted by computing the p -value and rejecting H_0 only if the p -value is less than α .

By the duality between hypothesis testing and confidence intervals (CI), we can equivalently construct a $1 - \alpha$ CI for θ and reject H_0 if the CI does not include the θ value(s) under H_0 . This duality applies equally to the standard two-sided CIs that correspond to a two-sided null

hypothesis, as well as one-sided CIs —where one of the CI bounds are $\pm\infty$ — that correspond to a one-sided null hypothesis.

While prediction intervals are the analogs of confidence intervals in estimation, there is no clearly defined analog of hypothesis testing as applied to prediction. We aim to fill this gap here by defining the new notion of *conjecture* testing. It all starts by formulating two conjectures (i.e., hypotheses) regarding the value of the future response Y_f , namely the null conjecture $C_{null} : Y_f \in S_{null}$ vs. the alternative conjecture $C_{alt} : Y_f \notin S_{null}$ where S_{null} is a set of interest.

By analogy with hypothesis testing, the simplest choices for S_{null} are:

1. $S_{null} = \{y_0\}$
2. $S_{null} = [y_0, \infty)$
3. $S_{null} = (-\infty, y_0]$

where y_0 is some given value of interest. Case 1 is a point null (with a two-sided alternative), while Cases 2 and 3 are one-sided tests. The aim is to try to make a decision between the two conjectures C_{null} and C_{alt} while controlling the probability of false rejection of C_{null} to be not more than α . The probability of false rejection can be measured with probability $\mathbb{P}_k$ for some appropriate k .

As in the parameter estimation paradigm, there is also a natural duality between conjecture testing and prediction intervals (PI). For example, in Case 1 of the point null, we could construct a $1 - \alpha$ PI for Y_f , and reject C_{null} at level α if $y_0 \notin \text{PI}$. If the PI's $1 - \alpha$ level is measured with probability $\mathbb{P}_k$ (for some k), the size of the test will also be according to the same $\mathbb{P}_k$.

It is quite common in practice to be interested in the behavior of a future response with respect to the predicted value in a particular direction, i.e., leading to one-sided conjectures and PIs. For example, when excessive risk is associated with response having higher (say) values than the projection; in this case, the practitioner would like to find a region of probable values for Y_f

that exclude high-risk extreme values. Applications can be found in many fields, such as climate control, system risk monitoring, Value at Risk (VaR), etc.

The above discussion motivates the one-sided tests of Cases 2 and 3. Focusing on Case 2—the other case being similar—we may start by constructing a $1 - \alpha$ lower one-sided PI of the type $(-\infty, c_f]$ for Y_f , such that $\mathbb{P}_k(Y_f \notin (-\infty, c_f]) \approx \alpha$. We can then reject $C_{null} : Y_f \in [y_0, \infty)$ when

$$y_0 \notin (-\infty, c_f]. \quad (1.7.1)$$

Note that in such a case, $\mathbb{P}_k(Y \geq y_0) \leq \mathbb{P}_k(Y_f \geq c_f) = \alpha$, i.e., the probability of false rejection of C_{null} will not exceed α as required.

Statistically speaking, under the setup of (1.3.8) with i.i.d. data and future covariate of interest x_f , to construct a one-sided PI we need to find $c \in \mathbb{R}$ such that

$$\mathbb{P}_k(Y_f - \hat{Y}_f \leq c) \approx 1 - \alpha \quad (1.7.2)$$

for either $k \in \{1, 2, 3\}$ depending on the appropriate data scenario. The inclusion of $\hat{Y}_f$ in (1.7.2) rules out using the QE and CP approaches. On the other hand, the Model-free bootstrap works readily since we can estimate c by the $1 - \alpha$ quantile of the bootstrap predictive root R_f^* , i.e., $\hat{c}^* = \hat{D}_{x_f}^{*-1}(1 - \alpha)$. Hence, the threshold c_f appearing in the rejection region (1.7.1) can be estimated by $\hat{c}_f = \hat{Y}_f + \hat{c}^* = \hat{Y}_f + \hat{D}_{x_f}^{*-1}(1 - \alpha)$.

Constructing a one-sided PI – and inverting it to obtain a one-sided conjecture test – is also possible using QE. To do that, we can simply use the left/right quantiles of the estimated conditional CDF; for example, a $(1 - \alpha)100\%$ one-sided PI based on QE reads: $(\hat{F}_{y|x_f}^{-1}(\alpha), \infty)$ or $(-\infty, \hat{F}_{y|x_f}^{-1}(1 - \alpha))$.

Remark 1.7.1. Interestingly, the popular CP methodology is not designed to yield one-sided PIs. One might then think that CP can not handle one-sided conjecture tests. Delving deeper, however, it becomes apparent that this limitation is only due to the two-sided nature of the

popular conformity score $\hat{V}_i(y) = |\hat{U}_i'(X_f, y) - 1/2|$. In order to handle one-sided PIs and one-sided conjecture tests via CP we may instead adopt a one-sided conformity measure, such as $\hat{V}_i(y) = -\hat{U}_i'(X_f, y)$ under which the produced PI will have the form (c, ∞) . We used this choice in our second simulation in Section 1.8.

In Section 1.8, we give a practical example of applying one-sided conjecture testing to examine accuracy of VaR prediction for high-frequency stock returns data.

1.8 Numerical experiments

1.8.1 Synthetic data

Experiment setup

We consider the following model for simulations. Let $\{(X_i, Y_i)\}_{i=1}^N$ be i.i.d. samples where

$$\begin{aligned} X &\sim \text{Unif}(0, 1), \\ Y &= \sin(\pi X) + \sigma \sqrt{1 + 2X} \varepsilon \end{aligned} \tag{1.8.1}$$

and $\varepsilon \sim T_5$. We set the future covariate $x_f = 0.5$ to guarantee data balance on the two sides of x_f , such that potential boundary effect will not appear in the simulations. The following procedure is conducted to estimate the conditional coverage probability(CVP) of size $1 - \alpha$ prediction intervals averaged on data, i.e., an estimate for $\mathbb{P}_1(Y_f \in \mathcal{C}(x_f))$: for each $k \in \{1, \dots, K\}$, generate $\{(X_i, Y_i)\}_{i=1}^N$ according to (1.8.1); then apply QE, CP and MFB on the data to form three $1 - \alpha$ prediction intervals $\hat{\mathcal{C}}_k^{(QE)}(x_f)$, $\hat{\mathcal{C}}_k^{(CP)}(x_f)$ and $\hat{\mathcal{C}}_k^{(MFB)}(x_f)$. Also, after fixing $X_f = x_f$, sample from (1.8.1) M times to get $\{(Y_f)_i\}_{i=1}^M$, and estimate the k -th coverage probability by the fraction of Y_f that are in the prediction interval, that is

$$\widehat{CVP}(\hat{\mathcal{C}}_k) = \frac{\sum_{i=1}^M \mathbb{1}\left((Y_f)_i \in \hat{\mathcal{C}}_k\right)}{M} \quad (1.8.2)$$

The collection of $\widehat{CVP}(\hat{\mathcal{C}}_k)$ forms an empirical distribution for the coverage probability under model (1.8.1). Then we can estimate $\mathbb{P}_1(Y_f \in \mathcal{C}(x_f))$ by the sample average $\frac{1}{K} \sum_{k=1}^K \widehat{CVP}(\hat{\mathcal{C}}_k)$.

Parameter tuning

The aforementioned parameters are set to be the following: $\sigma = 0.2$, $\alpha = 0.05$, $K = 200$ and $M = 3000$. Both the nonparametric and quantile regression based CDF estimators are considered in our simulations. For the nonparametric estimator, the bandwidth parameters h and h_0 need to be selected. We apply the following criteria for selection: within a grid of candidate bandwidths, the ones with which the uniformized series $\{U_i\}_{i=1}^N$ has the largest p -value in terms of the Kolmogorov-Smirnov test for uniformity are selected. As for the quantile regression based CDF, we apply the Barrodale and Roberts algorithm to estimate $F_{y|x}^{-1}(\tau)$ for τ in the grid $\{0.01i, 0 \leq i \leq 100\}$ and then calculate $F_{y|x}(u)$ using a discretized version of (1.3.12). In the bootstrap algorithm, we set $B = 1000$.

Asymptotic performance

We plot the estimated coverage probabilities, variances of the estimation $\text{Var}(\widehat{CVP}(\hat{\mathcal{C}}_k))$ (suitably scaled), as well as the average lengths of prediction intervals for different values of $N \in \{50i, i = 1, \dots, 8\}$ in Figure 1.1.

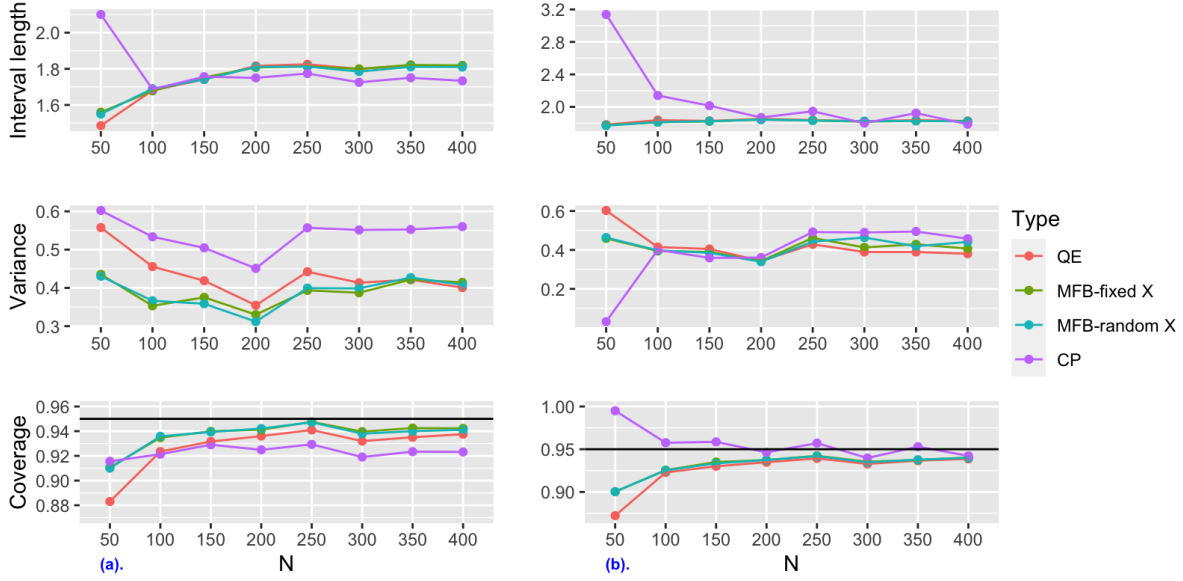


Figure 1.1: Asymptotic results for prediction intervals. (a): results based on the nonparametric CDF estimator; (b): results based on the quantile regression based CDF estimator.

We analyze the results in the following aspects.

1. Asymptotic validity. For the nonparametric CDF estimator, QE and MFB exhibit asymptotic validity as the coverage probabilities converge to 95% when N increases; while convergence for CP seems to stabilize around 92%. For the quantile regression based CDF estimator, all three algorithms converge to the nominal level.

2. Coverage comparison. The prediction intervals of MFB have higher coverage than QE for both CDF estimators, especially when the sample size N is small. This agrees with our result in Theorem 1.6.3, and also shows that MFB is able to alleviate undercoverage issue under limited data scenario. The behavior of CP varies based on the choice of CDF estimator: when the nonparametric CDF is used, CP performance is not as good as the other methods; while for the quantile regression CDF, CP in general has good performance except for the case $N = 50$, which we will discuss in further details in the next section.

3. Quality of estimation. The quality of prediction intervals is assessed by the variances of coverage probabilities and also the lengths of the prediction intervals. Differences between the

algorithms are most observable from the results that use the nonparametric CDF estimator. In terms of the variances, the MFB is consistently lower than QE and CP, while CP has the largest variance among the three methods. This is related to the pertinence property of the MFB that helps tracking the nominal level more consistently than the other methods. As for the lengths of prediction intervals, the MFB has equal length compared with QE while delivering higher coverage. On the contrary, CP has an anomaly at $N = 50$ where the length is significantly larger; as N increases, the lengths drop below those of QE and MFB, as do the coverage probabilities.

Coverage behaviors under limited data scenario

The behavior of CP under the limited data scenario ($N = 50$) is abnormal: for both CDF estimators, CP's lengths of prediction intervals are much higher than competing methods; for the quantile regression based estimator, the coverage is even close to 100%. Hence here, we further investigate the results for $N = 50$. In Figure 1.2 we plot the histograms of the coverage probabilities. As we see from the graphs, CP will "overcorrect" the prediction intervals to reach the nominal level 95%. In the case of using the nonparametric CDF estimator, the average coverage of CP is almost identical to that of MFB, but many of the prediction intervals are corrected to reach a 100% coverage, which also cause the interval lengths of CP to be much higher than those of MFB. In the case of using the quantile regression based CDF, this behavior is even more severe, causing the averaged coverage to be almost 100%. Therefore, the reliability of CP to generate conditionally valid prediction intervals is questionable under limited data setting.

1.8.2 VaR prediction for intraday stock returns data

We demonstrate an application of conjecture testing on high-frequency intraday stock returns data. To fix notations, let S_t denote the price process of a stock; $X_t = \log \frac{S_t}{S_{t-1}}$ denotes the log-return $\log \frac{S_t}{S_{t-1}}$ at time t . The concept of Value-at-Risk (VaR) is a risk measure in the following sense: for a finite time period m and a pre-specified level α , the VaR is the threshold c

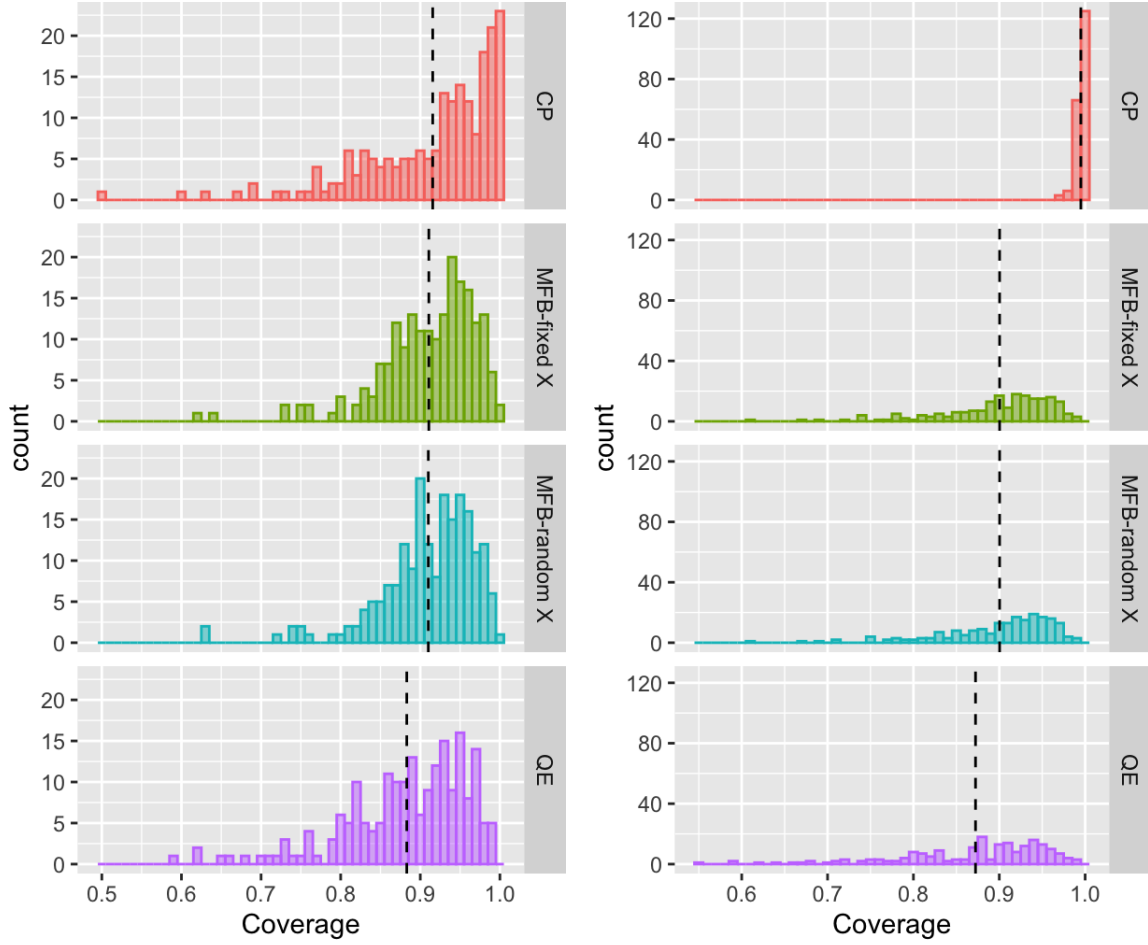


Figure 1.2: Histogram of coverage probabilities at $N = 50$. Left: results based on the nonparametric CDF estimator; right: results based on the quantile regression based CDF estimator.

such that the m -period return $T_m = \sum_{t=1}^m X_t$ is guaranteed to surpass c with probability $1 - \alpha$. In other words, $Var(\alpha) = F_m^{-1}(\alpha)$, where F_m is the CDF of T_m .

This concept can be extended in various ways. First of all, we can extend the statistic of interest beyond the case of total return $\sum_{k=t+1}^{t+m} X_k$, such as the form $T_m = f(X_1, \dots, X_m)$ where f is a continuous function. Examples include the worst m -period cumulative return, $T_m = \min_{1 \leq t \leq m} R_t$, where $R_t = \sum_{k=1}^t X_k$, and the worst one-period return $T_m = \min_{k=1}^m X_k$. These risk-type statistics will induce different forms of F_m . These examples are considered in Bertail et al. [2004].

Another extension is to find appropriate covariate that can have predictive power for

future VaR. For this purpose, we introduce additional dependence on the time parameter t for T_m : $T_{t+1,m} = f(X_{t+1}, \dots, X_{t+m})$. Also, let $V_{t,m}$ be the covariate of interest that is $\sigma(X_{t-m+1}, \dots, X_t)$ measurable. Then given $V_{t,m}$ we want to find the α -lower quantile for the conditional random variable $T_{t+1,m}|V_{t,m}$ as our VaR prediction for the future m -period returns. This is analogous to the one-sided prediction bounds described in Section 1.7, and we can use either QE or MFB to estimate the quantile of interest.

In our experiment, we consider 1-minute level return data during trading hours for our simulation. When extracting data, we remove those that are gathered from the first and last 5-minute in each trading day for quality control. We use the past m -period realized volatility $V_{t,m} = \sum_{k=t-m+1}^t X_k^2$ as the covariate and study its impact on future $T_{t+1,m}$ via VaR. We take $T_{t+1,m}$ to be the worst m -period cumulative return for the future data $(X_{t+1}, \dots, X_{t+m})$, and our experiments show that by choosing appropriate m there exhibits clear heteroscedascity in the paired data $(V_{t,m}, T_{t+1,m})$. Being able to capture the relationship between them will help individual traders to control their risk exposure, especially for day trading practitioners.

We formally setup the experiment as in the following. Let $\{X_n\}_{n=1}^N$ be the observed data up to time N ; our goal is to predict the VaR of $T_{N+1,m}$ conditioning on $V_{N,n}$. The regression model is trained on observed data from the past $\{(V_{i,m}, T_{i+1,m})\}_{i \in \mathcal{I}}$ where $\mathcal{I} \subset \{m, \dots, N-m\}$ is some index set. Because both $V_{i,m}$ and $T_{i+1,m}$ are aggregated statistics over m -periods, they each tends to have strong serial dependence for adjacent time indices which violates the i.i.d. assumption in our regression setup. To fix this issue, we only consider data from the time grid $\mathcal{I} = \{(2k+1)m, k \in \mathbb{N}\}$, such that $(V_{(2k+1)m,m}, T_{(2k+1)m+1,m})$ are based on non-overlapping blocks of data $\{(X_{2mk+1}, \dots, X_{2m(k+1)})\}, k \in \mathbb{N}\}$, thereby reducing the sample size by a factor of $2m$.

A new challenge arises from this procedure. On the one hand, as the dynamics of price process is time-varying, to ensure the joint distribution of $(V_{i,m}, T_{i+1,m})$ does not change over time, the past data $\{X_n\}_{n=1}^N$ needs to be from a short time frame, e.g., within 2 weeks so that the data size is less than $N_0 = (6.5 \times 60 - 10) \times 10 = 3800$. On the other hand, through our observation,

the heteroscedasticity phenomena are most pronounced when m is larger than some threshold m_0 , e.g. $m_0 = 30$, although the choice of m_0 varies for different stocks. By our data trimming procedure above, the sample size of the paired data is at most $|\mathcal{I}| = \lfloor \frac{N_0}{2m_0} \rfloor = 63$. This situation naturally falls into our limited data scenario introduced in Section 1.5.

We apply QE, MFB as well as CP using the nonparametric CDF estimator to predict the future VaR on two stocks, AMC and GME listed on NYSE; these two stocks have recently been very popular with day traders, and have exhibited abnormal stock price upsurges despite drops in revenue reported in their financial statements. Coupled with this phenomenon, the volatility and fluctuations in the returns of these two stocks are higher than market average. It is therefore of practical value to devise a more formal testing procedure.

We gathered data from April 10th to July 10th 2021 using the AlphaVantage stock APIs and performed post processing outlined in the above paragraphs to get pairs of (V, T) data. Our VaR predictions are calculated for $\alpha \in \{0.01, 0.05, 0.1\}$ on $(V_{t,m}, T_{t+1,m})$ with $m = 45$ for GME and $m = 30$ for AMC, where the training data used for testing either have $|\mathcal{I}| = 34$ or $|\mathcal{I}| = 68$ prior to time t . We perform conjecture testing based on predicted VaRs and future values $T_{t+1,m}$, with null hypothesis being $T_{t+1,m} > VaR_{pred}$. This testing procedure is carried out consecutively with respect to t so that we have a series of accept-reject results. Lastly, we calculate the empirical acceptance rate based on the results and compare it with the nominal level, $1 - \alpha$. We benchmark the methods by closeness between the empirical acceptance rate and the nominal level.

The results of our experiments are presented in Table 1.3 and 1.4. The methods used for comparison are: naïve quantile estimation(QE); Model-free bootstrap with L_1 and L_2 predictors (MFB- L_1 and MFB- L_2 respectively); and (one-sided) distributional conformal prediction(CP).

Based on the empirical results, both QE and MFB still exhibit undercoverage issue, while we find that overall MFB is able to boost the coverage of QE for both $|\mathcal{I}| = 34$ and $|\mathcal{I}| = 68$, exemplifying the result of Theorem 1.6.4; when sample size increases to $|\mathcal{I}| = 68$, the undercoverage issue is relieved to some level for both methods.

Measured by closeness to the nominal level, for the AMC stock the performance of CP is superior to other participating methods, while for GME the MFB has the edge. Notice that for the GME stock and some datapoints of the AMC stock, there are overcoverage issues for CP, echoing our analysis for the behavior of CP under limited-data scenario in the previous numerical experiment.

Table 1.3: Empirical acceptance rate results for GME.

Ticker: GME				
Sample size	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
$ \mathcal{I} = 34$	QE	95.9	94.1	88.2
	MFB- L^1	97.2	94.1	89.1
	MFB- L^2	97.7	94.1	88.6
	CP	99.5	97.7	91.4
$ \mathcal{I} = 68$	QE	96.7	94.1	89.8
	MFB- L^1	97.3	95.7	89.8
	MFB- L^2	98.6	95.1	89.8
	CP	99.4	96.7	92.5

Table 1.4: Empirical acceptance rate results for AMC.

Ticker: AMC				
Sample size	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
$ \mathcal{I} = 34$	QE	95.1	90.8	86.5
	MFB- L^1	95.1	91.6	87.6
	MFB- L^2	95.4	91.4	86.8
	CP	98.2	97.1	90.5
$ \mathcal{I} = 68$	QE	95.8	91.7	87.9
	MFB- L^1	97.1	92.6	88.2
	MFB- L^2	96.8	92.3	87.9
	CP	99.0	95.8	89.8

1.9 Additional proofs

1.9.1 Asymptotic analysis for CP

Denote $\hat{U}(x, y) = \hat{F}(y|x)$, $\hat{U}_i = \hat{U}(X_i, Y_i)$, and $\hat{U}_{(\alpha/2)}$, $\hat{U}_{(1-\alpha/2)}$ the $\alpha/2$ sample quantiles of $\{\hat{U}_i\}$. Note that in Chernozhukov et al. [2019], the conditional CDF is estimated with both $\{(X_i, Y_i)_{i=1}^n\}$ and (X_f, y) , which results in an augmentation to $\hat{U}_i$ that is denoted by $\hat{U}'_i(X_f, y) = \hat{F}_{n+1}(Y_i|X_i)$.

We assert that $\hat{U}_i$ and $\hat{U}'_i(X_f, y)$ are asymptotically equivalent. Specifically, let (X_f, Y_f) take on arbitrary values (x, y) ; then, $\sup_i |\hat{U}_i - \hat{U}'_i(x, y)| = \mathcal{O}_p(1/n)$. As both $\hat{U}_{(\alpha)}$ and $\hat{U}'_{(\alpha)}$ converge to α in probability, $\hat{U}_{(\alpha)} - \hat{U}'_{(\alpha)} = o_p(1)$. Therefore, we can use $\hat{U}_i$ instead for simplified asymptotic analysis.

Although in Chernozhukov et al. [2019], the estimated ranks $\hat{U}'_i$ are further transformed

to $\hat{V}_i$ with which the p -values are calculated, the key to generating prediction interval still relies on the ranks themselves. With this in mind, the conditional coverage probability approximately equals to

$$\mathbb{P}\left(\hat{U}(X_f, Y_f) \in \left(\hat{U}_{(\alpha/2)}, \hat{U}_{(1-\alpha/2)}\right) | X_f = x_f\right). \quad (1.9.1)$$

Let $\mathbb{P}_n$ and $\mathbb{E}_n$ denote the probability measure and expectation respectively associated with data $(\mathbf{X}_n, \mathbf{Y}_n)$. Then, by iterated expectation, (1.9.1) equals to

$$\begin{aligned} & \mathbb{E}_{y_f \sim F(\cdot|x_f)} \mathbb{P}_n \left(\hat{U}(x_f, y_f) \in \left(\hat{U}_{(\alpha/2)}, \hat{U}_{(1-\alpha/2)} \right) | X_f = x_f \right) \\ &= \mathbb{E}_{y_f \sim F(\cdot|x_f)} \mathbb{E}_n \mathbb{1} \left\{ \hat{U}(x_f, y_f) \in \left(\hat{U}_{(\alpha/2)}, \hat{U}_{(1-\alpha/2)} \right) \right\} \\ &= \mathbb{E}_n \mathbb{E}_{y_f \sim F(\cdot|x_f)} \mathbb{1} \left\{ y_f \in \left(\hat{F}_{y|x_f}^{-1}(\hat{U}_{(\alpha/2)}), \hat{F}_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)}) \right) \right\} \\ &\approx \mathbb{E}_n F_{y|x_f}(\hat{F}_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)})) - F_{y|x_f}(\hat{F}_{y|x_f}^{-1}(\hat{U}_{(\alpha/2)})) \end{aligned}$$

Similarly as before, denote $A'_n = \hat{F}_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)}) - F_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)})$ and $B'_n = \hat{F}_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)}) - F_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)})$ the errors of estimated quantiles; using Taylor expansion we get a similar expansion as in Equation (1.4.4), i.e.,

$$\begin{aligned} & \mathbb{E}_n F_{y|x_f}(\hat{F}_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)})) - F_{y|x_f}(\hat{F}_{y|x_f}^{-1}(\hat{U}_{(\alpha/2)})) \\ &= (1 - \alpha) + \mathbb{E}_n(f_{y|x_f}(F_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)}))) \mathbb{E}_n(B'_n) - \mathbb{E}_n(f_{y|x_f}(F_{y|x_f}^{-1}(\hat{U}_{(\alpha/2)}))) \mathbb{E}_n(A'_n) \\ &+ \frac{1}{2} \mathbb{E}_n(f'_{y|x_f}(F_{y|x_f}^{-1}(\hat{U}_{(1-\alpha/2)}))) \text{Var}_n(B'_n) - \frac{1}{2} \mathbb{E}_n(f'_{y|x_f}(F_{y|x_f}^{-1}(\hat{U}_{(\alpha/2)}))) \text{Var}_n(A'_n) \\ &+ \text{higher-order terms.} \end{aligned} \quad (1.9.2)$$

Hence, the bias in coverage probability is directly due to the bias and variance of the quantile estimator; in that sense, QE and CP perform very similarly.

The most notable difference between QE and CP is the addition of a new data point (X_f, y) in the original samples $\{(X_i, Y_i)\}_{i=1}^n$ to estimate $\hat{F}$, which results in the augmentation $\hat{U}'_i(X_f, y)$

from $\hat{U}_i$. An intuitive interpretation of this is that the new procedure "overfits" on a future observation (X_f, y) and may lead to better coverage with finite sample size. With exchangeability holding, the procedure has guaranteed unconditional validity, while conditional validity holds asymptotically.

1.9.2 Asymptotic analysis for MFB

We adopt the usual notation, where $\mathbb{P}$ denotes the probability measure for the original probability space, and $\mathbb{P}^*$ the probability measure in the bootstrap world, i.e., conditioning on the data $(\mathbf{X}_n, \mathbf{Y}_n)$. Given future covariate x_f , let D_{x_f} denote the distribution of the predictive root $R_f = Y_f - \hat{Y}_f$ under $\mathbb{P}$, and let $\hat{D}_{x_f}^*$ and $\hat{D}_{x_f}^{*-1}$ denote the distribution and quantile respectively of the bootstrap version R_f^* under $\mathbb{P}^*$; also assume that both are continuous functions, i.e., that R_f is a continuous random variable. The bootstrap prediction interval $\hat{\mathcal{C}}^{MF}(X_f)$ is $(\hat{Y}_f + \hat{D}_{x_f}^{*-1}(\alpha/2), \hat{Y}_f + \hat{D}_{x_f}^{*-1}(1 - \alpha/2))$, and is considered a random element under $\mathbb{P}$. The randomness of $\hat{\mathcal{C}}^{MF}(X_f)$ is due to variability across samples; given the observed sample, $\hat{\mathcal{C}}^{MF}(X_f)$ is a fixed quantity.

For the moment, let us focus our interest on the coverage probability without conditioning on the observed samples; hence the governing law is the original probability measure $\mathbb{P}$. In addition, we have the following lemma concerning the distribution of bootstrap samples.

Lemma 1.9.1. *1. Under $\mathbb{P}^*$, the distribution of X_i^* conditioning on $(\mathbf{X}_n, \mathbf{Y}_n)$ is $\frac{1}{n} \sum \mathbb{1}\{\cdot \leq X_i\}$;*

The distribution of $Y_i^* | X_i^* = X_k$ is $\hat{F}_{y|X_k}$.

2. Under $\mathbb{P}$, the distributions of X_i^* and Y_i^* are the expectations of their distributions under $\mathbb{P}^*$.

i.e., the distribution of X_i^* is F_X , and $Y_i^* | X_i^* = x$ follows $\mathbb{E}_n \hat{F}_{y|x}$.

3. Under $\mathbb{P}^*$, Y_f^* is independent of $\hat{Y}_f^*$, so the distribution of $R_f^* = Y_f^* - \hat{Y}_f^*$ is the convolution between $Y_f^* \sim \hat{F}_{y|x_f}$ and the distribution of $\hat{Y}_f^* = g(\{X_i^*, Y_i^*\}_{i=1}^n)$.

4. Under $\mathbb{P}$, Y_f is independent of $\hat{Y}_f$ as well, and the distribution D_{x_f} is the convolution of $F_{y|x_f}$ and the distribution of $\hat{Y}_f$.

Let $T_z f(x) = f(x+z)$ be the translation operator. For a CDF F , $(T_z F)^{-1}(u) = F^{-1}(u) - z$. Denote $\Delta_1^{(u)} = \hat{D}_{x_f}^{*-1}(u) - \left(T_{\mathcal{J}(\hat{F}_{y|x_f}^*)} \hat{F}_{y|x_f} \right)^{-1}(u)$; $\Delta_2 = \mathcal{J}(\hat{F}_{y|x_f}^*) - \mathcal{J}(\hat{F}_{y|x_f})$. Also, A_n and B_n are the same as in section 1.4.

Note that $\hat{Y}_f$ and R_f^* are dependent as they both depend on the data, so the previous route for coverage calculation can not be directly applied here. To carry out some asymptotic analysis, we make the following extra assumptions: $\hat{Y}_f$ and $\hat{Y}_f^*$ converge (respectively) to $\mathcal{J}(F_{y|x_f})$ and $\mathcal{J}(\hat{F}_{y|x_f}^*)$, which are estimands of the two, in probability and in L^1 . Also assume D_{x_f} and $\hat{D}_{x_f}^*$ are uniformly continuous, so that by Taylor expansion (δ -method), the distributions $D_{x_f}(x)$ and $\hat{D}_{x_f}^*(x)$ converge (respectively) to $T_{\mathcal{J}(F_{y|x_f})} F_{y|x_f}(x)$ and $T_{\mathcal{J}(\hat{F}_{y|x_f}^*)} \hat{F}_{y|x_f}(x)$, uniformly in x . As a consequence, $\forall u \in (0, 1)$, $\hat{D}_{x_f}^{*-1}(u) \xrightarrow{P^*} \left(T_{\mathcal{J}(\hat{F}_{y|x_f}^*)} \hat{F}_{y|x_f} \right)^{-1}(u) = \hat{F}_{y|x_f}^{-1}(u) - \mathcal{J}(\hat{F}_{y|x_f})$, and thus $\Delta^{(u)} \xrightarrow{P^*} 0$.

Consider the following decomposition holds:

$$\begin{aligned} \hat{Y}_f + \hat{D}_{x_f}^{*-1}(\alpha/2) &= F_{y|x_f}^{-1}(\alpha/2) + A_n + \Delta_1^{(\alpha/2)} + \Delta_2, \\ \hat{Y}_f + \hat{D}_{x_f}^{*-1}(1 - \alpha/2) &= F_{y|x_f}^{-1}(1 - \alpha/2) + B_n + \Delta_1^{(1-\alpha/2)} + \Delta_2 \end{aligned} \quad (1.9.3)$$

Denote the total error in (1.9.3) $e_{total} = A_n + B_n + \Delta_1^{(\alpha/2)} + \Delta_1^{(1-\alpha/2)} + \Delta_2$. Now by iterated expectation and Taylor expansion, the coverage probability for the PI (1.3.11) is

$$\begin{aligned} &\mathbb{E}_n \left(F_{y|x_f}(\hat{Y}_f + \hat{D}_{x_f}^{*-1}(1 - \alpha/2)) - F_{y|x_f}(\hat{Y}_f + \hat{D}_{x_f}^{*-1}(\alpha/2)) \right) \\ &= \mathbb{E}_n F_{y|x_f} \left(F_{y|x_f}^{-1}(1 - \alpha/2) + A_n + \Delta_1^{(1-\alpha/2)} + \Delta_2 \right) - \mathbb{E}_n F_{y|x_f} \left(F_{y|x_f}^{-1}(\alpha/2) + B_n + \Delta_1^{(\alpha/2)} + \Delta_2 \right) \\ &= (1 - \alpha) + O(\mathbb{E}_n e_{total} + \text{Var}_n(e_{total})) + \text{higher-order terms}. \end{aligned} \quad (1.9.4)$$

Despite different formalization, the above discussion shows the three methods have similar asymptotic behavior for their respective (unconditional) coverage probability, in the sense that the

source of coverage bias is from the first two moments of the total estimation error in each method.

1.9.3 Proof of Lemma 1.6.1

Proof. First of all, we assume (i) holds for all $u \in \mathbb{R}$. By Taylor's Theorem,

$$f(u+z) = f(u) + f'(u)z + \frac{f''(u)}{2}z^2 + R(z),$$

where the remainder $R(z)$ has the form

$$R(z) = \frac{f^{(3)}(\xi(z))}{3!}z^3$$

for some $\xi(z) \in (u, u+z)$. Since $|f^{(3)}(\xi(z))|$ is bounded, we have

$$\int_{\mathbb{R}} R(z)\phi_{\sigma}(z) dz = o(\sigma^2),$$

where $\phi_{\sigma}(z)$ is the density function for the normal distribution $\mathcal{N}(0, \sigma^2)$.

Denote the density of $Y + \sigma Z$, where Y has density $f(u)$ and $\sigma Z \sim \mathcal{N}(0, \sigma^2)$ as $\tilde{f}$. Then by convexity, $\forall |u| > u_0$

$$\begin{aligned} \tilde{f}(u) &= (f \star \phi_{\sigma})(u) = \int_{\mathbb{R}} f(u+z)\phi_{\sigma}(z) dz \\ &= \int_{\mathbb{R}} \left(f(u) + f'(u)z + \frac{f''(u)}{2}z^2 + R(z) \right) \phi_{\sigma}(z) dz \\ &= f(u) + \frac{f''(u)}{2}\sigma^2 + o(\sigma^2) \\ &> f(u) \end{aligned} \tag{1.9.5}$$

for small enough σ^2 . As a consequence, $\forall u > u_0$, $\overline{F \star \phi_{\sigma}}(u) = \int_u^{\infty} \tilde{f}(x) dx > \int_u^{\infty} f(x) dx = \bar{F}(u)$.

Analogously, $F \star \phi_{\sigma}(u) > F(u)$ for $u < -u_0$.

In the more general case where (i) is satisfied on the restricted domain $|u| > u_0, \forall \varepsilon > 0$, we can choose σ small enough where $\mathbb{P}(|\sigma Z| > u_0) < \varepsilon$, so that $\forall |u| > 2u_0, |u + \sigma Z|$ is concentrated within (u_0, ∞) with probability greater than $1 - \varepsilon$; therefore $f(u + \sigma Z)$ satisfies (i) and (ii) for $|u| > 2u_0$ with probability $1 - \varepsilon$. By the same procedure of Equation (1.9.5), this gives

$$\tilde{f}(u) > f(u)(1 - \varepsilon) + \frac{f''(u)}{2} \left(\sigma^2 - \int_{z > c_{\sigma, \varepsilon/2}} z^2 \phi_{\sigma}(z) dz \right) + o(\sigma^2)$$

where $c_{\sigma, \varepsilon/2}$ is the upper $\varepsilon/2$ quantile for the $\mathcal{N}(0, \sigma^2)$ distribution. Since ε is arbitrarily small and f is convex for $|u| > 2u_0$, we conclude that $\tilde{f}(u) > f(u)$ for $|u| > 2u_0$. The remainder of the proof is the same as above. $\square$

Note that we may relax condition (i) in lemma 1.6.1 where f'' exists and is absolutely continuous on $u > |u_0|$. Then $f^{(3)}$ exists and the remainder $R(z)$ has an integral form $R(z) = \int_{(u, u+z)} \frac{f^{(3)}(t)}{2} (u + z - t)^2 dt$, and with bounded $f^{(3)}$, $\int_{\mathbb{R}} R(z) \phi_{\sigma}(z) dz$ is also $o(\sigma^2)$.

1.9.4 Proofs of Theorems 1.6.2, 1.6.3 and 1.6.4

Proof of Theorem 1.6.2. By Lemma 1.2.1 of Politis et al. [1999], under (C1) with $x = x_f$, $\hat{F}_{y|x_f}^{-1}(\alpha) \xrightarrow{P} F_{y|x_f}^{-1}(\alpha), \forall \alpha \in (0, 1)$. As a consequence, the boundary points of $\hat{\mathcal{C}}_{1-\alpha}^{(QE)}(x_f)$ converges to those of $\mathcal{C}_{1-\alpha}(x_f) = \left(F_{y|x_f}^{-1}(\alpha/2), F_{y|x_f}^{-1}(1 - \alpha/2) \right)$, then we have that $\mathbb{P}_3 \left(Y_f \in \hat{\mathcal{C}}_{1-\alpha}^{(QE)}(X_f) \right) \xrightarrow{P} 1 - \alpha$.

With further assumption that (C1) holds for all x , then $\hat{U}(Y_i|X_i)$ (and $\hat{U}'(Y_i|X_i)$) has asymptotic uniform distribution in $[0, 1]$. To see this, note that for $X_i = x$ fixed, $\hat{U}(Y_i|X_i = x) - U(Y_i|X_i = x) \xrightarrow{P} 0$, then $\mathbb{P}(\hat{U}(Y_i|X_i = x) \leq u) \rightarrow u, \forall u \in [0, 1]$. Now

$$\mathbb{P}(\hat{U}(Y_i|X_i) \leq u) = \mathbb{E}_{X_i} \mathbb{E}_{Y_i|X_i=x} \mathbb{E}_{n-1} I(\hat{U}(Y_i|X_i = x) \leq u) = \mathbb{E}_{X_i} \mathbb{P}(\hat{U}(Y_i|X_i = x) \leq u),$$

then by dominated convergence theorem, $\mathbb{P}(\hat{U}(Y_i|X_i) \leq u) \rightarrow u$. Then for CP, the analysis in

section 1.4 shows that $\hat{\mathcal{C}}_{1-\alpha}^{(CP)}(x_f)$ is asymptotically equivalent to $\hat{\mathcal{C}}_{1-\alpha}^{(QE)}(x_f)$.

With the additional assumption $\hat{Y}_f^* - \hat{Y}_f \xrightarrow{P} 0$, $\hat{\mathcal{C}}_{1-\alpha}^{(MF)}(x_f)$ is asymptotically equivalent to $\hat{\mathcal{C}}_{1-\alpha}^{(QE)}(x_f)$, thus MFB is asymptotically valid.

Proof of Theorem 1.6.3. We decompose R_f and R_f^* as follows.

$$R_f = Y_f - \hat{Y}_f = \left(Y_f - \mathcal{J}(F_{y|x_f}) \right) - \left(\hat{Y}_f - \mathcal{J}(F_{y|x_f}) \right) := \varepsilon_f + e_f; \quad (1.9.6)$$

$$R_f^* = Y_f^* - \hat{Y}_f^* = \left(Y_f^* - \mathcal{J}(\hat{F}_{y|x_f}) \right) - \left(\hat{Y}_f^* - \mathcal{J}(\hat{F}_{y|x_f}) \right) := \varepsilon_f^* + e_f^*. \quad (1.9.7)$$

with $\varepsilon_f = Y_f - \mathcal{J}(F_{y|x_f})$, $\varepsilon_f^* = Y_f^* - \mathcal{J}(\hat{F}_{y|x_f})$; and $e_f = \mathcal{J}(F_{y|x_f}) - \hat{Y}_f$, $e_f^* = \mathcal{J}(\hat{F}_{y|x_f}) - \hat{Y}_f^*$.

Under assumption (C1), (2) in definition 1.6.1 is satisfied. Under assumption (C2), both $\tau_n e_f$ and $\tau_n e_f^*$ converge in distribution to normal distributions with equal variances, which shows (3) in definition 1.6.1.

Observe that ε_f only depends on Y_f , and the randomness of e_f depends on the data $(\mathbf{X}_n, \mathbf{Y}_n) \perp\!\!\!\perp Y_f$, therefore $\varepsilon_f \perp\!\!\!\perp e_f$. Similarly, conditioning on the data, $\varepsilon_f^* \perp\!\!\!\perp e_f^*$ in the bootstrap world. This shows (4).

Proof of Theorem 1.6.4. Under scenario (A3), assumption (C2) and previous result, the distribution of the bootstrap root (R_f^*) is asymptotically equivalent to the convolution between $\hat{F}_{y|x_f}$ and a Gaussian distribution $\mathcal{N}(0, \tau_n^2)$. Then assumption (C3) guarantees that $\exists \alpha_0 \in (0, 1)$, $n_0 \in \mathbb{N}$ such that $\forall \alpha < \alpha_0$ and $n > n_0$, $\hat{\mathcal{C}}_{1-\alpha}^{(QE)}(x_f) \subseteq \hat{\mathcal{C}}_{1-\alpha}^{(MF)}(x_f)$. This shows Equation (1.6.1) for $k = 3$. By Lemma 1.5.1, we can show (1.6.1) for the other two scenarios.

1.10 Acknowledgement

Chapter 1, in full, is a reprint of the materials in the paper “Model-free Bootstrap and Conformal Prediction in Regression: Conditionality, Conjecture Testing, and Pertinent Prediction Intervals”, Yiren Wang, Dimitris N. Politis. It has been submitted for consideration of publication

in a major statistics journal. The dissertation author is the primary investigator and author of this material.

Chapter 2

Model-free Bootstrap for a General Class of Stationary Time Series

2.1 Abstract

A model-free bootstrap procedure for a general class of stationary time series is introduced. The theoretical framework is established, showing asymptotic validity of bootstrap confidence intervals for many statistics of interest. In addition, asymptotic validity of one-step ahead bootstrap prediction intervals is also demonstrated. Finite-sample experiments are conducted to empirically confirm the performance of the new method, and to compare with popular methods such as the block bootstrap and the autoregressive (AR)-sieve bootstrap.

Keywords: m-approximation, bootstrap validity, transformation function, confidence intervals.

2.2 Introduction

The bootstrap, since its introduction by Efron [1979], has been an invaluable tool for statistical inference with independent data. Resampling for time series has also been a flourishing topic since the late 1980s. However, there are a plethora of ways to resample a stationary time series. It is always important to validate the correctness of such bootstrap procedures, i.e., to show their asymptotic validity and range of applicability with respect to common statistics. These problems have been well studied for popular methods like the block bootstrap and the autoregressive (AR)-sieve bootstrap. For a summary of the state-of-the-art, see McElroy and Politis [2019] and Kreiss and Paparoditis [2020].

In a dependent setup, the main purpose of bootstrap is two-fold: One is to obtain confidence intervals for a parameter of interest and/or conduct a hypothesis test. Another important aspect of time series analysis is forecasting. A standard setup is the following: Given the time series data $\{Y_t\}_{t=1}^n$, the goal is h -step ahead prediction, i.e., predicting Y_{t+h} for some integer $h \geq 1$. An optimal h -step ahead point predictor $\hat{Y}_{t+h}$ should minimize the expected loss between the true Y_{t+h} and itself, conditioned on the current data $\{Y_1, \dots, Y_n\}$. The most widely used loss functions are L^1 and L^2 . The L^2 loss, $E \left(\left(\hat{Y}_{t+h} - Y_{t+h} \right)^2 | Y_1, \dots, Y_n \right)$ is minimized by the conditional expectation $\hat{Y}_{t+h} = E(Y_{t+h} | Y_1, \dots, Y_n)$. The L^1 loss, $E \left(|\hat{Y}_{t+h} - Y_{t+h}| | Y_1, \dots, Y_n \right)$ is minimized by the conditional median $\text{med}(Y_{t+h} | Y_1, \dots, Y_n)$ instead.

Besides point predictors, prediction intervals and joint prediction regions are quite useful; since any point predictor will invariably incur an error, it is important to provide a range of values where the future point Y_{t+h} will be found with high probability. Prediction intervals can be constructed by approximating the distribution of the so-called *predictive root*, i.e. $Y_{t+h} - \hat{Y}_{t+h}$, and using the respective quantiles to produce upper and lower bounds. Approximating this distribution typically requires one to fit a specific model to the data, which enables a model-based resampling for Y_{t+h} and $\hat{Y}_{t+h}$ separately; see Pan and Politis [2016a] for a review.

However, model-fitting and prediction are two separate notions with very different objective functions. Cross-validation ideas that are currently popular attempt to link the two notions, in choosing a model that is actually good for predictive purposes. Nevertheless, it is possible for the practitioner to proceed directly to prediction without the intermediate step of model-fitting; this is the essence of the *model-free* prediction principle of Politis [2013], Politis [2015]. To describe it, the goal is to find an invertible transformation that transforms the data vector $(Y_1, \dots, Y_n)'$ to a new data vector $(e_1, \dots, e_n)'$ whose entries are independent and identically distributed (i.i.d.). One can then employ the i.i.d. bootstrap on the $e_1, \dots, e_m$ to generate $e_1^*, \dots, e_m^*$, and use the inverse transform to get bootstrap samples $Y_1^*, \dots, Y_m^*$ in the domain of the original data. Using $m = n$ is the standard framework for estimation and confidence intervals; interestingly, using $m = n + h$ allows us to equally address the problem of forecasting Y_{t+h} with prediction intervals.

Under regularity conditions, such a transformation always exists but is not unique; see Ch. 2.3.3 of Politis [2015]. The challenge for the practitioner is to use the structure of the data at hand in order to devise a transformation that works in the given setting, having features that can be estimated from the data. In a model-based approach, these steps are analogous to choosing a model, and then fitting the model using the data. Indeed, any model driven by i.i.d. errors can be used to define a transformation of the data towards the i.i.d. target; however, the power of the model-free approach is that it can work without restricting oneself to a model equation.

To elaborate, if the data arise as a stretch of a strictly stationary time series $\{Y_t\}$ with (absolutely) continuous distributions, then the Rosenblatt transformation (Rosenblatt [1952]) can be used to transform $Y_1, \dots, Y_n$ to a set of n i.i.d. Uniform random variables. In general, this application of the Rosenblatt transformation can not be implemented in practice because it involves n unknown conditional distribution functions. However, if additional structure is assumed, e.g., when $\{Y_t\}$ is a stationary Markov sequence, then this approach is feasible; see Pan and Politis [2016b] and Ch. 8 of Politis [2015].

To describe a different approach, recall the Linear Process Bootstrap (LPB) of McMurry

and Politis [2010] which essentially transforms the data vector $(Y_1, \dots, Y_n)'$ to a data vector $(W_1, \dots, W_n)'$ that has uncorrelated entries, i.e., $\{W_t\}$ is a ‘white noise’. If $\{Y_t\}$ is a linear time series, then $\{W_t\}$ can further be claimed to be i.i.d. (under some conditions). The LPB parallels with the AR-sieve bootstrap since both are applicable to nonlinear time series as long as the statistic of interest has a large-sample distribution that only depends on the first and second order moment structure of the data; see Kreiss et al. [2011] and Jentsch and Politis [2015].

Nevertheless, in the search of a transformation that renders the data i.i.d., it may be helpful to first devise a transformation into Gaussianity; see e.g., Ch. 2.3.2 of Politis [2015]. For example, we can use a version of the Probability Integral Transform (PIT) in order to transform our time series data to Gaussian; the latter can then be transformed to i.i.d. by a decorrelating/whitening operation as in the LPB. This approach was first suggested in Ch. 9 of Politis [2015], and was practically implemented to the setting of a locally stationary time series by Das and Politis [2021].

In the paper at hand, we focus on stationary time series data, with the goal of establishing the realm of applicability of the above mentioned procedure which, for lack of a better word, we will call the *model-free bootstrap* (MFB). We will show asymptotic validity of the MFB for a general class of stationary processes, and for many types of statistics of interest. We will also establish MFB’s validity for the construction of one-step-ahead prediction intervals, i.e., to fix ideas we will focus on the case $h = 1$ in the above.

The remaining of the paper is organized as follows. Section 2 restates the MFB algorithm carefully. Section 3 introduces some necessary tools and assumptions to be used, and summarizes some useful preliminary results for our proofs. Section 4 proves MFB’s asymptotic validity for various estimation problems, while Section 5 shows its validity for prediction intervals. Numerical experiments that back up our asymptotic results are presented in Section 6. Technical proofs are given in the supplementary material.

2.3 Model-free bootstrap algorithm

2.3.1 The MFB algorithm

Here we describe the model-free bootstrap (MFB) algorithm for inference and prediction as proposed in Chapter 9 of Politis [2015]. Given a time series $\{Y_t\}_{t \in \mathbb{Z}}$ that is strictly stationary, let F_Y be the cumulative distribution function (CDF) of Y_0 . The PIT defined by

$$U_t = F_Y(Y_t)$$

implies that U_t is uniformly distributed on $[0, 1]$, assuming F_Y is continuous. See Angus [1994]. Let Φ be the CDF of standard normal distribution and $\Phi^{-1}(p) = \inf\{x \in \mathbb{R} : \Phi(x) \geq p\}$ be the quantile function; then, $Z_t = \Phi^{-1}(U_t)$ is $\mathcal{N}(0, 1)$ distributed. Also, stationarity is preserved for $\{U_t\}$ and $\{Z_t\}$.

Let Σ_n denote the covariance matrix of $\underline{Z}_n = (Z_1, \dots, Z_n)$, and denote by $\Sigma_n^{-\frac{1}{2}}$ the lower triangular matrix from the Cholesky decomposition of Σ_n^{-1} . Then, $\underline{\xi}_n = \Sigma_n^{-\frac{1}{2}} \underline{Z}_n$ is a vector of i.i.d. $\mathcal{N}(0, 1)$ entries, provided $Z_1, \dots, Z_n$ are *jointly* normal.

Suppose we use a resampling scheme to create the i.i.d. bootstrap sample $\xi_1^*, \xi_2^*, \dots, \xi_n^*$. Then, letting $\underline{Z}_n^* = \Sigma_n^{\frac{1}{2}} \underline{\xi}_n^*$ where $\underline{\xi}_n^* = (\xi_1^*, \xi_2^*, \dots, \xi_n^*)'$, and $Y_t^* = F_Y^{-1}(\Phi(Z_t^*))$, then $\{Y_t^*\}$ is our bootstrapped sample.

Moreover, ξ_{n+1}^* can also be generated through i.i.d. sampling, and Z_{n+1}^* can be generated through the relation $(\underline{Z}_n, Z_{n+1}^*) = \Sigma_{n+1}^{1/2}(\underline{\xi}_n, \xi_{n+1}^*)$. Using the inverse of the previously mentioned transforms, the next bootstrap value can be generated by $Y_{n+1}^* = F_Y^{-1}(\Phi(Z_{n+1}^*))$. It can be shown that by using these theoretical transforms, $Y_{n+1}^* | \underline{Y}_n$ has the same distribution as $Y_{n+1} | \underline{Y}_n$.

Nevertheless, to use the above steps for practical purposes, each transform must also be estimated in a consistent manner from the data at hand. Furthermore, the validity of the bootstrap procedure has to be investigated, both for estimation and prediction. Thus, several questions arise:

- Under what circumstance are the entries of $\underline{Z}_n$ jointly normal?
- What estimators for F_Y and Σ_n should we use so that the above steps lead to validity of the bootstrap?
- How should we create the i.i.d. bootstrap values $\{\xi_t^*\}$?

The first two points will be addressed in the following paragraphs. For the third point, Politis [2015] has proposed two ways to do it. One way is sampling with replacement from $\{\hat{\xi}_t\}_{t=1}^n$, with $\hat{\xi}_t$ calculated from Y_t using estimated transform functions. A second way is to generate ξ_t^* as i.i.d. $\mathcal{N}(0, 1)$, which is presumably the limiting distribution of $\hat{\xi}_t$. The first method is called **model-free** (MF), and the second is referred to as **limit model-free** (LMF) since the limit distribution is used.

Frequently used notations include the following. Let $\hat{F}$ and $\hat{\Sigma}_n$ denote general estimators for F_Y and Σ_n respectively. The subscript Y is dropped from $\hat{F}$ for simplicity. Φ is the CDF of a standard normal distribution with Φ^{-1} its quantile function. Let $\tilde{\Phi}$ be the CDF of a thresholded standard normal distribution: suppose $X \sim \mathcal{N}(0, 1)$, $X_c = X$ for $|X| \leq c$ and $X_c = \text{sgn}(X)c$ for $|X| > c$, where $\text{sgn}(\cdot)$ is the sign function. Then $\tilde{\Phi}$ denotes the CDF of X_c and its inverse $\tilde{\Phi}^{-1}$ the quantile function. We omit c in the notation for simplification. Asymptotically we also require $c \rightarrow \infty$ such that $\tilde{\Phi}^{-1}$ converges to Φ^{-1} . The reason of this augmentation is provided in Section 2.4 and asymptotic details are explained in Section 3.

By using these practical transforms, we can calculate $\hat{U}_t = \hat{F}(Y_t)$, and $\tilde{Z}_t = \tilde{\Phi}^{-1}(\hat{U}_t)$, which are the estimations for the latent series $\{U_t\}$ and $\{Z_t\}$ respectively. Since $\{Z_t\}$ is latent, $\hat{\Sigma}_n$ can not be directly calculated. Instead, we use $\hat{\tilde{\Sigma}}_n$ which is the same estimator calculated based on $\{\tilde{Z}_t\}$.

Let $\sigma_Z(k) = EZ_0Z_k$ be the lag- k autocovariance of Z_t , $\hat{\sigma}_Z(k)$ be its estimator and $\hat{\tilde{\sigma}}_Z(k)$ be the estimator calculated from $\tilde{Z}_t$. Let $\|\cdot\|_p = E(|\cdot|^p)^{\frac{1}{p}}$ denote the p -norm of a random variable; $\|\cdot\|_{op}$ denotes the operator norm of a matrix, i.e., $\|M\|_{op} = \sup_{x \in \mathbb{R}^n, \|x\|_2=1} \|Mx\|_2$ where M is a $n \times n$ square matrix. Relative quantities in the bootstrap world will be denoted by a superscript $*$.

Given the above introduction, we can now describe the model-free bootstrap algorithm.

2.3.2 MFB for confidence intervals

Let θ_0 be a population parameter of interest, $\hat{\theta}_n$ an estimator of θ_0 from data $\{Y_t\}_{t=1}^n$, and $\hat{\theta}_n^*$ the same estimator from bootstrapped data $\{Y_t^*\}_{t=1}^n$. Define the real-world *root* $r = \theta_0 - \hat{\theta}_n$; let R denote its CDF. Then, a $(1 - \alpha)100\%$ equal-tailed confidence interval (CI) for θ_0 is

$$(\hat{\theta}_n + R^{-1}(\alpha/2), \hat{\theta}_n + R^{-1}(1 - \alpha/2)),$$

where $R^{-1}(x) = \inf\{r \in \mathbb{R} : R(r) \geq x\}$ denotes the quantile function of R . The distribution R could be approximated through bootstrap simulations.

Algorithm 1 (Model-free bootstrap for parameter inference)

1. Given data $\{Y_t\}_{t=1}^n$, let $\hat{U}_t = \hat{F}(Y_t)$; $\tilde{Z}_t = \tilde{\Phi}^{-1}(\hat{U}_t)$; $\hat{\xi}_{\underline{n}} = \hat{\Sigma}_n^{-\frac{1}{2}} \tilde{Z}_{\underline{n}}$.
2. (MF) Let ξ_t^* be i.i.d. samples from $\bar{F}_{\hat{\xi}}$, where $\bar{F}_{\hat{\xi}}$ is the empirical CDF of $\{\hat{\xi}_t\}_{t=1}^n$, and $\underline{Z}_n^* = \hat{\Sigma}_n^{\frac{1}{2}} \underline{\xi}_n^*$. Let $Y_t^* = \hat{F}^{-1}(\Phi(Z_t^*))$. Calculate the bootstrap root $r^* = \hat{\theta}_n - \hat{\theta}_n^*$.
3. Do the above step B times to form an empirical CDF $\bar{R}$ based on the B replicates of r^* . $\bar{R}$ is used to approximate R ; hence, an approximate $(1 - \alpha)100\%$ CI for θ_0 is

$$(\hat{\theta}_n + \bar{R}^{-1}(\alpha/2), \hat{\theta}_n + \bar{R}^{-1}(1 - \alpha/2)).$$

For the limit model-free (LMF) bootstrap, replace step 2 with following:

2. (LMF) Let ξ_t^* be i.i.d. samples from $\mathcal{N}(0, 1)$, and $\underline{Z}_n^* = \hat{\Sigma}_n^{\frac{1}{2}} \underline{\xi}_n^*$. Let $Y_t^* = \hat{F}^{-1}(\Phi(Z_t^*))$. Calculate the bootstrap root $r^* = \hat{\theta}_n - \hat{\theta}_n^*$. This is equivalent to sample $\underline{Z}_n^*$ from $\mathcal{N}(\mathbf{0}, \hat{\Sigma}_n)$ distribution.

Below are some simple examples of statistics of interest.

- The mean: $\theta_0 = E(Y_0)$; $\hat{\theta}_n = \frac{\sum_{t=1}^n Y_t}{n}$.
- Autocovariance: $\theta_0 = \gamma_Y(k)$; $\hat{\theta}_n = \hat{\gamma}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (Y_t - \bar{Y}_n)(Y_{t+k} - \bar{Y}_n)$.
- Autocorrelation: $\theta_0 = \rho(k) = \frac{\gamma(k)}{\gamma(0)}$; $\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}$.

And additional examples will be addressed in Section 4.

2.3.3 MFB for prediction intervals

For prediction problems, we want to use the bootstrap to simulate the distribution of Y_{n+1} **conditional** on past values $\{Y_t\}_{t=1}^n$. For this purpose, we use bootstrap to approximate the conditional distribution of the predictive root $Y_{n+1} - \hat{Y}_{n+1}$, where $\hat{Y}_{n+1}$ is a predictor chosen by the practitioner. Let G denote the conditional distribution of the predictive root defined above. Then a $(1 - \alpha)100\%$ equal tailed prediction interval for Y_{n+1} is :

$$(\hat{Y}_{n+1} + G^{-1}(\alpha/2), \hat{Y}_{n+1} + G^{-1}(1 - \alpha/2))$$

Model-free bootstrap algorithm for one-step ahead prediction is the following:

Algorithm 2 (Model-free bootstrap for one-step ahead prediction)

1. Given data $\{Y_t\}_{t=1}^n$, let $\hat{U}_t = \hat{F}(Y_t)$, $\tilde{Z}_t = \tilde{\Phi}^{-1}(\hat{U}_t)$.
2. Denote $\hat{\Sigma}_{11} = \hat{\Sigma}_n$, $\hat{\Sigma}_{12} = \begin{bmatrix} \hat{\sigma}(n) \\ \vdots \\ \hat{\sigma}(1) \end{bmatrix}$, and $\hat{\Sigma}_{22} = \hat{\sigma}(0)$. Let $\hat{\Sigma}_{n+1} = \begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^T & \hat{\Sigma}_{22} \end{bmatrix}$.

Let $\{\xi_t^*\}_{t=1}^{n+1}$ be drawn randomly with replacement from $\{\xi_t\}_{t=1}^n$, and $\underline{Z}_n^* = \hat{\Sigma}_n^{-\frac{1}{2}} \xi_n^*$. Let Z_{n+1}^* be the $(n+1)^{th}$ element of the vector $\hat{\Sigma}_{n+1}^{-\frac{1}{2}}(\xi_n^*, \xi_{n+1}^*)$. Denote the distribution of Z_{n+1}^* as $\hat{F}_Z^{(n+1)}$. This is also the estimated conditional distribution of $Z_{n+1}|Y_n$. The form of this distribution is conditional on our data $\underline{Y}_n = (Y_1, \dots, Y_n)$. Let $Y_{n+1}^* = \hat{F}^{-1}(\Phi(Z_{n+1}^*))$.

3. Choose a predictor $\hat{Y}_{n+1}$ for Y_{n+1} based on $\underline{Y}_n$. For example, the L^2 optimal predictor as mentioned in Section 1 is the expectation of Z_{n+1} conditioning on $\underline{Y}_n$ that can be approximated by

$$\hat{Y}_{n+1} = \int \hat{F}^{-1}(\Phi(z)) d\hat{F}_Z^{(n+1)}(z).$$

The above integral can be evaluated through Monte-Carlo simulation. The chosen predictor will be used as the center of our prediction interval, and the bootstrap procedure will be used to capture the distribution of the predictive root in the next steps.

4. Re-estimate all the transforms, matrices and the distribution $\hat{F}_Z^{(n+1)}$ used in the above calculation, with bootstrapped data $\underline{Z}_n^* = (Z_1^*, \dots, Z_n^*)$ and $Y_t^* = \hat{F}^{-1}(\Phi(Z_t^*))$. Let $(\hat{F}_Z^{(n+1)})^*$ denote the re-estimated distribution function for Z_{n+1}^* with bootstrap data $\underline{Y}_n^* = (Y_1^*, \dots, Y_n^*)$. Let $(Z_{n+1}|Y_n)^*$ denote the random variable with estimated conditional distribution $(\hat{F}_Z^{(n+1)})^*$.

Let $\hat{Y}_{n+1}^*$ denote the one-step ahead predictor with re-estimated transforms based on the bootstrap pseudo data. In the L^2 -optimal setting,

$$\hat{Y}_{n+1}^* = \int (\hat{F}^*)^{-1}(\Phi(z)) d(\hat{F}_Z^{(n+1)})^*(z).$$

5. The bootstrapped L^2 - optimal predictive root is:

$$Y_{n+1}^* - \hat{Y}_{n+1}^*$$

6. Denote the empirical CDF of bootstrapped predictive roots as $\bar{G}$. The approximate $(1 - \alpha)$ prediction interval for Y_{n+1} is

$$\left(\hat{Y}_{n+1} + \bar{G}^{-1}(\alpha/2), \hat{Y}_{n+1} + \bar{G}^{-1}(1 - \alpha/2) \right).$$

Algorithm 3 (Limit model-free bootstrap for 1 step ahead prediction)

1. Given data $\{Y_t\}_{t=1}^n$, let $\hat{U}_t = \hat{F}(Y_t)$, $\tilde{Z}_t = \tilde{\Phi}^{-1}(\hat{U}_t)$.
2. (LMF) Denote $\hat{\hat{\Sigma}}_{11} = \hat{\hat{\Sigma}}_n$, $\hat{\hat{\Sigma}}_{12} = \begin{bmatrix} \hat{\hat{\sigma}}(n) \\ \vdots \\ \hat{\hat{\sigma}}(1) \end{bmatrix}$. Let $\underline{Z}_n^* \sim \mathcal{N}(0, \hat{\hat{\Sigma}}_n)$;
 $Z_{n+1}^* \sim \mathcal{N}(\hat{\hat{\Sigma}}_{21} \hat{\hat{\Sigma}}_{11}^{-1} \tilde{Z}_n, \hat{\hat{\Sigma}}_{22} - \hat{\hat{\Sigma}}_{21} \hat{\hat{\Sigma}}_{11}^{-1} \hat{\hat{\Sigma}}_{12})$; $Y_{n+1}^* = \hat{F}^{-1}(\Phi(Z_{n+1}^*))$
3. Choose a predictor for Y_{n+1} based on $\underline{Y}_n$. For example, the L^2 optimal predictor is

$$\hat{Y}_{n+1} = E(\hat{F}^{-1}(\Phi(Z_{n+1}))) | \underline{Y}_n.$$

Where $Z_{n+1} | \underline{Y}_n \sim \mathcal{N}(\hat{\hat{\Sigma}}_{21} \hat{\hat{\Sigma}}_{11}^{-1} \tilde{Z}_n, \hat{\hat{\Sigma}}_{22} - \hat{\hat{\Sigma}}_{21} \hat{\hat{\Sigma}}_{11}^{-1} \hat{\hat{\Sigma}}_{12})$.

4. *Re-estimate all the transforms and matrices used in the above calculation, with $\underline{Z}_n^*$ and $Y_{n+1}^* = \hat{F}^{-1}(\Phi(Z_{n+1}^*))$. The one-step ahead predictor in the bootstrap world is $\hat{Y}_{n+1}^* = E^*((\hat{F}^*)^{-1}(\Phi(Z_{n+1}))|\underline{Y}_n)$, where the expectation in the bootstrap world is calculated through the distribution of $(Z_{n+1}|\underline{Y}_n)^*$ that is $\mathcal{N}(\hat{\Sigma}_{21}^*(\hat{\Sigma}_{11}^*)^{-1}\hat{\underline{Z}}_n, \hat{\Sigma}_{22}^* - \hat{\Sigma}_{21}^*(\hat{\Sigma}_{11}^*)^{-1}\hat{\Sigma}_{12}^*)$ distributed.*
5. *The bootstrapped L^2 -optimal predictive root is:*

$$Y_{n+1}^* - \hat{Y}_{n+1}^*$$

6. *Denote the empirical CDF of bootstrapped predictive root as $\bar{G}$. The approximate $(1 - \alpha)$ prediction interval for Y_{n+1} is*

$$\left(\hat{Y}_{n+1} + \bar{G}^{-1}(\alpha/2), \hat{Y}_{n+1} + \bar{G}^{-1}(1 - \alpha/2) \right).$$

Here we provide an explanation to step 5 above. The bootstrap is supposed to capture the distribution of the predictive root $Y_{n+1} - E(\hat{F}^{-1}(\Phi(Z_{n+1}))|\underline{Y}_n)$, where for $Y_{n+1} = F_Y^{-1}(\Phi(Z_{n+1}))$, Z_{n+1} has the conditional distribution $\mathcal{N}(\Sigma_{21}\Sigma_{11}^{-1}\underline{Z}_n, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$; and in the expectation $Z_{n+1}|\underline{Y}_n \sim \mathcal{N}(\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\underline{Z}}_n, \hat{\Sigma}_{22} - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12})$ is estimated from data. Clearly, the randomness in the distribution of the predictive root not only comes from the randomness of the series, but also randomness in the estimation. Thus in the bootstrap world, we should replace theoretical transforms with their data-dependent analogue, and also account for all the errors arising from estimation, i.e. replace F_Y with $\hat{F}$, $\hat{F}$ with $\hat{F}^*$, and $\hat{\Sigma}_n$ with $\hat{\Sigma}_n^*$, resulting in the formula in step 5. In this way, the bootstrap procedure can capture all the randomness in the distribution of predictive root, which helps relieve potential *undercoverage* issue for finite data. See also Ch. 9 of Politis [2015].

The above algorithms require the practitioner to select the parameter c in the threshold normal distribution in order to calculate the inverse $\tilde{\Phi}^{-1}$. Section 2.4 gives a detailed reasoning for this procedure. In the simulation section, we provide an easy to implement rule for c based on the number of data and choice of CDF estimator $\hat{F}$, which is used to generate the numerical results. As we will show in Section 4, to establish asymptotic results, we also need c to diverge at a rate that depends on the sample size n .

Computationally, the MFB has two procedures that require intensive computational resources. First of all, the time complexity of finding the Cholesky decomposition is $O(n^2)$ for Toeplitz matrices. Secondly, because $\hat{F}^{-1}$ usually does not have a closed-form formula, for example, when $\hat{F}$ is the inverse kernel CDF estimator, computing the value $\hat{F}^{-1}(U_t^*)$ is usually performed through numerical interpolation of a range of sample points $\{(\hat{F}(x_i), x_i)\}$. This procedure is time-costly as sample size grow larger. It is important to mention that MFB for prediction intervals is more time consuming than for confidence intervals by a factor of B , since we need to re-estimate the transforms and autocovariance matrix (and its Cholesky decomposition) for each bootstrap sample.

2.3.4 Appropriate estimators for $\hat{F}$, $\tilde{\Phi}^{-1}$ and $\hat{\Sigma}_n$

Now we discuss what should be the appropriate estimators $\hat{F}$, $\hat{\Sigma}_n$ and also why an augmented version of Φ^{-1} might be needed. Firstly, it is necessary that $\hat{F}$, $\hat{F}^{-1}$, $\hat{\Sigma}_n$ should be consistent in certain forms for F_Y , F_Y^{-1} and Σ_n respectively. For $\hat{F}$, the first idea is to use the empirical CDF $\bar{F}(y) = \frac{1}{n} \sum_{t=1}^n I\{Y_t \leq y\}$, where $I\{\cdot\}$ is the indicator function, and its inverse $\bar{F}^{-1}(p) = \inf\{y \in \mathbb{R} : \bar{F}(y) \geq p\}$. Under moment and short-range dependence assumptions, consistency of $\bar{F}$ and $\bar{F}^{-1}$ can be established by looking into the empirical process and quantile process. Details are in later sections and will play an important role in our proofs.

Another natural candidate is the kernel smoothed CDF estimator $\hat{F}_h(y) = \frac{1}{n} \sum_{t=1}^n K_h(y - Y_t)$, where $K_h(y - Y_t) = K(\frac{y - Y_t}{h})$ and K is a smooth CDF function with additional assumptions. The obvious advantage of $\hat{F}_h$ is that it is continuous, which is a property $\bar{F}$ is lacking. An additional implication of using $\bar{F}$ is the resulting $\hat{U}_t = \bar{F}(Y_t)$ only takes value in $\{\frac{1}{n}, \frac{2}{n}, \dots, 1\}$ and $Y_t^* = \bar{F}^{-1}(U_t^*)$ only takes value in $\{Y_t\}_{t=1}^n$. But by using the kernel estimator $\hat{F}_h$ and its inverse, Y_t^* can take values that did not appear in the original series. If the data size n is large, the influence of this is minimal; whereas if n is small, $\hat{F}_h$ is a better estimator because of its ability to interpolate unseen values compared to the coarse behavior of $\bar{F}$. It is also worth mentioning that when n is

large, using $\bar{F}$ and its inverse will save computational time comparing to $\hat{F}_h$.

Nevertheless, using $\bar{F}$ in the first step of the Model-free procedure, i.e., letting $\hat{U}_t = \bar{F}(Y_t)$ will result in one of the variables $\hat{U}_1, \dots, \hat{U}_n$ taking the value 1 in which case the inverse transform $\Phi^{-1}(1)$ in the second step is not well defined. To address this issue, we can use an augmented version of the normal quantile denoted by $\tilde{\Phi}^{-1}$, which is the inverse CDF of a thresholded standard normal $\mathcal{N}_c(0, 1)$ as defined in Section 2.1. By doing this, $\tilde{\Phi}^{-1}$ is bounded on $[0, 1]$, which relieves this problem.

The purpose of using $\tilde{\Phi}^{-1}$ goes beyond being an ad-hoc fix. It further plays an important rule in asymptotic analysis of the bootstrap algorithm, in that the thresholding helps control the fast diverging behavior of Φ^{-1} near the endpoints 0 and 1, thereby bounding the extreme values of $\tilde{Z}_t$ by a controllable rate of divergence. In turn, this enables the key step of analyzing convergence of the covariance estimator $\hat{\Sigma}_n$ derived from $\tilde{Z}_t$, for both $\bar{F}$ and $\hat{F}_h$ scenarios.

Consistent estimator $\hat{\Sigma}_n$ of the autocovariance matrix Σ_n has been well studied. Wu and Pourahmadi [2009] established the first result on consistency of a banded matrix estimator. Here we shall use the more general flat-top estimators of McMurry and Politis [2010]. Let $\kappa(x)$ be the tapering weight function:

$$\kappa(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ g(|x|) & \text{if } 1 < |x| \leq c_\kappa \\ 0 & \text{if } |x| > c_\kappa \end{cases} \quad (2.3.1)$$

where $|g(x)| < 1$. The most commonly used flat-top kernel is defined by

$$\kappa(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ 2 - |x| & \text{if } 1 < |x| \leq 2 \\ 0 & \text{if } |x| > 2; \end{cases} \quad (2.3.2)$$

see Politis and Romano [1995]. Let l be the bandwidth of choice and $\kappa_l(x) = \kappa(x/l)$. The tapered

estimator $\hat{\Sigma}_n = \hat{\Sigma}_n^{(\kappa, l)}$ at entry (i, j) has value:

$$\hat{\Sigma}_n^{(\kappa, l)}(i, j) = \kappa_l(k) \left(\frac{1}{n} \sum_{t=1}^{n-k} Z_t Z_{t+k} \right), \quad (2.3.3)$$

where $|i - j| = k$.

Remark 2.3.1. It is worth noting that the above estimator is not guaranteed to be positive definite (PD) for finite samples, but can be corrected towards PD by looking into its Cholesky decomposition; for details see McMurry and Politis [2010], McMurry and Politis [2015]. Asymptotically, the estimator is PD with probability tending to one, and the corrected estimator enjoys the same rate of convergence as the original estimator. Either way, it is not a problem in asymptotic studies.

While convergence results exist for $\hat{\Sigma}_n$, the true series $\{Z_t\}$ is latent and can not be used in the calculation of $\hat{\Sigma}_n$. We can only use the estimator with the estimated $\tilde{Z}_t$ defined in Section 2.1 that are calculated from original data Y_t , resulting in an estimator $\hat{\hat{\Sigma}}_n$, of which the entries are given by

$$\hat{\hat{\Sigma}}_n^{(\kappa, l)}(i, j) = \kappa_l(k) \left(\frac{1}{n} \sum_{t=1}^{n-k} \tilde{Z}_t \tilde{Z}_{t+k} \right). \quad (2.3.4)$$

Consistency of $\hat{\hat{\Sigma}}_n$ to Σ_n will be shown in the following.

2.3.5 Connection with previous work on MFB for time series

The idea of the model-free bootstrap for time series was first proposed in the monograph Politis [2015], as well as in Das and Politis [2021]. The emphasis of Das and Politis [2021] is on applying the model-free Prediction Principle to provide a novel statistical procedure for prediction inference of locally stationary data. It has more emphasis on methodology, simulation studies, and real data. By contrast, our emphasis is to build a framework for establishing theoretical properties for the MFB under the stationarity assumption, both for inference (i.e., confidence intervals) as well as for prediction purposes (i.e., prediction intervals). This provides grounds and

theoretical guarantees for the MFB algorithm, which has not been addressed in previous works. While the assumptions made in this paper do not cover the locally stationary scenario in Das and Politis [2021], we are actively working on establishing theoretical results for locally stationary time series with the help of recent advances in this field.

2.4 Assumptions and preliminary results

2.4.1 Acceptable forms of $\{Y_t\}$

The stationary series $\{Y_t\}$ in our setting will be assumed to have the form:

$$Y_t = f(W_t) \tag{2.4.1}$$

where f is some continuously differentiable function such that the CDF F_Y of Y_t is strictly increasing and continuously differentiable, and $\{W_t\}$ is a strictly stationary Gaussian process. Without loss of generality, we may assume that $EW_t = 0$ since the mean of W_t can be incorporated into the function f .

Equation (2.4.1) is a common form of extension from Gaussian series to the non-Gaussian case. It has been used in the study of long range dependence, as well as in analyzing time series with heavy tails; see Samorodnitsky and Taqqu [1994]. This assumption also figures in a completely different setting, namely that of Bayesian machine learning; see Snelson et al. [2004].

By the Wold decomposition (see Brockwell and Davis [1991]) coupled with Gaussianity, W_t admits the expansion

$$W_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} + V_t \tag{2.4.2}$$

where ε_t are i.i.d. $\mathcal{N}(0, 1)$ distributed, and V_t is a deterministic process independent from $\{\varepsilon_t\}$. If we assume W_t is purely nondeterministic, then V_t vanishes, and it is clear that Y_t is of the form

$Y_t = h(\cdots, \varepsilon_{t-1}, \varepsilon_t)$ for some function h . Such a representation naturally appears in many time series and dynamical system models, and is also a common form used for developing short range dependence conditions; see Wu [2005a] for details.

The importance and necessity of equation 2.4.1 and the conditions on f and F_Y can be explained by the following lemma:

Lemma 2.4.1. *Let $Z_t = \Phi^{-1} \circ F_Y(Y_t)$. Then $Z_1, \cdots, Z_n$ are jointly normal if and only if Y_t admits the representation in equation (2.4.1).*

Proof. If (2.4.1) holds, $Z_t = \Phi^{-1} \circ F_Y \circ f(W_t)$. Since both W_t and Z_t are normally distributed, $\Phi^{-1} \circ F_Y \circ f$ is a normality preserving, continuously differentiable transform. Therefore the transform is linear by Corollary 2 of Mase [1977]. As a result, each Z_t is linearly transformed from W_t , $t \in \{1, 2, \cdots, n\}$. Collecting the n instantaneous linear transforms, we obtain a transform of the vector $(W_1, \cdots, W_n)$ to $(Z_1, \cdots, Z_n)$ that is linear. Since $(W_1, \cdots, W_n)$ are jointly normal and a linear transform preserves normality, $(Z_1, \cdots, Z_n)$ is multivariate normal; see Lemma 3.1 of Das and Politis [2021].

Conversely, given $Z_t = \Phi^{-1} \circ F_Y(Y_t)$ is jointly normal and strictly stationary, we have $Y_t = F_Y^{-1} \circ \Phi(Z_t)$, which is of the form (2.4.1). $\square$

Remark 2.4.1. The above lemma essentially clarifies the class of time series for which the model-free algorithm is applicable. Under condition (2.4.1), we can recover the standardized underlying Gaussian process by the transform $Z_t = \Phi^{-1} \circ F_Y(Y_t)$ that results in $Z_t = \frac{W_t - \mu_W}{\sigma_W}$, where μ_W and σ_W are the mean and standard deviation of the hidden Gaussian process W_t .

Under some assumptions, it is indeed possible to estimate the monotone transfer function f , see e.g. Zhang et al. [2019]. We could then estimate the unobserved process W_t , in an analogous way that regression residuals are estimates of the underlying (and unobserved) errors. From the estimates of W_t , we can construct estimates of its statistical features can be estimated, e.g., its covariance structure. However, consistency of these estimates would not automatically imply the

bootstrap validity we need. Since knowledge of f and of the estimates of W_t is not really needed in performing the MFB procedures, we have chosen to work with the transformations detailed here, and focus on proving bootstrap validity instead.

Since $F_Y^{-1} \circ \Phi$ is monotone, it is reasonable to believe that the function f is monotone. In fact, $f(\cdot)$ is equivalent to $F_Y^{-1} \circ \Phi(\cdot)$ modulo affine differences in their arguments. Thus we have the following lemma:

Lemma 2.4.2. *Under the setup in equation (2.4.1), given F_Y is strictly increasing, then f is a strictly monotone function.*

One large subclass of strictly monotone, continuously differentiable functions is functions with derivatives bounded away from 0. As it turns out, this subclass can simplify our short range dependence assumption to be introduced next.

2.4.2 Short range dependence (SRD) assumptions on $\{Y_t\}$

The SRD assumption is a necessary tool typically assumed to establish statistical results under dependent settings. In the context of our bootstrap scheme, the following types of results are necessary for carrying out proof of bootstrap validity:

1. Consistency of estimated transforms can be established in proper mathematical forms.
2. Certain central limit theorems for the statistic of interest can be established.
3. Consistency of banded/tapered sample autocovariance matrix and CLT variance in the bootstrap world, as further explained in Section 4.

One of the most widely used SRD assumptions for strictly stationary time series is the strong mixing condition introduced by Rosenblatt [1956] and extensively studied since. Many useful results are available for series under certain mixing rates. However such conditions are

often hard to verify for general time series models; moreover, the mixing conditions were not sufficient for the proof of banded/tapered autocovariance matrix estimator. In Wu [2005a], a new type of dependence measure named physical dependence measure was introduced which is as follows. Assume $Y_t = h(\cdots, \varepsilon_{t-1}, \varepsilon_t) \in L^p$ with $\varepsilon_t \stackrel{i.i.d.}{\sim} F_\varepsilon$, then the physical dependence measure is defined by

$$\delta_p(j) = \|Y_j - Y'_j\|_p,$$

where $Y'_j = h(\cdots, \varepsilon_{-1}, \varepsilon'_0, \cdots, \varepsilon_j)$ with $\varepsilon'_0 \sim F_\varepsilon$ independent of $\{\varepsilon_t\}$. The series Y_t is called strongly p -stable if

$$(C0) \quad \Delta_p = \sum_{j=0}^{\infty} \delta_p(j) < \infty.$$

The p -stable assumption was able to provide the first results on the convergence of banded and/or tapered autocovariance matrix estimators by Wu and Pourahmadi [2009] and McMurry and Politis [2010].

Alas, functional central limit theorem results for the empirical process—which relates to uniform consistency of CDF estimators cannot be readily established with the physical dependence measure, and require further conditions. To simplify assumptions, we hope for a condition that offers more flexibility than what is mentioned above. One short range dependence measure that is gaining interest is the m -approximation assumption developed in a series of papers in Berkes et al. [2009], Berkes et al. [2011] and Hörmann and Kokoszka [2010].

The following related conditions are used for proving different properties and are very important in our proofs.

Definition 2.4.1. The following different m -approximation conditions are proposed. Let $\{Y_t\}$ be a strictly stationary time series, there exists an m -dependent series $\{Y_t^{(m)}\}$ that

$$(C1) \quad \forall t \in \mathbf{Z}, P(|Y_t - Y_t^{(m)}| > \gamma_m) < \delta_m, \text{ for some sequence } \gamma_m \rightarrow 0 \text{ and } \delta_m \rightarrow 0.$$

$$(C2) \quad \text{both } Y_t \text{ and } Y_t^{(m)} \text{ are in } L^p; \exists \delta(m) : \mathbb{N} \rightarrow \mathbb{R}_+, \text{ satisfying } \delta(m) \ll m^{-A} \text{ for some}$$

$A > 0$, and such that $\left\| Y_t - Y_t^{(m)} \right\|_p \leq \delta(m)$.

(C3) both Y_t and $Y_t^{(m)}$ are in L^p ; $\sum_{m=0}^{\infty} \left\| Y_t - Y_t^{(m)} \right\|_p < \infty$.

Here, the notation " $a_n \ll b_n$ " means $\limsup \left| \frac{a_n}{b_n} \right| \rightarrow 0$.

Assumption (C1) was introduced in Berkes et al. [2009] to establish asymptotic behavior of empirical process. (C2) was introduced in Berkes et al. [2011] to establish an invariance principle for partial sums. Both are relevant in our setting. (C3) appeared in Hörmann and Kokoszka [2010] in the context of functional time series. While it is not directly related to the setup here, there is an important relation between (C3) and the physical dependence measure (C0), under which consistency of tapered autocovariance matrix estimator can be established. A detailed layout of the relative theorems mentioned above that are necessary for our proofs can be found in section A of the supplementary material.

There are various methods to construct the m -dependent sequence $Y_t^{(m)}$. Typically one can use truncation, substitution and coupling methods on the representation $Y_t = h(\cdots, \varepsilon_{t-1}, \varepsilon_t)$; for details refer to Berkes et al. [2009]. The coupling construction is the most desirable: we replace ε by i.i.d. independent copies of ε' for times that are at least m steps away from current time t , i.e. let $Y_t^{(m)} = h(\cdots, \varepsilon'_{t-m-1}, \varepsilon'_{t-m}, \varepsilon_{t-m+1}, \cdots, \varepsilon_t)$. The resulting $Y_t^{(m)}$ is m -dependent, and also has the advantage that it has the same distribution and moments as Y_t . This construction along with (C2) were also used in Wu [2005b] where it was called the geometric-moment contracting property, assuming a faster geometric decay rate. We will use the coupling construction in what follows.

The above SRD conditions are related; the following lemma clarifies.

Lemma 2.4.3. *Suppose that $\{Y_t^{(m)}\}$ is constructed by coupling as defined above, then:*

1. Assume (C2) with $A = C(\frac{1}{p} + \frac{1}{\theta})$ for some $C > 0$, $\theta \in (0, 1)$; then (C1) holds with

$$P(|Y_t - Y_t^{(m)}| > m^{-C/\theta}) < m^{-C}.$$

2. Assume (C2) with $A > 1$; then (C3) holds.

3. Assume (C3); then (C0) holds.

4. (C2) is preserved (with a new rate) under θ -Lipschitz transforms. To elaborate, let g be a θ -Lipschitz function, i.e., for some constant K and $\theta \in (0, 1]$, $|g(x) - g(y)| \leq K|x - y|^\theta$. If $\|Y_t - Y_t^{(m)}\|_p \leq \delta(m)$ with $\delta(m) \ll m^{-A}$, then

$$\|g(Y_t) - g(Y_t^{(m)})\|_p \leq K\delta(m)^\theta \ll m^{-\theta A}.$$

Based on the above lemma, it is convenient that we only assume condition (C2) with the appropriate rate, such that the remaining dependence measures and the results they imply can be derived from it. We summarize our first set of assumptions as follows.

Assumption Set 2.4.1. Model setup and weak dependence assumptions

(A1) $Y_t = f(W_t)$, where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a continuously differentiable function, and W_t is a (zero mean) stationary Gaussian process with spectral density bounded and strictly bounded away from 0. Also assume W_t is purely nondeterministic.

(A2) The CDF $F_Y(\cdot)$ is strictly increasing, θ -Lipschitz and continuously differentiable with density function $f_Y(\cdot) > 0$.

(A3) $Y_t \in L^p$ satisfies (C2) with $p > 2$, and A to be specified later. Also, $\exists c > 0$, $|f'| \geq c > 0$. The following assumption is a slight generalization of (A3):

(A4) W_t satisfies (C2) with $p > 2$, and A to be specified later. Also, f preserves the (C2) property.

Remark 2.4.2. (a). Our simulations in Section 6 suggest that we may relax (A1) and (A2) where f is a possibly discontinuous but strictly monotone function, and F_Y is an absolutely continuous function. It can be shown that Lemma 2.4.1 still holds under this assumption; see Corollary 3.1 of Das and Politis [2021]. However some extra constraints are necessary so that our asymptotic results are still true. To avoid making additional assumptions for this edge case, we use stronger assumptions (A1) and (A2).

(b). The assumption (A3) or (A4) establishes SRD for both $\{Y_t\}$ and $\{Z_t\}$. (A3) is a convenient condition under which the inverse of transfer function f is Lipschitz, so that SRD of Y_t is inherited by other series, see Lemma 2.4.4; meanwhile, (A4) is a more general condition for this matter, and it is implied by (A3).

Lemma 2.4.4. *Assume (A1)–(A3), then $\{W_t\}$ and $\{Z_t\}$ satisfy (C2).*

Proof. Under (A1), (A2) and by Lemma 2.4.2, f is continuously differentiable and strictly monotone. Therefore it is invertible and $|(f^{-1})'| = |\frac{1}{f'}|$ is bounded, which means f^{-1} is a Lipschitz function. Since (C2) is preserved under Lipschitz transform and $W_t = f^{-1}(Y_t)$, $\{W_t\}$ also satisfies (C2) with same rate $\delta(m)$ as $\{Y_t\}$. By Lemma 3.1, the transform from W_t to Z_t is linear; thus $\{Z_t\}$ also satisfies (C2) with same rate $\delta(m)$. □

We now proceed to show bootstrap validity for parameter estimation and prediction. The technical proofs can be found in section B of the supplementary material.

2.5 Model-free bootstrap validity for estimation

Let P^* denote the probability measure conditioning on the data. In this section, we show that the MFB is asymptotically valid for constructing confidence intervals of a parameter of interest θ_0 , where there exists an estimator $\hat{\theta}_n$ having the form of the so-called functions of linear statistics, i.e.,

$$\hat{\theta}_n = q\left(\frac{1}{n-k+1} \sum_{t=1}^{n-k+1} g(Y_t, \dots, Y_{t+k-1})\right) \quad (2.5.1)$$

where $q : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and $g : \mathbb{R}^k \rightarrow \mathbb{R}^d$ are smooth functions. The most important case is the mean where $\theta_0 = E(Y_t)$ and $\hat{\theta}_n = \frac{1}{n} \sum_{t=1}^n Y_t$. In fact, the class of statistics of the type (2.5.1) includes a wide range of estimators beyond the sample mean, such as sample autocovariances and autocorrelations, to name a few. Some other literature that investigated (2.5.1) include Künsch [1989], Politis and Romano [1992], Bühlmann [1997], Lahiri [2003] and Kreiss et al. [2011].

2.5.1 Bootstrap validity for the mean

As a demonstration of our approach, we focus on bootstrap validity for the mean $\theta_0 = E(Y_t)$. We look to prove that the following relation holds:

$$\sup_{x \in \mathbb{R}} |P^*(\sqrt{n}(\bar{Y}_n^* - E^*(Y_t^*)) \leq x) - P(\sqrt{n}(\bar{Y}_n - E(Y_t)) \leq x)| \xrightarrow{P} 0. \quad (2.5.2)$$

Convergence of the type (2.5.2) is usually proved in a two step procedure. First of all, a central limit theorem holds for both $\sqrt{n}(\bar{Y}_n - E(Y_t))$, as well as its analogue in the bootstrap world $\sqrt{n}(\bar{Y}_n^* - E^*(Y_t^*))$, where the latter distribution is conditional on the data $\{Y_t\}_{t=1}^n$. This relation holds true under a wide range of SRD conditions including our (C2); see e.g., lemma 2.5.1. Secondly, we need to show that the variance of the limiting normal distribution, namely the long run variance of the stationary time series denoted by σ_∞^2 and $(\sigma_\infty^*)^2$, are equal in probability as

$n \rightarrow \infty$. In other words, the proposed bootstrap should correctly mimic the second-order moments of the original time series model in the limit.

Following the two step procedure, the following lemma provides a central limit result:

Lemma 2.5.1. *Let $\{Y_t\}$ satisfy assumptions (A1)-(A4) with $p > 2$, $\eta \in (0, 1)$, $A > \frac{p-2}{2\eta}(1 - \frac{1+\eta}{p}) \vee 1$ and $(1 + \eta)/p - 1/2 < 0$. Then $\gamma_Y(k)$ is absolutely summable. Also,*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^n (Y_k - EY_0) \xrightarrow{d} \mathcal{N}(0, \sigma_\infty^2) \quad (2.5.3)$$

with $\sigma_\infty^2 = \sum_{k \in \mathbb{Z}} \gamma_Y(k)$.

Proving second moment consistency requires some additional work. The formulas for estimated transforms $\hat{F}$ and $\hat{F}^{-1}$ involve the whole range of data and complicate the expression as well as the dependence structure of the estimated series $\hat{Z}_t$, and we do not have an explicit expression for $(\sigma_\infty^*)^2$ that enables us to analyze its convergence. Alternatively, we have to prove that each step of our transform is consistent; furthermore, a uniform rate of convergence for the forward transform is indispensable for the consistency of autocovariance estimation to that of the hidden Gaussian process. See Remark 2.5.1 for detailed explanation.

Lemma 2.5.2. *Under the assumptions of Lemma 2.5.1, and the additional assumption $\frac{A}{\frac{1}{p} + \frac{1}{\theta}} > 4$, we have*

$$P(|Y_t - Y_t^{(m)}| > m^{-C/\theta}) < m^{-C} \quad (2.5.4)$$

with $C > 4$, and

$$\sup_{t \in [n]} |\hat{U}_t - U_t| = O_p(n^{-1/2}). \quad (2.5.5)$$

Remark 2.5.1. The above lemma argues that the uniform error for the first step transform $\hat{U}_t = F_Y(Y_t)$ is of $O_p(n^{-1/2})$ rate. This is important, as consistency of $\hat{\Sigma}_n$ to Σ_n depends on the consistency of the estimated $\hat{Z}_t$ to Z_t , which itself depends on the consistency of $\hat{U}_t$ to U_t .

Lemma 2.5.2 depends on a crucial rate of approximation of the associated empirical process of Y_t towards its Gaussian process limit. The use of kernel estimators complicates this issue, as there are no known results on the rates of approximation for empirical processes indexed by kernel functions under the m -approximation condition. To achieve the same result as in Lemma 2.5.2 with kernel estimators $\hat{F}_h$, we need the following high level assumptions:

Assumption Set 2.5.1. Additional assumptions for using the kernel CDF estimator:

(A5) F_Y is 2^{nd} order continuously differentiable such that $F_Y^{(2)}(x) = \frac{d^2 F_Y(x)}{dx^2}$ is bounded and Hölder continuous.

(A6) The kernel K is a distribution function with density k satisfying:

1. $k(x)$ is bounded.
2. $k(x) = k(-x)$.
3. $\int |x|^p k(x) dx < \infty$.

(A7) The class of functions $\mathcal{F}_K = \{K(x_0 - x) : x_0 \in \mathbb{R}\}$ is Donsker with respect to $\{Y_t\}$ under condition (C1). In other words, there exists a tight Gaussian process $G_h(x), x \in \mathbb{R}$, such that

$$\sup_{x \in \mathbb{R}} \left| \frac{1}{\sqrt{n}} \sum_{k=1}^n \left(K\left(\frac{x - Y_k}{h}\right) - E\left(K\left(\frac{x - Y_0}{h}\right)\right) \right) - G_h(x) \right| = o_p(n^{1/2}). \quad (2.5.6)$$

In the above, the covariance of $G_h(x)$ is given by

$$\Gamma_{G_h}(x, x') = \sum_{k \in \mathbb{Z}} E \left(K\left(\frac{x - Y_0}{h}\right) - E\left(K\left(\frac{x - Y_0}{h}\right)\right) \right) \left(K\left(\frac{x' - Y_k}{h}\right) - E\left(K\left(\frac{x' - Y_k}{h}\right)\right) \right).$$

(A8) $h = o(n^{-1/4})$.

Remark 2.5.2. The purpose of assumption (A7) is to help establish tightness for the supremum of the empirical process, which helps establish the uniform $O_p(\frac{1}{\sqrt{n}})$ error rate for the first transform.

It is not yet clear if the m -approximable condition can lead to (A7); however under similar dependence conditions, (A7) is shown to be correct. For example Theorem 2.1 in Arcones and Yu [1994] established such a result for β -mixing stationary series with $\mathcal{F}_K$ a V-C subgraph class of functions.

Lemma 2.5.3. *Let $\hat{U}_t^{(h)} = \hat{F}_h(Y_t)$. With assumptions (A1)–(A8) holding,*

$$\sup_{t \in [n]} |\hat{U}_t^{(h)} - U_t| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

With assumptions (A1)–(A8) along with Lemma 2.5.2 and 2.5.3 holding, we can prove MFB validity results using either the empirical CDF estimator $\bar{F}$ or the kernel CDF estimator $\hat{F}_h$. Therefore in the following results, we only consider a general CDF estimator $\hat{F}$ that can represent either $\bar{F}$ or $\hat{F}_h$.

The next theorem states that by choosing appropriate threshold c and bandwidth parameter l , we can indeed estimate the autocovariance structure of the hidden Gaussian process with the estimated series $\tilde{Z}_t$.

Theorem 2.5.4. *Assume the conditions of Lemma 2.5.2 or 2.5.3, with appropriate rate for $l(n) \rightarrow \infty$ and $c(n) \rightarrow \infty$. Then*

$$\sum_{k=0}^{\lfloor c\kappa l \rfloor} |\hat{\sigma}(k) - \sigma(k)| \xrightarrow{P} 0 \quad (2.5.7)$$

which implies

$$\left\| \hat{\Sigma}_n - \Sigma_n \right\|_{op} \xrightarrow{P} 0 \quad (2.5.8)$$

and

$$\left\| \hat{\Sigma}_n^{-1} - \Sigma_n^{-1} \right\|_{op} \xrightarrow{P} 0. \quad (2.5.9)$$

The rate of divergence for $l(n)$ and $c(n)$ can be found in the proof of Theorem 2.5.4.

With the above results, we can show that the pseudo data generated by MF and LMF

bootstrap procedure converge in distribution to the truth. Before proceeding, we state one more assumption:

Assumption. *Consistency of matrix square root estimator:*

$$(A9) \quad \left\| \widehat{\Sigma}_n^{1/2} - \Sigma_n^{1/2} \right\|_{op} = o_p(1) \text{ and } \left\| \widehat{\Sigma}_n^{-1/2} - \Sigma_n^{-1/2} \right\|_{op} = o_p(1).$$

Assumption (A9) holds if we have an additional mild convergence rate for the estimator towards the true covariance matrix, which can be achieved by choosing an appropriate tapering parameter $l(n) \rightarrow \infty$ and normal thresholding parameter $c(n) \rightarrow \infty$. By Theorem 2.1 of Drmac et al. [1994], if

$$(\log n)^2 \left\| \widehat{\Sigma}_n - \Sigma_n \right\|_{op} \xrightarrow{P} 0 \quad (2.5.10)$$

then (A9) holds; see also Jentsch and Politis [2015] and the Corrigendum of McMurry and Politis [2010]).

Lemma 2.5.5. *Assume (A1)-(A8). (1) For LMF bootstrap, $\forall d \in \mathbb{N}$,*

$$(Y_{t_1}^*, Y_{t_2}^*, \dots, Y_{t_d}^*) \xrightarrow{d^*} (Y_{t_1}, Y_{t_2}, \dots, Y_{t_d}), \text{ in probability.} \quad (2.5.11)$$

(2) Further assume (A9). Then the model-free procedure is asymptotically equivalent to the limit model-free procedure. I.e., the infinite sequence

$$(\xi_1^*, \xi_2^*, \dots) \text{ converges in distribution to } (\zeta_1, \zeta_2, \dots) \text{ in probability} \quad (2.5.12)$$

where $(\zeta_1, \zeta_2, \dots)$ is an infinite sequence with entries being i.i.d. $\mathcal{N}(0, 1)$. Furthermore, equation (2.5.11) holds for the model-free bootstrap.

Let $\gamma_Y^*(k)$ and Γ_n^* denote the bootstrap analogues of $\gamma_Y(k)$ and Γ_n . The long-run variance in the bootstrap world is then $(\sigma_\infty^*)^2 = \sum_{k \in \mathbb{Z}} \gamma_Y^*(k)$. We now establish our main theorem as in

(2.5.2) :

Theorem 2.5.6. *Under (A1)-(A9). For both MF bootstrap and LMF bootstrap*

$$(\sigma_{\infty}^*)^2 \rightarrow \sigma_{\infty}^2, \text{ in probability.} \quad (2.5.13)$$

and

$$\sup_{x \in \mathbb{R}} \left| P^*(\sqrt{n}(\bar{Y}_n^* - E^*(Y_t^*)) \leq x) - P(\sqrt{n}(\bar{Y}_n - E(Y_0)) \leq x) \right| \xrightarrow{P} 0. \quad (2.5.14)$$

2.5.2 Bootstrap validity for smooth functions of linear statistics

Now that the model-free bootstrap validity for the mean has been shown, we can extend the result for statistics of the form (2.5.1). For convenience, denote $X_t = (Y_t, \dots, Y_{t+k-1})$, $\hat{\theta}_n = \frac{1}{n-k+1} \sum_{t=1}^{n-k+1} g(X_t)$ and $\theta_0 = E(g(X_t))$ with bootstrap analogue $\hat{\theta}_n^* = \frac{1}{n-k+1} \sum_{t=1}^{n-k+1} g(X_t^*)$ and $\theta_0^* = E^*(g(X_t^*))$, respectively. In order to extend previous results using same proof strategy, we need assumptions such that the key points in the proof are checked:

Assumption Set 2.5.2. *Properties for functions of linear statistics*

(A10) $q(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{\tilde{d}}$ has continuous partial derivatives in a neighborhood of θ_0 , and $\sum_{i=1}^d (\partial q / \partial x_i)|_{x=\theta_0} x_i$ does not vanish.

(A11) $g = (g_1, \dots, g_d)$ is a continuous function and $\forall i \leq d$, $g_i(X_t)$ is $L^p - m$ -approximable in the sense of (C2) with $p > 2$ and $g_i(X_t^{(m)})$ its m -approximation, and appropriate constant A such that Lemma 2.5.1 holds for all $g_i(X_t)$.

Theorem 2.5.7. *With (A1) - (A11) hold, for MF bootstrap and LMF bootstrap with empirical*

CDF $\bar{F}$ and kernel smoothed CDF $\hat{F}_h$,

$$\sup_{x \in \mathbb{R}^d} |P^*(\sqrt{n}(q(\hat{\theta}_n^*) - q(\theta_0^*)) \leq x) - P(\sqrt{n}(q(\hat{\theta}_n) - q(\theta_0)) \leq x)| \xrightarrow{P} 0. \quad (2.5.15)$$

The proof is sketched as follows. First of all, by checking the key points in the proof of Theorem 2.5.6, we can show that every linear combination of $\{\hat{\theta}_{n,i}\}_{i=1}^d$ and its bootstrap analogue satisfy a CLT, and their limiting variances equal in probability. By the Cramér-Wold device, central limit theorem holds for the vector form $\hat{\theta}_n$. Then by the δ -method, we can prove CLT for the smooth function transform $q(\hat{\theta}_n)$.

The mean is now a special case of Theorem 2.5.7. Moreover, it can be applied to more general statistics such as the autocovariance and/or autocorrelation. We will address the autocovariance case in the following.

Autocovariances

For simplicity, let $X_t = (Y_t, Y_{t+1}, \dots, Y_{t+k}) \in \mathbb{R}^{k+1}$, $g(X_t) = (Y_t Y_{t+1}, \dots, Y_t Y_{t+k})$ and q is the identity map. Then $\frac{1}{n-k} \sum_{t=1}^{n-k} g(X_t)$ is a version of an estimator that estimates autocovariances up to lag k , assuming $EY_0 = 0$. Note that we can also design g and q more carefully such that we get the usual autocovariance estimators. Then the following corollary holds:

Corollary 2.5.7.1. *Assume $Y_t \in L^p$, $p > 4$ such that (C2) is satisfied with rate A . Then $g(X_t)$ as defined above also satisfies (C2) with $p' = \frac{p}{2} > 2$. Then by Theorem 2.5.7 the model-free bootstrap procedure is asymptotically correct for bootstrapping a vector of autocovariances.*

Proof. Let $Y_t^{(m)}$ be the approximation series generated by coupling. Then

$$\begin{aligned}
\left\| Y_t Y_{t+k} - Y_t^{(m)} Y_{t+k}^{(m)} \right\|_{p'} &= \left\| Y_t Y_{t+k} - Y_t^{(m)} Y_{t+k} + Y_t^{(m)} Y_{t+k} - Y_t^{(m)} Y_{t+k}^{(m)} \right\|_{p'} \\
&\leq \|Y_{t+k}\|_p \left\| Y_t - Y_t^{(m)} \right\|_p + \left\| Y_t^{(m)} \right\|_p \left\| Y_{t+k} - Y_{t+k}^{(m)} \right\|_p \\
&= 2 \|Y_t\|_p \left\| Y_t - Y_t^{(m)} \right\|_p \\
&= o(m^{-A})
\end{aligned} \tag{2.5.16}$$

To complete proof, we apply Theorem 2.5.7. □

Remark 2.5.3. (a). The usual assumption of finite sum of 4^{th} order cumulants of Y_t that ensures finiteness of asymptotic variance can be dropped since it can be deduced by checking assumption (A11) along with Lemma 2.5.1.

(b). Notably, Kreiss et al. [2011] showed that the autoregressive(AR) sieve bootstrap procedure does not work in general for bootstrapping autocovariances of strictly stationary series (even for linear series that are not Gaussian) because of inconsistency of the limiting variance associated the companion AR process. In essence, the AR-sieve bootstrap (and the previously mentioned Linear Process bootstrap) mimic correctly the first and second order moment structure of Y_t ; if the statistic of interest has a large-sample distribution that depends on higher order moments, the AR-sieve bootstrap (and the LPB) may fail. However, bootstrap validity holds for the MFB due to the assumption that $Y_t = f(W_t)$ where W_t is a stationary Gaussian process whose covariance structure can be consistently estimated (up to affine transform). Since a Gaussian process is fully described by its second order moment property, consistency of higher moments can also be obtained by the bootstrap procedure. Therefore the model-free bootstrap also works for higher order statistics.

2.5.3 Approximately linear statistics and sample quantiles

Now consider a statistic $\hat{\eta}_n$ that can be expressed as

$$\hat{\eta}_n = \frac{1}{n} \sum_{t=1}^n g(Y_t) + o_p(1/\sqrt{n}). \quad (2.5.17)$$

Under aforementioned conditions, the model-free bootstrap can be applied to the linear statistic $\hat{\theta}_n = \frac{1}{n} \sum_{t=1}^n g(Y_t)$. Since $\hat{\theta}_n$ will be $\sqrt{n}$ -consistent (under the required conditions), it follows that $\hat{\eta}_n$ and $\hat{\theta}_n$ are asymptotically equivalent, i.e., $\sqrt{n}(\hat{\eta}_n - \theta_0)$ has the same asymptotic distribution with $\sqrt{n}(\hat{\theta}_n - \theta_0)$ where $\theta_0 = E(g(Y_t))$. Therefore, since the model-free bootstrap works to approximate the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$, it will also work to approximate the distribution of $\sqrt{n}(\hat{\eta}_n - \theta_0)$.

Focusing on sample quantiles, we have the following result:

Lemma 2.5.8. *Assume (C2) for $\{Y_t\}$ with appropriate rate A as well as (A2). Let $u_{n,p} = \bar{F}^{-1}(p)$; $u_p = F_Y^{-1}(p)$. Then for the Bahadur-Kiefer process mentioned in Wu [2005b]:*

$$\alpha_n = |f_Y(u_p)(u_{n,p} - u_p) - (p - \bar{F}(u_p))|. \quad (2.5.18)$$

We have $\forall p \in (0, 1)$, $\alpha_n = O_p(r_n)$ where $r_n = o(n^{-1/2})$. As a result,

$$u_{n,p} - u_p = \frac{p - \bar{F}(u_p)}{f_Y(u_p)} + o_p(n^{-1/2}). \quad (2.5.19)$$

Under the additional assumptions of (A5)-(A8), the above equation also holds for $u_{n,p} = \hat{F}_h^{-1}(p)$, i.e.

$$u_{n,p} - u_p = \frac{p - \hat{F}_h(u_p)}{f_Y(u_p)} + o_p(n^{-1/2}). \quad (2.5.20)$$

Remark 2.5.4. As shown in Wu [2005b], for the empirical CDF estimator $\bar{F}$, a stronger version of equation (2.5.19) holds provided a faster geometric convergence rate for the m - approximation

assumption (C2).

In equation (2.5.19) and (2.5.20), $\frac{p - \hat{F}(u_p)}{f_Y(u_p)}$ is of the linear form in equation (2.5.17). For equation (2.5.19),

$$\frac{p - \bar{F}(u_p)}{f_Y(u_p)} = \frac{1}{n} \sum_{k=1}^n \frac{F_Y(u_p) - I\{Y_k \leq u_p\}}{f_Y(u_p)},$$

the same holds for equation (2.5.20) by the relation

$$\frac{p - \hat{F}_h(u_p)}{f_Y(u_p)} = \frac{1}{n} \sum_{k=1}^n \frac{F_Y(u_p) - K\left(\frac{u_p - Y_k}{h}\right)}{f_Y(u_p)}.$$

By the previous analysis, we have the following theorem.

Theorem 2.5.9. *Under assumptions of Theorem 2.5.7 with respect to the linear part of $u_{n,p} - u_p$ and with Lemma 2.5.8 holding, both MF bootstrap and LMF bootstrap are asymptotically valid for sample quantiles.*

A special case of interest is to let $p = \frac{1}{2}$. Then, the above discussion shows validity of the model-free bootstrap for the sample median.

2.5.4 Bootstrap validity for kernel smoothed spectral density

Another important parameter of interest is the spectral density evaluated at some frequency ω . For this subsection, let $f_{sp.d}(\omega)$ denote the spectral density of $\{Y_t\}$ which is defined as $f_{sp.d}(\omega) = \frac{1}{2\pi} \sum_{k \in \mathbb{Z}} \gamma_Y(k) e^{-ik\omega}$. The kernel smoothed estimator is a widely used estimator of $f_{sp.d}(\omega)$ in this setting. Let $I_n(\omega_j) = \frac{1}{2\pi n} \left| \sum Y_k e^{-ik\omega_j} \right|^2$ denote the periodogram of $\{Y_t\}$ where $\omega_j \in \{\frac{2\pi j}{n}, j \in [n]\}$ are the Fourier frequencies. The kernel smoothed spectral density estimator is defined as

$$\hat{f}_{sp.d}(\omega) = \sum_{j \in [n]} \tilde{\kappa}_h(\omega - \omega_j) I_n(\omega_j), \quad (2.5.21)$$

where $\tilde{\kappa}(\omega)$ is a kernel function under certain assumptions and $\tilde{\kappa}_h(\cdot) = h^{-1} \tilde{\kappa}(\cdot/h)$. Let $\hat{f}_{sp.d}^*(\omega)$ be the kernel smoothed estimator based on the bootstrap samples $\{Y_t^*\}_{t=1}^n$ generated through

model-free bootstrap. We would like to show validity for this procedure. Besides key assumptions for previous theorems to hold, we also need the following:

Assumption Set 2.5.3. Additional assumptions for bootstrapping spectral density

(A12) $\{Y_t\}$ satisfies (C2) with $p = 4$; $\inf_{\omega \in [-\pi, \pi]} f_{sp.d}(\omega) > 0$.

(A13) The kernel $\tilde{\kappa}(\cdot)$ is a symmetric and bounded square integrable function. Also, $\int \tilde{\kappa}(u) du = 1$ and $\int u^2 \tilde{\kappa}(u) du < \infty$.

(A14) $h \rightarrow 0$ and $nh \rightarrow \infty$.

Then the following holds.

Theorem 2.5.10. Under (A1)-(A9) and (A12)-(A14), for any fixed $\omega \in [-\pi, \pi]$

$$\sqrt{nh} \left(\hat{f}_{sp.d}(\omega) - E \hat{f}_{sp.d}(\omega) \right) \xrightarrow{d} \mathcal{N}(0, \sigma_\omega^2). \quad (2.5.22)$$

In the above, $\sigma_\omega^2 = f_{sp.d}^2(\omega) \int \tilde{\kappa}^2(u) du$ for $\omega/\pi \notin \mathbb{Z}$ and $\sigma_\omega^2 = 2f_{sp.d}^2(\omega) \int \tilde{\kappa}^2(u) du$ for $\omega/\pi \in \mathbb{Z}$.

Also, for both model-free and limit model-free bootstrap

$$\sup_{x \in \mathbb{R}} \left| P^*(\sqrt{nh} (\hat{f}_{sp.d}^*(\omega) - E^* \hat{f}_{sp.d}^*(\omega)) \leq x) - P(\sqrt{nh} (\hat{f}_{sp.d}(\omega) - E \hat{f}_{sp.d}(\omega)) \leq x) \right| \xrightarrow{P} 0. \quad (2.5.23)$$

As is well-known, the bandwidth h governs the trade-off between the bias and variance of $\hat{f}_{sp.d}(\omega)$. If we choose h in such a way that *undersmoothing* occurs, i.e., when the bias of $\hat{f}_{sp.d}(\omega)$ is of a smaller order of magnitude than its standard deviation, then equation (2.5.23)

holds true with $E\hat{f}_{sp.d}(\omega)$ replaced by $f_{sp.d}(\omega)$, i.e., we have

$$\sup_{x \in \mathbb{R}} \left| P^*(\sqrt{nh}(\hat{f}_{sp.d}^*(\omega) - E^*\hat{f}_{sp.d}^*(\omega)) \leq x) - P(\sqrt{nh}(\hat{f}_{sp.d}(\omega) - f_{sp.d}(\omega)) \leq x) \right| \xrightarrow{P} 0. \quad (2.5.24)$$

Equation (2.5.24) can then be used to construct model-free confidence intervals for $f_{sp.d}(\omega)$ based on the quantiles of the bootstrap distribution of $\sqrt{nh}(\hat{f}_{sp.d}^*(\omega) - E^*\hat{f}_{sp.d}^*(\omega))$.

2.6 Model-free bootstrap for prediction

For simplicity, we focus on proving model-free bootstrap validity for one-step ahead prediction; generalizing to the h -step ahead case is also possible. Let Y_{n+1} be the 1- step ahead value of the time series dataset $\{Y_t\}_{t=1}^n$ we wish to predict. We call the conditional distribution D_{n+1} of $Y_{n+1}|\underline{Y}_n$ the predictive distribution. Naturally, the quantiles of D_{n+1} can help us construct prediction interval of Y_{n+1} .

Oftentimes, we also want the prediction interval to be centered at a certain meaningful point predictor, which brings the notion of **predictive root**. Let $\hat{Y}_{n+1}$ be a one-step ahead predictor of Y_{n+1} based on the original series $\{Y_t\}_{t=1}^n$, i.e., $\hat{Y}_{n+1}$ depends entirely on the data $\{Y_t\}_{t=1}^n$. The predictive root, analogous to the case of confidence intervals, is defined as the residual between the data and predictor

$$Y_{n+1} - \hat{Y}_{n+1}. \quad (2.6.1)$$

Let $L_{\alpha/2}$ and $R_{\alpha/2}$ be the lower $\alpha/2$ quantile and upper $\alpha/2$ quantile of the (conditional on $\underline{Y}_n$) distribution of the predictive root. Then, an exact two-sided $1 - \alpha$ prediction interval for Y_{n+1} is:

$$(\hat{Y}_{n+1} + L_{\alpha/2}, \hat{Y}_{n+1} + R_{\alpha/2}); \quad (2.6.2)$$

in other words,

$$P(\hat{Y}_{n+1} + L_{\alpha/2} \leq Y_{n+1} \leq \hat{Y}_{n+1} + R_{\alpha/2} | \underline{Y}_n) = 1 - \alpha. \quad (2.6.3)$$

We wish to approximate the quantiles of the predictive root by the bootstrap procedure. Let Y_{n+1}^* be the one-step ahead value generated through the bootstrap; also let $\hat{Y}_{n+1}^*$ be the one-step ahead predictor of Y_{n+1} conditioning on $\underline{Y}_n$, where its formula is estimated by the bootstrap samples $\{Y_t^*\}_{t=1}^n$. We say that bootstrap validity for prediction intervals holds if for any $\alpha \in (0, 1)$:

$$L_{\alpha/2}^* \rightarrow L_{\alpha/2} \text{ and } R_{\alpha/2}^* \rightarrow R_{\alpha/2} \quad (2.6.4)$$

in probability. Here $L_{\alpha/2}^*$ and $R_{\alpha/2}^*$ denote the relative $\alpha/2$ quantiles with respect to the conditional distribution of the bootstrap predictive root $Y_{n+1}^* - \hat{Y}_{n+1}^*$.

In other words, the two-sided approximate $1 - \alpha$ bootstrap prediction interval for Y_{n+1} is

$$(\hat{Y}_{n+1} + L_{\alpha/2}^*, \hat{Y}_{n+1} + R_{\alpha/2}^*), \quad (2.6.5)$$

while equation (2.6.4) would imply

$$P(\hat{Y}_{n+1} + L_{\alpha/2}^* \leq Y_{n+1} \leq \hat{Y}_{n+1} + R_{\alpha/2}^* | \underline{Y}_n) \rightarrow 1 - \alpha. \quad (2.6.6)$$

The following theorem states that bootstrap validity holds for one-step ahead prediction intervals.

Theorem 2.6.1. (*Model-free bootstrap validity for one-step ahead prediction*) Assume that the conditional distribution of the predictive root $Y_{n+1} - \hat{Y}_{n+1}$ is continuous. With assumptions (A1) - (A9), **Algorithms 2 and 3** are asymptotically valid in the sense of equation (2.6.4).

Unlike the bootstrap for parameter inference, the conditional distribution of the predictive root is not asymptotically degenerate. Hence, no central limit theorems are required for parameter estimates since their associated variability vanishes asymptotically.

Asymptotic validity of bootstrap prediction intervals is of less importance than that of

confidence intervals. A good finite-sample prediction interval should incorporate the variance of all estimated features, or it will result in *undercoverage*. However, as stated above, the property of asymptotic validity has nothing to do with capturing finite-sample variability in estimation; see also Ch. 2.4.1 of Politis (2015). In this respect, the performance of prediction intervals must be quantified in finite-sample numerical experiments as in the following section.

2.7 Numerical results

We look into three data generating models and evaluate the coverage performance of our proposed MFB methods with respect to both confidence intervals and prediction intervals. All three models satisfy equation (2.4.1) with

$$f(x) = \begin{cases} -\sqrt{-x} & x < 0 \\ \frac{(x+1)^2}{10} & x \geq 0 \end{cases} \quad (2.7.1)$$

and W_t generated by an autoregressive moving average (ARMA) model, namely

$$W_t - \phi_1 W_{t-1} - \cdots - \phi_p W_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

with $\varepsilon_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. The three ARMA models considered are as follows: in one of the following ways:

1. MA model of order one, with $\theta_1 = -0.5$ and all other parameters zero.
2. AR model of order one, with $\phi_1 = 0.5$ and all other parameters zero.
3. MA model of order 30, with $\theta_1 = 2$, $\theta_2 = 1$, $\theta_k = \frac{10}{k^2}$ when $3 \leq k \leq 30$, and all other parameters zero.

We analyze the performance of model-free and limit model-free bootstrap with both the empirical CDF estimator and the kernel CDF estimator.

2.7.1 Performance for bootstrap confidence intervals

For confidence intervals, the parameter of interests are: the mean for all three models; lag 1 autocovariance for model 1 and 2; lag 2 autocovariance for model 3. As a comparison, we benchmark our methods with the two most popular alternative methods, i.e. block bootstrap and AR-sieve bootstrap.

The mean is, of course, a linear statistic. The lag- k autocovariance can be written in the form (2.5.1) with

$$g(Y_t, \dots, Y_{t+k}) = \begin{bmatrix} Y_t Y_{t+k} \\ Y_t \end{bmatrix}$$

and

$$q\left(\frac{1}{n-k} \sum_{t=1}^{n-k} g(Y_t, \dots, Y_{t+k})\right) = \frac{1}{n-k} \sum_{t=1}^{n-k} Y_t Y_{t+k} - \left[\frac{1}{n-k} \sum_{t=1}^{n-k} Y_t\right]^2.$$

Although we can do block bootstrap on the Y_t and recompute $\hat{\gamma}_k$ on the bootstrap data to construct confidence intervals, this procedure will suffer from end effects resulting into an estimator that is biased towards 0. As a remedy, let $X_t = (Y_t, \dots, Y_{t+k})$ as we did in Section 2.5.2, then a block bootstrap on the X_t data will relieve this problem and it is equivalent to the so-called blocks-of-blocks bootstrap on the Y_t data; see Politis and Romano [1992], as well as Paradigm 12.8.11 of the book by McElroy and Politis [2019]. For our purposes, we used first level of blocking with $k = 5$ that can be used to capture autocovariances up to lag 4.

For the block bootstrap on the X_t data, we can choose the block size b as $b = \text{const} * n^{1/3}$ where const is selected according to Patton et al. [2009]. As for the AR-sieve bootstrap, the order p of the fitted AR model is selected through minimizing the AIC; see Bühlmann [1997].

For simulation purpose, in the model-free bootstrap algorithm 1 we need to choose the

threshold parameter c . Under a finite sample setting, the main goal is to resolve the issue of calculating $\Phi^{-1}(1)$. The following simple empirical rule is used: for $\bar{F}$ scenario, we adjust the value 1 in $\{\hat{U}_t\}_{t=1}^n$ to $\frac{n-1}{n}$ to avoid the problem of $\Phi^{-1}(1)$, which is equivalent to setting $c = \Phi^{-1}(\frac{n-1}{n})$. In the case of the kernel CDF $\hat{F}_h$, besides applying the above rule, we can also avoid having 1 in $\{\hat{U}_t\}_{t=1}^n$ by choosing an appropriate kernel function(e.g., the Gaussian CDF) and the bandwidth, in which case choosing c is no longer necessary.

The metric we use for comparison is the empirical coverage rate(empirical CVR) for the bootstrap intervals. To elaborate, let $N = 1000$ be the number of replicated experiments, $n \in \{100, 200, 500, 1000\}$ be the length of samples, $\alpha = 0.05$ and $B = 250$ the number of bootstrap replications. So for experiment $i \in \{1, \dots, N\}$, we generate a time series sample of length n , and use different methods to construct a $1 - \alpha$ bootstrap confidence interval (L_i, R_i) based on the B bootstrap replicates. Let θ_0 denote the true parameter of interest; then the empirical coverage rate for the whole experiment is:

$$CVR_{\theta_0} = \frac{1}{1000} \sum_{i=1}^{1000} I_{\theta_0 \in (L_i, R_i)} \quad (2.7.2)$$

The purpose of looking at CVR is two fold: first, comparing CVR values in a finite-sample size setting will tell which method has better performance. Second, as the sample size n gets large, asymptotic validity can be observed by checking whether the empirical CVR converges to the nominal $100(1 - \alpha)\%$ percent, which will provide numerical grounds to our previous conclusions.

The following tables contain CVR results for different settings all with nominal $\alpha = 0.05$ (95% confidence interval). The acronyms in the "Method" column of Table 1 have the following meaning. "MF" for model-free bootstrap; "LMF" for limit model-free bootstrap; "emp" for empirical CDF estimator used in the bootstrap procedure; "ker" for kernel CDF estimator; "BB" for block bootstrap; "AR-sieve" for the autoregressive sieve bootstrap.

For the mean parameter (see Table 1), all the methods being compared are theoretically valid for all three models. We can observe that for each method, as n increases the coverage rate approaches 95% albeit the speed of convergence can be different. The AR sieve holds an

Table 2.1: Empirical CVR for the mean parameter across 3 models.

Method	n=100	200	500	1000	2000
model 1					
MF-ker	92.1	92.5	93.2	93.6	94.8
LMF-ker	92.3	92.5	93.0	94.7	93.9
MF-emp	91.6	92.4	93.1	94.0	94.4
LMF-emp	92.0	92.9	93.5	93.8	94.0
BB	90.4	93.1	94.1	95.0	95.0
AR-sieve	94.0	93.6	94.0	95.2	95.6
model 2					
MF-ker	91.4	92.4	93.2	93.2	93.7
LMF-ker	88.5	92.1	93.3	93.3	94.4
MF-emp	88.7	90.9	92.8	94.1	92.9
LMF-emp	85.3	88.4	92.4	93.0	92.9
BB	85.2	89.6	91.3	92.3	93.0
AR-sieve	90.6	92.9	92.8	93.6	94.2
model 3					
MF-ker	87.9	89.3	90.3	91.7	93.3
LMF-ker	86.9	88.1	90.5	92.1	93.1
MF-emp	82.7	85.5	89.6	91.4	92.7
LMF-emp	80.9	85.8	89.5	91.4	92.8
BB	82.2	85.4	87.3	89.8	91.4
AR-sieve	83.6	88.1	90.0	91.9	92.2

obvious advantage over other methods for model 1 with coverage close to 95% even at $n = 100$, whereas the proposed model-free bootstrap works on par with the block bootstrap. However for the other two models, MF and LMF bootstrap with kernel CDF estimator has a noticeable advantage over the AR-sieve and block bootstrap, especially for model 3, which has a more complex data generating process.

Interestingly, the behavior of the model-free bootstrap using the empirical CDF is almost the same as that of the block bootstrap. Fortunately enough, the coverage advantage of MF based on kernel CDF over empirical CDF goes away for large n . As a result, we can use the empirical CDF approach under a large sample scenario, to avoid calculating the quantile inverse of a kernel CDF, which becomes computationally expensive as sample size grows.

Table 2.2: Empirical CVR for lag 1 autocovariance for the first 2 models.

Method	n=100	200	500	1000	2000
model 1					
MF-ker	82.0	88.2	90.5	92.2	91.1
LMF-ker	89.6	94.3	94.0	94.8	94.1
MF-emp	80.0	88.3	90.1	91.6	92.0
LMF-emp	89.2	93.5	94.0	95.0	93.3
BB	86.0	88.8	91.5	91.1	93.0
AR-sieve	94.3	94.8	94.6	96.3	95.0
model 2					
MF-ker	75.0	86.5	94.2	94.6	94.6
LMF-ker	92.2	94.2	95.4	94.0	93.4
MF-emp	68.6	82.9	90.2	91.9	94.2
LMF-emp	76.9	82.5	89.0	91.6	94.4
BB	82.3	86.2	89.0	90.5	91.5
AR-sieve	88.3	91.8	95.1	95.5	95.5

Tables 2 and 3 present the empirical CVR for the autocovariance with different models. Both lag 1 and lag 2 autocovariances were considered; for conciseness, we present lag 1 results from models 1 and 2 (see Table 2), and lag 2 results from model 3 (see Table 3). Recall that the AR-sieve bootstrap is not asymptotically valid in general for the autocovariance. Comparing Tables 2 and 3, we see that the AR-sieve bootstrap works well for the autocovariance in models 1 and 2 but not with data from model 3; see Table 3 where the AR-sieve CVR appears to converge to around 86% instead of the 95% nominal level.

We can also observe the asymptotic validity of model-free bootstrap methods manifested in Table 2 and 3. Furthermore, model-free methods appear to enjoy a faster convergence towards nominal compared to block bootstrap. Interestingly, limit model-free with kernel CDF estimator works significantly better than other methods across all 3 models.

Table 2.3: Empirical CVR for lag 2 autocovariance for model 3.

Method	n=100	200	500	1000	2000
model 3					
MF-ker	76.9	87.8	91.7	93.0	94.1
LMF-ker	77.7	86.0	91.3	92.4	93.8
MF-emp	68.2	80.9	86.8	91.6	93.5
LMF-emp	65.1	77.5	85.8	89.5	93.5
BB	64.0	73.9	83.2	87.3	89.4
AR-sieve	69.4	76.4	84.9	86.1	86.1

2.7.2 Performance of bootstrap prediction intervals

We now move on to the prediction performance for the three models. The empirical coverage rate is defined similarly as

$$CVR_{Y_{n+1}} = \frac{1}{1000} \sum_{i=1}^{1000} I_{Y_{n+1}^{(i)} \in (\hat{Y}_n^{(i)} + L_i, \hat{Y}_n^{(i)} + R_i)} \quad (2.7.3)$$

where (L_i, R_i) are sample quantiles of the bootstrap predictive root generated by bootstrap and $Y_{n+1}^{(i)}$ is the $n+1$ value of the time series sample from the i^{th} experiment.

Since the block bootstrap is not a viable method for generating one-step ahead prediction value, we benchmark predictive performance of the model-free methods by comparing them with the AR-sieve bootstrap. The bootstrap samples are generated in a forward bootstrap manner as described in Algorithm 3.1 of Pan and Politis [2016a]; see also Alonso et al. [2002]. Table 4 provides the empirical CVR with $n \in \{100, 200, 300, 500\}$; the samples sizes are smaller compared to our previous simulations because of the increasing computational cost for prediction. However, it is reassuring that all methods considered, i.e., the 4 model-free variations as well as the AR-sieve bootstrap, produce prediction intervals with CVR close to the nominal 95% even with n as low as 200. It is difficult to do a finer comparison of these 5 bootstrap methods since, for computational reasons, we had to choose a small number of bootstrap replications ($B = 250$). Ongoing work includes devising an analog of the ‘Warp-Speed’ method of Giacomini et al. [2013] that will

speed up Monte Carlo experiments involving bootstrap in the case of prediction intervals.

Remark 2.7.1. An important question is the performance difference of the MFB algorithm proposed here and that proposed in Das and Politis [2021]. The difference is two-fold: First of all, the main thrust of prediction based on locally stationary data is that only local estimates (of the CDF of Y_t and autocovariance, etc.) can be used. If the local window is of size $b \ll n$, the estimates of Das and Politis [2021] would be less accurate than ours that are based on sample size n . As a consequence, our algorithm will perform better than Das and Politis [2021] in the case of globally stationary data. On the other hand, if global stationarity fails, our procedure will no longer be consistent, whereas that of Das and Politis [2021] would still perform well.

In addition, in order to handle boundary effects, Das and Politis [2020] developed a local linear distribution estimator which was then employed in Das and Politis [2021]. The local linear distribution estimator is not needed in the case of global stationarity which is the focus of the paper at hand.

2.8 Additional proofs

2.8.1 Preliminary results under m -approximation assumptions

Theorem 2.8.1. (*Berkes, Hörmann, Schauer (2009)*)

Define $R(s, t) = \sum_{1 \leq k \leq t} (I\{Y_k \leq s\} - F_Y(s))$. Assume F_Y is θ -Lipschitz continuous, and (C1) holds with $\gamma_m = m^{-C/\theta}$, $\delta_m = m^{-C}$ and some $C > 4$. Then there exists a two parameter Gaussian process $K(s, t)$ with $E(K(s, t)K(s', t')) = (t \wedge t')\Gamma(s, s')$, such that

$$\sup_{s, t} |R(s, t) - K(s, t)| = o(n^{1/2}(\log n)^{-\alpha}), \text{ a.s.} \quad (2.8.1)$$

Table 2.4: Empirical CVR for prediction intervals across the three models.

Method	n=100	200	300	500
model 1				
MF-ker	95.0	94.6	91.6	91.8
LMF-ker	94.1	96.6	93.9	93.0
MF-emp	94.0	94.2	92.2	94.2
LMF-emp	92.2	93.9	94.0	93.8
AR-sieve	94.4	94.6	94.0	93.2
model 2				
MF-ker	93.6	94.0	92.2	93.2
LMF-ker	90.2	95.3	96.4	92.8
MF-emp	94.1	93.4	95.0	94.2
LMF-emp	89.2	90.7	95.0	93.6
AR-sieve	94.0	93.2	94.0	94.4
model 3				
MF-ker	93.8	94.4	95.4	94.8
LMF-ker	91.8	96.7	95.6	93.6
MF-emp	92.2	96.2	96.4	95.2
LMF-emp	92.8	94.8	96.4	93.2
AR-sieve	91.4	94.8	93.2	93.0

for some $\alpha > 0$. In addition,

$$\Gamma(s, s') = \sum_{k \in \mathbb{Z}} E(I(Y_0 \leq s) - F_Y(s))(I(Y_k \leq s') - F_Y(s')) \quad (2.8.2)$$

is absolutely convergent for all choices of s, s' .

We are more interested in the case $t = n$ such that the above theorem reduces to a 1-dimensional centered Gaussian process K indexed by $s \in \mathbb{R}$. It turns out that tightness of K plays an important role in the accuracy of the estimated $\hat{U}_t$. Although Theorem 2.8.1 does not guarantee the limiting Gaussian process to be tight, it is reasonable to believe so and also it can be shown that for series of the form (3.1) tightness can be shown through continuity and boundedness of the covariance structure.

Theorem 2.8.2. (Berkes, Hörmann, Schauer (2011))

Assume (C2) with $p > 2$, $\eta \in (0, 1)$, $A > \frac{p-2}{2\eta}(1 - \frac{1+\eta}{p}) \vee 1$. Then $\sigma_\infty^2 = \sum_{k \in \mathbb{Z}} \gamma_Y(k)$ is absolutely convergent. Also,

$$\sum_{k=1}^n (Y_k - EY_0) = W_1(s_n^2) + W_2(t_n^2) + O(n^{(1+\eta)/p}), \text{ a.s.} \quad (2.8.3)$$

where W_1 and W_2 are two Brownian motions and $s_n^2 \sim \sigma_\infty^2 n$, $t_n^2 \sim cn^\gamma$ with $\gamma \in (0, 1)$. Here $a_n \sim b_n$ means $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1$; $a \vee b = \max\{a, b\}$.

Theorem 2.8.3. (McMurry & Politis(2010))

For the tapered estimator defined in equation (2.3) with tapering function (2.1), assume (C0) with $2 < p \leq 4$ for $\{Z_t\}$ and $\sum_{k \in \mathbb{Z}} \sigma(k) < \infty$, with $l = o(n^{\frac{p-2}{p}})$ then,

$$\sum_{k=0}^{\lfloor c_\kappa l \rfloor} |\hat{\sigma}(k) - \sigma(k)| \xrightarrow{L^{p/2}} 0, \quad (2.8.4)$$

and

$$\left\| \hat{\Sigma}_n^{(\kappa, l)} - \Sigma_n \right\|_{op} \xrightarrow{L^{p/2}} 0. \quad (2.8.5)$$

Furthermore, assuming Z_t has spectral density f_Z satisfying $0 < c_1 \leq f_Z(w) \leq c_2 < \infty, \forall w \in [0, 2\pi]$, $\hat{\Sigma}_n^{(\kappa, l)}$ is positive definite with probability tending to 1, and

$$\left\| (\hat{\Sigma}_n^{(\kappa, l)})^{-1} - \Sigma_n^{-1} \right\|_{op} = o_p(1) \quad (2.8.6)$$

2.8.2 Proof of relative lemmas and theorems

Proof of Lemma 3.3. 1. Let $\gamma_m = m^{-C/\theta}$ and $\delta_m = m^{-C}$. By Markov's inequality,

$$\begin{aligned} P(|Y_t - Y_t^{(m)}| > m^{-C/\theta}) &= P(|Y_t - Y_t^{(m)}|^p > m^{-Cp/\theta}) \\ &\leq \frac{E(|Y_t - Y_t^{(m)}|^p)}{m^{-Cp/\theta}} \\ &\ll \frac{m^{-pA}}{m^{-Cp/\theta}} = m^{-C} \end{aligned} \tag{2.8.7}$$

2.

$$\sum_{m=0}^{\infty} \|Y_t - Y_t^{(m)}\|_p \leq \sum_{m=0}^{\infty} \delta(m).$$

Since $\delta(m) \ll m^{-A}$ with $A > 1$, $\delta(m)$ is summable. Thus (C3) holds.

3. Let ε'_0 be the new independent sample to be used in $\delta_p(t)$. Let $\varepsilon''_0, \varepsilon''_{-1}, \dots$ be an infinite sequence of i.i.d. samples from F_ε . By triangle inequality we have:

$$\begin{aligned} \delta_p(m) &= \|h(\dots, \varepsilon_{-1}, \varepsilon_0, \dots, \varepsilon_m) - h(\dots, \varepsilon_{-1}, \varepsilon'_0, \dots, \varepsilon_m)\|_p \\ &\leq \|h(\dots, \varepsilon_{-1}, \varepsilon_0, \dots, \varepsilon_m) - h(\dots, \varepsilon''_{-1}, \varepsilon''_0, \varepsilon_1, \dots, \varepsilon_m)\|_p \\ &\quad + \|h(\dots, \varepsilon''_{-1}, \varepsilon''_0, \varepsilon_1, \dots, \varepsilon_m) - h(\dots, \varepsilon_{-1}, \varepsilon'_0, \dots, \varepsilon_m)\|_p \\ &= 2 \|Y_m - Y_m^{(m)}\|_p \end{aligned} \tag{2.8.8}$$

Thus (C3) implies (C0).

4. See page 2443 in Berkes et al. [2011]. □

Proof of Lemma 4.1. By Theorem 2.8.1, $\gamma_Y(k)$ is absolutely summable for $k \in \mathbb{Z}$. Then

$$\begin{aligned}\sqrt{n}(\bar{Y}_n - EY_0) &= \frac{1}{\sqrt{n}} \left(W_1(s_n^2) + W_2(t_n^2) + O_{a.s.}(n^{(1+\eta)/p}) \right) \\ &= W_1(s_n^2/n) + W_2(t_n^2/n) + O_{a.s.}(n^{(1+\eta)/p-1/2}) \\ &\xrightarrow{d} \mathcal{N}(0, \sigma_\infty^2).\end{aligned}\tag{2.8.9}$$

since $n^{\gamma-1} \rightarrow 0$ and $n^{(1+\eta)/p-1/2} \rightarrow 0$.

□

Proof of Lemma 4.2. Equation (4.5) is deduced by result in Lemma 3.3. Then by Theorem 2.8.1

$$\begin{aligned}\sup_{t \in [n]} |\hat{U}_t - U_t| &\leq \frac{1}{n} \sup_{s \in \mathbb{R}} \left| \sum_{k=1}^n (I(Y_k \leq s) - F_Y(s)) \right| \\ &= \frac{1}{n} \sup_s |R(s, n)| \\ &\leq \frac{1}{\sqrt{n}} \sup_s |K(s, 1)| + o(n^{-1/2}(\log n)^{-\alpha}), a.s.\end{aligned}\tag{2.8.10}$$

where $K(s, 1)$ is a centered Gaussian process with covariances

$$\Gamma(s, s') = \sum_{k \in \mathbb{Z}} E(I(Y_0 \leq s) - F_Y(s))(I(Y_k \leq s') - F_Y(s'))\tag{2.8.11}$$

absolutely convergent $\forall s, s' \in \mathbb{R}$. We also need the limiting Gaussian process $K(s, 1)$ to be tight such that $\sup_s |K(s, 1)| = O_p(1)$, which usually is true but not mentioned in Theorem 2.8.1. Another way to show this is to prove certain continuity and boundedness condition for the covariance $\Gamma(s, s')$.

Note that $K(s, 1)$ can be reparametrized as $K'(u), u = F_Y(s)$ such that the centered gaussian process is now living in a bounded domain $T = [0, 1]$ (reparametrization of the empirical process). Also, since F_Y is absolutely continuous and strictly increasing, $\sup_{s \in \mathbb{R}} |K(s, 1)| = \sup_{t \in [0, 1]} |K'(t)|$.

Let $\Gamma'(t, t')$ denote the covariance kernel of $K'(u)$. Obviously $\Gamma'(t, t') = \Gamma(F_Y^{(-1)}(t), F_Y^{(-1)}(t'))$, for which we have $\Gamma'(0, 0) = \Gamma'(1, 1) = 0$. Thus the Gaussian process $K'(u)$ is pinned to 0 a.s. at endpoints 0 and 1.

We claim that

- (1) $\Gamma'(t, t')$ is uniformly bounded for $t, t' \in [0, 1]^2$.
- (2) $\Gamma'(t, t')$ is continuous in t and t' .
- (3) K' satisfies the Kolmogorov-Čentsov condition: $\exists \alpha, \beta, C > 0$, such that

$$E|K'(t) - K'(s)|^\alpha \leq C|t - s|^{1+\beta}$$

(2) is a consequence of (1). To see this, note that $\forall t, \forall t_1 < t_2$,

$$\begin{aligned} |\Gamma'(t_1, t) - \Gamma'(t_2, t)| &= \left| \sum_{k \in \mathbb{Z}} E(I(U_0 \leq t_1) - t_1)(I(U_k \leq t) - t) - E(I(U_0 \leq t_2) - t_2)(I(U_k \leq t) - t) \right| \\ &= E[I(t_1 \leq U_0 \leq t_2) - (t_2 - t_1)] \left[\sum_{k \in \mathbb{Z}} (I(U_k \leq t) - t) \right] \end{aligned}$$

By previous theorem, the expectation exists $\forall t$ and $t_1 < t_2$, and uniformly bounded by

$2 \sup_{t \in [0, 1]} |\Gamma'(t, t)|$. As $|t_2 - t_1| \rightarrow 0$, the random variable inside the expectation converges to 0 a.s. Thus by dominated convergence theorem we have $|\Gamma'(t_1, t) - \Gamma'(t_2, t)| \rightarrow 0$.

For (1), by positive definiteness of covariance kernel we have that

$$\begin{aligned}
\sup_{(t,t') \in [0,1]^2} |\Gamma'(t,t')| &= \sup_{t \in [0,1]} |\Gamma'(t,t)| = \sup_{t \in [0,1]} \left| \sum_{k \in \mathbb{Z}} E(I(U_0 \leq t) - t)(I(U_k \leq t) - t) \right| \\
&= \sup_{t \in [0,1]} \left| \sum_k E \left[I(U_0 \leq t) I(U_k \leq t) - t^2 \right] \right| \\
&= \sup_{t \in [0,1]} \left| \sum_k \int_{(-\infty, g^{-1}(t)]^2} f_W^{(k)}(x,y) - f_W(x)f_W(y) dx dy \right| \\
&\leq \int_{\mathbb{R}^2} \sum_k |f_W^{(k)}(x,y) - f_W(x)f_W(y)| dx dy \quad \dots (*)
\end{aligned} \tag{2.8.12}$$

Where $g = F_Y \circ f$ is a monotone (increasing) function; $f_W^{(k)}$ is the bivariate normal density of (W_0, W_k) ; f_W is the normal density of W_t . We show that $(*)$ is bounded.

Rewrite $f_W^{(k)}(x,y)$ as $f_{xy}(a_k, b_k) = \frac{1}{2\pi} a_k^{-1/2} \exp\{-\frac{1}{2} a_k^{-1} b_k\}$, with $a_k = \sigma_W^2(0) - \sigma_W^2(k)$ and $b_k = \sigma_W(0)(x^2 + y^2) - 2\sigma_W(k)xy$, where $\sigma_W(k)$ is the lag- k autocovariance for W_t .

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial a_k} \\ \frac{\partial f}{\partial b_k} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} a_k^{-\frac{3}{2}} (a_k^{-1} b_k - 1) e^{-\frac{1}{2} a_k^{-1} b_k} \\ -\frac{1}{2} a_k^{-\frac{3}{2}} e^{-\frac{1}{2} a_k^{-1} b_k} \end{bmatrix} \tag{2.8.13}$$

Then

$$\begin{aligned}
&\int_{\mathbb{R}^2} |f_{xy}(a_k, b_k) - f_{xy}(\sigma_W^2(0), \sigma_W(0)(x^2 + y^2))| dx dy \\
&= \int_{\mathbb{R}^2} |\nabla f^T(\sigma_W^2(0), \sigma_W(0)(x^2 + y^2)) \begin{bmatrix} -\sigma_W^2(k) \\ -2\sigma_W(k)xy \end{bmatrix}| dx dy + o(\sigma_W(k) + \sigma_W^2(k))
\end{aligned} \tag{2.8.14}$$

and

$$\begin{aligned}
(*) &= \int_{\mathbb{R}^2} \sum_k |f_{xy}(a_k, b_k) - f_{xy}(\sigma_W^2(0), \sigma_W(0)(x^2 + y^2))| dx dy \\
&= O \left(\int_{\mathbb{R}^2} \sum_k |\nabla f^T(\sigma_W^2(0), \sigma_W(0)(x^2 + y^2)) \begin{bmatrix} -\sigma_W^2(k) \\ -2\sigma_W(k)xy \end{bmatrix}| dx dy \right) \\
&= O \left(\sum_k \sigma_W(k) + \sum_k \sigma_W^2(k) \right)
\end{aligned} \tag{2.8.15}$$

With $\{W_t\}$ satisfying (C2), $\sum_k \sigma_W(k)$ is absolutely convergent, then so is $\sum_k \sigma_W^2(k)$, the above quantity is bounded.

For (3), Since K' is a centered gaussian process, We have for $s < t$

$$E|K'(t) - K'(s)|^\alpha = \sqrt{\frac{2}{\pi}} \int_{x \in (0, \infty)} x^\alpha \frac{1}{\sigma_{s,t}} e^{-\frac{x^2}{2\sigma_{s,t}^2}} dx = C_3 \sigma_{s,t}^{\alpha-1} \tag{2.8.16}$$

Where $\sigma_{s,t}^2 = \sum_{k \in \mathbb{Z}} E(I(U_0 \in (s, t]) - (t-s))(I(U_k \in (s, t]) - (t-s)) \dots (**)$. It suffice to show that $\exists \alpha > 1, \beta, C > 0$ s.t. $\sigma_{s,t}^2 \leq C(t-s)^{\frac{2(1+\beta)}{\alpha-1}}$. Note that

$$(**) \leq \int_{(x,y) \in A_{s,t}^2} \sum_k |f_W^{(k)}(x,y) - f_W(x)f_W(y)| dx dy \tag{2.8.17}$$

Where $A_{s,t} \subseteq \mathbb{R}$ is such that $\int_{A_{s,t}} f_W(x) dx = t-s$. With same Taylor expansion arguments, there exists positive constants C_i such that the right hand side of equation (2.8.17) is bounded by $C_1 E(W^2 I(W \in A_{s,t})) + C_2 E(|W| I(W \in A_{s,t})) + C_3 E(I(W \in A_{s,t}))$, where W is a normal random variable with density f_W . Since in the calculation of the expectations the dominating term is the exponential decay, $\forall \varepsilon > 0, E(W^2 I(W \in A_{s,t})) = o((t-s)^{1-\varepsilon})$, and so is $E(|W| I(W \in A_{s,t}))$. Thus $\exists \varepsilon_0 \in (0, 1), (**) \leq C(t-s)^{\varepsilon_0}$. The Kolmogorov continuity condition is satisfied. By Borell's inequality (Chapter 2.1, Adler [1990]), the tail probability of the supremum is bounded by the tail probability of the Gaussian distribution with variance $\sup_{t \in [0,1]} \Gamma'(t, t) < \infty$. Thus we have

proved the desired result.

□

Proof of Lemma 4.3. By calculations in Theorem 1.2 in Li and Racine [2006],

$$E(\hat{F}_h(x)) - F(x) = \frac{c_2}{2} h^2 F^{(2)}(x) + o(h^2). \quad (2.8.18)$$

where $c_2 = \int x^2 k(x) dx < \infty$. Thus with $F^{(2)}(x)$ bounded and $h = o(n^{-1/4})$,

$$\sup_{x \in \mathbb{R}} |E(\hat{F}_h(x)) - F(x)| = o(n^{-1/2}). \quad (2.8.19)$$

Also by (A7) we have

$$\sqrt{n}(\hat{F}_h(x) - E(\hat{F}_h(x))) = G_h(x) + o_p(1) \quad (2.8.20)$$

and $G_h(x)$ is a tight centered Gaussian process with covariance

$$\Gamma_{G_h}(x, x') = \sum_{k \in \mathbb{Z}} E \left(K\left(\frac{x - Y_0}{h}\right) - E(K(\frac{x - Y_0}{h})) \right) \left(K\left(\frac{x' - Y_k}{h}\right) - E(K(\frac{x' - Y_k}{h})) \right) \rightarrow \Gamma(x, x')$$

as $h(n) \rightarrow 0$, $\forall x, x' \in \mathbb{R}$, where Γ is given in equation (2.8.2). As shown in Lemma 4.2, the limit Gaussian process is tight. This is sufficient for

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_h(x) - E(\hat{F}_h(x)) \right| = O_p\left(\frac{1}{\sqrt{n}}\right). \quad (2.8.21)$$

Together with equation (2.8.19) we have $\sup_{x \in \mathbb{R}} |\hat{F}_h(x) - F(x)| = O_p(\frac{1}{\sqrt{n}})$. Therefore

$$\sup_{t \in [n]} |\hat{U}_t^{(h)} - U_t| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

□

Proof of Theorem 4.4. Note that

$$\sum_{k=0}^{\lfloor c_{\kappa} l \rfloor} |\hat{\sigma}(k) - \sigma(k)| \leq \underbrace{\sum_{k=0}^{\lfloor c_{\kappa} l \rfloor} |\hat{\sigma}(k) - \hat{\sigma}(k)|}_{(1)} + \underbrace{\sum_{k=0}^{\lfloor c_{\kappa} l \rfloor} |\hat{\sigma}(k) - \sigma(k)|}_{(2)} \quad (2.8.22)$$

First of all, for (2), by Lemma 3.4, $\{Z_t\}$ is (C2) with $A > 1$ and therefore satisfies (C0). Then by Theorem 2.8.3, with $l = o(n^{\frac{p-2}{p}})$, (2) $\xrightarrow{P} 0$.

For (1), let $\tilde{Z}_t = \tilde{\Phi}^{-1}(U_t)$ and $\tilde{\tilde{Z}}_t = \tilde{\Phi}^{-1}(\hat{U}_t)$, where $\tilde{\Phi}$ is defined in Section 2. Then $\tilde{\Phi}^{-1}$ is a function bounded by c and $-c$.

Now consider:

$$\begin{aligned} \sum_{k=0}^{\lfloor c_{\kappa} l \rfloor} |\hat{\sigma}_Z(k) - \hat{\sigma}(k)| &= \sum_{k=0}^{\lfloor c_{\kappa} l \rfloor} \frac{1}{n} \left| \sum_{t=1}^{n-k} (Z_t Z_{t+k} - \tilde{\tilde{Z}}_t \tilde{\tilde{Z}}_{t+k}) \right| \\ &= \sum_{k=0}^{\lfloor c_{\kappa} l \rfloor} \frac{1}{n} \left| \sum_{t=1}^{n-k} (Z_t Z_{t+k} - \tilde{Z}_t \tilde{Z}_{t+k} + \tilde{Z}_t \tilde{Z}_{t+k} - \tilde{\tilde{Z}}_t \tilde{\tilde{Z}}_{t+k}) \right| \\ &\leq \sum_{k=0}^{\lfloor c_{\kappa} l \rfloor} \frac{1}{n} \sum_{t=1}^{n-k} |Z_t Z_{t+k} - \tilde{Z}_t \tilde{Z}_{t+k}| \quad (T_1) \\ &\quad + \sum_{k=0}^{\lfloor c_{\kappa} l \rfloor} \frac{1}{n} \sum_{t=1}^{n-k} |\tilde{Z}_t \tilde{Z}_{t+k} - \tilde{\tilde{Z}}_t \tilde{\tilde{Z}}_{t+k}| \quad (T_2) \end{aligned} \quad (2.8.23)$$

Let $X \lesssim Y$ denote $\exists c > 0, X \leq cY$.

For T_2 ,

$$\begin{aligned}
T_2 &= \sum_{k=0}^{\lfloor c\kappa l \rfloor} \left| \frac{1}{n} \sum_{t=1}^{n-k} \tilde{Z}_t \tilde{Z}_{t+k} - \tilde{Z}_t \tilde{\tilde{Z}}_{t+k} + \tilde{Z}_t \tilde{\tilde{Z}}_{t+k} - \tilde{\tilde{Z}}_t \tilde{\tilde{Z}}_{t+k} \right| \\
&= \sum_{k=0}^{\lfloor c\kappa l \rfloor} \left| \frac{1}{n} \sum_{t=1}^{n-k} \tilde{Z}_t \left(\tilde{Z}_{t+k} - \tilde{\tilde{Z}}_{t+k} \right) + \tilde{\tilde{Z}}_{t+k} \left(\tilde{Z}_t - \tilde{\tilde{Z}}_t \right) \right| \\
&\lesssim \sum_{k=0}^{\lfloor c\kappa l \rfloor} \frac{c}{\phi(c)} \sup_{t \in [n]} |U_t - \hat{U}_t| \\
&= O_p(lce^{\frac{c^2}{2}} n^{-1/2}).
\end{aligned} \tag{2.8.24}$$

Choose $c(n) \rightarrow \infty$ such that $lce^{\frac{c^2}{2}} = o(\sqrt{n})$ makes $T_2 \rightarrow 0$ in probability.

Meanwhile,

$$\begin{aligned}
ET_1 &\leq \sum_{k=0}^{\lfloor c\kappa l \rfloor} \frac{1}{n} E \left| \sum_{t=1}^{n-k} Z_t Z_{t+k} - \tilde{Z}_t \tilde{Z}_{t+k} \right| \\
&= \sum_{k=0}^{\lfloor c\kappa l \rfloor} \frac{n-k}{n} E |Z_t Z_{t+k} - Z_t \tilde{Z}_{t+k} + Z_t \tilde{Z}_{t+k} - \tilde{Z}_t \tilde{Z}_{t+k}| \\
&\lesssim lE |Z_t (Z_{t+k} - \tilde{Z}_{t+k})| \\
&\leq l \|Z_t\|_2 \|Z_{t+k} - \tilde{Z}_{t+k}\|_2 \\
&= O(l(ce^{-c^2/2})^{1/2})
\end{aligned} \tag{2.8.25}$$

with specific calculations for $E(Z_{t+k} - \tilde{Z}_{t+k})^2$ using $Z_{t+k} \sim \mathcal{N}(0, 1)$. Thus by choosing $l, c \rightarrow \infty$ such that $lce^{c^2/2} = o(\sqrt{n})$ and $lc^{1/2}e^{-c^2/4} = o(1)$ we have both $T_1 \xrightarrow{P} 0$ and $T_2 \xrightarrow{P} 0$.

Equation (4.9) is a direct consequence of equation (4.8). Note that both $\hat{\hat{\Sigma}}_n$ and Σ_n are Toeplitz, then:

$$\begin{aligned}
\left\| \hat{\hat{\Sigma}}_n - \Sigma_n \right\|_{op} &\leq \sqrt{\left\| \hat{\hat{\Sigma}}_n - \Sigma_n \right\|_1 \left\| \hat{\hat{\Sigma}}_n - \Sigma_n \right\|_\infty} = \sum_{k \in \mathbb{Z}} |\hat{\hat{\sigma}}(k) - \sigma(k)| \\
&= \sum_{|k| \leq [c_\kappa l]} |\hat{\hat{\sigma}}(k) - \sigma(k)| + \sum_{|k| > [c_\kappa l]} |\sigma(k)|
\end{aligned} \tag{2.8.26}$$

Where the 1-norm and ∞ -norm of a $n \times n$ matrix A is defined by $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{i,j}|$ and $\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{i,j}|$ respectively. If A is Toeplitz, then $\|A\|_1 = \|A\|_\infty$ by definition, thus the first equality holds.

The First term of above is shown to converge to 0 in probability. For second term, note that $\sigma(k)$ is absolutely summable by (C2) condition on Z_t , then with $l(n) \rightarrow \infty$, second term converge to 0. Since the spectral density of W_t is both bounded and bounded away from 0, with Z_t a linear transform from W_t , same holds for Z_t . Then by Theorem 2 of McMurtry and Politis [2010], $\left\| \hat{\hat{\Sigma}}_n^{-1} - \Sigma_n^{-1} \right\|_{op} \xrightarrow{P} 0$. $\square$

Proof of Lemma 4.5. Before proving the assertion, some preliminary lemmas are required and listed below:

Lemma 2.8.4. *Given that $\hat{F}$ is uniformly consistent for F_Y and F_Y is continuous and strictly increasing, then $\forall p \in (0, 1)$, $\hat{F}^{-1}(p) \xrightarrow{P} F_Y^{-1}(p)$.*

Proof. See Lemma 1.2.1 in Politis et al. [1999] $\square$

Lemma 2.8.5. *Let $U \sim \text{Unif}(0, 1)$. Then $\hat{Y} = \hat{F}^{-1}(U) \sim \hat{F}$ and $Y = F_Y^{-1}(U) \sim F_Y$, With $d_\infty(\hat{F}, F_Y) = \sup_x |\hat{F}(x) - F_Y(x)| \xrightarrow{P} 0$, $\hat{Y} \xrightarrow{d} Y$ in probability.*

Proof. For the first part, see Angus [1994]. The second part is straightforward. $\square$

By Lemma 2.8.5 and Theorem 4.4, also by continuous mapping theorem, (1) in Lemma 4.5 is straightforward. We focus on proving (2).

Note that for both empirical CDF and kernel CDF with appropriate bandwidth choice, $\sup_x |\hat{F}(x) - F_Y(x)| \xrightarrow{P} 0$ with relative assumptions. Then by continuous mapping theorem, $\forall k \in \mathbb{N}$, $\tilde{\Phi}^{-1}(\hat{F}_Y(Y_k))$ converges in probability to $Z_k = \Phi^{-1}(F(Y_k))$. Specifically, $\forall d \leq n$, $\hat{\underline{Z}}_d \xrightarrow{P} \underline{Z}_d$. Then with assumption (A9)

$$\hat{\underline{\xi}}_d \xrightarrow{P} \underline{\xi}_d \sim \mathcal{N}(0, I_d) \quad (2.8.27)$$

Thus by Lemma 3.4 of Kallenberg [1997], sequence $\hat{\xi} = (\hat{\xi}_1, \dots)$ converges in probability to $\xi = (\xi_1, \dots)$ with respect to the metric $\rho(\hat{\xi}, \xi) = \sum_{k=1}^{\infty} 2^{-k} |\hat{\xi}_k - \xi_k|$, where $\xi_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Let $\bar{F}_{\hat{\xi}}, \bar{F}_{\xi}$ be the empirical CDF of $\hat{\xi}$ and ξ respectively, i.e. $\bar{F}_{\hat{\xi}}(x) = \frac{\sum_{k=1}^n I(\hat{\xi}_k \leq x)}{n}$ and $L_{n,x}(\xi) = \bar{F}_{\xi}(x)$, $L_x(\xi) = \lim_{n \rightarrow \infty} L_{n,x}(\xi)$. By using continuous mapping theorem (Theorem 18.11, Van Der Vaart [1998]) on $L_{n,x}(\cdot)$ and $L_x(\cdot)$, we have for x a.e., $\bar{F}_{\hat{\xi}}(x) - \bar{F}_{\xi}(x) \xrightarrow{P} 0$. Since $\sup_x |\bar{F}_{\xi}(x) - \Phi(x)| \rightarrow 0, a.s.$, along with continuity of Φ and by Pólya's theorem, $\sup_x |\bar{F}_{\hat{\xi}}(x) - \Phi(x)| \xrightarrow{P} 0$. Since for any finite dimensional vector $(\xi_{t_1}^*, \dots, \xi_{t_d}^*)$, each of the elements are i.i.d. sampled from $\bar{F}_{\hat{\xi}}$, then $(\xi_{t_1}^*, \dots, \xi_{t_d}^*)$ converge in distribution (in bootstrap world) to a d-vector with i.i.d. standard normals in probability. By Theorem 3.29 of Kallenberg [1997], assertion in equation (4.13) holds. The remaining proof goes back to the setting in (1). $\square$

Proof of Theorem 4.6. The following several key components constitutes the proof:

1. Finite dimensional convergence in distribution shown in Lemma 4.5.
2. The bootstrap sample Y_t^* are $\lfloor c_K l \rfloor$ -dependent with respect to P^* due to *i.i.d.* sampling of the ξ_t^* s and bandedness of $\hat{\Sigma}_n$. Also $\exists \delta > 0, \forall t, E^*(Y_t^*)^{2+\delta} = O_p(1)$ from $L^p - m$ approximable assumption of Y_t (with $p > 2$) which implies uniform integrability of $(Y_t^*)^2$ (with respect to n). Also (C2) holds for Y_t^* in probability.

3. Central limit theorem holds by Lemma 4.1 for $\bar{Y}_n^*$ in probability.

The consequences of 1 & 2 above are two fold. One, the long run variance of Y_t^* can be written as $(\sigma_\infty^*)^2 = \sum_{k=-\lfloor c_{\kappa l} \rfloor}^{\lfloor c_{\kappa l} \rfloor} \gamma_Y^*(k)$ which is absolutely convergent in probability as $n \rightarrow \infty$ by Lemma 2.5.1. Second, by similar arguments to the proof of Theorem 3.3 in Bühlmann [1997], we have $\forall M > 0, \sum_{k=-M}^M |\gamma_Y^*(k) - \gamma_Y(k)| = o_p(1)$. Note that $\sum_{k \in \mathbb{Z}} \gamma_Y(k)$ is also absolutely convergent. All together, they imply $\sum_{k=-\lfloor c_{\kappa l} \rfloor}^{\lfloor c_{\kappa l} \rfloor} |\gamma_Y^*(k) - \gamma_Y(k)| = o_p(1)$, and as a consequence, $(\sigma_\infty^*)^2 \xrightarrow{P} \sigma_\infty^2$.

Since CLT holds for both $\bar{Y}_n$ and $\bar{Y}_n^*$ (in probability), and the limiting variance of $\bar{Y}_n^*$ is consistent to that of $\bar{Y}_n$, we finished proving Theorem 4.6.

□

Proof of Theorem 4.10. The kernel smoothed spectral density estimator

$$\hat{f}_{sp.d}(\omega) = \sum_{j \leq n} \tilde{\kappa}_h(\omega - \omega_j) I_n(\omega_j) \quad (2.8.28)$$

can be rewritten as the lag-window spectral density estimator with asymptotically negligible error (even after $\sqrt{nh}$ - scaling), i.e.:

$$\hat{f}_{sp.d}(\omega) = \frac{1}{2\pi} \sum_{k \in \mathbb{Z}} K_h(k) \hat{\gamma}(k) e^{-i\omega k} + O\left(\frac{1}{n}\right). \quad (2.8.29)$$

with $\hat{\gamma}(k) = \frac{1}{n} \sum_{t=1}^n (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})$ and $K_h(k) = \sum_{j=1}^n \tilde{\kappa}_h(\omega - \omega_j) e^{i(\omega_j - \omega)k}$. See Politis and Romano [1992]. Under assumption (A12) and Lemma 3.3, also with previous results holding, $\{Y_t\}$ and $\{Y_t^*\}$ satisfy (C0) with $p = 4$; with additional assumptions (A13) and (A14) holding, conditions in Theorem 2, Liu and Wu [2010] are satisfied. Therefore asymptotic normality in the sense of equation (4.23) holds for both $\hat{f}_{sp.d}(\omega)$ and $\hat{f}_{sp.d}^*(\omega)$ with limiting variance $\sigma_\omega^2 = (\eta_\omega + 1) f^2(\omega) \int \tilde{\kappa}^2(u) du$ and $(\sigma_\omega^*)^2 = (\eta_\omega + 1) (f^*(\omega))^2 \int \tilde{\kappa}^2(u) du$ respectively. Since $|f_{sp.d}(\omega) - f_{sp.d}^*(\omega)| \leq \frac{1}{2\pi} \sum_{k \in \mathbb{Z}} |(\gamma(k) - \gamma_Y^*(k)) e^{i\omega k}| \leq \sum_{k \in \mathbb{Z}} |\gamma(k) - \gamma_Y^*(k)| \xrightarrow{P} 0$ and both $f_{sp.d}(\omega)$ and

$f_{sp,d}^*(\omega)$ are bounded and bounded a.s. , $|f_{sp,d}^2(\omega) - (f_{sp,d}^*(\omega))^2| \xrightarrow{P} 0$, which implies $(\sigma_\omega^*)^2 \xrightarrow{P} \sigma_\omega^2$. $\square$

Proof of Theorem 5.1. We prove it for $\hat{F} = \bar{F}$. The case for the kernel estimator $\hat{F}_h$ is similar with relative assumptions assumed. We mainly show that the distribution of bootstrap predictive root converges in probability to the true distribution in a conditional sense. i.e., $Y_{n+1}^* - \hat{Y}_{n+1}^* \xrightarrow{d^*} Y_{n+1} - \hat{Y}_{n+1}$ in probability, conditioning on past values $\underline{Y}_n$.

Firstly, conditioning on $\underline{Y}_n$, $Y_{n+1}^* \xrightarrow{d^*} Y_{n+1}$ in probability:

$$Y_{n+1}^* = \bar{F}_Y^{-1}(\Phi(Z_{n+1}^*)) = (\bar{F}_Y^{-1}(\Phi(Z_{n+1}^*)) - F_Y^{-1}(\Phi(Z_{n+1}^*))) + F_Y^{-1}(\Phi(Z_{n+1}^*)) \quad (2.8.30)$$

We show that first term converges to 0 in probability and the distribution of second term converges to that of Y_{n+1} , in probability. Let $U = \Phi(Z_{n+1}^*)$. Then $\bar{F}_Y^{-1}(U) = Y_{(i)}$, where $\frac{i-1}{n} < U \leq \frac{i}{n}$. It is equivalent to show $Y_{(i)} - F_Y^{-1}(U) \xrightarrow{P} 0$. Consider $F_Y(Y_{(i)}) - U = (F_Y(Y_{(i)}) - \bar{F}_Y(Y_{(i)})) + (\bar{F}_Y(Y_{(i)}) - U)$. The first term is $O_p(\frac{1}{\sqrt{n}})$ by Lemma 4.2. For the second term, $\bar{F}_Y(Y_{(i)}) - U = \frac{i}{n} - U$, where $U \in (\frac{i-1}{n}, \frac{i}{n}]$. Hence the second term goes to 0. Therefore $F_Y(Y_{(i)}) - U \rightarrow 0$ in probability. Since F_Y^{-1} is continuous, $Y_{(i)} - F_Y^{-1}(U) \xrightarrow{P} 0$.

To show $F_Y^{-1}(\Phi(Z_{n+1}^*)) \xrightarrow{d^*} Y_{n+1} = F_Y^{-1}(\Phi(Z_{n+1}))$, we need to show conditioning on $\underline{Y}_n$, $Z_{n+1}^* \xrightarrow{d^*} Z_{n+1}$ in probability. For MF bootstrap, let $L_n(\cdot, \cdot) : (\underline{Y}_n, \xi_{n+1}^*) \rightarrow Z_{n+1}^*$ with its theoretical analogue $L(\cdot, \cdot) : (\underline{Y}_n, \xi_{n+1}) \rightarrow Z_{n+1}$. The formula of $L_n(\cdot, \cdot)$ is estimated from $\underline{Y}_n$ and therefore depends on the past values $\underline{Y}_n$; While $L(\cdot, \cdot)$ is the theoretical data-generating mechanism that links the past values $\underline{Y}_n$ and innovation ξ_{n+1} so it does not depend on $\underline{Y}_n$. By results in Lemma 4.5 and assumption (A9), $L_n(x, y)$ converges to $L(x, y)$ in probability; Also the distribution of ξ_{n+1}^* converges to Φ for MF bootstrap. As a result of continuous mapping theorem, $L_n(\underline{Y}_n, \xi_{n+1}^*) \xrightarrow{d^*} L(\underline{Y}_n, \xi_{n+1})$ in probability.

For LMF bootstrap, we only need to show the normal parameters in step 2 of **Algorithm**

3 converge to their theoretical analogue, in probability. As an example, we show $\hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1} \tilde{Z}_n \rightarrow \Sigma_{21} \Sigma_{11}^{-1} Z_n$ in probability. Let $c_n = \Sigma_{21} \Sigma_{11}^{-1}$ & $\hat{c}_n = \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1}$. we have

$$\begin{aligned} |\hat{c}_n \tilde{Z}_n - c_n Z_n| &\leq \sum_{i=1}^n |\hat{c}_{n,i} \tilde{Z}_i - c_{n,i} Z_i| \\ &\leq \sum_{i=1}^n |\hat{c}_{n,i}| |\tilde{Z}_i - Z_i| + \left| \sum_{i=1}^n (c_{n,i} - \hat{c}_{n,i}) Z_i \right| \end{aligned} \quad (2.8.31)$$

The first term is bounded by $\left(\sum_{i=1}^n |\hat{c}_{n,i}| \right) \sup_i |\tilde{Z}_i - Z_i|$, which is $o_p(1)$ by continuous mapping theorem and $\sum_{i=1}^n |\hat{c}_{n,i}|$ is bounded in probability for large n . The second term has $\mathcal{N}(0, (c_n - \hat{c}_n) \Sigma_n (c_n - \hat{c}_n)^T)$ distribution, where the variance is bounded by $\lambda_{\Sigma_n}^{max} \|c_n - \hat{c}_n\|_2^2$. Thus if $\|c_n - \hat{c}_n\|_2 \xrightarrow{P} 0$, then second term converge to 0 in probability.

$$\|c_n - \hat{c}_n\|_2 \leq \left\| \hat{\Sigma}_n^{-1} - \Sigma_n^{-1} \right\|_{op} \left\| \hat{\Sigma}_{21} \right\|_2 + \left\| \hat{\Sigma}_{21} - \Sigma_{21} \right\|_2 \left\| \Sigma_n^{-1} \right\|_{op} \quad (2.8.32)$$

Since $\left\| \hat{\Sigma}_n^{-1} - \Sigma_n^{-1} \right\|_{op} \xrightarrow{P} 0$, by previous results, $\left\| \hat{\Sigma}_{21} - \Sigma_{21} \right\|_2 \xrightarrow{P} 0$. Then $\|c_n - \hat{c}_n\|_2 \xrightarrow{P} 0$. The proof for consistency of the second parameter follows a similar fashion.

What remains to be shown is conditioning on $\underline{Y}_n$,

$$\hat{Y}_{n+1}^* \xrightarrow{P^*} \hat{Y}_{n+1}. \quad (2.8.33)$$

(In probability), where $\hat{Y}_{n+1}$ is the self-chosen predictor. In L^2 optimal sense, $\hat{Y}_{n+1} = E(Y_{n+1} | \underline{Y}_n) = E(F_Y^{-1}(\Phi(Z_{n+1})) | \underline{Y}_n)$. Therefore it is sufficient to show that:

$$E^*((\bar{F}^*)^{-1}(\Phi(Z_{n+1})) | \underline{Y}_n) \xrightarrow{P^*} E(F_Y^{-1}(\Phi(Z_{n+1})) | \underline{Y}_n) \quad (2.8.34)$$

Where the expectation is taken with respect to $(Z_{n+1} | \underline{Y}_n)^*$ and $(Z_{n+1} | \underline{Y}_n)$ as defined in **Algorithm 2** and **3**. Since directly calculating the expectation in our setup is impossible, we use the same

approach as previously used in the proof of Theorem 4.6 by showing:

1. $(Z_{n+1}|\underline{Y}_n)^* \xrightarrow{d^*} (Z_{n+1}|\underline{Y}_n)$.
2. Consistency of $(\bar{F}^*)^{-1}$ to F_Y^{-1} .
3. Uniform integrability of $(\bar{F}^*)^{-1}(\Phi(Z_{n+1}))$ and $F_Y^{-1}(\Phi(Z_{n+1}))$ conditioning on $\underline{Y}_n$.

First of all, we need to show that $(Z_{n+1}|\underline{Y}_n)^* \xrightarrow{d^*} Z_{n+1}|\underline{Y}_n$ in probability. In LMF bootstrap setup as in **Algorithm 3**, this can be checked by showing convergence of the bootstrapped autocovariance matrix $\hat{\Sigma}_n^*$ to Σ_n in probability. Since $\{Y_t^*\}_{t=1}^n$ are m -dependent and have finite p^{th} , ($p > 2$) moment in probability, Theorem 2.8.3 applies to $\{Y_t^*\}_{t=1}^n$ and thus

$$\left\| \hat{\Sigma}_n^* - \Sigma_n^* \right\|_{op} \xrightarrow{P^*} 0 \quad (2.8.35)$$

where Σ_n^* is the autocovariance matrix of the bootstrap samples $\{Y_t^*\}$. Also by Lemma 4.5 and uniform integrability of $(Y_t^*)^2$, We have $\sum_{k \in \mathbb{Z}} |\sigma^*(k) - \sigma(k)| \xrightarrow{P^*} 0$. This implies $\|\Sigma_n^* - \Sigma_n\|_{op} \xrightarrow{P} 0$. Therefore by triangle inequality,

$$\left\| \hat{\Sigma}_n^* - \Sigma_n \right\|_{op} \xrightarrow{P^*} 0. \quad (2.8.36)$$

In MF bootstrap setup as in **Algorithm 2**, we need to show $\forall z \in \mathbb{R}$, $(\hat{F}_Z^{(n+1)})^*(z) \xrightarrow{P^*} \hat{F}_Z^{(n+1)}(z)$. This can also be shown through equation (2.5.11) in Lemma 2.5.5 and a slightly stronger result than equation (2.8.36) above, i.e.: we assume $(\log n)^2 \left\| \hat{\Sigma}_n^* - \Sigma_n \right\|_{op} \xrightarrow{P^*} 0$.

Since $\bar{F}_Y^*$ is the empirical CDF of $\{Y_t^*\}$ for which the CDF F_Y^* satisfies $\sup_{y \in \mathbb{R}} |\bar{F}_Y^*(y) - F_Y^*(y)| \xrightarrow{P} 0$ by Lemma 4.5, also along with m -dependence of $\{Y_t^*\}$, $\sup_{y \in \mathbb{R}} |\bar{F}_Y^*(y) - F_Y^*(y)| \rightarrow 0$ almost surely with respect to P^* . Combining these two we have $\sup_{y \in \mathbb{R}} |\bar{F}_Y^*(y) - F_Y(y)| \rightarrow 0$ in probability. By Lemma 2.8.4 and continuous mapping theorem used in Lemma 4.5, $(\bar{F}_Y^*)^{-1}(\Phi(Z_{n+1}))$ with Z_{n+1} having the distribution of $(Z_{n+1}|\underline{Y}_n)^*$ converges in distribution to $F_Y^{-1}(\Phi(Z_{n+1}))$ with Z_{n+1} following the conditional distribution of $Z_{n+1}|\underline{Y}_n$, in probability. Also by finite p^{th} moment of Y_t we have uniform integrability of $(\bar{F}_Y^*)^{-1}(\Phi(Z_{n+1}))$ (in probability) and $F_Y^{-1}(\Phi(Z_{n+1}))$.

Thus equation (2.8.34) holds. For the L^1 optimal predictor mentioned in Section 1, proving $\hat{Y}_{n+1}^* \xrightarrow{P^*} \hat{Y}_{n+1}$ follows in a similar fashion but may require additional assumptions; for example, continuity and strict monotonicity of the conditional distribution $Y_{n+1}|\underline{Y}_n$ is necessary to ensure consistency of $\hat{Y}_{n+1}$, which is the conditional sample median.

Summing up the previous results and by Slutsky's theorem, the bootstrap predictive root converges in distribution to the true predictive root(in probability). Since the distribution of the true predictive root is continuous, we have consistency of the bootstrap quantiles. $\square$

2.9 Acknowledgement

Chapter 2, in full, is a reprint of the materials in the paper “Model-free Bootstrap for a General Class of Stationary Time Series”, Yiren Wang, Dimitris N. Politis. It has been accepted for publication in Bernoulli. The dissertation author is the primary investigator and author of this material.

Chapter 3

Model-free Bootstrap Prediction Regions for Multivariate Time Series

3.1 Abstract

In Das and Politis [2021], a model-free bootstrap(MFB) paradigm was proposed for generating prediction intervals of univariate, locally stationary time series. Theoretical guarantees for this algorithm was resolved in Wang and Politis [2021b] under stationarity and weak dependence conditions. Following this line of work, here we extend MFB for predictive inference under a multivariate time series setup. We describe two algorithms, the first one works for a particular class of time series under any fixed dimension d ; the second one works for a more generalized class of time series under low-dimensional setting. We justify our procedures through theoretical validity and simulation performance.

3.2 Introduction

Time series forecasting(prediction) is widely applicable in many different fields where the prediction is carried out by conditioning on previous observations. Traditional approaches for time series prediction often rely on a parametric model assumption that can describe the explicit dependence relations of the data. The space of viable time series models that have been studied in the past is huge, therefore the search of good models that can describe the data requires both good understanding of domain knowledge as well as statistical expertise in order to execute model fitting and goodness-of-fit tests.

In a recent monograph, Politis [2015] proposed the idea of *model-free* prediction in the context of regression and time series problems. The idea behind it is to find a one-to-one transform H that takes the original data $\{X_t\}_{t=1}^n$ into a new sequence of data that are i.i.d.(independent and identically distributed), then the (inverse) transform along with resampling in the i.i.d. world can be combined to construct both valid point predictors and prediction intervals. Following this, Das and Politis [2021] proposed a model-free bootstrap(MFB) algorithm for generating prediction intervals for locally stationary time series that outperforms model-based approach. This algorithm was further studied in Wang and Politis [2021b] under stationary weakly-dependent setting, where bootstrap validity was proved for both prediction interval and confidence intervals for linear statistics and spectral density. In this paper, we extend the model-free bootstrap algorithm of Wang and Politis [2021b] for prediction regions of multivariate time series. We demonstrate the effectiveness of the algorithm by both showing theoretical bootstrap validity and also numerical simulations.

3.3 Description of algorithm

3.3.1 Models of interest

Let $\underline{Y}_t = (Y_{1,t}, \dots, Y_{d,t})^\top$ represent a d -dimensional time series. We consider $\underline{Y}_t$ to be from one of the following models:

Model 1: Let $\{f_i : \mathbb{R} \rightarrow \mathbb{R}\}_{i=1}^n$ be strictly monotone, continuous functions. $\underline{W}_t \in \mathbb{R}^d$ is a purely non-deterministic, strictly stationary Gaussian process. $\underline{Y}_t$ satisfies

$$Y_{i,t} = f_i(W_{i,t}), \quad i = 1, \dots, d. \quad (3.3.1)$$

Model 3.3.1 is nothing more than a multivariate analog of the class of time series models studied in Wang and Politis [2021b], see Equation (2.4.1) in Chapter 2. Interestingly, it also has connection with copula models for multivariate time series, which we will investigate later. Let us introduce a second class of model:

Model 2: Let $f_i(\cdot; \underline{Y}_{1:i-1,t}) : \mathbb{R} \rightarrow \mathbb{R}$ be strictly monotone(increasing) functions whose parameters also depend on the previous $i - 1$ dimensional entries of $\underline{Y}_t$: $\underline{Y}_{1:i-1,t}$, with the exception of $i = 1$, where f_1 is a deterministic function. The time series $\underline{Y}_t$ are defined sequentially in dimensional order in the following manner:

$$\begin{aligned} Y_{1,t} &= f_1(W_{1,t}); \\ Y_{i,t} &= f_i(W_{i,t}; \underline{Y}_{1:i-1,t}), i \geq 2. \end{aligned} \quad (3.3.2)$$

Model 2 is more complicated than model 1 in that we allow for the parameters of subsequent transfer functions $f_i, i \geq 2$ to be dependent on previous entries of $\underline{Y}_t$. Model 1 can be recovered by setting $f_i(\cdot; \underline{Y}_{1:i-1,t}) = f_i$.

Because of the monotonicity of transfer functions, the map $\underline{W}_t \rightarrow \underline{Y}_t$ is invertible for both models, thus the information set generated by $\underline{Y}_t$ is equivalent to the one generated by $\underline{W}_t$, i.e.,

$$\mathcal{F}_t = \sigma(\underline{Y}_s, s \leq t) = \sigma(\underline{W}_s, s \leq t).$$

3.3.2 Connection to copula-based models

An active line of research for forecasting multivariate time series involves utilizing a copula representation. By Skyler's theorem, an arbitrary time series $\underline{Y}_t$ can be fully described by the conditional distributions of $Y_{i,t}|\mathcal{F}_{t-1} \sim F_i(|\mathcal{F}_{t-1})$, and a conditional copula function $C(\cdot|\mathcal{F}_{t-1}) : [0, 1]^d \rightarrow \mathbb{R}$ such that $\forall \underline{y} = (y_1, \dots, y_d)^\top$,

$$F(\underline{y}|\mathcal{F}_{t-1}) = C(F_1(y_1|\mathcal{F}_{t-1}), \dots, F_d(y_d|\mathcal{F}_{t-1})|\mathcal{F}_{t-1}).$$

The copula representation offers an approach to separate the modeling of joint dependence into the modeling of marginal CDFs and the copula function. Note that in the econometrics literature, the conditional distribution of each dimension are often assumed to be fixed, i.e., $F_i(|\mathcal{F}_{t-1}) = F_i$, to guarantee effectiveness of valid statistical procedures, see Patton [2013]. One of the most famous copulas is the Gaussian copula: a centered d -dimensional Gaussian random vector $\underline{Z}$ with correlation matrix Σ has distribution function

$$\Phi_{\Sigma,d}(\underline{z}) = C_{\Sigma}(\underline{u}), \tag{3.3.3}$$

where $\underline{u} = (u_1, \dots, u_d)^\top$ with $u_i = \Phi(z_i)$; and $C_{\Sigma}(\underline{u}) = \Phi_{\Sigma,d}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))$ is the Gaussian copula with correlation matrix Σ . Generalizing this to the case of a stationary Gaussian processes denoted by $\underline{Z}_t$, since the conditional distribution of $\underline{Z}_t|\mathcal{F}_{t-1}$ is multivariate normal, it can be fully represented by the Gaussian copula (3.3.3) above, where $\Sigma = \Sigma_{t|t-1}$ will be the correlation matrix of the conditional normal distribution. Thus $C(|\mathcal{F}_{t-1}) = C_{\Sigma_{t|t-1}}$.

As is well known, the copula of a joint distribution stays unchanged under monotone function transformations of each marginal random variable. To put it under the setup of model

(3.3.1) and assume that $F_i(|\mathcal{F}_{t-1}) = F_i$,

$$F(\underline{y}|\mathcal{F}_{t-1}) = C_\Sigma(F_1(y_1), \dots, F_d(y_d)),$$

where C_Σ is the conditional Gaussian copula of $\underline{W}_t|\mathcal{F}_{t-1}$. Thus, the time series $\underline{Y}_t$ of model 3.3.1 inherits the same conditional copula as the Gaussian process $\underline{W}_t$. $\underline{Y}_t$ is also referred to as the Gaussian copula process, see Wilson and Ghahramani [2010]. However, this does not apply to the case of model 3.3.2, where the transfer functions' parameters depends on other indices of $\underline{Y}_t$. Therefore, model 3.3.2 can be more complicated in that the conditional copula function can go beyond the Gaussian copula.

Notably, the model-free bootstrap to be introduced in the next section shares the same spirit of the semi-parametric method (cf. Patton [2013] for more details) for copula process prediction. To elaborate, the model-free bootstrap under model (3.3.1) requires both estimation of the marginal CDFs via nonparametric method, as well as consistent estimation for the autocovariance structure of the underlying Gaussian process. Similarly, the semi-parametric approach in Patton [2013] assumes general distributions for the CDF which are estimated via nonparametric methods, while the copula function is assumed to belong to a particular parametric family(e.g, Gaussian), and then estimation for the copula is carried out parametrically.

3.3.3 The model-free bootstrap algorithm

The model-free bootstrap of Politis [2015] offers a general principle for conducting bootstrap resampling procedure for parameter inference and predictive inference under regression or time series setups. In the univariate time series setting, the algorithm relies on an invertible transform H_n between a length n time series data $\{X_t\}_{t=1}^n$, and a set of n i.i.d. random variables $\{\xi_t\}_{t=1}^n$, where H_n is sample size adaptive. For parameter inference, let θ be the parameter of interest which is estimated by some statistic $\hat{\theta}_n$. The model-free bootstrap is first performed in the

space of i.i.d. random variables to generate $\{\xi_t^*\}_{t=1}^n$, and then we use the H_n^{-1} to get bootstrap samples $\{X_t^*\}_{t=1}^n$. Then the distribution of $\hat{\theta}_n - \theta$ can be approximated by $\hat{\theta}_n^* - \hat{\theta}_n$, where $\hat{\theta}_n^*$ is calculated based on the bootstrap sample.

Predictive inference is in a more difficult situation, as valid predictions should be carried out conditioning on the entire observed series $\{X_t\}_{t=1}^n$. First of all, we need a notion called the predictive root, an analog to the pivot random variable $\hat{\theta}_n - \theta$ in previous setting, which is defined as

$$r_{n+1} = X_{n+1} - \hat{X}_{n+1},$$

where $X_{n+1} \sim F_{n+1|n} := F(\cdot | \mathcal{F}_n)$ is the next unobserved data, and $\hat{X}_{n+1}$ is a sample point predictor for the future observation, such that

$$\hat{X}_{n+1} = \arg \min_{x \in \mathbb{R}} \mathbb{E}_{X_{n+1} \sim \hat{F}_{n+1|n}} \mathcal{L}(x, X_{n+1}).$$

Here $\hat{F}_{n+1|n}$ is the estimated conditional CDF based on the estimated transform $\hat{H}_n$. Examples for the loss function $\mathcal{L}$ include the L^1 and L^2 loss minimizers, for which we have L^1/L^2 -optimal predictors, respectively. Essentially, $\hat{X}_{n+1}$ is the sample estimator for the actual 1-step ahead predictor:

$$\mathcal{P}(X_{n+1} | \mathcal{F}_n) = \arg \min_{x \in \mathbb{R}} \mathbb{E}_{X_{n+1} \sim F_{n+1|n}} \mathcal{L}(x, X_{n+1}).$$

By approximating the conditional distribution of r_{n+1} via bootstrap, we can construct a two-sided prediction interval of size $1 - \alpha$ around the point predictor as follows:

$$\left(\hat{X}_{n+1} + L_{\alpha/2}^*, \hat{X}_{n+1} + R_{\alpha/2}^* \right),$$

where $L_{\alpha/2}^*$ and $R_{\alpha/2}^*$ are the lower/higher $\alpha/2$ -quantiles for the distribution of r_{n+1}^* , the bootstrap version of r_{n+1} .

The foremost problem is how to construct the invertible transform H_n towards i.i.d.-ness.

As Politis [2015] points out, such an invertible transform always exists. However, H_n needs to be estimated based on data, and further structural assumptions both simplifies the estimation procedure and also guarantees certain level of efficiency required for the bootstrap to be valid. Das and Politis [2021] proposed the following H_n based on the probability integral transform(PIT) under the structural assumption that X_t is a monotone transform of a stationary Gaussian process:

1. Let F_X be the CDF of X_t ; Let $U_t = F_X(X_t)$.
2. Let $Z_t = \Phi^{-1}(U_t)$. Das and Politis [2021] showed that Z_t is a Gaussian process. Let $\underline{Z}_n = (Z_1, \dots, Z_n)^\top$
3. Let Ξ_n be the n -dimensional autocovariance matrix of Z_t , and $\Xi_n^{1/2}$ the upper Cholesky decomposition matrix of Ξ . Then $\underline{\xi}_n = \Xi_n^{-1/2} \underline{Z}_n$ consists of n i.i.d. standard normal random variables.

H_n^{-1} will map n i.i.d. normal random variables $\underline{\xi}_n$ back to $\{X_t\}_{t=1}^n$ in the following way:

1. Let $\underline{Z}_n = \Xi_n^{1/2} \underline{\xi}_n$;
2. let $U_t = \Phi(Z_t)$, $X_t = F_X^{-1}(U_t)$.

Wang and Politis [2021b] showed that under certain weakly dependence assumption of X_t , one can efficiently estimate the CDF F_X with a nonparametric estimator. By using an augmented version of Φ in the transform, Ξ_n and its inverse can also be efficiently estimated, thereby showing consistency of the estimated transforms $\hat{H}_n$ and $\hat{H}_n^{-1}$, and also validity of the model-free bootstrap.

To create the predictive root r_{n+1}^* , both X_{n+1} and $\hat{X}_{n+1}$ need to be resampled separately. To sample $\hat{X}_{n+1}$, with $\underline{\xi}_n^*$ being sampled from $\underline{\xi}_n$, we can use the above $\hat{H}_n^{-1}$ to get the bootstrap samples $\underline{X}_n^*$, which is used to re-estimate $\hat{H}_n$ to get $\hat{F}_{n+1|n}^*$. Then

$$\hat{X}_{n+1}^* = \arg \min_{x \in \mathbb{R}} \mathbb{E}_{X_{n+1} \sim \hat{F}_{n+1|n}^*} \mathcal{L}(x, X_{n+1}).$$

As for X_{n+1}^* , we first extend $\hat{H}_n^{-1}$ to $\hat{H}_{n+1}^{-1}$, which is then used to map the vector $(\underline{\xi}_n, \xi_{n+1}^*)$ back to $(\underline{X}_n, X_{n+1}^*)$. It is easy to see that by doing this, $X_{n+1}^* \sim \hat{F}_{n+1|n}$.

We next extend the MFB to multivariate time series under model 3.3.1 or 3.3.2 following a similar route as above. Specifically, we first transform $\underline{Y}_t$ to a centered multivariate Gaussian process $\underline{Z}_t$ using the PIT, and then further whiten it to get i.i.d. normal vectors by decorrelating $\underline{Z}_t$ with its covariance structure. The inverse of the above transforms are used to construct the transform bootstrap samples from the i.i.d. space back to the space of time series.

In order to whiten the Gaussian process $\underline{Z}_t$ which has $n \times d$ observations, we need the following technique adapted from Jentsch and Politis [2015]. The entire sequence $\mathbf{Z} = [\underline{Z}_1, \dots, \underline{Z}_n]$ can be flattened by stacking the observations into one row vector: $\underline{Z}_{dn} = \text{vec}(\mathbf{Z}) = [\underline{Z}_1^\top, \dots, \underline{Z}_n^\top]$. $\underline{Z}_{dn}$ is a multivariate normal vector, whose covariance matrix is symmetric block Toeplitz, and has the following form:

$$\Gamma_{dn} = \begin{bmatrix} \Gamma_0 & \Gamma_1 & \cdots & \Gamma_{n-1} \\ \Gamma_1^\top & \Gamma_0 & \cdots & \Gamma_{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \Gamma_{n-1}^\top & \Gamma_{n-2}^\top & \cdots & \Gamma_0 \end{bmatrix}, \quad (3.3.4)$$

where $\Gamma_h = \text{Cov}(\underline{Z}_0, \underline{Z}_h) = \mathbb{E}\underline{Z}_0\underline{Z}_h^\top$ is the $d \times d$ lag- h autocovariance matrix of $\underline{Z}_t$. Next, $\underline{Z}_{dn}$ can be whitened through left-multiplying $\Gamma_{dn}^{-1/2}$, which results in i.i.d. standard normal variables.

To consistently estimate Γ_{dn} , we require to use the flat-top estimator introduced in McMurry and Politis [2010]. The autocovariance matrix at lag $h > 0$ can be estimated through the usual estimator

$$\hat{\Gamma}_h = \frac{1}{n} \sum_{t=1}^{n-h} \underline{Z}_t \underline{Z}_{t+h}^\top.$$

While for $h < 0$ we can use $\hat{\Gamma}_h = \hat{\Gamma}_{|h|}^\top$. Let κ_l be the flat-top kernel with base function κ and

bandwidth parameter l , such that $\kappa_l(x) = \kappa(x/l)$. The new estimator is defined as

$$\hat{\Gamma}_{\kappa,l} = \left(\kappa_l(i-j) \hat{\Gamma}_{|i-j|} \right)_{1 \leq i,j \leq n}, \quad (3.3.5)$$

i.e., the autocovariance matrices at large lags are shrunk towards $\mathbf{0}$. It is well known that with appropriate rate of divergence for l as $n \rightarrow \infty$, the flat-top estimator is consistent to Γ_{dn} in operator norm:

$$\left\| \hat{\Gamma}_{\kappa,l} - \Gamma_{dn} \right\|_{op} \xrightarrow{P} 0.$$

A common tapering function is the following trapezoid function:

$$\kappa(x) = \begin{cases} 1, & |x| \leq 1 \\ 2 - |x|, & 1 < |x| \leq 2 \\ 0, & \text{otherwise} \end{cases} \quad (3.3.6)$$

We also extend the concept of (1-step ahead) predictive root to the multivariate scenario. Let $\underline{R}_{n+1} = \underline{Y}_{n+1} - \hat{\underline{Y}}_{n+1}$, where $\underline{Y}_{n+1}$ is the future observation conditioning on $\{\underline{Y}_t\}_{t=1}^n$; and $\hat{\underline{Y}}_{n+1}$ is the 1-step ahead predictor that satisfies

$$\hat{\underline{Y}}_{n+1} = \arg \min_{\underline{y} \in \mathbb{R}^d} \mathbb{E}_{\underline{Y}_{n+1} \sim \hat{F}_{n+1|n}} \mathcal{L}(\underline{y}, \underline{Y}_{n+1}), \quad (3.3.7)$$

where $\hat{F}_{n+1|n}$ is the estimated conditional CDF. Similarly, $\hat{\underline{Y}}_{n+1}$ is an estimator for the true 1-step ahead predictor

$$\mathcal{P}(\underline{Y}_{n+1} | \mathcal{F}_n) = \arg \min_{\underline{y} \in \mathbb{R}^d} \mathbb{E}_{\underline{Y}_{n+1} \sim F_{n+1|n}} \mathcal{L}(\underline{y}, \underline{Y}_{n+1}).$$

Algorithm 3.3.1. *Bootstrap algorithm for 1-step ahead prediction region under model 3.3.1.*

1. For each dimension i , estimate the marginal CDFs F_i via empirical or the nonparametric

CDF estimator, denoted by $\hat{F}_i$.

2. Let $\hat{U}_{i,t} = \hat{F}_i(Y_{i,t})$; let $\tilde{\Phi}_c^{-1}$ be the quantile function of a thresholded normal distribution (see Wang and Politis [2021b] for further details). Let $\tilde{Z}_{i,t} = \tilde{\Phi}_c^{-1}(U_{i,t})$ which are estimations for the destination Gaussian process. Estimate the covariance structure of $\tilde{\underline{Z}}_{dn}$, denoted by $\hat{\tilde{\Gamma}}_{dn}$, with the tapered covariance matrix estimator (3.3.5).
3. Flatten $\tilde{\underline{Z}} = [\tilde{\underline{Z}}_1, \dots, \tilde{\underline{Z}}_n]$ as $\tilde{\underline{Z}}_{dn} = \text{vec}(\tilde{\underline{Z}}) = [\tilde{\underline{Z}}_1^\top : \dots : \tilde{\underline{Z}}_n^\top]^\top$. Let $\hat{\underline{\xi}}_{dn} = \hat{\tilde{\Gamma}}_{dn}^{-1/2} \tilde{\underline{Z}}_{dn}$. Also, based on the estimations in previous steps, calculate $\hat{\underline{Y}}_{n+1}$ by equation (3.3.7).
4. (bootstrap)
 - (a) Let $\underline{\xi}_{dn}^*$ be a vector of i.i.d. random variables uniformly sampled with replacement from the entries of $\hat{\underline{\xi}}_{dn}$. Let $\underline{Z}_{dn}^* = \hat{\tilde{\Gamma}}_{dn}^{1/2} \underline{\xi}_{dn}^*$, based on which we get $\underline{Z}_t^*$, $t = 1, \dots, n$. Then $Y_{i,t}^* = \hat{F}_i^{-1}(\Phi(Z_{i,t}^*))$. Use $\mathbf{Y}_n^* = [\underline{Y}_1^*, \dots, \underline{Y}_n^*]$ to re-estimate the transforms above to get $\hat{F}_{n+1|n}^*$, then get $\hat{\underline{Y}}_{n+1}^*$ by equation (3.3.7).
 - (b) Let $\underline{\xi}_d^* = (\xi_{dn+1}^*, \dots, \xi_{d(n+1)}^*)$, with indices uniformly sampled from $\hat{\underline{\xi}}_{dn}$, and $\underline{\xi}_{d(n+1)}^* = (\hat{\underline{\xi}}_{dn}^\top, \underline{\xi}_d^*)^\top$. Then $\hat{\tilde{\Gamma}}_{d(n+1)}^{1/2} \underline{\xi}_{d(n+1)}^* := (\tilde{\underline{Z}}_{dn}^\top, \underline{Z}_{n+1}^*)$, where $\underline{Z}_{n+1}^*$ is the bootstrap sample for the 1-step ahead future observation for the Gaussian process $\underline{Z}_t$. Let $\underline{Y}_{i,n+1}^* = \hat{F}_i^{-1}(\Phi(Z_{i,n+1}^*))$ for $i = 1, \dots, d$.
 - (c) Let $\underline{R}_{n+1}^* = \underline{Y}_{n+1}^* - \hat{\underline{Y}}_{n+1}^*$.
5. Use step 4 to bootstrap $\underline{R}_{n+1}^*$ B times. Let $r_p^{*(b)} = \left\| \underline{R}_{n+1}^{*(b)} \right\|_p$, where $\|\cdot\|_p$ is the l_p -norm of a d -dimensional vector; and q_α^* the upper α -quantile for $\{r_p^{*(b)}\}_{b=1}^B$. The L^p -norm based $1 - \alpha$ prediction region for $\underline{Y}_{n+1}$ is

$$\{\underline{y} \in \mathbb{R}^d : \left\| \underline{y} - \hat{\underline{Y}}_{n+1} \right\|_p \leq q_\alpha^*\}.$$

Remark 3.3.1. (Choice of l_p -norm) Different p values will affect the shape of the prediction

region. Some common choices for p include $p = 1$, $p = 2$, and $p = \infty$. For example, using $p = 2$ will produce a d -dimensional ball; while using $p = \infty$ will produce a d -dimensional rectangle.

Remark 3.3.2. (Prediction region based on studentized root) Step 5 of algorithm 3.3.1 can be augmented to produce predictive region based on studentized predictive root. Let $\hat{\mathbf{V}}_n$ be the estimated covariance matrix of $\underline{R}_{n+1}$, and $\hat{\mathbf{V}}_n^*$ the estimated covariance matrix of $\underline{R}_{n+1}^*$, then the studentized root $\hat{\mathbf{V}}_n^{*-1/2} \underline{R}_{n+1}^*$ can be used to replace 4(c). The corresponding prediction region is then

$$\{\underline{y} \in \mathbb{R}^d : \left\| \hat{\mathbf{V}}_n^{-1/2} (\underline{y} - \hat{\underline{Y}}_{n+1}) \right\|_p \leq q_{\alpha}^* \},$$

where q_{α}^* is the analog of q_{α} for the studentized root.

Remark 3.3.3. (Limit model-free bootstrap) The entries of $\underline{\xi}_{dn}^*$ and $\underline{\xi}_d^*$ can also be sampled from $\mathcal{N}(0, 1)$, which will be the limiting distribution for the entries of $\hat{\underline{\xi}}_{dn}$. This is called the limit model-free bootstrap in Politis [2015] and has superior performance under certain scenarios.

Remark 3.3.4. (Generalization to h -step ahead prediction region) The above bootstrap algorithm can be generalized to replicate the h -step ahead predictive root

$$\underline{R}_{n+h} = \underline{Y}_{n+h} - \hat{\underline{Y}}_{n+h},$$

based on which h -step ahead prediction region can be constructed as well. This is useful for the next section, where we propose a new approach for constructing joint prediction bands for univariate time series.

Remark 3.3.5. (MFB with fixed predictor) Another variant to bootstrap $\underline{R}_{n+1}$ is to generate

$$\hat{\underline{R}}_{n+1}^* = \underline{Y}_{n+1}^* - \hat{\underline{Y}}_{n+1},$$

for faster execution time, more stability, and (possibly) better performance. We compare the bootstrap performance of both procedures in Section 3.5.

Algorithm 3.3.1 will work under model assumption (3.3.1). In order to perform bootstrap for the more complicated model (3.3.2), certain changes need to be applied to the above algorithm detailed below:

Algorithm 3.3.2. *Bootstrap algorithm under model 3.3.2:*

Replace each F_i, F_i^{-1} by $F_{i,i-1}(\cdot) = \mathbb{P}(Y_{i,t} \leq \cdot | \underline{Y}_{1:i-1,t})$ and its inverse. Possible estimator choices for $F_{i,i-1}(\cdot)$ include the nonparametric conditional CDF estimator, and estimator based on quantile/distributional regression.

3.3.4 Generating joint prediction band for univariate time series

Apart from producing prediction regions for multivariate time series, the above algorithm can also be used under a univariate time series setup. For this section, consider $\{Y_t\}_{t=1}^n$ to be realizations from model 3.3.1 with $d = 1$. The problem of interest now is to generate a joint prediction band(JPB) for observations from time $n + 1$ up until $n + h$.

First of all, the algorithm described in Wang and Politis [2021b] can be used to generate prediction intervals(PI) of level $1 - \alpha$ for observations up to h -step ahead, $\{\hat{C}_{1-\alpha}(j)\}_{j=n+1}^{n+h}$, such that asymptotic validity holds for all the PIs: as $n \rightarrow \infty$,

$$\sup_{n+1 \leq j \leq n+h} |\mathbb{P}\left(Y_{t+j} \in \hat{C}_{1-\alpha}(j) | \{Y_t\}_{t=1}^n\right) - (1 - \alpha)| \rightarrow 0.$$

A straightforward solution to adapting towards a JPB $\hat{\mathbf{C}}_{n+1:n+h}$ is through the Bonferroni correction: let

$$\hat{\mathbf{C}}_{n+1:n+h}^{(Bon)} = \bigtimes_{j=n+1}^{n+h} \hat{C}_{1-\alpha/h}(j)$$

Then by simple union bounds, for large enough n ,

$$\mathbb{P}\left(\underline{Y}_{n+1:n+h} \in \widehat{\mathbf{C}}_{n+1:n+h}^{(Bon)} | \{Y_t\}_{t=1}^n\right) \geq 1 - \alpha. \quad (3.3.8)$$

However, equation (3.3.8) does not guarantee an exact $1 - \alpha$ coverage; it is also well recognized that Bonferroni correction is a conservative method that produces prediction regions with coverage much larger than nominal levels, particularly so when the prediction intervals $\widehat{C}_{1-\alpha/h}(j)$ are correlated, which is indeed the case under our time series setting.

A different perspective for generating valid JPBs is through controlling errors arising from multiple testing, such as family-wise errors(FWE) or false discovery rate. For example, Wolf and Wunderli [2015] proposed general bootstrap methods for producing JPBs with guaranteed control for k -FWE: the error associated with false coverage for at least k future observations.

Here, we offer a new approach for generating JPB with guaranteed coverage level utilizing the model-free bootstrap algorithm described above, through *stacking* the univariate time series. With a slight abuse of notation, for this section, let

$$\underline{Y}_t = (Y_{t-h+1}, \dots, Y_t)^\top, t = h, \dots, n.$$

Then $\{\underline{Y}_t\}_{t=h}^n$ are current observations stacked into vectors of dimension h , and the h -step ahead data vector $\underline{Y}_{n+h} = (Y_{n+1}, \dots, Y_{n+h})^\top$ consists of the next h future observations. Since Y_t follows model 3.3.1 with $d = 1$, the stacked version $\underline{Y}_t$ also satisfies model 3.3.1 with $d = h$. This facilitates using algorithm 3.3.1 to construct a $1 - \alpha$ prediction region for $\underline{Y}_{n+h}$, which is also the JPB for the next h observations of Y_t .

3.4 Theoretical results

In this section, we prove bootstrap prediction region validity of algorithm 3.3.1 by using the same technique of Wang and Politis [2021b]. The proof mostly follows from the proof of Theorem 5.1 in Wang and Politis [2021b]. The main challenge is to show

$$\left\| \hat{\Gamma}_{dn} - \Gamma_{dn} \right\|_{op} \xrightarrow{P} 0,$$

under certain assumptions. The additional difference under multivariate setting is that Γ_{dn} is no longer a Toeplitz matrix, as is the case in Wang and Politis [2021b], but rather a block Toeplitz matrix.

We list the following general assumptions:

Assumption 3.4.1. (A1). $\underline{Y}_t \in \mathbb{R}^d$ follows model (3.3.1), with each f_i continuously differentiable and strictly monotone.

(A2). For each $1 \leq i \leq d$, the estimator $\hat{F}_i$ satisfies a uniform consistency condition with $\mathcal{O}_p(1/\sqrt{n})$ rate:

$$\sup_{y \in \mathbb{R}} |\hat{F}_i(y) - F_i(y)| = \mathcal{O}_p(1/\sqrt{n}). \quad (3.4.1)$$

(A3). $\sum_{h=0}^{\infty} |\Gamma_h|_1 < \infty$.

(A4). $\exists M > 0$, such that for all $i, j = 1, \dots, d$, and all $|h| < n$,

$$\mathbb{E} \left| \sum_{t=1}^{n-|h|} Z_{i,t} Z_{j,t+|h|} - n \Gamma_{|h|}(i, j) \right|_1 \leq M \sqrt{n}.$$

(A5). $\exists \lambda_0, \varepsilon_0 > 0, n_0 \in \mathbb{N}$, such that the eigenvalues of Γ_{dn} are uniformly bounded and

bounded away from 0 for all $n > n_0$, i.e.

$$\lambda_0 \geq \lambda_{\max}(\Gamma_{dn}) \geq \lambda_{\min}(\Gamma_{dn}) \geq \varepsilon_0 > 0.$$

(A6). Both l and c diverge to infinity as $n \rightarrow \infty$, such that $1/l + l/\sqrt{n} = o(1)$, $lce^{c^2/2} = o(\sqrt{n})$, and $lc^{1/2}e^{-c^2/4} = o(1)$.

Proposition 3.4.1. Under assumptions (A1) and (A2), $\underline{Z}_t$ is a multivariate Gaussian process, and

$$\sup_{1 \leq t \leq n} |\hat{U}_{i,t} - U_{i,t}| = \mathcal{O}_p(1/\sqrt{n}).$$

Lemma 3.4.1. Under assumptions (A1)-(A5), as $n \rightarrow \infty$, $\hat{\Gamma}_{dn}$ is positive definite in probability; also, both $\left\| \hat{\Gamma}_{dn} - \Gamma_{dn} \right\|_{op}$ and $\left\| \hat{\Gamma}_{dn}^{-1} - \Gamma_{dn}^{-1} \right\|_{op}$ converge to 0 in probability.

Proof. First of all, since $\hat{\Gamma}_{dn} - \Gamma_{dn}$ is symmetric,

$$\begin{aligned} \left\| \hat{\Gamma}_{dn} - \Gamma_{dn} \right\|_{op} &\leq \sqrt{\left\| \hat{\Gamma}_{dn} - \Gamma_{dn} \right\|_1 \left\| \hat{\Gamma}_{dn} - \Gamma_{dn} \right\|_{\infty}} \\ &= \left\| \hat{\Gamma}_{dn} - \Gamma_{dn} \right\|_{\infty} \\ &= \max_{1 \leq i \leq dn} \sum_{j=1}^{dn} |(\hat{\Gamma}_{dn} - \Gamma_{dn})(i, j)|. \end{aligned}$$

By block Toeplitz property,

$$\begin{aligned} \max_{1 \leq i \leq dn} \sum_{j=1}^{dn} |(\hat{\Gamma}_{dn} - \Gamma_{dn})(i, j)| &\leq \max_{0 \leq k \leq n-1} \sum_{i=kd+1}^{(k+1)d} \sum_{j=1}^{dn} |(\hat{\Gamma}_{dn} - \Gamma_{dn})(i, j)| \\ &\leq \sum_{h=1-n}^{n-1} \sum_{i=1}^d \sum_{j=1}^d |(\hat{\Gamma}_h - \Gamma_h)(i, j)|. \end{aligned} \tag{3.4.2}$$

Notice that $\sum_{i=1}^d \sum_{j=1}^d |(\hat{\Gamma}_h - \Gamma_h)(i, j)| = |\hat{\Gamma}_h - \Gamma_h|_1$ is the entry-wise l_1 norm, thus the RHS of equation (3.4.2) equals

$$\begin{aligned} \sum_{h=1-n}^{n-1} |\hat{\Gamma}_h - \Gamma_h|_1 &= \sum_{h=1-n}^{n-1} |\hat{\Gamma}_h - \hat{\Gamma}_h + \hat{\Gamma}_h - \Gamma_h|_1 \\ &\leq \sum_{h=1-n}^{n-1} |\hat{\Gamma}_h - \hat{\Gamma}_h|_1 + \sum_{h=1-n}^{n-1} |\hat{\Gamma}_h - \Gamma_h|_1. \end{aligned}$$

The second sum was proved to converge to 0 in probability by Theorem 2.1, Jentsch and Politis [2015] under (A3), (A4) and (A6). We only need to show the first sum $\sum_{h=1-n}^{n-1} |\hat{\Gamma}_h - \hat{\Gamma}_h|_1$ converges to 0 as well.

$$\begin{aligned} \sum_{h=1-n}^{n-1} |\hat{\Gamma}_h - \hat{\Gamma}_h|_1 &= \sum_{1 \leq i, j \leq d} \sum_{h=1-n}^{n-1} |\hat{\Gamma}_h(i, j) - \hat{\Gamma}_h(i, j)| \\ &= \sum_{1 \leq i, j \leq d} \sum_{h=1-n}^{n-1} \left| \frac{1}{n} \sum_{t=1}^{n-|h|} (\tilde{Z}_{i,t} \tilde{Z}_{j,t+h} - Z_{i,t} Z_{j,t+h}) \right|, \end{aligned} \tag{3.4.3}$$

As we assume the dimension d is fixed, we only need to show the second sum

$\sum_{h=1-n}^{n-1} \left| \frac{1}{n} \sum_{t=1}^{n-|h|} (\tilde{Z}_{i,t} \tilde{Z}_{j,t+h} - Z_{i,t} Z_{j,t+h}) \right|$ converges to 0 for all pairs of (i, j) . Under the result of proposition 3.4.1 and (A6), Wang and Politis [2021b] shows the sum does converge to 0 in probability. Thus $\sum_{h=1-n}^{n-1} |\hat{\Gamma}_h - \hat{\Gamma}_h|_1 \xrightarrow{P} 0$, and $\left\| \hat{\Gamma}_{dn} - \Gamma_{dn} \right\|_{op} \xrightarrow{P} 0$. $\square$

Remark 3.4.1. For the next theorem, we need a slightly stronger result of Lemma 3.4.1, which is

$$\frac{1}{(\log n)^2} \left\| \hat{\Gamma}_{dn} - \Gamma_{dn} \right\|_{op} \xrightarrow{P} 0. \tag{3.4.4}$$

This guarantees $\left\| \hat{\Gamma}_{dn}^{1/2} - \Gamma_{dn}^{1/2} \right\|_{op} \xrightarrow{P} 0$ as well as $\left\| \hat{\Gamma}_{dn}^{-1/2} - \Gamma_{dn}^{-1/2} \right\|_{op} \xrightarrow{P} 0$. Note that (3.4.4) is achievable if we choose appropriate rates for l and c .

Theorem 3.4.2. *Under (A1) - (A6) such that (3.4.4) also holds, then the predictive distribution*

$F_{n+1|n}$ is continuous, and

$$\sup_{\underline{y} \in \mathbb{R}^d} |\hat{F}_{n+1|n}(\underline{y}) - F_{n+1|n}(\underline{y})| \xrightarrow{P} 0. \quad (3.4.5)$$

Assume now that under equation (3.4.5), $\hat{\underline{Y}}_{n+1} \rightarrow \mathcal{P}(\underline{Y}_{n+1}|\mathcal{F}_n)$ in probability. Then the prediction region generated by the MFB algorithm 3.3.1 is asymptotically valid.

The proof of Theorem 3.4.2 proceeds the same as that of Theorem 5.1 of Wang and Politis [2021b].

3.5 Numerical results

3.5.1 Synthetic data experiment

Experiment setup. Our main focus in this section is to numerically demonstrate the coverage performance of the multivariate MFB. Consider the following nonlinear time series model with dimension $d = 2$:

$$Y_{i,t} = f_i(W_{i,t}),$$

where $f_1(x) = f_2(x) = \text{sgn}(x)\sqrt{|x|}$, and $\underline{W}_t$ is a VAR(1) process with Gaussian innovations:

$$\underline{W}_t = A\underline{W}_{t-1} + \underline{\varepsilon}_t,$$

where we set $A = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.6 \end{pmatrix}$, $\underline{\varepsilon}_t \stackrel{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, B)$ with $B = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}$. Since the f_i s are continuous monotone functions and $\underline{W}_t$ is a Gaussian process, $\underline{Y}_t$ satisfies model 3.3.1.

Let $\mathbf{Y}_n = \{\underline{Y}_t\}_{t=1}^n$ be an arbitrary sample path drawn from above. By applying algorithm 3.3.1 we then generate a two sided prediction region: $\hat{\mathbf{C}}_{1-\alpha}(\mathbf{Y}_n)$. Let $\{\underline{Y}_{n+1}^{(m)}\}_{m=1}^M$ be a set of future observations sampled directly from the above model conditioning on $\mathbf{Y}_n$, namely $\underline{Y}_{n+1}^{(m)} \stackrel{i.i.d.}{\sim} F_{n+1|n}$. The coverage probability of $\hat{\mathbf{C}}_{1-\alpha}(\mathbf{Y}_n)$ is estimated by the empirical coverage rate for $\{\underline{Y}_{n+1}^{(m)}\}_{m=1}^M$,

i.e.,

$$\widehat{CVR}(\mathbf{Y}_n) = \frac{1}{M} \sum_{m=1}^M I\left(\underline{Y}_{n+1}^{(m)} \in \hat{\mathbf{C}}_{1-\alpha}(\mathbf{Y}_n)\right).$$

To achieve stable estimations for the coverage probability, we generate multiple realizations of $\mathbf{Y}_n$ and calculate the average of the CVR estimations as our metric.

Parameter selection. The following parameters need to be properly selected in algorithm 3.3.1: the bandwidths $\{b_i\}_{i=1}^d$ for the nonparametric CDFs $\{\hat{F}_i\}_{i=1}^d$; the banding parameter l in $\hat{\Gamma}_{dn}$; the choice of optimal predictor as well as the L_p norm when evaluating the root $\underline{R}_{n+1}$.

While there are both sample-based rules as well as cross-validation for selecting b_i and l , see e.g, Das and Politis [2021] and McMurry and Politis [2010], in this experiment we take another approach. We use a fixed set of parameters for the bandwidths $\{b_i\}_{i=1}^d$, and evaluate the performance of the MFB algorithm with various combinations of the remaining parameters. This way, we can see the effect of each parameter on the performance of MFB. The coverage metrics are obtained for a range of sample sizes $n \in \{100, 200, 300, 400, 500\}$. We set $\alpha = 0.05$ which means the nominal coverage should be 95%. Results of our simulations are plotted below in Figure 3.1 to 3.5 .

As expected, as n increases, the coverage probability converges to the nominal $1 - \alpha$, showing asymptotic validity. For this particular example, the choice of L^p norm and predictor type has effect on the relative advantage between MFB with resampled/fixed predictor. Note that if we compare the best performance of the two algorithms according to closeness of coverage probability with respect to the nominal level, both seem to work very well under certain parameter combinations.

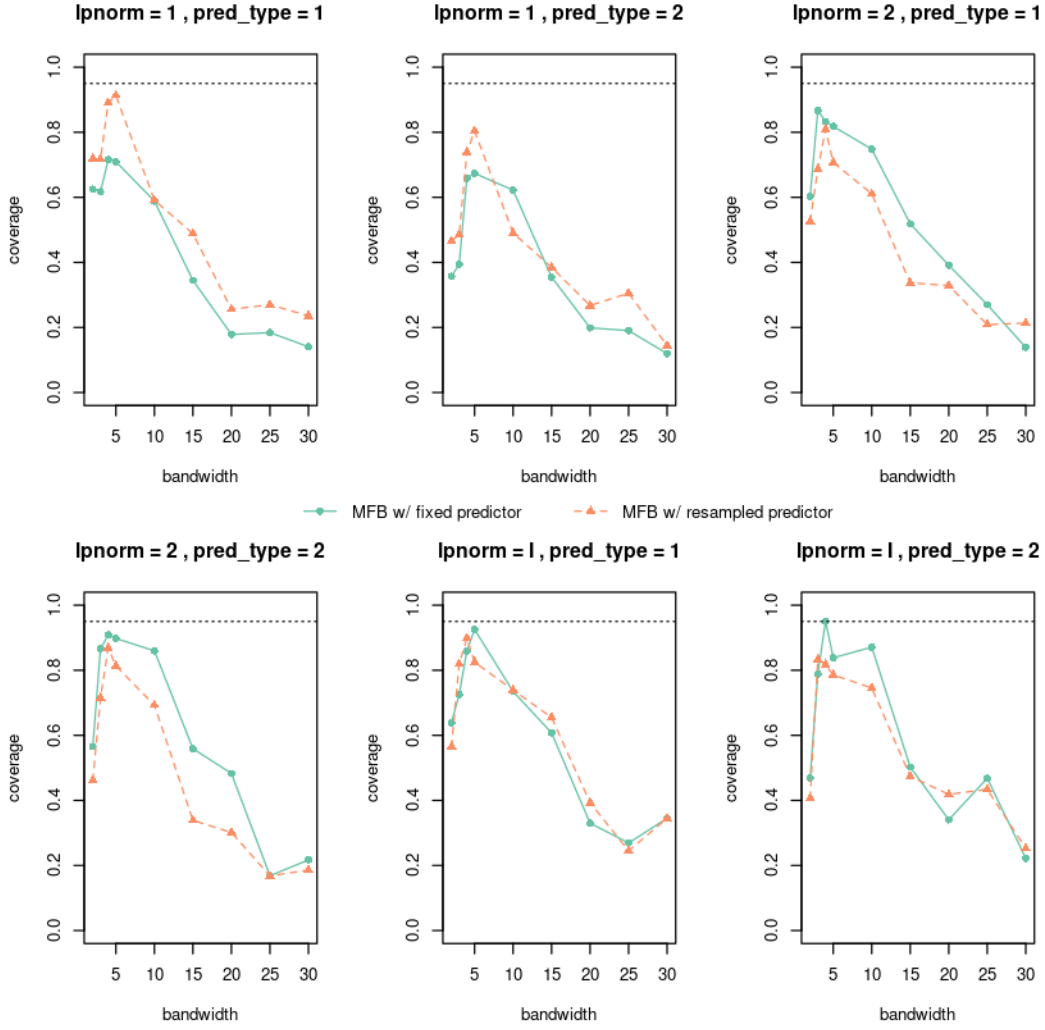


Figure 3.1: CVR plots with $n = 100$.

3.5.2 Joint prediction bands for heavy-tailed financial returns

A lot of real world time series data exhibit heavy-tailed phenomena, particularly exemplified under the realm of financial time series, wherein a common characteristic is the so-called leptokurtosis, i.e.,

$$Kurt = \frac{\mu_4}{\sigma^4} > 3,$$

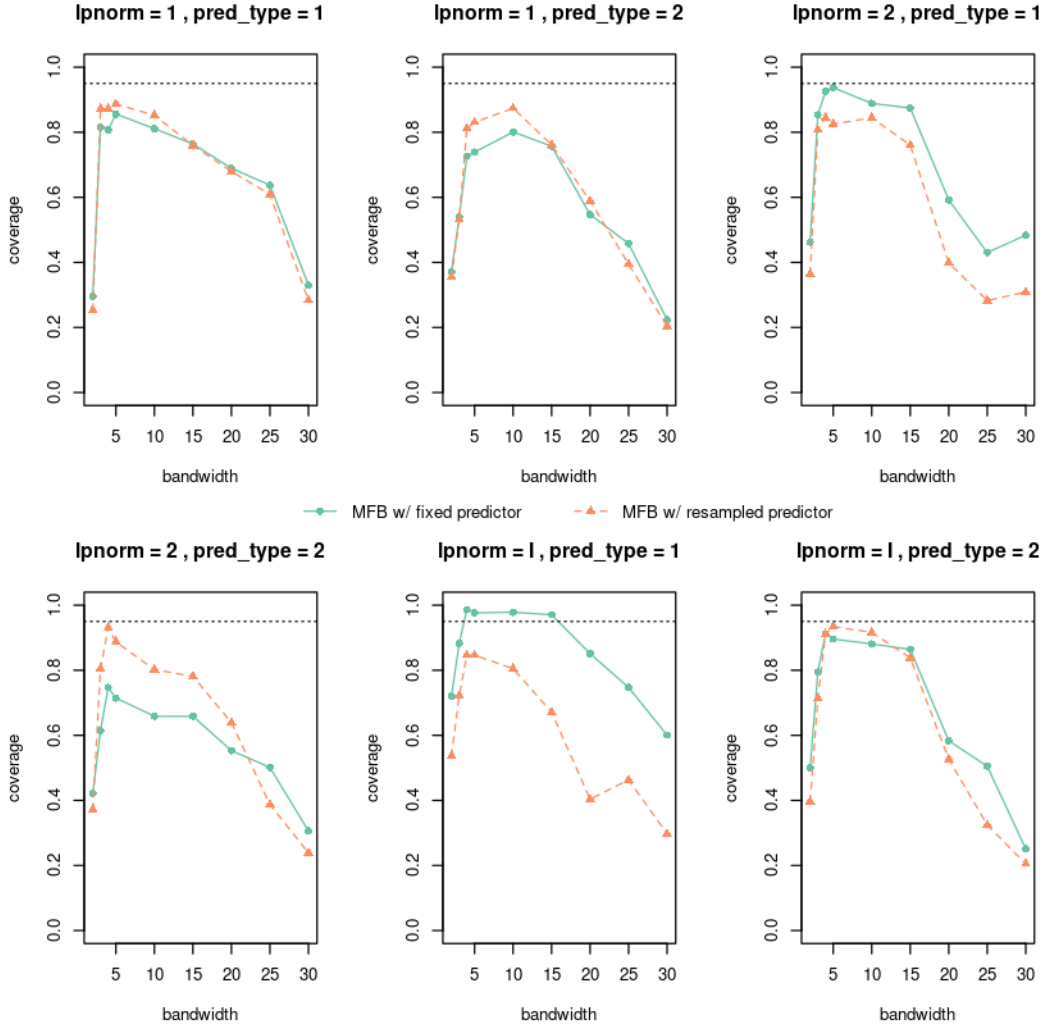


Figure 3.2: CVR plots with $n = 200$.

where $\mu_4 = \mathbb{E}(Y_t - \mu)^4$ is the centered fourth moment of the univariate time series Y_t , and $\sigma = \sqrt{\mathbb{E}(Y_t - \mu)^2}$ is the standard deviation. Under a leptokurtic situation, the marginal distribution of the time series has heavier tail than the normal distribution. Accurate modeling for heavy-tailedness is of great importance in the financial market. For example, Value-at-Risk(VaR) is a fundamental metric commonly used to measure the risk associated with an asset, defined as

$$VaR(\alpha) = \inf\{y \in \mathbb{R} : F(y) \geq \alpha\}.$$

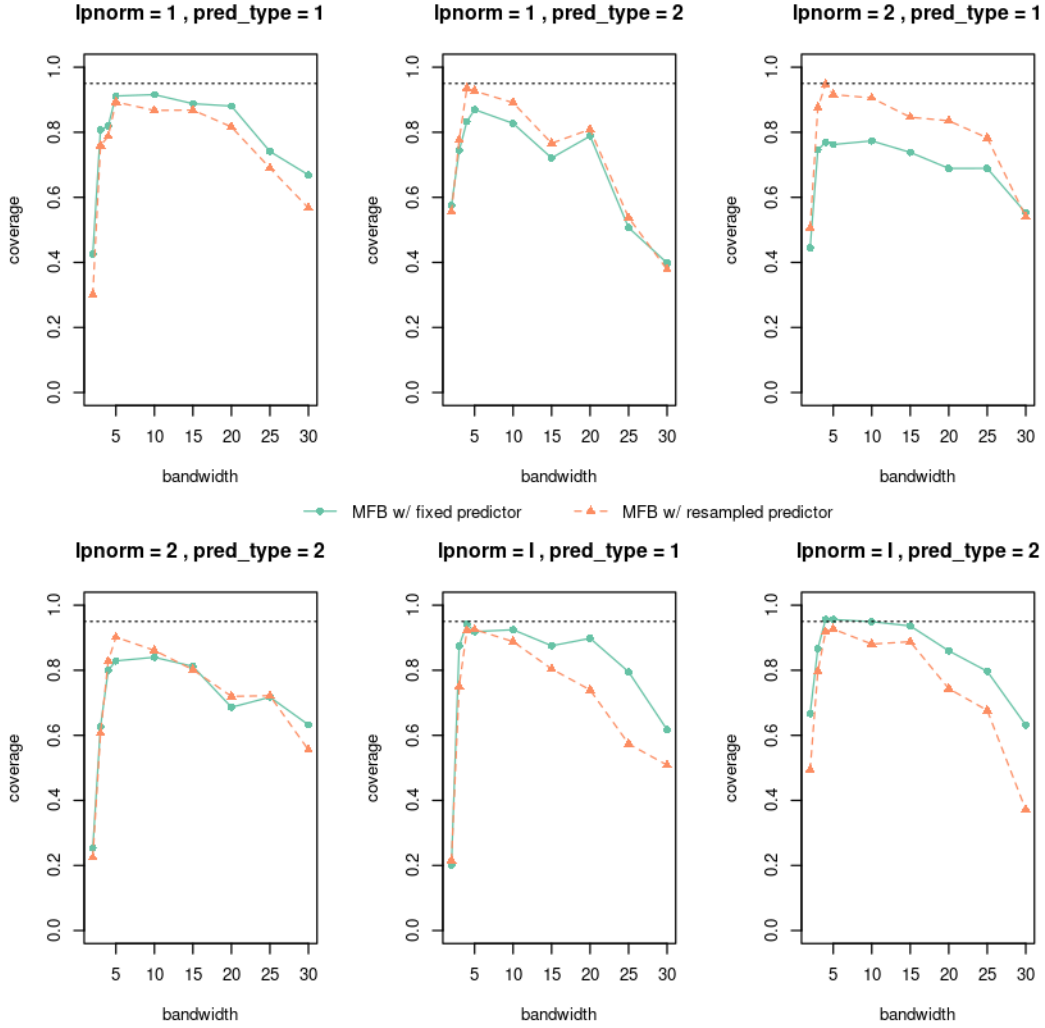


Figure 3.3: CVR plots with $n = 300$.

Clearly, VaR captures the (left) heavy-tailedness of the CDF F .

A common approach for modeling heavy-tailedness of financial data is through a Gaussian mixture model: Let $Y_t = \sigma_t z_t$, where σ_t is random and $\mathcal{F}_{t-1}$ -measurable; $z_t \sim \mathcal{N}(0, 1)$. Then Y_t is distributed according to a Gaussian mixture, thus leptokurtic. The famous ARCH/GARCH model, stochastic volatility model and the more recent NoVaS transformation of Politis [2007] are both paragons of this approach.

A different approach is to model heavy-tailedness via nonlinear transformation models.

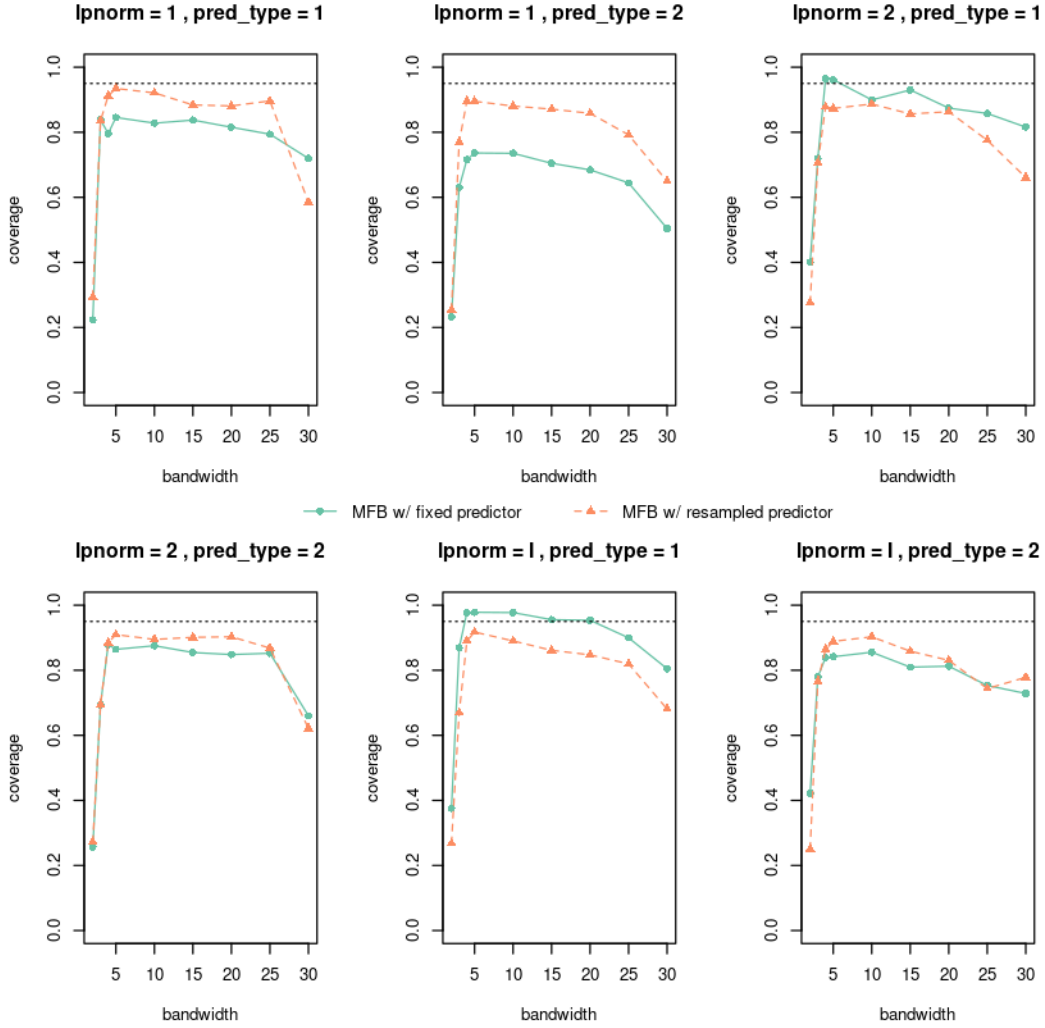


Figure 3.4: CVR plots with $n = 400$.

In particular,

$$Y_t = f(W_t)$$

of model 3.3.1 is a classic nonlinear model studied by many, both from a theoretical perspective (cf. Samorodnitsky and Taqqu [1994], Breuer and Major [1983]) and also an applied perspective (cf. Hull and White [1998]). As previously discussed in Section 3.3.4, the multivariate MFB algorithm can be used to generate JPB under this setup. The JPB will be useful to describe probable region of future paths and thus provide more information for trading. In this section, we present some

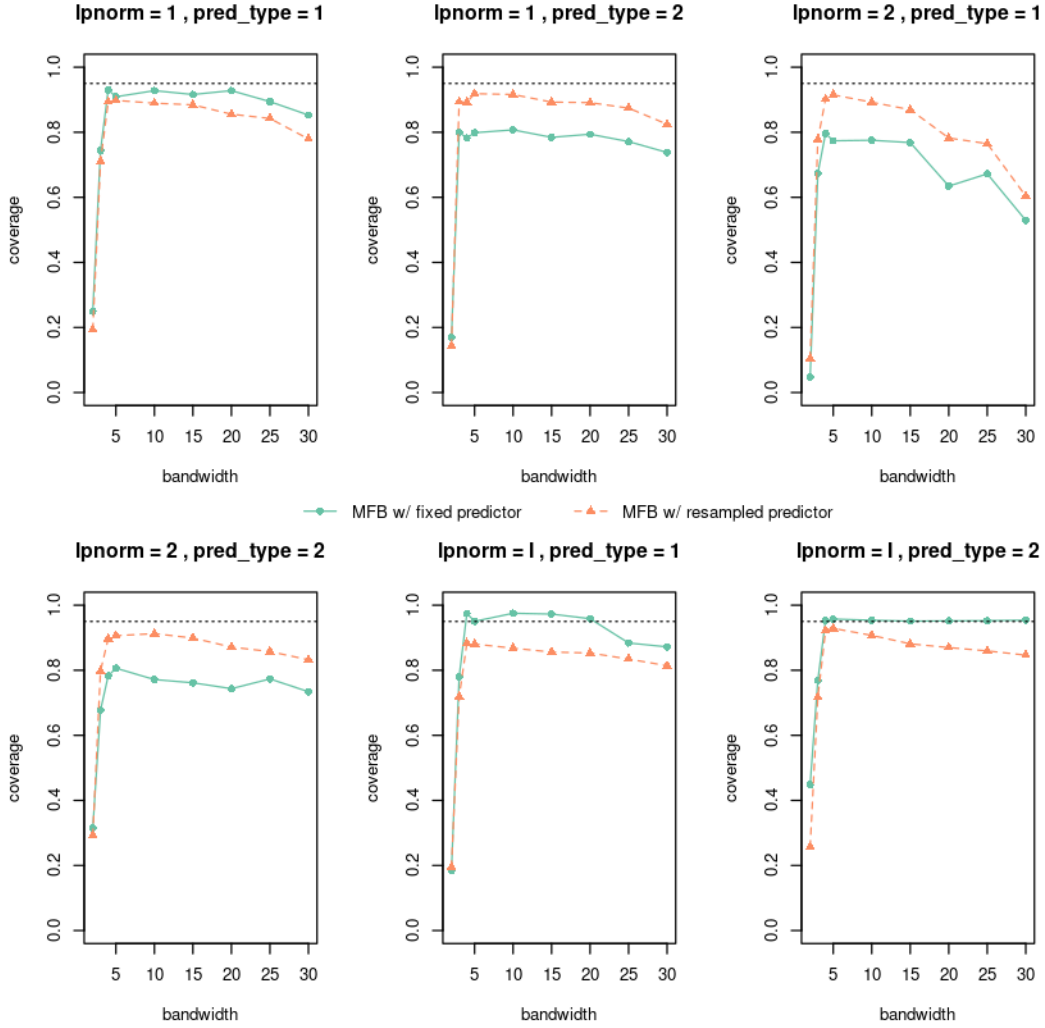


Figure 3.5: CVR plots with $n = 500$.

numerical results on the empirical coverage performance of the MFB-based JPB for daily stock returns. We also compare them with a benchmark method, which is JPB based on a garch(1,1)-bootstrap – see Chen and Politis [2019] for details.

The details of our experiments goes as follows. We pick the following stocks: AAPL, AMZN, TSLA, GME as candidate datasets, where we gathered daily stock returns from 2018-01-01 to 2021-08-31, denoted as $\{Y_t\}_{t=1}^n$ with $n = 922$. We then sequentially generate a collection of

(*past, future*) pairs:

$$\left\{ \left(\{Y_i\}_{i=t-n_0+1}^t, \{Y_j\}_{j=t+1}^{t+h} \right) : t = n_0 + kh, 0 \leq k \leq \left\lfloor \frac{n-n_0}{h} \right\rfloor \right\}. \quad (3.5.1)$$

The parameter n_0 represents the number of days used to backtrack past data for prediction purpose; and h represents the dimension of future data we try to predict. By using the data generation scheme in (3.5.1), the future observations are non-overlapping and are better suited for calculating the empirical coverage. To simplify notations, we let $\underline{X}_{k,1} = (Y_{kh+1}, \dots, Y_{kh+n_0})^\top$ and $\underline{X}_{k,2} = (Y_{n_0+kh+1}, \dots, Y_{n_0+(k+1)h})^\top$. The JPB calculated from the data $\underline{X}_{k,1}$ is denoted by $\hat{\mathbf{C}}(\underline{X}_{k,1})$. Finally, the empirical coverage rate (ECVR) based on data Y_t and parameters n_0, h is calculated by

$$ECVR(\{Y_t\}_{t=1}^n, n_0, h) = \frac{1}{\left\lfloor \frac{n-n_0}{h} \right\rfloor + 1} \sum_{k=0}^{\left\lfloor \frac{n-n_0}{h} \right\rfloor} \mathbf{1}_{\underline{X}_{k,2} \in \hat{\mathbf{C}}(\underline{X}_{k,1})}.$$

We plot the ECVR against a range of n_0 values for different stocks and h . In the MFB algorithm, we choose to use the L^2 -optimal predictor and the L^1 norm for root evaluation; the bandwidth $b = 0.01$ and banding parameter $l = 0.4$ are determined via cross validation. The results of our experiments are presented in Figure 3.6 - 3.9.

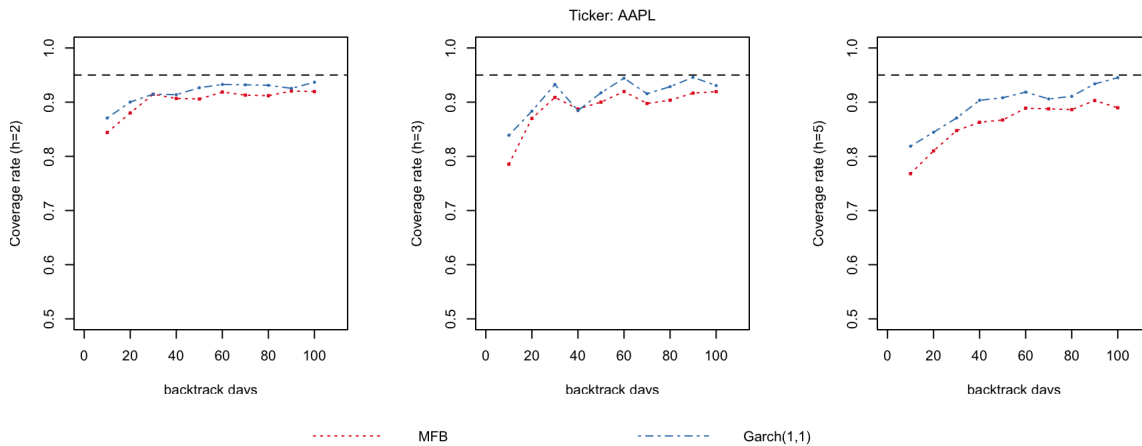


Figure 3.6: ECVR plots for AAPL.

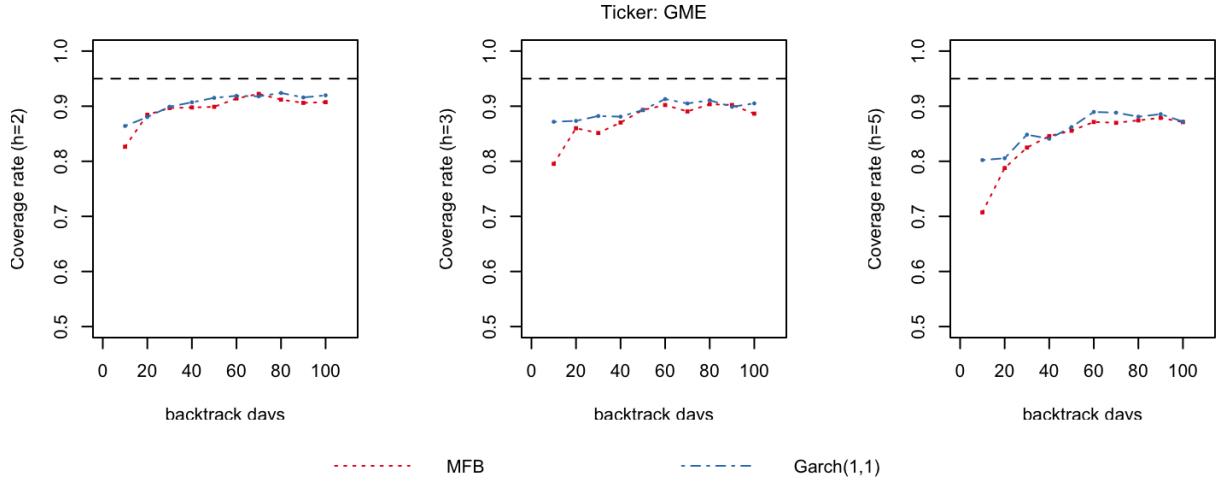


Figure 3.7: ECVR plots for GME

We can see that as the number of backtracking days increases, the empirical coverage also grows towards 95% because of asymptotic validity. However, both methods still have under coverage issue, partially due to the fact that the stationarity assumption does not perfectly fit the stock market. We also observe that for $h = 2$ and 3, the performance of the two methods are on par with each other. However, for $h = 5$ the garch-based bootstrap has superior performance comparing with the MFB. This is due to the more efficient parameter searching scheme in garch fitting, whereas for the MFB, the process of parameter searching mostly relies on cross validation, which can be inefficient and suboptimal.

3.6 Acknowledgement

Chapter 3, in full, is a version of the paper “Model-free Bootstrap Prediction Regions for Multivariate Time Series”, Yiren Wang, Dimitris N. Politis. It is currently being prepared for publication in a major statistics journal. The dissertation author is the primary investigator and author of this material.

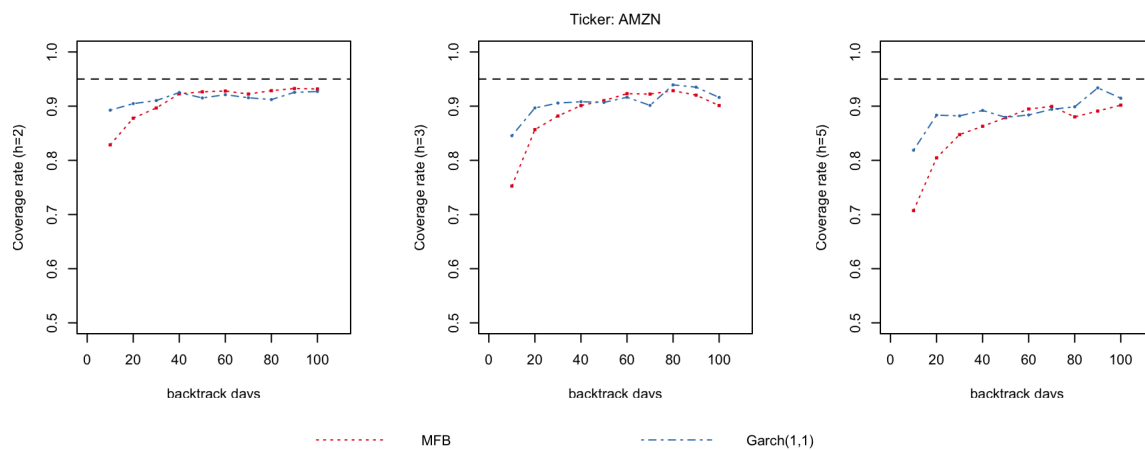


Figure 3.8: ECVR plots for AMZN.

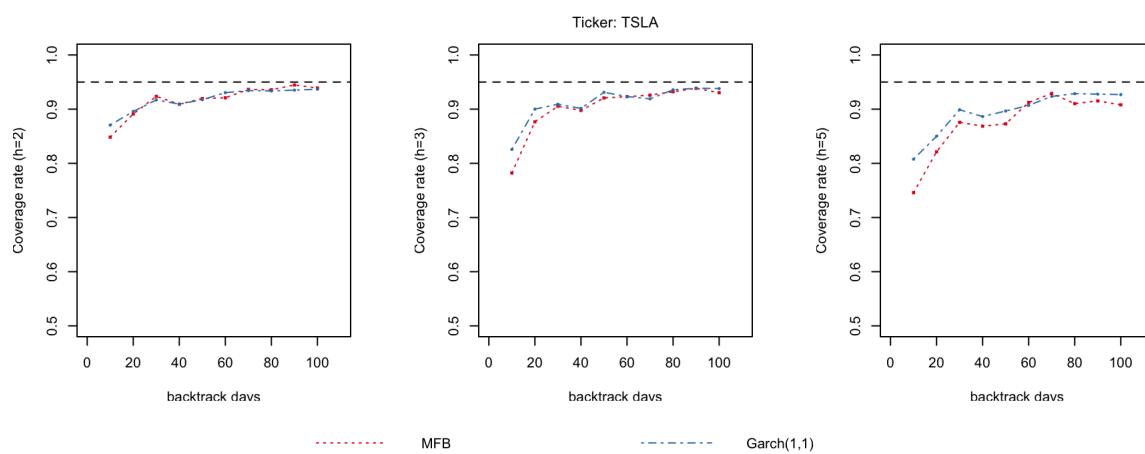


Figure 3.9: ECVR plots for TSLA.

Chapter 4

Bootstrap Methods for Quantile Autoregressive Time Series

4.1 Abstract

Since the introduction of Koenker and Bassett [1978], quantile regression has received wide recognition for its superior efficiency for conditional quantile estimation than traditional methods. The quantile autoregression(QAR) model introduced by Koenker and Xiao [2006] successfully extended the use of quantile regression into time series literature. In this paper, we propose to use the forward bootstrap of Pan and Politis [2016b] and the Model-free bootstrap of Politis [2015] to construct both confidence intervals for parameter inference and prediction intervals for the future observations. We introduce a novel approach of discretization and interpolation for estimating the quantile regression coefficients, which guarantees bootstrap validity of the proposed procedure.

4.2 Introduction

Our interest of this paper is to introduce a set of bootstrap algorithms for both inference and prediction problems for a class of time series models. Consider the following univariate random coefficient model for $\{X_t\}_{t=1}^n$:

$$X_t = \theta_0(U_t) + \sum_{k=1}^p \theta_k(U_t)X_{t-k} \quad (4.2.1)$$

where U_t are i.i.d. uniform $[0, 1]$ distributed; $\theta_i(\cdot) : [0, 1] \rightarrow \mathbb{R}$ are functions such that the right hand side of (4.2.1) is monotone increasing with respect to U_t . The model (4.2.1) implies that the quantile function Q of X_t conditioning on past information $\mathcal{F}_{t-1} = \sigma(X_1, \dots, X_{t-1})$ satisfies the linear equation:

$$Q_{t|t-1}(\tau) := Q_{X_t|\mathcal{F}_{t-1}}(\tau) = \theta_0(\tau) + \sum_{k=1}^p \theta_k(\tau)X_{t-k}, \forall \tau \in [0, 1]. \quad (4.2.2)$$

Such model is also called the quantile autoregression(QAR) model.

The linear structure in (4.2.2) suggests that one can use quantile regression(QR) to estimate $Q_{X_t|\mathcal{F}_{t-1}}(\tau)$. As illustrated in Koenker and Xiao [2006], let $\rho_\tau(u) = u(\tau - I(u < 0))$ be the check function, $\mathbf{Y}_t = (1, X_{t-1}, \dots, X_{t-p})^\top$, then

$$\theta(\tau) = (\theta_0(\tau), \dots, \theta_k(\tau))^\top = \arg \min_{\beta \in \mathbb{R}^{p+1}} \mathbb{E} \rho_\tau(X_t - \beta^\top \mathbf{Y}_t).$$

We estimate $\theta(\tau)$ through minimizing the empirical loss, i.e.

$$\hat{\theta}(\tau) = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_t \rho_\tau(X_t - \beta^\top \mathbf{Y}_t). \quad (4.2.3)$$

Let $\hat{Q}_{t|t-1} = \hat{Q}_{X_t|\mathcal{F}_{t-1}}(\tau) = \hat{\theta}(\tau)^\top \mathbf{Y}_t$ for $t > p$. When $t \leq p$, $\hat{Q}_{t|t-1}$ is constructed by fitting the

quantile regression model with order $t - 1$.

We introduce the following notations. Let $\mu_0 = \mathbb{E}X_t$, $\gamma_k = \mathbb{E}X_t X_{t+k}$, and $\Omega_0 = \mathbb{E}\mathbf{Y}_t \mathbf{Y}_t^\top = \begin{bmatrix} 1 & \mu_0 \\ \mu_0 & \Gamma_p \end{bmatrix}$, where Γ_p is the p -dimensional autocovariance matrix for X_t . Let

$$\Omega_1(\tau) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n f_{t|t-1}(F_{t|t-1}^{-1}(\tau)) \mathbf{Y}_t \mathbf{Y}_t^\top,$$

where $F_{t|t-1}$ is the conditional CDF of X_t on $\mathcal{F}_{t-1}$, and $f_{t|t-1}$ is the density of $F_{t|t-1}$. Denote $\Sigma(\tau) = \Omega_1^{-1}(\tau) \Omega_0 \Omega_1^{-1}(\tau)$.

For the purpose of building large sample theory for the bootstrap methods, we list one important result regarding the invariance principle of the QR process:

Theorem 4.2.1. *(Theorem 3.2, Koenker and Xiao [2006]) Under conditions A.1 - A.3 listed in Koenker and Xiao [2006],*

$$\sqrt{n} \Sigma^{-1/2}(\tau) \left(\hat{\theta}(\tau) - \theta(\tau) \right) \Rightarrow \mathbf{B}_{p+1}(\tau), \quad (4.2.4)$$

as $n \rightarrow \infty$, where $\mathbf{B}_{p+1}(\tau)$ is the standard $p + 1$ dimensional Brownian bridge.

The above theorem, under mild additional assumptions, can be used to show the following:

$$\sup_{\tau \in [0,1]} \left\| \hat{\theta}(\tau) - \theta(\tau) \right\|_1 = \mathcal{O}_p(1/\sqrt{n}). \quad (4.2.5)$$

Parallely, we can derive the asymptotic performance for the quantile process as well. i.e.,

$$\begin{aligned} \sqrt{n} \left(\hat{Q}_{t|t-1}(\tau) - Q_{t|t-1}(\tau) \right) &= \sqrt{n} \left(\hat{\theta}(\tau)^\top \mathbf{Y}_t - \theta(\tau)^\top \mathbf{Y}_t \right) = \mathbf{Y}_t^\top \sqrt{n} \left(\hat{\theta}(\tau) - \theta(\tau) \right) \\ &\Rightarrow \mathbf{Y}_t^\top \Sigma^{1/2}(\tau) \mathbf{B}_{p+1}(\tau) \end{aligned}$$

The limiting 1-dimensional Gaussian process $G(\tau) = Y_t^\top \Sigma^{1/2}(\tau) \mathbf{B}_{p+1}(\tau)$ has continuous sample

path on $[0, 1]$ inherited from $\mathbf{B}_{p+1}(\tau)$, since F_{t-1} has continuous density, which guarantees entry-wise continuity of $\Sigma(\tau)$. The covariance of $G(\tau)$ is

$$u(1-v)\mathbf{Y}_t^T \Sigma^{1/2}(u) \Sigma^{1/2}(v) \mathbf{Y}_t, \forall 0 \leq u < v \leq 1.$$

Assume that f_{t-1} has bounded support and also bounded away from 0, then we have $\|\mathbf{Y}_t\|_2 < \infty$ and also $\sup_{\tau} \lambda_{\max} \Sigma(\tau) < \infty$. Therefore

$$\sup_{\tau \in [0,1]} \mathbb{E} G(\tau)^2 < \infty.$$

These results guarantee that $\sup_{\tau} G(\tau)$ is $\mathcal{O}_p(1)$. As a result, the estimated quantiles are uniformly consistent to the truth at $\mathcal{O}_p(1/\sqrt{n})$ rate:

$$\sup_{\tau \in [0,1]} \left(\hat{Q}_{t|t-1}(\tau) - Q_{t|t-1}(\tau) \right) = \mathcal{O}_p(1/\sqrt{n}). \quad (4.2.6)$$

Both Equation (4.2.5) and (4.2.6) are useful in later analysis.

We have the following relation between the conditional CDF $F_{t|t-1}$ and quantile function $Q_{t|t-1}$, which is relevant for constructing the bootstrap procedure:

$$F_{t|t-1}(x) = \int_0^1 I(Q_{t|t-1}(u) \leq x) du = \sup\{\tau \in [0, 1] : Q_{t-1}(\tau) \leq x\}. \quad (4.2.7)$$

By replacing the theoretical quantile function $Q_{t|t-1}$ with $\hat{Q}_{t|t-1}$, the above equation (4.2.7) can also be used to construct $\hat{F}_{t|t-1}$.

It is worth noting that in the bootstrap algorithms to be introduced, we assume the dimension p in 4.2.1 is fixed and known *a priori*. In practice, the dimension is often estimated via penalization methods, e.g., the Lasso, Akaike information criterion(AIC), etc.

4.3 Bootstrap algorithms

4.3.1 Model-based forward bootstrap

The QAR model (4.2.1) suggests that we can devise a forward bootstrap algorithm(cf. Pan and Politis [2016b]) based on the estimated quantile function vector $\hat{\theta}(\tau)$. To generate confidence region for $\theta(\tau)$, the general forward bootstrap procedure proceeds as following:

Algorithm 4.3.1. *Model-based bootstrap for the QAR(p) model (4.2.1):*

1. Choose $(X_1^*, \dots, X_p^*)$ uniformly from all blocks of size p in the original data $(X_1, \dots, X_n)$.
2. Let U_t^* be i.i.d. $\text{Uniform}(0, 1)$ samples. Sequentially generate $X_t^* = \hat{\theta}(U_t^*)^\top \mathbf{Y}_t^\top$ for $t = p+1, \dots, n$. (Forward bootstrap)
3. Step 2 is repeated for B times to get B bootstrap samples. Let $\hat{\theta}^{*(b)}(\tau)$ be the quantile regression estimator on the b^{th} sample $\{X_t^{*(b)}\}_{t=1}^n$.

The quantile regression process in the bootstrap world also satisfies an invariance principle, i.e.

$$\sqrt{n}\hat{\Sigma}(\tau)^{-1/2} \left(\hat{\theta}^{*(b)}(\tau) - \hat{\theta}(\tau) \right) \Rightarrow \mathbf{B}_{p+1}(\tau), \text{ in probability,} \quad (4.3.1)$$

where $\hat{\Sigma}(\tau)$ is of the same form as $\Sigma(\tau)$ based on the estimated model parameters $\hat{\theta}$. In order for the bootstrap procedure to be valid, we need $\hat{\Sigma}(\tau) \xrightarrow{P} \Sigma(\tau)$, such that $\sqrt{n} \left(\hat{\theta}^{*(b)}(\tau) - \hat{\theta}(\tau) \right)$ has the same limiting variance as $\sqrt{n} \left(\hat{\theta}(\tau) - \theta(\tau) \right)$, thereby validating the bootstrap procedure for inference of $\theta(\tau)$.

However, to ensure consistency of matrix estimator $\hat{\Sigma}$, we need an additional discretization step in Algorithm 4.3.1. The main difficulty for showing consistency lies in the form of $\Omega_1(\tau)$,

which equals

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n f_{t|t-1}(F_{t|t-1}^{-1}(\tau)) \mathbf{Y}_t \mathbf{Y}_t^\top, \quad (4.3.2)$$

and is dependent on the conditional densities $f_{t|t-1}$; in parallel,

$$\hat{\Omega}_1(\tau) = \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \hat{f}_{t|t-1}(\hat{F}_{t|t-1}^{-1}(\tau)) \mathbf{Y}_t^* \mathbf{Y}_t^{*\top} \text{ in the bootstrap world, where } \hat{f}_{t|t-1} \text{ and } \hat{F}_{t|t-1} \text{ are conditional density/CDF based on the estimated parameter } \hat{\theta}. \text{ Correspondingly, } \hat{\Omega}_0 = \mathbb{E}^* \mathbf{Y}_t^* \mathbf{Y}_t^{*\top} = \begin{bmatrix} 1 & \hat{\mu}_0 \\ \hat{\mu}_0 & \hat{\Gamma}_p \end{bmatrix}.$$

While $\hat{\Omega}_0 \xrightarrow{P} \Omega_0$ is straightforward via consistency of moments, showing $\hat{\Omega}_1(\tau) \xrightarrow{P} \Omega_1(\tau)$ requires consistency of conditional densities. This, however, is not guaranteed from the QR coefficient estimator. To see this, note that we have

$$f_{t|t-1}(F_{t|t-1}^{-1}(\tau)) = \frac{1}{dQ_{t|t-1}(\tau)/d\tau} = \frac{1}{\theta'(\tau)^\top \mathbf{Y}_t},$$

with $\theta'(\tau)$ the vector of entry-wise derivatives of $\theta(\tau)$. Assuming the series $f_{t|t-1}(F_{t|t-1}^{-1}(\tau)) \mathbf{Y}_t \mathbf{Y}_t^\top$ is ergodic, then Equation (4.3.2) converges in probability to $\mathbb{E} \frac{\mathbf{Y}_t \mathbf{Y}_t^\top}{\theta'(\tau)^\top \mathbf{Y}_t}$. Similarly, we have that $\hat{\Omega}_1(\tau) \xrightarrow{P^*} \mathbb{E}^* \frac{\mathbf{Y}_t^* \mathbf{Y}_t^{*\top}}{\hat{\theta}'(\tau)^\top \mathbf{Y}_t^*}$. Therefore, in order for consistency of $\hat{\Omega}_1(\tau)$ to hold, we need convergence of derivatives, i.e. $\hat{\theta}'(\tau) \xrightarrow{P} \theta'(\tau)$.

Unfortunately, this is not guaranteed in general. To begin with, differentiability of $\hat{\theta}$ is questionable without further assumptions. Besides, by Theorem 4.2.1, the limit on the right hand side is a Brownian bridge, which is known to be non-differentiable almost everywhere; this forbids us to further take a path based analysis.

Now that it is clear that the issue lies in inconsistency of conditional density $\hat{f}_{t|t-1}$ implied by the quantile regression estimator $\hat{\theta}$, towards the truth $f_{t|t-1}$, we can bypass it by designing an alternative estimator, where the implied conditional density will be consistent. Consider the following choice. Let $\mathcal{T}_h = \{\tau_k = \min(kh, 1) : k = 0, \dots, \lceil 1/h \rceil\}$ be a discrete grid on $[0, 1]$ with

mesh size h . Our new proposal for estimating the QR coefficients proceeds as below. Let

$$\hat{\theta}_{new}(\tau) = \begin{cases} \hat{\theta}(\tau_k) & \tau = \tau_k \\ \frac{\tau - \tau_k}{\tau_{k+1} - \tau_k} \hat{\theta}(\tau_{k+1}) + \frac{\tau_{k+1} - \tau}{\tau_{k+1} - \tau_k} \hat{\theta}(\tau_k) & \tau_k < \tau < \tau_{k+1} \end{cases} \quad (4.3.3)$$

i.e., we only use the estimated QR coefficients on the grid $\mathcal{T}_h$, while outside the grid we linearly interpolate the values.

The main benefit of the new estimator is its almost-everywhere differentiability on $[0, 1]$, except for the grid points. This is sufficient for conducting inference with $\tau \in [0, 1] \setminus \mathcal{T}_h$. If our interest lies inside $\mathcal{T}_h$, we can also make minimal adjustment by excluding the point of interest from $\mathcal{T}_h$; asymptotically, this adjustment is not necessary.

Remark 4.3.1. While there are other popular methods for interpolation that guarantees differentiability everywhere – such as fitting splines from nonparametric function estimation, (4.3.3) is an easier approach for later analysis.

While $\hat{\theta}_{new}(\tau)$ is differentiable, in order to validate bootstrap, we need to show convergence of derivatives. This can only be true when h satisfies certain rate of convergence. Note that without loss of generality,

$$\hat{\theta}'_{new}(\tau) = \frac{\hat{\theta}(\tau_{k+1}) - \hat{\theta}(\tau_k)}{\tau_{k+1} - \tau_k} = \frac{\theta(\tau_k + h) - \theta(\tau_k)}{h} + \frac{\hat{\theta}(\tau_{k+1}) - \theta(\tau_{k+1})}{h} + \frac{\theta(\tau_k) - \hat{\theta}(\tau_k)}{h} \quad (4.3.4)$$

The first term converges to $\theta'(\tau)$ as long as $h \rightarrow 0$. Using Theorem 4.2.1, the second and third term in (4.3.4) satisfy

$$\left\| \frac{\hat{\theta}(\tau_k) - \theta(\tau_k)}{h} \right\|_1 \leq \sup_{\tau \in [0, 1]} \left\| \frac{\hat{\theta}(\tau) - \theta(\tau)}{h} \right\|_1 = \mathcal{O}_p(1/h\sqrt{n}).$$

So in order for the two terms to be asymptotically negligible, we also need $h\sqrt{n} \rightarrow \infty$. Similar

analysis on choice of rate for h can also be found in the literature, see e.g., Chapter 15, Koenker et al. [2017].

We now state the large sample asymptotic result using $\hat{\theta}_{new}$ in the bootstrap procedure of Algorithm 4.3.1. With the previous defined new QR coefficient estimator:

$$\sqrt{n}\hat{\Sigma}_{new}^{-1/2}(\tau) \left(\hat{\theta}^{*(b)}(\tau) - \hat{\theta}_{new}(\tau) \right) \Rightarrow \mathbf{B}_{p+1}(\tau), \text{ in probability;}$$

in addition, $\hat{\Sigma}_{new}(\tau) \xrightarrow{P} \Sigma(\tau)$ uniformly for $\tau \in [0, 1]$.

Algorithm 4.3.1 can be used to perform linear hypothesis testing on θ of the following form:

$$H_0 : \mathbf{R}\theta(\tau) - \mathbf{r} = \Psi(\tau) \quad \text{vs.} \quad H_1 : \mathbf{R}\theta(\tau) - \mathbf{r} \neq \Psi(\tau) \quad (4.3.5)$$

with $\mathbf{R} \in \mathbb{R}^{q \times p+1}$, $q \leq p+1$, and $\mathbf{r} \in \mathbb{R}^q$; $\Psi(\tau)$ is a known continuous function. Such hypothesis is studied in Koenker and Xiao [2002] and Chernozhukov and Fernández-Val [2005]. Under H_0 , $\sqrt{n}\mathbf{R}(\hat{\theta}(\tau) - \theta(\tau)) = \sqrt{n}\mathbf{R}(\hat{\theta}(\tau) - \Psi(\tau) + \mathbf{r})$, which is approximated by the distribution of $\sqrt{n}\mathbf{R}(\hat{\theta}^{*(b)}(\tau) - \hat{\theta}(\tau))$. If we let $d_{1-\alpha}$ be the $1 - \alpha$ quantile of $\sup_{\tau} \|\mathbf{R}(\hat{\theta}^{*(b)}(\tau) - \hat{\theta}(\tau))\|_2$, then $\mathbb{P}_{H_0} \left(\sup_{\tau} \|\mathbf{R}(\hat{\theta}(\tau) - \Psi(\tau) + \mathbf{r})\|_2 \leq d_{1-\alpha} \right) \rightarrow 1 - \alpha$. Therefore, the null hypothesis should be rejected if $\sup_{\tau} \|\mathbf{R}(\hat{\theta}(\tau) - \Psi(\tau) + \mathbf{r})\|_2 > d_{1-\alpha}$.

4.3.2 Model-free bootstrap

The model representation of (4.2.1) suggest that X_t is a Markov process of order p . The work of Pan and Politis [2016b] proposed a general procedure to conduct Model-free bootstrap by utilizing the conditional distribution function as the transformation towards i.i.d.ness, a key ingredient for the model-free bootstrap principle. The conditional CDF in Pan and Politis [2016b] is estimated by the kernel Nadaraya-Watson estimator, which is inefficient when the dimension of the conditioned variable p is large, as the rate of consistency is at $O(h^2 + 1/\sqrt{nh^p})$. See Li and Racine [2006]. As a remedy, we can use the quantile regression estimator, which is consistent

and can achieve higher efficiency at $O(1/\sqrt{n})$ under model(4.2.1).

For inference problem, the following bootstrap algorithm is an alternative to algorithm 4.3.1.

Algorithm 4.3.2. *Model-free bootstrap for inference under (4.2.1).*

1. Use $\hat{\theta}_{new}$ as the QR coefficient estimator. Construct the conditional CDF estimator, $\hat{F}_{t|t-1}(x)$, via Equation (4.2.7). Sequentially estimate the transformed series $\hat{u}_t = \hat{F}_{t|t-1}(X_t)$, $t = p+1, \dots, n$.
2. (bootstrap)
 - (a) Choose $(X_1^*, \dots, X_p^*)$ uniformly from all blocks of size p in the original data $(X_1, \dots, X_n)$.
 - (b) (Model-free) Uniformly sample with replacement $u_{p+1}^*, \dots, u_n^*$ from $\{\hat{u}_t\}_{t=p+1}^n$. Let $X_t^* = \hat{Q}_{t|t-1}(u_t^*)$, $t = p+1, \dots, n$.
 - (c) Step 2(b) is repeated for B times to get B bootstrap samples. Let $\hat{\theta}^{*(b)}$ be the quantile regression estimator on the b^{th} sample $\{X_t^{*(b)}\}_{t=1}^n$.

Remark 4.3.2. Step 2(b) can be replaced by the following alternative:

2(b'): Uniform sample $u_{p+1}^*, \dots, u_n^*$ from the $Unif(0, 1)$ distribution.

Algorithm 4.3.2 with step 2(b') is called the limit model-free bootstrap in Politis [2015], since u^* is sampled from the limit distribution of $\hat{u}_t$. In fact, the limit model-free bootstrap is equivalent to algorithm 4.3.1 under model assumption (4.2.1).

Additionally, we can also use model-free bootstrap to generate one-step ahead prediction intervals. First of all, we introduce the one-step ahead predictor in the following. $\hat{X}_{n+1} = g(\mathbf{Y}_n; \hat{\theta})$ is called a one-step ahead predictor for X_{n+1} if it minimizes

$$\mathbb{E}_{X_{n+1} \sim \hat{F}_{n+1|n}}(\mathcal{L}(X_{n+1}, \hat{X}_{n+1}) | \mathbf{Y}_n)$$

for some loss function $\mathcal{L}$. Common choice for the loss function includes the L^1 and L^2 loss, under which the predictor $\hat{X}_{n+1}^{L_1} = \hat{Q}_{n+1|n}(\frac{1}{2})$ and $\hat{X}_{n+1}^{L_2} = \int x d\hat{F}_{n+1|n}(x)$ respectively.

The predictor can be used to construct the predictive root:

$$R_{n+1} = X_{n+1} - \hat{X}_{n+1} \quad (4.3.6)$$

whose (conditional) distribution will be approximated by the bootstrap in the following algorithm.

Algorithm 4.3.3. *Model-free bootstrap for prediction.*

1. Use $\hat{\theta}_{new}$ as the QR coefficient estimator. Construct the conditional CDF estimator $\hat{F}_{t|t-1}(x)$ via Equation (4.2.7). Sequentially estimate the transformed series $\hat{u}_t = \hat{F}_{t|t-1}(X_t)$, $t = p+1, \dots, n$.
2. Based on the estimation and observations, construct the 1-step ahead predictor $\hat{X}_{n+1} = g(\mathbf{Y}_n; \hat{\theta})$.
3. (bootstrap)
 - (a) Choose $(X_1^*, \dots, X_p^*)$ uniformly from all blocks of size p in the original data $(X_1, \dots, X_n)$.
 - (b) (Model-free) Uniformly sample with replacement $u_{p+1}^*, \dots, u_n^*$ from $\{\hat{u}_t\}_{t=p+1}^n$. Let $X_t^* = \hat{Q}_{t|t-1}(u_t^*)$, $t = p+1, \dots, n$.
 - (c) Repeat the above step for B times to get B bootstrap samples. Let $\hat{\theta}^{*(b)}$ be the quantile regression estimator on the b^{th} sample $\{X_t^{*(b)}\}_{t=1}^n$. Calculate the b^{th} predictor $\hat{X}_{n+1}^{*(b)} = g(\mathbf{Y}_n, \hat{\theta}^{*(b)})$.
 - (d) Sample from $\hat{F}_{n+1|n}$ B times to generate $X_{n+1}^{*(b)}$, $b = 1, \dots, B$. Calculate $R_{n+1}^{*(b)} = X_{n+1}^{*(b)} - \hat{X}_{n+1}^{*(b)}$. Let $\hat{L}^*(\alpha/2)$ and $\hat{R}^*(\alpha/2)$ be the left and right $\alpha/2$ quantile for the bootstrap predictive roots $R_{n+1}^{*(b)}$, $b = 1, \dots, B$.

4. The $1 - \alpha$ prediction interval for X_{n+1} is constructed by

$$\left(\hat{X}_{n+1} + \hat{L}^*(\alpha/2), \hat{X}_{n+1} + \hat{R}^*(\alpha/2) \right).$$

Algorithm 4.3.3 creates bootstrap replicas for the predictive root R_{n+1} , which involves resampling both X_{n+1}^* and $\hat{X}_{n+1}^*$. In this way, the full randomness in the predictive root is accounted for. In Chapter 1, it is shown that bootstrap procedures of this form satisfy the so-called pertinence property, a term coined by Politis [2015] that attempts to explain the better coverage phenomena of bootstrap comparing with naïve methods, such as using the relative quantiles of the fitted conditional CDF.

4.4 Simulations

4.4.1 Asymptotic performance of bootstrap for inference

This section mainly focuses on demonstrating the bootstrap validity for inference problem via simulation. We consider the following quantile autoregressive processes of order 1. Let $\{X_t\}_{t=1}^N$ be time series samples drawn from model (4.2.1) with

$$\begin{aligned} \theta_0(\tau) &= \sin(\tau - 0.5); \text{ and} \\ \theta_1(\tau) &= \sqrt{\tau}. \end{aligned} \tag{4.4.1}$$

We use this true model and the bootstrap algorithm 4.3.2 to generate $\{(\hat{\theta}^{(b)}, \hat{\theta}^{*(b)})\}_{b=1}^B$ with $B = 500$. Note that $\{\hat{\theta}^{*(b)}\}_{b=1}^B$ are generated via bootstrap based on a fixed time series sample. Then, We plot the empirical cumulative distribution functions(eCDF) of $\sqrt{n}(\hat{\theta}(\tau_0) - \theta(\tau_0))$ as well as $\sqrt{n}(\hat{\theta}^*(\tau_0) - \hat{\theta}_{new}(\tau_0))$, so as to demonstrate the validity of bootstrap via closeness of the two functions. We also provide p -values of the Kolmogorov-Smirnov goodness-of-fit tests.

In our experiments, we fix $p = 1$; choose $\tau_0 = 0.5$, $n \in \{1000, 5000, 10000\}$, and let $h = \frac{1}{4n^{1/4}}$. The $(\sqrt{n}$ -scaled) eCDF of $\sqrt{n}(\hat{\theta}(\tau_0) - \theta(\tau_0))$ reflects the distribution in real world, while the eCDF of $\sqrt{n}(\hat{\theta}^*(\tau_0) - \hat{\theta}_{new}(\tau_0))$ approximates the bootstrap distribution.

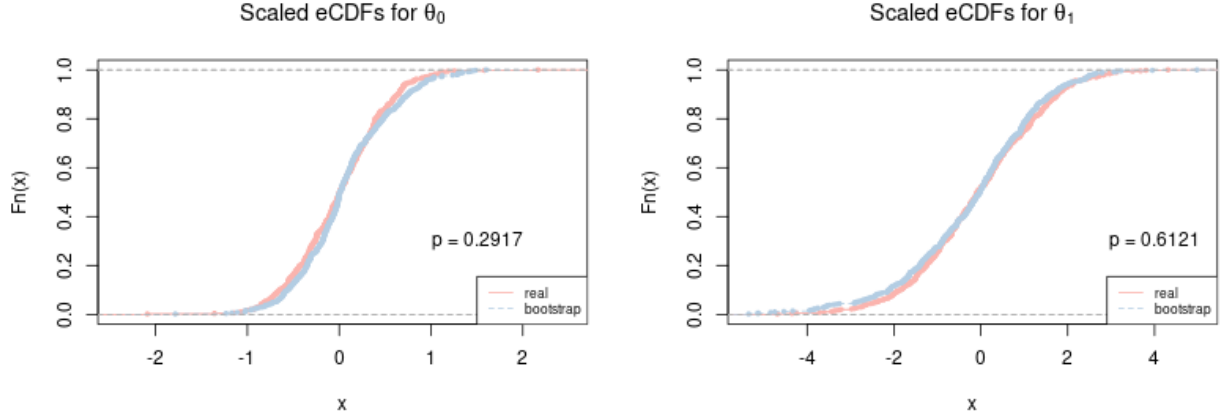


Figure 4.1: Approximation performance of the bootstrap. ($n = 1000$)

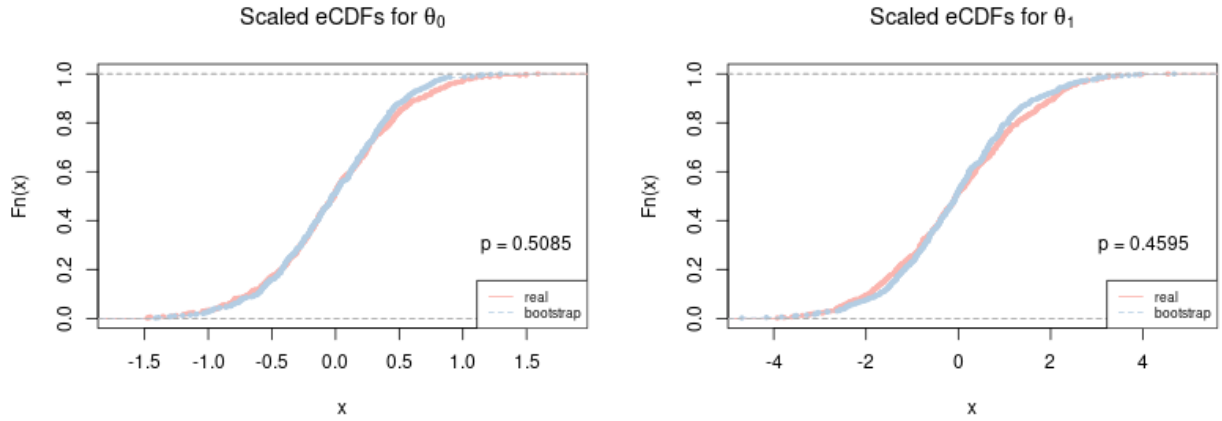


Figure 4.2: Approximation performance of the bootstrap. ($n = 5000$)

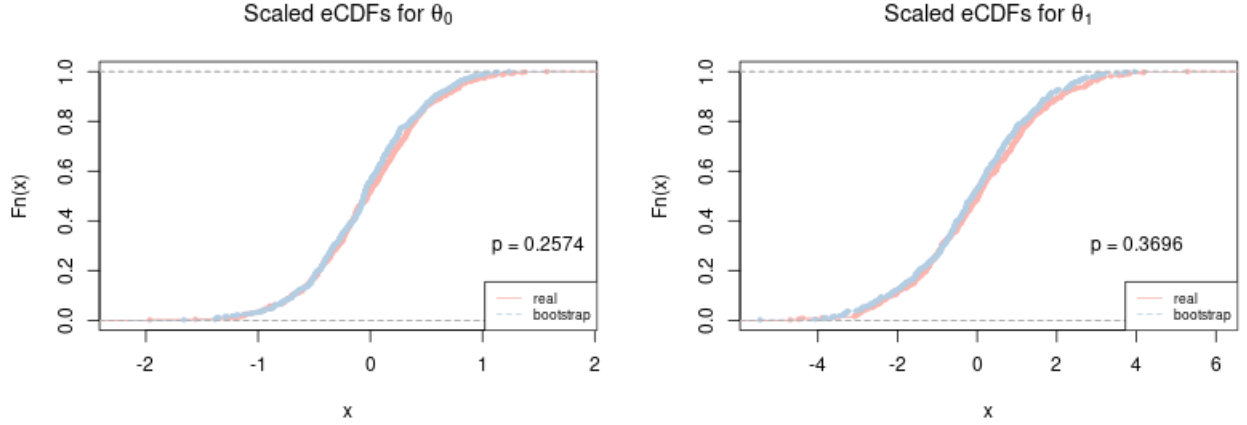


Figure 4.3: Approximation performance of the bootstrap. ($n = 10000$)

As figures 4.1 to 4.3 show, the eCDF in the bootstrap world converges to that in the real world as $n \rightarrow \infty$, and both converge to a Gaussian CDF. Hence, bootstrap validity holds for conducting inference with respect to $\theta(\tau)$.

4.4.2 Coverage performance of prediction intervals

This section aims at demonstrating the superior performance of algorithm 4.3.3 comparing with the naïve quantile regression method. In particular, we evaluate the coverage probability of the model-free bootstrap PI as well as PI based on the estimated quantiles for 1-step ahead prediction.

Let $\{X_t\}_{t=1}^T$ be a realization from model (4.4.1), T_0 denotes the size of past observations based on which prediction intervals are generated. The empirical coverage(eCVR) of a size $(1 - \alpha) \times 100\%$ PI $\hat{C}_{1-\alpha}$ for $\{X_t\}_{t=1}^T$ is evaluated as

$$\text{eCVR} = \frac{1}{T - T_0} \sum_{t=T_0}^{T-1} \mathbb{I} \left(X_{t+1} \in \hat{C}_{1-\alpha}(\{X_k\}_{k=t-T_0+1}^t) \right).$$

We choose $T = T_0 + 1000$. The following table provides eCVR results for various choices of T_0 .

Table 4.1: Empirical coverage performance of 1–step ahead PIs between different algorithms.

Methods	$T_0 = 100$	$T_0 = 200$	$T_0 = 500$	$T_0 = 1000$
MFB- L_1	95.2	95.6	96.3	96.0
MFB- L_2	94.8	94.6	95.0	96.1
QR	93.3	94.2	93.4	95.5

4.5 Acknowledgement

Chapter 4, in full, is a version of the paper “Bootstrap Methods for Quantile Autoregressive Time Series”, Yiren Wang, Dimitris N. Politis. It is currently being prepared for publication in a major statistics journal. The dissertation author is the primary investigator and author of this material.

Appendix A

MFB under misspecification

This appendix studies the behavior of MFB of Chapter 2 under model misspecification. Although model 2.4.1 is fairly non-restrictive, it can still be further generalized, such as when f is no longer assumed monotone. It was conjectured that the MFB algorithm for the sample mean will still work, and can be proved so by utilizing the concept of companion process in Kreiss et al. [2011]. However, our investigation of this approach here shows that the assumption of monotonicity for f is necessary. In brief, a non-monotone transfer function causes the limiting variance in the bootstrap CLT to be generally different from that in the CLT for the sample mean in the original probability space.

Let $Y_t = f(W_t)$ where W_t is a standard Gaussian process with autocovariance $\gamma_W(k)$, and f is a non-monotone transform function such that $\mathbb{E}(Y_t) = 0$. Denote $Z_t = H^{-1}(Y_t)$, where $H(\cdot) = F_Y^{-1} \circ \Phi(\cdot)$. We have that $Z_t \sim \mathcal{N}(0, 1), \forall t$. Contrary to previous paper, $\{Z_t\}$ is no longer a Gaussian process. Let $\gamma_Z(k)$ denote the autocovariance of Z_t .

Let $\tilde{Z}_t$ be a Gaussian process with mean 0 and autocovariance $\gamma_Z(k)$, and $\tilde{Y}_t = H(\tilde{Z}_t)$ be the companion process of Y_t , which has the same dependence structure than the series Y_t^* produced by MFB. Assuming $f \in L^2(\mathbb{R}, e^{-x^2/2}dx)$, then so is H and $\tilde{Y}_t \in L^2$.

We have that H can be expanded in terms of Hermite polynomials $\{H_q\}_{q \geq 1}$ which is an orthonormal basis in $L^2(\mathbb{R}, \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx)$. More precisely, let $H = \sum_{q \geq 1} a_q H_q$. Let σ_∞^2 be the limiting variance of CLT for Y_t with respect to sample mean. We have that

$$\sigma_\infty^2 = \sum_{k \in \mathbb{Z}} \gamma_Y(k) = \sum_{k \in \mathbb{Z}} \text{Cov}(H(Z_0), H(Z_k))$$

With $H = \sum_{q \geq 1} a_q H_q$, the above expression can be further expanded as:

$$\begin{aligned} \sigma_\infty^2 &= \sum_{k \in \mathbb{Z}} \text{Cov}\left(\sum a_q H_q(Z_0), \sum a_q H_q(Z_k)\right) \\ &= \sum_{k \in \mathbb{Z}} \sum_{i,j=1}^{\infty} a_i a_j \text{Cov}(H_i(Z_0), H_j(Z_k)) \\ &= \sum_{k \in \mathbb{Z}} \left[\sum_{q=1}^{\infty} a_q^2 \text{Cov}(H_q(Z_0), H_q(Z_k)) + \sum_{i \neq j} a_i a_j \text{Cov}(H_i(Z_0), H_j(Z_k)) \right] \end{aligned} \quad (\text{A.0.1})$$

Similarly, for the companion process, we have

$$\tilde{\sigma}_\infty^2 = \sum_{k \in \mathbb{Z}} \left[\sum_{q=1}^{\infty} a_q^2 \text{Cov}(H_q(\tilde{Z}_0), H_q(\tilde{Z}_k)) + \sum_{i \neq j} a_i a_j \text{Cov}(H_i(\tilde{Z}_0), H_j(\tilde{Z}_k)) \right] \quad (\text{A.0.2})$$

Given that $\{\tilde{Z}_t\}$ is a standard Gaussian process, by the orthogonality of Hermite polynomials, we have that:

$$\text{Cov}(H_i(\tilde{Z}_0), H_j(\tilde{Z}_k)) = \begin{cases} q! a_q^2 \gamma_Z(k)^q & i = j = q \\ 0 & i \neq j \end{cases} \quad (\text{A.0.3})$$

Note that this does not hold for Z_t since joint normality does not hold for it.

Thus simplifying equation (A.0.2) into

$$\tilde{\sigma}_{\infty}^2 = \sum_{q=1}^{\infty} a_q^2 q! \sum_{k \in \mathbb{Z}} \gamma_Z(k)^q \quad (\text{A.0.4})$$

Comparing to what appeared at the end of equation (A.0.2), there are mainly two differences. Firstly, the off diagonal sum of covariances is 0 for the companion process, while it is in general not the case for the original process Y_t , unless Z_t is a Gaussian process.

Secondly, the diagonal sum for the original process is

$$\sum_{q=1}^{\infty} a_q^2 \sum_{k \in \mathbb{Z}} \text{Cov}(H_q(Z_0), H_q(Z_k))$$

Which is also only equal to

$$\sum_{q=1}^{\infty} a_q^2 q! \sum_{k \in \mathbb{Z}} \gamma_Z(k)^q$$

under the condition that Z_t is a Gaussian process.

Therefore, we conclude that the limiting variances do not equal in general. Only when Z_t is a Gaussian process does equality hold, this requires f to be a monotone function.

Bibliography

- R. J. Adler. *An introduction to continuity, extrema, and related topics for general Gaussian processes*, volume Volume 12 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1990.
- A. M. Alonso, D. Peña, and J. Romo. Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference*, 100(1):1 – 11, 2002.
- J. E. Angus. The probability integral transform and related results. *SIAM Review*, 36(4):652–654, 1994.
- M. A. Arcones and B. Yu. Central limit theorems for empirical and u-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7(1):47–71, 1994.
- S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv:2104.00673*, 2021.
- I. Berkes, S. Hörmann, and J. Schauer. Asymptotic results for the empirical process of stationary sequences. *Stochastic Processes and their Applications*, 119(4):1298 – 1324, 2009.
- I. Berkes, S. Hörmann, and J. Schauer. Split invariance principles for stationary processes. *Ann. Probab.*, 39(6):2441–2473, 2011.
- P. Bertail, C. Haefke, D. N. Politis, and H. White. Subsampling the distribution of diverging statistics with applications to finance. *Journal of Econometrics*, 120(2):295–326, 2004.
- A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.
- P. Breuer and P. Major. Central limit theorems for non-linear functionals of gaussian fields. *Journal of Multivariate Analysis*, 13(3):425–441, 1983.
- P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer New York, New York, NY, 1991.
- P. Bühlmann. Sieve bootstrap for time series. *Bernoulli*, 3(2):123–148, 1997.
- J. Chen and D. N. Politis. Optimal multi-step-ahead prediction of arch/garch models and novas transformation. *Econometrics*, 7(3), 2019.

- V. Chernozhukov and I. Fernández-Val. Subsampling inference on quantile regression processes. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, 67(2):253–276, 2005.
- V. Chernozhukov, K. Wüthrich, and Y. Zhu. Distributional conformal prediction. *arXiv:1909.07889*, 2019.
- D. R. Cox. Some Problems Connected with Statistical Inference. *The Annals of Mathematical Statistics*, 29(2):357 – 372, 1958.
- S. Das and D. N. Politis. Nonparametric estimation of the conditional distribution at regression boundary points. *The American Statistician*, 74(3):233–242, 2020.
- S. Das and D. N. Politis. Predictive inference for locally stationary time series with an application to climate data. *Journal of the American Statistical Association*, 116(534):919–934, 2021.
- Z. Drmac, M. Omladic, and K. Veselic. On the perturbation of the cholesky factorization. *SIAM J. Matrix Anal. Appl.*, 15(4):1319–1332, 1994.
- B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- S. Geisser. *Predictive Inference: An Introduction*. Chapman and Hall, New York, 1993.
- R. Giacomini, D. N. Politis, and H. White. A warp-speed method for conducting monte carlo experiments involving bootstrap estimators. *Econometric Theory*, 29(3):567–589, 2013.
- J. C. Hull and A. D. White. Value at risk when daily changes in market variables are not normally distributed. *The Journal of Derivatives*, 5(3):9–19, 1998.
- S. Hörmann and P. Kokoszka. Weakly dependent functional data. *Ann. Statist.*, 38(3):1845–1884, 2010.
- C. Jentsch and D. N. Politis. Covariance matrix estimation and linear process bootstrap for multivariate time series of possibly increasing dimension. *Ann. Statist.*, 43(3):1117–1140, 2015.
- O. Kallenberg. *Foundations of Modern Probability*. Springer, New York, 1997.
- R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.
- R. Koenker and Z. Xiao. Inference on the quantile regression process. *Econometrica*, 70(4):1583–1612, 2002.
- R. Koenker and Z. Xiao. Quantile autoregression. *Journal of the American Statistical Association*, 101(475):980–990, 2006.
- R. Koenker, V. Chernozhukov, X. He, and L. Peng, editors. *Handbook of quantile regression*. CRC Press, Boca Raton, 2017.

- J.-P. Kreiss and E. Paparoditis. *Bootstrap for Time Series: Theory and Applications*. Springer New York, 2020.
- J.-P. Kreiss, E. Paparoditis, and D. N. Politis. On the range of validity of the autoregressive sieve bootstrap. *Ann. Statist.*, 39(4):2103–2130, 2011.
- H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17(3):1217–1241, 1989.
- S. Lahiri. *Resampling Methods for Dependent Data*. Springer, New York, 2003.
- J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):71–96, 2014.
- J. Lei, J. Robins, and L. Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.
- J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
- Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2006.
- W. Liu and W. B. Wu. Asymptoticis of spectral density estimates. *Econometric Theory*, 26(4): 1218–1245, 2010.
- S. Mase. Some theorems on normality-preserving transformations. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 39(2):186–190, 1977.
- T. S. McElroy and D. N. Politis. *Time Series: A First Course with Bootstrap Starter*. Chapman and Hall/CRC Press, Boca Raton, 2019.
- T. L. McMurry and D. N. Politis. Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, 31(6):471–482, 2010.
- T. L. McMurry and D. N. Politis. High-dimensional autocovariance matrices and optimal linear prediction. *Electron. J. Statist.*, 9(1):753–788, 2015.
- J. Neyman and H. Jeffreys. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937.
- L. Pan and D. N. Politis. Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. *Journal of Statistical Planning and Inference*, 177:1 – 27, 2016a.
- L. Pan and D. N. Politis. Bootstrap prediction intervals for markov processes. *Computational Statistics and Data Analysis*, 100:467 – 494, 2016b.

- J. K. Patel. Prediction intervals - a review. *Communications in Statistics - Theory and Methods*, 18(7):2393–2465, 1989.
- A. Patton. Chapter 16 - copula methods for forecasting multivariate time series. In G. Elliott and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 2 of *Handbook of Economic Forecasting*, pages 899–960. Elsevier, 2013.
- A. Patton, D. N. Politis, and H. White. Correction to automatic block-length selection for the dependent bootstrap by d. politis and h. white. *Econometric Reviews*, 28(4):372–375, 2009.
- D. N. Politis. Model-free versus Model-based Volatility Prediction. *Journal of Financial Econometrics*, 5(3):358–359, 2007.
- D. N. Politis. Model-free model-fitting and predictive distributions. *Test*, 22:183–221, 2013.
- D. N. Politis. *Model-Free Prediction and Regression: A Transformation-Based Approach to Inference*. Springer, New York, 2015.
- D. N. Politis and J. P. Romano. A general resampling scheme for triangular arrays of α -mixing random variables with application to the problem of spectral density estimation. *Ann. Statist.*, 20(4):1985–2007, 1992.
- D. N. Politis and J. P. Romano. Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis*, 16(1):67–103, 1995.
- D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer New York, New York, NY, 1999.
- S. I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York, 2007.
- J. Robins and L. Wasserman. Conditioning, likelihood, and coherence: A review of some foundational concepts. *Journal of the American Statistical Association*, 95(452):1340–1346, 2000.
- Y. Romano, E. Patterson, and E. J. Candès. Conformalized quantile regression, 2019.
- M. Rosenblatt. Remarks on a multivariate transformation. *Ann. Math. Statist.*, 23(3):470–472, 1952.
- M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences*, 42(1):43–47, 1956.
- S. Rosset and R. J. Tibshirani. From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 115(529):138–151, 2020.

- G. Samorodnitsky and M. S. Taqqu. *Stable Non-Gaussian Random Processes*. Routledge, New York, 1994.
- G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.
- B. W. Silverman and G. A. Young. The bootstrap: To smooth or not to smooth? *Biometrika*, 74(3):469–479, 1987.
- E. Snelson, Z. Ghahramani, and C. E. Rasmussen. Warped gaussian processes. In *Advances in Neural Information Processing Systems 16*, pages 337–344. MIT Press, 2004.
- R. J. Tibshirani, R. F. Barber, E. J. Candes, and A. Ramdas. Conformal prediction under covariate shift. *arXiv:1904.06019*, 2019.
- A. W. Van Der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
- V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005.
- L. Wang and D. N. Politis. Asymptotic validity of bootstrap confidence intervals in nonparametric regression without an additive model. *Electronic Journal of Statistics*, 15(1):392–426, 2021a.
- Y. Wang and D. N. Politis. Model-free bootstrap for a general class of stationary time series. *Bernoulli*, 2021b.
- A. G. Wilson and Z. Ghahramani. Copula processes. In *Advances in Neural Information Processing Systems(NIPS)*, volume 23. Curran Associates, Inc., 2010.
- M. Wolf and D. Wunderli. Bootstrap joint prediction regions. *Journal of Time Series Analysis*, 36(3):352–376, 2015.
- W. B. Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005a.
- W. B. Wu. On the bahadur representation of sample quantiles for dependent sequences. *Ann. Statist.*, 33(4):1934–1963, 2005b.
- W. B. Wu and M. Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, 19(4):1755–1768, 2009.
- Y. Zhang, J. Liu, Z. Pan, and D. N. Politis. Estimating transformation function. *Electronic Journal of Statistics*, 13(2):3095–3119, 2019.