6-24-2021

Essays on Statistical Issues in Finance

Haitao Huang

ESSAYS ON STATISTICAL ISSUES IN FINANCE

BY

HAITAO HUANG

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree
Of
Doctor of Philosophy
in the Robinson College of Business
Of
Georgia State University

GEORGIA STATE UNIVERSITY
ROBINSON COLLEGE OF BUSINESS
2021

ACCEPTANCE

This dissertation was prepared under the direction of the HAITAO HUANG Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business Administration in the J. Mack Robinson College of Business of Georgia State University.

Richard Phillips, Dean

DISSERTATION COMMITTEE

Dr. Liang Peng (Chair)
Dr. Stephen H. Shore
Dr. Ajay Subramanian
Dr. Lei Jiang (External – Tsinghua University)

ABSTRACT

ESSAYS ON STATISTICAL ISSUES IN FINANCE

BY

HAITAO HUANG

JUNE 24, 2021

Committee Chair: Dr. Liang Peng

Major Academic Unit: Risk Management and Insurance

Empirical finance has growingly relied on statistical methods to draw inferences. Such finance applications require tailoring the methods to particular problems, especially when the underlying assumptions are violated in the data. This dissertation studies the development and application of statistical methodologies to address empirical problems in the contexts of empirical asset pricing, household finance and investments.

The dissertation consists of four chapters. The first chapter gives an overview of the empirical problems and associated statistical issues for three different finance settings: stock return predictability, house price comovement and mutual fund performance. It also briefly outlines the main contribution of this dissertation in each setting. The second chapter develops a robust methodology of unit root testing and statistical inference for autoregressive processes when the errors are heteroscedastic and heavy-tailed. Applications of the robust test demonstrate that some commonly used financial ratios for stock return predictability are highly persistent with unit roots. The third chapter introduces a new nonparametric framework for estimating and testing comovements among U.S. regional home prices. Comovements are found to be strong in housing prices of four U.S. states, but there is little empirical support for asymmetric tail dependence. The fourth chapter comprehensively studies the bootstrap inference problem in fund performance evaluation. It shows the inadequate size and power properties of two existing bootstrap tests and develops the theory for a valid bootstrap Hotelling's T-squared test. The new bootstrap test, applied in a sequential testing procedure, identifies a small set of skilled funds. Skilled funds are more engaged in active management and hold stocks with higher expected anomalous returns.

ACKNOWLEDGEMENTS

I am immensely indebted to my advisor, Professor Liang Peng, for completion of this dissertation and my Ph.D. study, for guiding me on research in financial econometrics, for encouraging and pushing me to be a better scholar, and for the invaluable help with my career. For everything he has done that makes me where I am now, I am incredibly thankful. I am grateful to my dissertation committee, Professor Stephen Shore, Professor Ajay Subramanian and Professor Lei Jiang, for their relentless guidance and support. I would also like to thank Vincent W. Yao, Stefanos Orfanos, Alejandro Del Valle, Vikas Agarwal, Scott Murray and many other faculty members in the Robinson College of Business for their teachings and mentorship. Department staffs and Ph.D. cohorts make this journey much easier – I thank their friendship, help and kindness. I acknowledge the generous financial support from GSU Risk Management Foundation, Center for the Economic Analysis of Risk and the S.S. Huebner Foundation for Insurance Education.

I dedicate this dissertation to my father, Hong'en Huang, and my mother, Xinzhen Yu.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS vi

LIST OF TABLES x

LIST OF FIGURES xii

CHAPTER 1 INTRODUCTION 1

1.1 Stock Return Predictability 4

1.1.1 Empirical Problems and Statistical Issues 4

1.1.2 Contributions 5

1.2 House Price Comovement 6

1.2.1 Empirical Problems and Statistical Issues 6

1.2.2 Contributions 8

1.3 Mutual Fund Performance 9

1.3.1 Empirical Problems and Statistical Issues 9

1.3.2 Contributions 11

CHAPTER 2 ROBUST INFERENCE FOR AN AR PROCESS REGARDLESS OF FINITE OR INFINITE VARIANCE GARCH ERRORS 12

2.1 Introduction 13

2.2 Methodologies and Main Results 19

2.2.1 Assumptions 20

2.2.2 Empirical Likelihood Based Unit Root Test 21

2.2.3 Robust Estimation and Inference for Stationary AR Processes . . 30

2.2.4 Extension to GARCH Errors 32

2.3 Simulation Study 33

2.3.1 Unit Root Test 33

2.3.2 Estimation and Inference for Stationary AR Processes 37

2.4 Applications 38

2.4.1 Unit Root Testing in Predictive Variables 38

2.4.2 Inference for HKD/USD Exchange Rate 39

2.5 Conclusion 40

2.6 Proofs 41

CHAPTER 3 COMOVEMENTS AND ASYMMETRIC TAIL DEPENDENCE IN STATE HOUSING PRICES IN THE US . **64**

3.1 Introduction 65

3.2 Methodologies 69

3.2.1 AR-GARCH Estimation 69

3.2.2 Comovement Estimation and Hypothesis Tests 70

3.3 Results 71

3.4 Conclusions 73

Appendix 3.A Measuring Comovements 76

Appendix 3.B AR-GARCH Estimation 77

Appendix 3.C Comovement Estimation and Hypothesis Tests 78

Appendix 3.D Effects of Data Revision on Comovements 80

Appendix 3.E Value-at-Risk of Portfolios 81

Appendix 3.F Additional Tables and Figures 81

CHAPTER 4 BOOTSTRAP ANALYSIS OF MUTUAL FUND PERFORMANCE **94**

4.1 Introduction 94

4.2 Existing Bootstrap Methods 102

4.2.1 Measuring Fund Performance 105

4.2.2 Assessing Fund Performance 106

4.2.3 Size and Power Properties of the Kosowski et al. (2006) Bootstrap Test 108

4.2.4 Theoretical Difficulty of the Fama and French (2010) Bootstrap Test 114

4.3 Methodology 117

4.3.1 Bootstrap Hotelling's T-squared Test In the Ideal Setting 117

4.3.2 Bootstrap Hotelling's T-squared Test In A Practical Setting . . . 119

4.3.3 Sequential Testing and Thresholding 123

4.4 Empirical Applications 126

4.4.1 Mutual Fund Performance 126

4.4.2 Characteristics of Different Fund Groups 128

4.5 Conclusion 130

Appendix 4.A Theoretical Derivations 140

4.A.1 Proofs of Theorems 4.1–4.3 140

4.A.2 Proofs of Theorems 4.4–4.6 147

4.A.3 Proofs of Theorems 4.7 and 4.8 150

Appendix 4.B Simulation Studies 153

4.B.1 Empirical Size and Power of Existing Bootstrap Tests 153

4.B.2 Empirical Size and Power of Bootstrap Hotelling's T-squared Test 158

Appendix 4.C Additional Results for Empirical Applications 161

4.C.1 Data 161

4.C.2 Additional Results Based on $\widetilde{HT}$ Test 163

Bibliography **179**

LIST OF TABLES

2.1 Empirical size of several unit root tests for AR processes with ARCH errors 58

2.2 Empirical size and power of several unit root tests for AR processes with GARCH errors 59

2.3 Empirical comparison of several estimators for AR parameters in AR processes with ARCH errors 61

2.4 Empirical coverage probabilities of the empirical likelihood confidence region for AR parameters in AR processes with ARCH errors . . . 62

2.5 Unit root tests for monthly financial ratios of stock return predictability during the period 1953/01–2016/12 62

2.6 Unit root tests for monthly financial ratios of stock return predictability during the period 1976/01–2016/12 62

2.7 Estimation and inference for log-returns of the daily HKD/USD exchange rate 63

3.1 Summary statistics of estimates from AR(3)-GARCH(1,1) models 74

3.2 Test results for asymmetric tail dependence based on Eqs. (3.3) and (3.4) 74

3.3 Test results for asymmetric tail dependence based on Eqs. (3.5) and (3.6) 75

3.A.1 Test results for the difference in interstate housing price comovements between two state pairs based on residuals 90

3.A.2 Test results for the difference in interstate housing price comovements between two state pairs based on the original HPI series 90

3.A.3 Test results for effects of data revision on interstate housing price comovements based on Eq. (3.A.2) 91

3.A.4 Test results for effects of data revision on interstate housing price co-movements based on Eq. (3.A.3) 91

3.A.5 Test results for effects of data revision on interstate housing price co-movements based on Eq. (3.A.4) 92

3.A.6 Test results for effects of data revision on interstate housing price co-movements based on Eq. (3.A.5) 92

3.A.7 Value-at-Risk (VaR(p)) and DV(p) for equally weighted portfolios of house prices 93

3.A.8 Diversification benefits of equally weighted portfolios of house prices 93

4.1 Summary statistics of the cross section of mutual fund returns and alternative performance evaluation tests 134

4.2 Mutual fund characteristics based on HT test 135

4.A.1 Empirical size of bootstrap tests under normal fund residuals . . . 167

4.A.2 Empirical size of bootstrap tests under moderately skewed fund residuals 168

4.A.3 Empirical size of bootstrap tests under heavily skewed fund residuals 169

4.A.4 Empirical power of existing bootstrap tests 170

4.A.5 Empirical size of existing bootstrap tests under cross-sectional dependence 171

4.A.6 Empirical size and power of bootstrap Hotelling's T-squared test . 172

4.A.7 Stock holding characteristics 173

4.A.8 Mutual fund characteristics based on $\widetilde{HT}$ test 175

LIST OF FIGURES

2.1 A large-sample comparison of the asymptotic validity of several unit root tests for model (2.1) with infinite variance GARCH errors. . 55

2.2 Residuals of financial ratios in the post-war period. 56

2.3 Hill estimates for residuals of financial ratios in the post-war period. 56

2.4 Residuals of financial ratios in the period after the oil shock recession. 57

2.5 Hill estimates for residuals of financial ratios in the period after the oil shock recession. 57

3.A.1 Hill's estimators for quarterly changes of HPI (1975:Q2 - 2017:Q1) 82

3.A.2 Autocorrelation functions of AR(1)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1) 82

3.A.3 Autocorrelation functions of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1) 83

3.A.4 Autocorrelation functions of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1) 83

3.A.5 Hill's estimators of AR(3)-GARCH(1,1) model residuals for the quarterly changes of HPI (1975:Q2 - 2017:Q1) 84

3.A.6 Autocorrelation functions of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1) 84

3.A.7 Autocorrelation functions of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1) 85

3.A.8 Hill's estimators of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1) 85

3.A.9 Upper tail dependence ($\hat{\gamma}^{+}_{i|j,e}$) and lower tail dependence ($\hat{\gamma}^{-}_{i|j,e}$) for AR(3)-GARCH(1,1) model residuals by SWQMELE 86

3.A.10 Upper tail dependence ($\hat{\gamma}^{+}_{i|j,X}$) and lower tail dependence ($\hat{\gamma}^{-}_{i|j,X}$) for the original HPI series 87

3.A.11 Upper tail dependence ($\hat{\gamma}^{*+}_{i|j,e}$) and lower tail dependence ($\hat{\gamma}^{*-}_{i|j,e}$) for AR(3)-GARCH(1,1) model residuals by SWQMELE 88

3.A.12 Upper tail dependence ($\hat{\gamma}^{*+}_{i|j,X}$) and lower tail dependence ($\hat{\gamma}^{*-}_{i|j,X}$) for the original HPI series 88

3.A.13 Upper tail dependence ($\hat{\gamma}^{*+}_{i|j,X}$) and lower tail dependence ($\hat{\gamma}^{*-}_{i|j,X}$) for portfolios of equally weighted original HPI series in two states . . 89

4.1 The cross section of monthly mutual fund returns. 132

4.2 Bootstrap Hotelling's T-squared test for fund selection – HT test. 133

4.A.1 Empirical accuracy of the sequential testing procedure for selecting skilled funds. 165

4.A.2 Bootstrap Hotelling's T-squared test for fund selection – $\widetilde{HT}$ test. 166

CHAPTER 1

Introduction

This dissertation is dedicated to developing statistical methodologies to understand some empirical problems in finance research. In classical areas ranging from asset pricing to household finance to active asset management, financial economists have increasingly embraced more sophisticated statistical tools to extract information from data to test empirical hypotheses and address economic questions. As a prime example, the mutual fund performance literature has evolved from simply looking at the histogram of fund alphas and statistical significance of individual t-statistics in Jensen (1968) to applying a multiple testing approach to control for false discoveries in Barras, Scaillet, and Wermers (2010). Lending further support to this burgeoning trend, as reviewed in Goldstein, Jiang, and Karolyi (2019) and Goldstein, Spatt, and Ye (2021), is the applications of machine learning and big data techniques in emerging fields such as financial technology and big data finance. The proliferation of data and rapid development of statistics have arguably shifted the research landscape in the finance profession. This dissertation is uniquely situated at the interface of statistics and finance. It tackles the challenges of validating and applying statistical approaches rigorous to distinctive properties of financial data and suited to the specific economic question.

The dissertation begins with robust inference for financial ratios relevant for stock return predictability in Chapter 2. It is onerous to design an efficient predictability test as the asymptotic property of the test depends on the properties of predictive variables. In particular, these variables are typically highly persistent and driven by heteroscedastic and heavy-tailed innovations. Chapter 2 provides a robust and efficient inferential framework for financial ratios following from possibly heavy-tailed AR-GARCH processes, including a system of unit root tests and weighted least-squares estimation for the stationary case. The methodology is applied to test the persistence of monthly financial ratios, confirming that several commonly used ratios are unit root (non-stationary).

Chapter 3 is concerned with estimating and testing comovements in regional house prices. Characterizations of house price comovements have important implications for the risk management and valuation of mortgage-related securities. Existing approaches either require restrictive parametric assumptions to estimate comovements or inefficient methods to make inference. The chapter first introduces new measures of comovements that are more coherent and interpretable than the existing one in literature. It then provides very flexible nonparametric procedures to estimate comovements and conduct formal hypothesis test of asymmetry in comovements. Empirically, the new measures and estimates reveal strong evidence of both upper-tail and lower-tail comovements among several state housing prices. On the contrary, the asymmetry test suggests very weak evidence of different degrees of comovements between the two extreme tails or between two state pairs.

Motivated by the debate over whether any mutual funds are skilled, Chapter 4 systematically studies how to apply the bootstrap technique to separate skill from luck in mutual funds. The chapter provides compelling insights into the inadequate size and power properties of existing bootstrap approaches in two influential finance studies. The main reason is that existing methods adopt an unconventional test statistic and fail to account for the unique features of mutual fund return data, in particular, a large cross-sectional dimension relative to a small time-series dimension and the majority of funds having negative alphas. Armed with the theoretical insights, the chapter further advances a valid bootstrap test for independent fund residuals and extends the test to a more realistic setting where fund returns are serially correlated and cross-sectionally dependent. The bootstrap test is then implemented in a sequential testing procedure to select skilled mutual funds. The proposed fund evaluation approach suggests that a minority of mutual funds have skills to deliver positive returns to investors, thus reconciling the opposing evidence in prior literature. This result shows that the debate over mutual fund performance could be an artifact of inadequate statistical methods in previous studies. Further analysis of portfolios formed by the selected funds indicates that skilled funds and unskilled funds differ dramatically both in fund attributes and stock holdings.

The dissertation highlights how rigorous methodological developments can lead to additional empirical insights and sometimes drastically different economic conclusions. The main message is that opportunities for and challenges with deploying statistical tools both abound in financial economics. On the one hand, the quest in classical areas and revolution in new frontiers provide fertile grounds for statistical methods to facilitate credible investigation and improve empirical understanding in finance research. The adoption of state-of-the-art techniques could lead to large economic gains where more primitive methods are inadequate. This has been well evidenced across such areas as empirical asset pricing and corporate finance (Loughran and McDonald, 2011; Barillas and Shanken, 2018; Chordia et al., 2020; Gu et al., 2020; Feng et al., 2020; Li et al., 2021). On the other hand, although it has become increasingly important to import statistical tools to finance research, many of the assumptions based on which traditional statistical theories work are violated in financial data. It is thus essential to examine the properties of data at hand and assess the applicability of statistical methods in the specific context. Equally important, it is unequivocal to bear in mind the economic problems to be addressed in designing or adapting new methods. Echoing the messages conveyed in some recent developments in the finance literature, such as Harvey and Liu (2020a) and Giglio, Liao, and Xiu (2020), ignoring these statistical issues could be economically costly in empirical research.

In the following, I provide an overview of the empirical problems and underlying statistical issues for the three finance fields and briefly outline the contribution of this dissertation to each individual field.

1.1 Stock Return Predictability

1.1.1 Empirical Problems and Statistical Issues

A central task in finance is to predict stock returns (equity premia) using available public information, such as financial ratios and macroeconomic variables. Despite significant research devoted to this task over the past decades, whether stock returns can be predicted remains highly debated. Some earlier studies, including Fama and French (1988) and Campbell and Shiller (1991), argue that long-horizon stock returns are highly predictable. Campbell and Yogo (2006) develop an efficient predictability test and find evidence for predictability with several financial variables, including dividend–price ratio, the smoothed earnings–price ratio, the short rate and the long-short yield spread. Ang and Bekaert (2007) empirically show that dividend yields together with the short rate predict excess returns at short horizons and the predictive power vanishes at long horizons. Welch and Goyal (2008) investigate both in-sample and out-of-sample predictability of predictors in linear regressions and bring disheartening evidence that the linear models have poor predictive ability and unstable performance. Lettau and Van Nieuwerburgh (2008) and Cochrane (2008) seek to reconcile the debate, who both defend return predictability. Using a novel methodology accounting for the time-series properties of financial variables, Kostakis, Magdalinos, and Stamatogiannis (2015) document short-horizon predictability, which disappears in more recent data. They further find that the predictability weakens as the predictive horizon is increased. Rapach, Ringgenberg, and Zhou (2016) argue that short interest, when aggregated across firms and detrended, is a very strong predictor of future stock returns. McLean and Pontiff (2016) suggest that investors learn about mispricing from academic research, which reduces stock return predictability.

Among many others, a common challenge confronting this literature is that the predictability is inferred from predictive regressions and thus depends on the reliability of subsequent hypothesis tests. In particular, a valid predictability test depends on the uncertainty in the degree of persistence in predictive variables. When the predictive variable is highly persistent with autoregressive roots extremely close to unity or being unity, conventional t-test leads to invalid inference. Recognzing this challenge, Campbell and Yogo (2006) propose a new test by taking into account the degree of persistence with a Bonferroni procedure and constructing the confidence intervals for both the predictive regression parameter and the persistence parameter. However, this method excludes the case where the predictor is stationary and only applies for a univariate predictor. As a response, Kostakis, Magdalinos, and Stamatogiannis (2015) develop a testing procedure that is robust to the degree of predictor's persistence and can test the joint predictive ability of multivariate predictors. The robustness of this test is achieved via an instrumental variable estimation (IVX) procedure. More recently, Liu et al. (2019) observe that even the construction of instrumental variables depends heavily on the degree of persistence, which could reduce the power of the IVX test. They further build the difference of the predicting variable into the simple linear predictive model and propose a unified predictability test regardless of the properties of the predicting variable. In short, statistical inference robust to the time-series properties of predictive variables is indispensable for empirically testing the predictability of stock returns. An immediate first step for designing such tests is then to investigate the properties of the predictive variables themselves, especially whether they are stationary or unit-root.

1.1.2 Contributions

Chapter 2 makes several contributions for robust inference on financial time series relevant to stock return predictability. On the methodological side, the paper develops a robust unit root test for an autoregressive process with heavy-tailed and heteroscedastic errors. The test applies the empirical likelihood method to weighted score equations and attains a chi-squared limiting distribution. The paper also provides a robust inference procedure when the AR process is stationary based on the weighted least-squares. In both cases, the robustness is achieved through using a data-driven weighting function such that the estimation and inference are valid without requiring prior knowledge on the moments of the errors. The efficiency comes from utilizing empirical likelihood to construct the test statistic or confidence region and conduct inference based on the asymptotic chi-squared distribution. Simulation studies confirm that the proposed methods perform well in finite samples. On the empirical side, the paper examines the time-series properties of several commonly used predictive variables, such as dividend-price ratio, dividend yield and term spread. As these variables exhibit heteroscedasticity and heavy-tails, the robust testing procedure is applied to pretest their degrees of persistence. The test does not reject the unit root null hypothesis for any of the variables. Although the paper does not directly test the predictability of these financial ratios, it is informative for properly formulating the models for such a study. For instance, the strong evidence of non-stationarity supports the adjustment to the linear predictive regression in Liu et al. (2019), as the traditional regression would imply that stock returns are nonstationary when the predictive variables are unit-root.

1.2 House Price Comovement

1.2.1 Empirical Problems and Statistical Issues

Comovements in regional house prices, or the phenomenon that house prices move in tandem across geographic areas, especially during extreme market upswings and downswings, is an important stylized fact in the housing market (Glaeser and Gyourko, 2006; Del Negro and Otrok, 2007; Shiller, 2007; Cotter et al., 2011; Kallberg et al., 2014; Landier et al., 2017; Cohen et al., 2021). The recent Great Recession is associated with the simultaneous booms and busts of the housing market. The four Sand States (Arizona, California, Florida, Nevada) with similar housing cycles are acutely impacted, accounting for over 40% of mortgage foreclosures initiated nationally.[1] The defining role of the housing bubble in triggering the financial crisis prompts academics and policymakers to take a second look at housing

[1] Olesiuk, S.M. and K. Kalser. (2009). *The Sand States: Anatomy of a Perfect Housing Market Storm.* FDIC Quarterly. https://www.fdic.gov/analysis/quarterly-banking-profile/fdic-quarterly/2009-vol3-1/vol3-1-sand-states.pdf

price comovements.

Comovements in regional housing markets have important implications for both policy design and risk management. Davis and Heathcote (2005) demonstrate that residential investment leads the business cycle. Leamer (2007, 2015) argue that housing is the leading precursor of the US business cycle. As regional recessions tend to comove across states and propagate into larger contraction (Hamilton and Owyang, 2012), comovements in regional housing markets must be studied to understand the transmission of business cycles and to make effective national policies. On the other hand, comovements in regional housing prices is particularly relevant for assessing the risk of mortgage-backed securities, such as collateralized debt obligations (CDOs), that package mortgages from different locations into tranches. Prior to the housing crisis, CDOs were perceived by rating agencies and investors alike as providing diversification benefits. The misbelief then was that house prices in noncontiguous regions are unlikely to experience simultaneous large declines. This underestimation of comovements led to substantial losses in CDOs during the housing crisis as house prices across states plunged around the same time. Coval, Jurek, and Stafford (2009) systematically explains how the pooling of mortgages and issuance of tranches result in significant exposure to and rating errors in the default risks, which are responsible for the rise and fall of CDOs. Therefore, accurate modeling and understanding of comovements are of paramount importance economically.

The modeling of comovements and evaluation of CDO ratings mainly relied on the Gaussian copula before the housing crisis, but its undesirable feature of tail independence precludes the interdependence of house prices in extreme housing downturns and underestimates the magnitude of comovements. Due to the inadequacy of the Gaussian copula to accommodate tail dependence, Zimmer (2012) proposes alternative copula specifications to model comovements in housing prices. Zimmer (2012) estimates upper and lower tail dependence through fitting parametric copulas (such as Clayton and Gumbel copulas) as the joint distributions and a parametric family of marginal distributions (normal and student t) to house price indices after filtering out the AR(1)-GARCH(1,1) components. A major drawback in this approach is that its success relies heavily on correct specifications of both the copula and marginal distributions. To overcome the restrictive parametric specifications, Ho, Huynh, and Jacho-Chávez (2016) propose to adopt a nonparametric copula estimator and nonparametric smoothing distribution estimators for the marginals and construct bootstrap confidence intervals to assess the asymmetry in tail dependence. While more flexible and robust, the confidence intervals used in their inference are too wide to reach a convincing conclusion surrounding the asymmetric tail dependence. A further concern for both studies lies with the measure of comovements, defined as conditional probabilities of house price changes in one location in relation to another. The measure, when used to test asymmetric tail dependence, is inherently biased in the event of asymmetric house price distributions. An earlier study by Croux, Forni, and Reichlin (2001) discusses the empirical relevance of defining appropriate comovement measures for economic variables.

1.2.2 Contributions

Chapter 3 makes the following contributions. Methodologically, it improves the AR-GARCH estimation procedure by taking into account the heavy-tailed feature of house price indices. It also provides a novel set of comovement measures by correcting the bias in the previous measure along with a very flexible nonparametric estimator. The measures are defined based on either original house price changes or filtered price changes with AR-GARCH estimation. Formal statistical tests for different degrees of tail dependence are further proposed based on distance-based test statistics and bootstrapped critical values. Empirically, the new comovement measures indicate that extreme house price movements exhibit strong upper-tail and lower-tail dependence among the Sand States, and lower-tail comovement dominates in most cases when the original house price changes are used. There is little evidence that the comovements in market upturns and downturns are significantly different, except for some neighboring states such as Arizona, California and Nevada. The asymmetric tail dependence is only revealed when using the original series. Finally, the chapter argues that the measures based on original observations instead of residuals are more advantageous for portfolio management. The state pairs with asymmetric tail dependence are associated with larger diversification benefits when built into portfolios.

1.3 Mutual Fund Performance

1.3.1 Empirical Problems and Statistical Issues

A chief question sought after by a large finance literature is whether actively managed mutual funds can create value to their clients. The literature stems from the foundational work of Jensen (1968), who finds that mutual funds on average do not outperform even gross of management expenses. The empirical evidence that mutual fund managers in general perform below the market is viewed as in favor of the efficient markets hypothesis in Fama (1965, 1970). This belief is further reinforced by the seminal work of Carhart (1997), who concludes that there is little evidence to support the existence of skilled or informed mutual fund managers. In a rational model, Berk and Green (2004) argue that, in equilibrium, the expected returns net of fees for investors are zero due to the competitive allocation of capital to mutual funds and decreasing returns to scale in managerial ability. The conventional wisdom has been challenged by a number of studies suggesting evidence of mutual fund skill. Berk and van Binsbergen (2015) and Cremers et al. (2019) provide extensive reviews on this literature. Despite this voluminous research, the debate over mutual fund performance remains unresolved for many. If actively managed mutual funds cannot add value to investors, the large active mutual fund industry would be only puzzling to comprehend. As of April 2021, a total of nearly 3000 US equity mutual funds manage over 20 trillion dollars of assets, accounting for around 40% of the total number and total net assets of US mutual funds.[2] On the contrary, the existence of skilled mutual fund managers would imply that these agents have access to information to allow them to earn returns above the markets, violating the market efficiency models.

[2] *Release: Trends in Mutual Fund Investing, April 2021.* (2021, May 27). ICI. Retrieved June 5, 2021, from https://www.ici.org/research/stats/trends_04_21

A major challenge in investigating fund performance is that skill is not directly observable and has to be estimated in a risk-return framework, thus plagued by estimation noises (luck). It is essential to distinguish whether a fund is genuinely skilled or appears to be so due to luck. To separate skill from luck in fund performance, Kosowski et al. (2006) and Fama and French (2010) apply bootstrap techniques to examine the cross-sectional distribution of fund's net alphas by simulating hypothetical funds with ex ante zero alpha. The bootstrap procedures in the two studies differ over how to handle the dependence in fund returns. The former is a standard fund-by-fund residual bootstrap whereas the latter bootstraps the factors and fund residuals simultaneously in the cross-section. In essence, the bootstrap approaches are used to conduct hypothesis tests on the statistical significance of the extreme alphas. The two studies arrive at quite opposing conclusions on the extent to which skill exists. Kosowski et al. (2006) conclude that a substantial number of fund managers have superior stock-picking abilities. To the contrary, Fama and French (2010) find little evidence of outperformance. Although the two intuitive bootstrap approaches have since gained wide popularity in the financial economics literature, the strikingly different conclusions from the studies fueled many to probe into comparing the two bootstrap tests. Recently, Harvey and Liu (2020a) find through a simulation study that the Fama and French approach lacks test power to detect skilled funds and suggest it as helping to reconcile the different findings in Kosowski et al. (2006) and Fama and French (2010). Harvey and Liu (2020b) further compare a variety of bootstrap implementations in terms of test size and power using simulations from mutual fund data. They argue that the Kosowski et al. (2006) approach is substantially over-sized while the Fama and French (2010) approach is under-sized and recommend adjusting the Fama and French (2010) approach for future research. While the studies by Harvey and Liu provide useful perspectives on thinking about the pitfalls of bootstrap approaches, they do not give theoretical insights into statistical inference based on bootstrap. A formal analysis is thus warranted for guiding future research in applying the bootstrap approach to the evaluation of fund performance.

1.3.2 Contributions

Chapter 4 contributes to the mutual fund performance literature in several aspects. It first shows that the two bootstrap tests have inadequate properties when applied to mutual fund data. Both tests have size distortions as the number of funds is much larger than the typical time-series length of monthly returns. They could suffer from low power due to the presence of a significant number of negative-alpha funds. The theoretical and analytical insights are confirmed by Monte Carlo simulations. The chapter further validates a zero-alpha test using Hotelling's T-squared statistic with bootstrap calibration. This new test is extended to the practical setting where fund residuals are serially correlated and cross-sectionally dependent. Implemented with a sequential testing procedure, the new bootstrap test is applied to select skilled funds. Empirical analyses indicate the existence of a small minority of skilled funds. Skilled funds are more engaged in active management and hold stocks with significantly different characteristics associated with higher expected anomalous returns.

CHAPTER 2

Robust Inference for an AR Process Regardless of Finite or Infinite Variance GARCH Errors[1]

Abstract

Statistical inference in finance often depends on certain moment conditions such as finite or infinite variance, yet it is practically challenging to disentangle these conditions. This article develops a class of unified unit root tests for AR(1) models and a weighted least squares estimator along with robust inference for a stationary AR(r) model regardless of finite or infinite variance GARCH errors. The inferential framework applies the empirical likelihood method to some weighted score equations without estimating the GARCH errors. In contrast to extant unit root tests relying on bootstrap or subsampling methods to approximate critical values, the proposed unit root tests can be easily implemented with critical values obtained directly from a chi-squared distribution using the Wilks theorem. Extensive simulation studies confirm the good finite sample performance of the proposed methods before we

[1] This chapter is based on the joint work: Huang, H., Leng, X., Liu, X., & Peng, L. (2020). Unified inference for an AR process regardless of finite or infinite variance GARCH errors. *Journal of Financial Econometrics*, 18(2), 425-470.

illustrate them empirically with financial ratios for stock return predictability and HKD/USD exchange rate returns.

2.1 Introduction

Start with the first-order autoregressive process (AR(1) process)

$$X_t = \phi X_{t-1} + e_t \quad \text{for} \quad t = 1, \cdots, n, \tag{2.1}$$

where e_t's are random errors with zero mean. Testing for a unit root (i.e., $H_0 : \phi = 1$) has a longstanding tradition in econometrics; see, for example, the recent review paper by Xiao (2014). The recent research efforts in predictive regression highlight the complications in deriving efficient tests of stock return predictability when the predictive variable is highly persistent as a local-to-unity process (Phillips and Lee, 2013; Kostakis, Magdalinos, and Stamatogiannis, 2015). A classical unit root test for model (2.1) is based on the least-squares estimator (LSE) of ϕ, and the asymptotic theory of such a test depends on whether e_t's are independent or dependent and whether e_t has finite or infinite variance. We first overview some existing studies on unit root testing, with particular focus on how the limit is affected by the dependence and tail heaviness of e_t.

- When $\{e_t\}$ is a stationary sequence with finite variance, Phillips (1987) derived the asymptotic distribution of a t-test based on the LSE of ϕ when $\phi = 1 - d/n$, which is non-normal.

- When e_t's are independent with infinite variance, Chan and Tran (1989) derived the asymptotic distribution of the LSE under the unit root null hypothesis, which has a non-normal limit distinct from the case of finite variance. Chan (1990) further derived the limit for the case of near unit root. Since the limit depends on the tail index of e_t and tabulating critical values is impossible, Jach and Kokoszka (2004) developed unit root tests using the subsampling method to approximate the null distribution of test statistics, but the tests critically hinge on a practical choice of the subsample size and can be quite over-sized in finite samples. Samarakoon and Knight (2009) considered Dickey-Fuller-type tests with infinite variance innovations based on M-estimators.

- When $\{e_t\}$ is a linear stationary sequence with infinite variance, Phillips (1990) derived the asymptotic distribution of the LSE, which is nonnormal and depends on the tail index of e_t.

- When $\{e_t\}$ is a stationary sequence with barely infinite variance in the sense that $\mathsf{E}\,|e_t|^\delta < \infty$ for any $\delta \in (0,2)$, Kourogenis and Pittis (2008) proposed a unit root test with a pivotal limit, but this test is not applicable to most cases of infinite variance, where $\mathsf{E}|e_t|^\delta = \infty$ for some $\delta \in (0,2)$.

- When $e_t = \sum_{j=0}^{\infty} c_j \sigma_{t-j} \epsilon_{t-j}$ with $\mathsf{E}\,\epsilon_t^4 < \infty$ and σ_t non-stochastic and strictly positive, Cavaliere and Taylor (2007) proposed a unit root test with its limiting null distribution depending on the behavior of σ_t, and the test requires simulation methods to operate since, unlike the test in Phillips and Perron (1988), critical values cannot be tabulated.

- When the dependence of $\{e_t\}$ follows some heteroscedastic time series model, Cavaliere and Taylor (2009) proposed a unit root test with an asymptotic distribution not requiring finite variance of e_t, but depending on both the standardized error process and the conditional volatility process. By assuming the independence between the conditional volatility process and the standardized error process, which excludes the well-known stationary GARCH models (defined in Equation (2.2) below) for e_t, Cavaliere and Taylor (2009) justified the applicability of a wild bootstrap scheme to obtain critical values.

- When $\{e_t\}$ is a GARCH(1,1) process, Chan and Zhang (2010) derived the asymptotic distribution of the LSE for ϕ under the unit root null hypothesis, which is nonnormal and different for the cases of finite and infinite variance of e_t.

- When $e_t = \sum_{i=0}^{\infty} \gamma_i \varepsilon_{t-i}$, ε_t has infinite variance and is in the domain of attraction of a stable law with index between zero and two, Cavaliere et al. (2018) derived the asymptotic null distributions of two augmented Dicky-Fuller (ADF) test statistics for unit root inference, which depend on the unknown stable law index, and provided a sieve wild bootstrap algorithm to approximate the critical values under the assumption that ε_t has a symmetric distribution.

It is clear from the above theoretical developments that dependence and infinite variance in innovations complicate a unit root test. Although the distinction between finite and infinite variance is typically ambiguous to practitioners, a unit root test heavily depends on how to handle extreme values when infinite variance innovations may be present. When a parametric distribution family is fitted to data, one can test for finite variance via parameter estimation. However, nonparametric test for finite variance is extremely challenging. In practice, one often assumes that the underlying distribution is heavy-tailed and employs tail index estimation in extreme value theory such as the pervasively used Hill estimator in Hill (1975). To further motivate the unit root test robust against dependence and infinite variance of innovations proposed in the present paper, we refer the readers to the real data analysis on financial ratios for stock return predictability in Section 2.4. There we employ the Hill estimator to estimate the tail indexes for several predictive variables using monthly data during the periods 1953–2016 and 1976–2016. The Hill estimates indicate that the data after 1976 may have infinite variance. Therefore, for robustness in theory and applicability in practice, there is certainly a need to develop valid unit root inference procedures without a prior on finite or infinite variance in innovations.

There have been some efforts in response to search for robust unit root tests suited to heavy-tailed errors. For unit root testing with independent e_t's, the m-out-of-n bootstrap method based on the LSE can be employed to obtain critical values under infinite variance (Ferretti and Romo, 1996). However, this test is not powerful and has difficulty in choosing the bootstrap sample size m, which satisfies $m = m(n) \to \infty$ and $m/n \to 0$ as $n \to \infty$. When e_t is a linear process with symmetrically distributed innovations, a sieve wild bootstrap approach can be validly applied in ignorance of whether the innovations display finite or infinite variance (Cavaliere et al., 2018). The sieve wild bootstrap ADF unit root test allows for AR(∞) errors, but cannot deal with cases involving GARCH errors. The restrictive symmetry assumption also hinders its applicability in practice. A common limitation these tests share is the necessity of computational intensive bootstrap simulations to compute critical values to conduct inference.

The first contribution of this paper is to provide an empirical likelihood based unit root test when $\{e_t\}$ follows from a GARCH(p, q) model in Engle (1982) and Bollerslev (1986), defined as

$$e_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i e_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{2.2}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$ $(i = 1, 2, ..., p)$, $\beta_j \geq 0$ $(j = 1, 2, ..., q)$ and $\{\epsilon_t\}$ is a sequence of independent and identically distributed random variables with zero mean and unit variance. The proposed unit root test applies the empirical likelihood method to some weighted score equations and attains a chi-squared limiting null distribution. As a consequence, it can be implemented without estimating the unknown parameters in e_t's and without using a bootstrap or simulation method to obtain critical values. Note that under some conditions given in Section 2.2, $\{e_t\}$ in model (2.2) may have infinite variance. Hence, the test is robust against infinite variance innovations.

The test for model (2.1) is extended to models where deterministic components may be present, in particular, the model with a constant term, i.e.,

$$X_t = \mu + \phi X_{t-1} + e_t \quad \text{for} \quad t = 1, \cdots, n, \tag{2.3}$$

and the model with both a constant and a time trend, i.e.,

$$X_t = \mu + \gamma t + U_t, \quad U_t = \phi U_{t-1} + e_t \quad \text{for} \quad t = 1, \cdots, n. \tag{2.4}$$

It is well-known that the test for a unit root based on the LSE has a different limit with a different rate of convergence depending on whether μ is zero or nonzero (Phillips, 1987). Importantly, unit root tests under our empirical likelihood framework for these models all have the Wilks-type asymptotic null distribution for both finite and infinite variance innovations without knowledge of whether $\mu = 0$. Unlike Cavaliere et al. (2018), the test developed in this paper does not impose the symmetry assumption on the innovations and allows for straightforward computation of the exact critical values from the chi-squared null distribution, but depends on the order in the GARCH errors and needs a pseudo sample when the AR(1) model includes a deterministic component. In cases of models (2.1) and (2.3), further extensions are made to deal with higher-order AR models admitting the ADF form.

Once the unit root hypothesis is rejected, one may be interested in inference for a stationary autoregressive process of order r (AR(r) process) driven by possibly infinite variance innovations, without using a bootstrap or simulation method to obtain critical values for interval estimation or hypothesis testing. Before providing such an inference procedure which fully exploits the heavy tails of the innovations, we again first review extant results on inferring a stationary AR(r) process. The recurring feature of this literature is that the limit theory is contingent on the dependence structure and heavy tails in innovations.

For a stationary AR(r) process with independent errors, the limiting distribution of the LSE is normal (nonnormal) for finite (infinite) variance errors (Davis and Resnick, 1985, 1986). Extension to infinite variance linear processes is given by Cavaliere et al. (2016). Statistical inference regardless of finite or infinite variance requires some computationally intensive method such as m-out-of-n bootstrap or wild bootstrap to obtain critical values when the errors are independent or follow a linear process.

For a stationary AR(r) process with G-GARCH noises, which includes GARCH(p, q) model as a special case, Zhang and Ling (2015) derived the asymptotic distribution of the LSE for the coefficients in the AR(r) part, which depends on whether the tail index of the G-GARCH noises belongs (or is equal) to $(0, 2)$, 2, $(2, 4)$, 4 or $(4, \infty)$. In particular, the LSE is inconsistent when the tail index is less than 2, which is different from the case of independent errors. Hence, the LSE cannot be used in inference for ϕ regardless of the tail heaviness of e_t when $\{X_t\}$ in model (2.3) is stationary with $\{e_t\}$ being a GARCH sequence. Note that, for model (2.3) with $|\phi| < 1$ and e_t's satisfying model (2.2), Lange (2011) showed X_t has the same tail index as e_t.

For a stationary ARMA process with GARCH errors, the asymptotic normality of the quasi-maximum likelihood estimator requires finite fourth moment for both the sequence itself and the standardized errors in the GARCH model, i.e., both $\mathsf{E}e_t^4 < \infty$ and $\mathsf{E}\epsilon_t^4 < \infty$ (Francq and Zakoian, 2004). For a stationary AR process with ARCH errors, Lange et al. (2011) proposed estimators with a normal limit when the noise in the ARCH errors has a symmetric distribution and the ARCH errors have finite variance. For a stationary ARMA process with GARCH errors, Hill (2015) proposed a trimmed estimator with a normal limit when the noise in the GARCH errors, i.e., ϵ_t in model (2.2), has a symmetric distribution and the density of the GARCH error, i.e., the density of e_t, is bounded, and Zhu and Ling (2015) proposed a self-weighted least absolute deviation estimator (SLADE) for the coefficients in the ARMA part without restriction on the moments of GARCH errors and without estimating the unknown parameters in GARCH errors when the innovations in the GARCH model have zero median instead of zero mean. Therefore, in order to employ this SLADE to perform inference for ϕ in model (2.3) regardless of the tail heaviness of e_t, a model transformation is indispensable to change the assumption of zero mean for ϵ_t in model (2.2) to that of zero median. This would be a significant change for skewed data. Some other issues on reparameterization for GARCH sequences are discussed in Fan et al. (2014).

The above review motivates the second contribution of this paper, which is a robust inference procedure for the parameters in a stationary AR process allowing for infinite variance GARCH errors. Specifically, since the LSE may be inconsistent as showed in Zhang and Ling (2015), we propose a weighted least squares estimator (WLSE) and an empirical likelihood method to construct a confidence region for AR parameters, which work without restriction on the moments of ARCH errors and under the assumption of a little more than finite first moment for GARCH errors. Like Zhu and Ling (2015), the new method does not need to estimate the unknown parameters in the GARCH model; hence it is robust and computationally convenient. But unlike Zhu and Ling (2015), we assume ϵ_t in model (2.2) has zero mean rather than zero median. An empirical comparison shows that the proposed WLSE performs better than the SLADE in Zhu and Ling (2015). Since the proposed estimator has an explicit formula, it requires neither an initial value nor an optimization procedure for implementation unlike the SLADE in Zhu and Ling (2015). We refer the readers to Section 2.3 for details.

We also remark that the proposed empirical likelihood method is totally different from the empirical likelihood inference in Hill and Prokhorov (2016) since we are interested in the AR part without estimating the GARCH part except exploiting the GARCH structure, while Hill and Prokhorov (2016) considered GARCH models rather than AR-GARCH models. Indeed, it is not straightforward to generalize the method in Hill and Prokhorov (2016) to AR-GARCH models, which will require trimming e_t and ϵ_t simultaneously.

We organize this article as follows. Section 2.2 presents the methodologies and main results for the proposed unit root tests via an empirical likelihood method, a weighted least squares estimator and the associated empirical likelihood inference for stationary AR processes. The results of simulation studies and comparison with existing methods in finite samples are summarized in Section 2.3. Applications to several data sets in finance are provided in Section 2.4. Section 2.5 concludes. All proofs are given in Section 2.6.

2.2 Methodologies and Main Results

This section proceeds as follows. Subsection 2.2.1 introduces the basic assumptions. Subsection 2.2.2 expounds the empirical likelihood based unit root test. Subsection 2.2.3 develops the robust estimation and inference for stationary AR processes through the weighted least squares and empirical likelihood approaches. While the results in these two subsections are derived when the AR processes are driven by ARCH errors, we demonstrate in Subsection 2.2.4 that they can be conveniently generalized to the case of GARCH errors.

2.2.1 Assumptions

Define the $(p+q-1)\times(p+q-1)$ matrix

$$\mathbf{A}_t = \begin{pmatrix} \alpha_1\epsilon_t^2+\beta_1 & \beta_2 & \ldots & \beta_{q-1} & \beta_q & \alpha_2 & \alpha_3 & \ldots & \alpha_p \\ 1 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & 0 & 0 & 0 & \ldots & 0 \\ \epsilon_t^2 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1 & 0 \end{pmatrix},$$

and denote the Euclidean norm in $\mathbb{R}^{p+q}$ by $|\cdot|$ and the operator norm for matrix $\mathbf{A}_t$ by $\|\mathbf{A}_t\| = \sup_{|x|=1}|\mathbf{A}_t x|$. Then the Lyapunov exponent for the sequence of random matrices $\{\mathbf{A}_t\}$ is given by

$$\gamma = \inf\left\{\frac{1}{n}\mathsf{E}\left(\ln\|\mathbf{A}_1\cdots\mathbf{A}_n\|\right),\ n\in\mathcal{N}\right\}.$$

When Equation (2.2) holds with $\alpha_0>0$ and $\gamma<0$, it follows from Theorem 3.1 of Basrak et al. (2002) that i) there exists a unique strictly stationary causal solution to Equation (2.2) if $\mathsf{E}\ \ln(\max(|\epsilon_1|,1))<\infty$; ii) $\{e_t\}$ is strongly mixing with geometric rate if ϵ_1 has a density positive in an interval containing zero; iii) e_t has a regularly varying tail under some conditions.

Hence, throughout the paper, we impose the following assumptions for the GARCH(p,q) model in Equation (2.2):

C1. $\alpha_0>0$, $\gamma<0$, and $\mathsf{E}\,|\epsilon_1|^{2+d^*}<\infty$ for some $d^*>0$;

C2. $\mathsf{E}\ \ln(\max(|\epsilon_1|,1))<\infty$;

C3. $\{\epsilon_t\}$ is a sequence of independent and identically distributed random variables with $\mathsf{E}\,\epsilon_t = 0$, $\mathsf{E}\,\epsilon_t^2 = 1$ and a density positive in an interval containing zero.

Note that **C1** still allows $\mathsf{E}e_t^2 = \infty$ and the distribution of ϵ_t can be asymmetric.

2.2.2 Empirical Likelihood Based Unit Root Test

To better appreciate the new methodologies, we first consider the AR(1) model without a constant term in Equation (2.1) driven by ARCH(p) errors, i.e., Equation (2.2) holds with $p \geq 1, q = 0$. In order to construct an interval for ϕ, Chan et al. (2012) proposed to apply the empirical likelihood method in Qin and Lawless (1994) to the weighted score equation

$$\sum_{t=1}^{n}\{X_t - \phi X_{t-1}\}\frac{X_{t-1}}{\sqrt{1+X_{t-1}^2}} = 0$$

when $\mathsf{E}\,e_t^2 < \infty$. We refer to Owen (2001) for an overview of the empirical likelihood method. As argued above, when e_t follows Equation (2.2), we could have $\mathsf{E}\,e_t^2 = \infty$. In this case, the methods in Chan et al. (2012) and Hill et al. (2016) fail. Since we still have $\mathsf{E}\,\epsilon_t^2 < \infty$, we could use another weight to bound σ_t by noting $X_t - \phi_0 X_{t-1} = \sigma_t \epsilon_t$ so as to apply the empirical likelihood method to the independent and identically distributed ϵ_t's, which have finite variance. Here and throughout, ϕ_0 denotes the true value of ϕ. Under $H_0 : \phi_0 = 1$, we could use the simple weight function $1 + \sum_{k=1}^{m}(X_{t-k} - X_{t-k-1})^2$ for some $m \geq p$ to bound σ_t^2 since

$$\begin{aligned}\sigma_t^2 &= \alpha_0 + \textstyle\sum_{k=1}^{p}\alpha_k(X_{t-k} - X_{t-1-k})^2 \\ &\leq \max\{\alpha_0, \alpha_1, \cdots, \alpha_p\}\{1 + \textstyle\sum_{k=1}^{m}(X_{t-k} - X_{t-1-k})^2\}.\end{aligned}$$

This motivates us to consider the following empirical likelihood method.

Put

$$Y_t(\phi) = \{X_t - \phi X_{t-1}\}\frac{X_{t-1}}{\sqrt{(1+X_{t-1}^2)\left\{1 + \sum_{k=1}^{m}(X_{t-k} - X_{t-k-1})^2\right\}}} \quad \text{for } t = 1, ..., n.$$

We then define the empirical likelihood function for ϕ as

$$L(\phi) = \sup\left\{\prod_{t=1}^{n}(np_t) : p_1 \geq 0, ..., p_n \geq 0, \sum_{t=1}^{n}p_t = 1, \sum_{t=1}^{n}p_t Y_t(\phi) = 0\right\}.$$

By the Lagrange multiplier technique, we obtain $p_t = \dfrac{1}{n\{1+\lambda Y_t(\phi)\}}$ and the log empirical likelihood ratio

$$l(\phi) = -2\log L(\phi) = 2\sum_{t=1}^{n}\log\{1+\lambda Y_t(\phi)\},$$

where $\lambda = \lambda(\phi)$ satisfies

$$\frac{1}{n}\sum_{t=1}^{n}\frac{Y_t(\phi)}{1+\lambda Y_t(\phi)} = 0. \tag{2.5}$$

The following theorem shows that the proposed empirical likelihood method gives a unit root test regardless of finite or infinite variance ARCH errors.

Theorem 2.1. *Suppose models (2.1) and (2.2) with $p \geq 1, q = 0$ satisfy conditions **C1**–**C3**. Choose $m \geq p$. Then, under $H_0 : \phi_0 = 1$, we have $l(1) \xrightarrow{d} \chi_1^2$ as $n \to \infty$, where χ_1^2 denotes a chi-squared random variable with one degree of freedom.*

Based on Theorem 2.1, we reject $H_0 : \phi_0 = 1$ at the significance level τ if $l(1) > \chi^2_{1,1-\tau}$, where $\chi^2_{1,1-\tau}$ denotes the $(1-\tau)$th quantile of a chi-squared distribution with one degree of freedom.

Note that the key idea in the above proposed test is to find a proper weight function to bound σ_t. Hence the proposed methodology works for other forms of σ_t as long as such a weight function is available.

To evaluate the power of the above proposed empirical likelihood test, we further assume that

> **C4.** There exist $\delta \in (1, 2]$ and a slowly varying function $L(n)$ (i.e., $L(nx)/L(n) \to 1$ for any $x > 0$ as $n \to \infty$) such that $\frac{\sum_{t=1}^{[ns]} e_t}{n^{1/\delta} L(n)}$ weakly converges to $\tilde{W}_\delta(s)$ in $D([0, 1])$ (the space of functions on $[0, 1]$ which are right-continuous and have left-hand limits, see Billingsley, 1999), where $\{\tilde{W}_\delta(s) : 0 < s \leq 1\}$ is a stable process for $\delta < 2$ and a Gaussian process for $\delta = 2$.

Remark 2.1. *When e_t has a heavy-tailed distribution with index δ, the above condition **C4** is true under some regularity conditions on the stationarity. For example, Theorem 2.1 of*

*Chan and Zhang (2010) shows that conditions **C1**–**C3** imply condition **C4**. Following the proofs of Lemmas 1–3 of Zhang and Ling (2015) in deriving the convergence of $\sum_{t=1}^{n} e_t e_{t-l}$, it is less complicated to derive the convergence of $\sum_{t=1}^{n} e_t$. Hence, we could also show that **C4** holds for G-GARCH errors with the same regularity conditions as in Zhang and Ling (2015).*

Theorem 2.2. *Suppose conditions of Theorem 2.1 and condition **C4** hold. Then under $H_a : \phi_0 = 1 - \frac{d_1}{n^{1/2+1/\delta}L(n)}$ for some constant $d_1 \in \mathbb{R}$, we have*

$$l(1) = \frac{\left\{ sgn(\tilde{J}_\delta(1)) \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{e_t}{\sqrt{1+\sum_{k=1}^{m} e_{t-k}^2}} - d_1 \Delta_2 \int_0^1 |\tilde{J}_\delta(s)|\, ds \right\}^2}{\Delta_1} + o_p(1),$$

where $\Delta_1 = \mathsf{E}\left(\frac{\sigma_t^2}{1+\sum_{k=1}^{m} e_{t-k}^2}\right)$, $\Delta_2 = \mathsf{E}\left(\frac{1}{\sqrt{1+\sum_{k=1}^{m} e_{t-k}^2}}\right)$, $\tilde{J}_\delta(s) = \tilde{W}_\delta(s) - d_1 \bar{d}_1 \int_0^s \tilde{W}_\delta(r) e^{-(s-r)d_1\bar{d}_1}\, dr$, $\bar{d}_1 = \lim_{n\to\infty} \frac{n}{n^{1/2+1/\delta}L(n)}$ and $sgn(x)$ is the sign function.

Remark 2.2. *When $\delta < 2$ or $\delta = 2$ but $\lim_{n\to\infty} L(n) = \infty$, we have $\bar{\delta}_1 = 0$. Note that*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{e_t}{\sqrt{1+\sum_{k=1}^{n} e_{t-k}^2}} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \epsilon_t \frac{\sigma_t}{\sqrt{1+\sum_{k=1}^{m} e_{t-k}^2}} \xrightarrow{d} N(0, \Delta_1).$$

Therefore, the power of the above empirical likelihood unit root test tends to one as $|d_1| \to \infty$.

Remark 2.3. *Let $\hat{\phi} = \sum_{t=1}^{n} X_t X_{t-1} / \sum_{t=1}^{n} X_{t-1}^2$ denote the least squares estimator. When $\{e_t\}$ is a GARCH(1,1) sequence, i.e., Equation (2.2) with $p = q = 1$, Chan and Zhang (2010) showed that $n(\hat{\phi} - \phi_0) = O_p(1)$ regardless of the tail heaviness of e_t. When e_t's are independent, we still have $n(\hat{\phi} - \phi) = O_p(1)$ for either finite variance (Phillips, 1987) or infinite variance (Chan, 1990). Therefore, a unit root test based on $\hat{\phi}$ has a nontrivial power and is strictly less than one only when $\phi_0 = 1 - d_0/n$ regardless of finite or infinite variance. In comparison, as showed in Theorem 2.2, the power of the proposed empirical likelihood unit root test tends to one when $\phi_0 = 1 - d_0/n$ and e_t has infinite variance (i.e., $\bar{d}_1 = 0$ and $d_1 = \infty$ in Theorem 2.2). In other words, the new unit root test is much more powerful than a test based on the least squares estimator when the errors have infinite variance. This is not surprising as the proposed test takes some information on the error structure into account.*

Next, we consider the AR(1) model with a constant term in Equation (2.3) with errors satisfying Equation (2.2) and $p \geq 1, q = 0$. As above, one may apply a similar empirical likelihood method to the following weighted score equations

$$\begin{cases} \sum_{t=1}^{n}\{X_t - \mu - \phi X_{t-1}\}\dfrac{1}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}} = 0, \\ \sum_{t=1}^{n}\{X_t - \mu - \phi X_{t-1}\}\dfrac{X_{t-1}}{\sqrt{(1+X_{t-1}^2)\{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2\}}} = 0. \end{cases} \tag{2.6}$$

It is easy to show that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{X_t - \mu_0 - \phi_0 X_{t-1}\}\frac{1}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}$$

has a normal limit with mean zero and variance $\mathsf{E}\frac{e_t^2}{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}$ under $H_0 : \phi_0 = 1$. Also, when $\phi_0 = 1$, we have $|X_t|/\sqrt{1+X_t^2} \xrightarrow{p} 1$ as $t \to \infty$, which implies that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{X_t - \mu_0 - \phi_0 X_{t-1}\}\frac{X_{t-1}}{\sqrt{1+X_{t-1}^2}\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}$$

converges in distribution to a normal distribution with mean zero and variance $\mathsf{E}\frac{e_t^2}{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}$. Unfortunately, when $\phi_0 = 1$ and $\mu_0 = 0$,

$$\frac{1}{n}\sum_{t=1}^{n}\{X_t - \mu_0 - \phi_0 X_{t-1}\}^2\frac{X_{t-1}}{\sqrt{1+X_{t-1}^2}\{1+\sum_{k=1}^{p}(X_{t-k}-X_{t-k-1})^2\}}$$

does not converge in probability since the normalized $X_{[ns]}$ converges in distribution, where $[\cdot]$ denotes the integer part and $s \in (0, 1]$. Therefore, the joint limit of the two normalized terms in the left hand sides of Equation (2.6) is not bivariate normal, which means that a direct application of empirical likelihood fails to achieve a chi-squared limit, i.e., the Wilks theorem does not hold. As in Li et al. (2014), to solve this difficulty, we employ the idea of adding a pseudo sample and changing the weight function $\sqrt{1+X_{t-1}^2}$ in the second equation to another weight function $\{1+X_{t-1}^2\}^{0.75}$, which has a faster rate of convergence to infinity in the case of unit root. More specifically, we define

$$\tilde{Y}_{t1}(\mu,\phi) = \{X_t - \mu - \phi X_{t-1}\}\frac{1}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}},$$

$$\tilde{Y}_{t2}(\mu,\phi) = \{X_t - \mu - \phi X_{t-1}\}\frac{X_{t-1}}{\{1+X_{t-1}^2\}^{0.75}\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}} + W_t,$$

and $\tilde{\boldsymbol{Y}}_t(\mu,\phi)=(\tilde{Y}_{t1}(\mu,\phi),\tilde{Y}_{t2}(\mu,\phi))^T$, where W_t's are simulated independent random variables from $N(0,\bar{\sigma}^2)$, and $\bar{\sigma}^2>0$ is chosen to be larger than $\mathsf{E}\{\tilde{Y}_{t2}(\mu_0,1)-W_t\}^2$ with μ_0 being the true value of μ. Note that, for a large n, $\mathsf{E}\{\tilde{Y}_{t2}(\mu_0,1)-W_t\}^2$ will be much smaller than $\mathsf{E}\,\tilde{Y}_{t1}^2(\mu_0,1)$ in the case of near unit root or unit root due to the fact that $X_{t-1}/(1+X_{t-1}^2)^{0.75}\xrightarrow{p}0$ as $n\to\infty$. For the choice of $\bar{\sigma}$, a large $\bar{\sigma}$ results in an accurate size, and a small $\bar{\sigma}$ leads to a good power. In our simulation study and data analysis, we choose $\bar{\sigma}=1.5\sqrt{\mathsf{E}\,\tilde{Y}_{t1}^2(\mu_0,1)}$, where $\mathsf{E}\,\tilde{Y}_{t1}^2(\mu_0,1)$ can be estimated by using the weighted least squares estimator $\tilde{\mu}$ for μ via solving

$$\sum_{t=1}^{n}\{X_t-\mu-X_{t-1}\}\frac{1}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}=0. \tag{2.7}$$

The consistency of $\tilde{\mu}$ under $H_0:\phi_0=1$ easily follows from the law of large numbers for martingale differences, the stationarity of $\{e_t\}$ and the fact that $\sigma_t/\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}$ is bounded by a constant uniformly for $t=1,\cdots,n$. The reason for choosing the weight $(1+X_{t-1}^2)^{0.75}$ in $\tilde{Y}_{t2}(\mu,\phi)$ is to have the first term disappear in the unit root case so that W_t dominates. On the other hand, we do not want the first term to disappear too fast for the purpose of power. More detailed explanations can be found in Li et al. (2014). Furthermore, in order to avoid the effect of a random seed in generating W_t's, we use $W_t=\frac{1}{\sqrt{10000}}\sum_{i=1}^{10000}W_{t,i}$ in our simulation study, where the $W_{t,i}$'s are independent random variables from $N(0,\bar{\sigma}^2)$. Based on $\{\tilde{\boldsymbol{Y}}_t(\mu,\phi)\}_{t=1}^n$, we define the empirical likelihood function for (μ,ϕ) as

$$\tilde{L}(\mu,\phi)=\sup\left\{\prod_{t=1}^{n}(np_t):\ p_1\geq 0,\cdots,p_n\geq 0,\ \sum_{t=1}^{n}p_t=1,\ \sum_{t=1}^{n}p_t\tilde{\boldsymbol{Y}}_t(\mu,\phi)=\mathbf{0}\right\}.$$

Since we are interested in ϕ, we consider the profile empirical likelihood function $\tilde{L}^P(\phi)=\max_\mu\tilde{L}(\mu,\phi)$ and put $\tilde{l}(\phi)=-2\log\tilde{L}^P(\phi)$. Note that $\tilde{\mu}$ in solving Equation (2.7) can be employed as an initial value for computing the above profile empirical likelihood function.

The following theorem shows that the proposed profile empirical likelihood method gives a unit root test for $H_0:\phi_0=1$ without restriction on the moments of the errors.

Theorem 2.3. *Suppose models (2.3) and (2.2) with* $p\geq 1, q=0$ *satisfy conditions* ***C1–C3****. Choose* $m\geq p$*. Then, under* $H_0:\phi_0=1$*, we have* $\tilde{l}(1)\xrightarrow{d}\chi_1^2$ *as* $n\to\infty$*.*

As before, a test for $H_0 : \phi_0 = 1$ at the level τ is to reject H_0 when $\tilde{l}(1) > \chi^2_{1,1-\tau}$. A similar power analysis as in Theorem 2.2 is presented in Theorem below, which clearly shows that the power depends on whether $\mu_0 = 0$ or not. For this task, we need the following assumption on the oscillation of a stable process:

> **C5.** There exists $\delta \in (1, 2]$ such that $S_n(s)/n^{1/\delta}$ weakly converges to $\tilde{W}_\delta(s)$ in $D([0,1])$, where $S_n(s) = \sum_{t=1}^{[ns]} e_t$ and $\{\tilde{W}_\delta(s) : 0 < s \leq 1\}$ is a stable process for $\delta < 2$ and a Gaussian process for $\delta = 2$. Moreover, for $\alpha \in (0,1)$, there exists a process $\Gamma_\alpha(s,t)$ such that
>
> $$\sup_{1 \leq s \leq 1, 0 \leq r \leq n^{1-\alpha}} \left| \frac{S_n(s - srn^{\alpha-1}) - S_n(s)}{n^{\alpha/\delta}} - \Gamma_\alpha(s,r) \right| = o_p(1).$$

Theorem 2.4. *Suppose conditions of Theorem 2.3 hold.*

*i) Assume $\mu_0 = 0$ and condition **C5** holds with $\alpha = \delta/(2\delta - 1)$. Then under $H_a : \phi_0 = 1 - \frac{d_1}{n^\alpha}$ for some constant $d_1 > 0$, we have*

$$\tilde{l}(1) = \frac{(n^{-1/2}\sum_{t=1}^n W_t + d_1 d_2 \Delta_2)^2}{\mathsf{E}\, W_1^2} + o_p(1),$$

where

$$d_2 = \left\{\int_0^1 L^*(s)\, ds\right\}\left\{\int_0^1 L^*(s)|L^*(s)|^{-3/2}\, ds\right\} - \int_0^1 |L^*(s)|^{1/2}\, ds$$

with $L^*(s) = -\int_0^\infty e^{-d_1 sr}\, d\Gamma_\alpha(s,r)$, and Δ_2 is defined in Theorem 2.2.

*ii) Assume $\mu_0 \neq 0$ and condition **C4** holds. Then under $H_a : \phi_0 = 1 - \frac{d_1}{n}$ for some constant $d_1 \in \mathbb{R}$, we have*

$$\tilde{l}(1) = \frac{(n^{-1/2}\sum_{t=1}^n W_t + |\mu_0|^{1/2} d_1 d_3)^2}{\mathsf{E}\, W_1^2} + o_p(1),$$

where

$$\begin{aligned} d_3 \;&=\; \frac{\mathsf{E}\int_0^1 \frac{f(s;d_1)}{\sqrt{1+\sum_{k=1}^m(-\mu_0 e^{-d_1 s}+e_{t-k})^2}}\, ds}{\mathsf{E}\int_0^1 \frac{1}{\sqrt{1+\sum_{k=1}^m(-\mu_0 e^{-d_1 s}+e_{t-k})^2}}}\mathsf{E}\int_0^1 \frac{f^{-1/2}(s;d_1)}{\sqrt{1+\sum_{k=1}^m(-\mu_0 e^{-d_1 s}+e_{t-k})^2}}\, ds \\ &\quad -\mathsf{E}\int_0^1 \frac{f^{1/2}(s;d_1)}{\sqrt{1+\sum_{k=1}^m(-\mu_0 e^{-d_1 s}+e_{t-k})^2}}\, ds \end{aligned}$$

with $f(s;d_1) = \{1 - e^{-sd_1}\}/d_1 \geq 0$ for $s \geq 0$.

Remark 2.4. *Theorem 2.4 shows that the power of the proposed empirical likelihood test goes to one as $|d_1| \to \infty$. By noting that $|1-\phi_0|$ in Theorem 2.2 is a smaller order than that in Theorem 2.4 for the case of $\mu_0 = 0$, we conclude that the empirical likelihood method for model (2.1) is more powerful than that for model (2.3), which is not surprising at all since the method for model (2.3) accommodates a nonzero drift in AR processes.*

The above proposed empirical likelihood methods can be extended straightforwardly to an AR(r) model in the so-called ADF form as follows:

$$X_t = \phi X_{t-1} + \sum_{j=1}^{r} \phi_j (X_{t-j} - X_{t-j-1}) + e_t \tag{2.8}$$

and

$$X_t = \mu + \phi X_{t-1} + \sum_{j=1}^{r} \phi_j (X_{t-j} - X_{t-j-1}) + e_t, \tag{2.9}$$

where e_t satisfies model (2.2).

For model (2.8), put $\boldsymbol{\theta} = (\phi_1, \cdots, \phi_r)^T$ and define for $i = 1, \cdots, r$ and $t = 1, \cdots, n$

$$Y^*_{t,1}(\phi, \boldsymbol{\theta}) = \left\{ X_t - \phi X_{t-1} - \sum_{j=1}^{r} \phi_j (X_{t-j} - X_{t-j-1}) \right\} \frac{X_{t-1}}{\sqrt{(1 + X^2_{t-1})\{1 + \sum_{k=1}^{m} (X_{t-k} - X_{t-k-1})^2\}}},$$

$$Y^*_{t,i+1}(\phi, \boldsymbol{\theta}) = \left\{ X_t - \phi X_{t-1} - \sum_{j=1}^{r} \phi_j (X_{t-j} - X_{t-j-1}) \right\} \frac{X_{t-i} - X_{t-i-1}}{\sqrt{1 + \sum_{k=1}^{m} (X_{t-k} - X_{t-k-1})^2}},$$

$$\boldsymbol{Y}^*_t(\phi, \boldsymbol{\theta}) = (Y^*_{t,1}(\phi, \boldsymbol{\theta}), \cdots, Y^*_{t,r+1}(\phi, \boldsymbol{\theta}))^T,$$

and the empirical likelihood function as

$$L^*(\phi, \boldsymbol{\theta}) = \sup \left\{ \prod_{t=1}^{n} (np_t) : p_1 \geq 0, \cdots, p_n \geq 0, \sum_{t=1}^{n} p_t = 1, \sum_{t=1}^{n} p_t \boldsymbol{Y}^*_t(\phi, \boldsymbol{\theta}) = \mathbf{0} \right\}.$$

Again, we consider the profile empirical likelihood function $L^{P*}(\phi) = \max_{\boldsymbol{\theta}} L^*(\phi, \boldsymbol{\theta})$ and put $l^*(\phi) = -2 \log L^{P*}(\phi)$.

Theorem 2.5. *Suppose models (2.8) and (2.2) with $p \geq 1, q = 0$ satisfy conditions **C1**–**C3**. Further assume $\{X_t - X_{t-1}\}$ is a strictly stationary sequence when $\phi_0 = 1$ and condition **C4** holds. Choose $m \geq p + r$. Then, under $H_0 : \phi_0 = 1$, we have $l^*(1) \xrightarrow{d} \chi^2_1$ as $n \to \infty$.*

For model (2.9), put $\boldsymbol{\theta} = (\mu, \phi_1, \cdots, \phi_r)^T$ and define for $i = 1, \cdots, r$ and $t = 1, \cdots, n$

$$\begin{aligned}
\tilde{Y}^*_{t,1}(\phi, \boldsymbol{\theta}) &= \left\{X_t - \mu - \phi X_{t-1} - \sum_{j=1}^{r} \phi_j (X_{t-j} - X_{t-j-1})\right\} \frac{1}{\sqrt{1 + \sum_{k=1}^{m} (X_{t-k} - X_{t-k-1})^2}}, \\
\tilde{Y}^*_{t,2}(\phi, \boldsymbol{\theta}) &= \left\{X_t - \mu - \phi X_{t-1} - \sum_{j=1}^{r} \phi_j (X_{t-j} - X_{t-j-1})\right\} \frac{X_{t-1}}{\{1 + X_{t-1}^2\}^{0.75} \sqrt{1 + \sum_{k=1}^{m} (X_{t-k} - X_{t-k-1})^2}} + W_t, \\
\tilde{Y}^*_{t,i+2}(\phi, \boldsymbol{\theta}) &= \left\{X_t - \mu - \phi X_{t-1} - \sum_{j=1}^{r} \phi_j (X_{t-j} - X_{t-j-1})\right\} \frac{X_{t-i} - X_{t-i-1}}{\sqrt{1 + \sum_{k=1}^{m} (X_{t-k} - X_{t-k-1})^2}}, \\
\tilde{\boldsymbol{Y}}^*_t(\phi, \boldsymbol{\theta}) &= (\tilde{Y}^*_{t,1}(\phi, \boldsymbol{\theta}), \cdots, \tilde{Y}^*_{t,r+2}(\phi, \boldsymbol{\theta}))^T,
\end{aligned}$$

and the empirical likelihood function as

$$\tilde{L}^*(\phi, \boldsymbol{\theta}) = \sup \left\{ \prod_{t=1}^{n} (np_t) : p_1 \geq 0, \cdots, p_n \geq 0, \sum_{t=1}^{n} p_t = 1, \sum_{t=1}^{n} p_t \tilde{\boldsymbol{Y}}^*_t(\phi, \boldsymbol{\theta}) = \mathbf{0} \right\}.$$

As before, we consider the profile empirical likelihood function $\tilde{L}^{P*}(\phi) = \max_{\boldsymbol{\theta}} \tilde{L}^*(\phi, \boldsymbol{\theta})$ and put $\tilde{l}^*(\phi) = -2 \log \tilde{L}^{P*}(\phi)$.

Theorem 2.6. *Suppose models (2.9) and (2.2) with* $p \geq 1, q = 0$ *satisfy conditions* ***C1****–****C3****. Further assume* $\{X_t - X_{t-1}\}$ *is a strictly stationary sequence when* $\phi_0 = 1$*. Choose* $m \geq p + r$*. Then, under* $H_0 : \phi_0 = 1$*, we have* $\tilde{l}^*(1) \xrightarrow{d} \chi^2_1$ *as* $n \to \infty$*.*

Remark 2.5. *As we have to treat* ϕ_i*'s as nuisance parameters, it remains unknown how to extend the proposed methods to the case of* $r = \infty$ *in the above two theorems. Also we fail to derive the asymptotic behavior of* $X_{[ns]}$*, which prevents us from analyzing the power of the proposed tests in Theorems 2.5 and 2.6.*

We now extend the unit root testing framework to the AR(1) model in Equation (2.4) with both a constant and a time trend, which implies that

$$X_t = (\mu - \mu\phi + \phi\gamma) + \gamma(1 - \phi)t + \phi X_{t-1} + e_t.$$

For testing $H_0 : \phi_0 = 1$, as before, we consider weighted scores with respect to $\mu^* := \mu - \mu\phi + \phi\gamma$, $\gamma^* := \gamma(1 - \phi)$ and ϕ. Specifically, define

$$\bar{Y}_{t1}(\mu^*,\gamma^*,\phi)=\{X_t-\mu+\mu\phi-\phi\gamma-\gamma(1-\phi)t-\phi X_{t-1}\}\frac{1}{\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}},$$
$$\bar{Y}_{t2}(\mu^*,\gamma^*,\phi)=\{X_t-\mu+\mu\phi-\phi\gamma-\gamma(1-\phi)t-\phi X_{t-1}\}\frac{t}{n\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}},$$
$$\bar{Y}_{t3}(\mu^*,\gamma^*,\phi)=\{X_t-\mu+\mu\phi-\phi\gamma-\gamma(1-\phi)t-\phi X_{t-1}\}\frac{X_{t-1}}{(1+X_{t-1}^2)^{0.75}\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}}+W_t,$$
$$\bar{\boldsymbol{Y}}_t(\mu^*,\gamma^*,\phi)=(\bar{Y}_{t1}(\mu^*,\gamma^*,\phi),\bar{Y}_{t2}(\mu^*,\gamma^*,\phi),\bar{Y}_{t3}(\mu^*,\gamma^*,\phi))^T,$$

where W_t's are simulated independent random variables from $N(0,\bar{\sigma}^2)$ given before. Then we define the empirical likelihood function as

$$\bar{L}(\mu^*,\gamma^*,\phi)=\sup\left\{\prod_{t=1}^n(np_t):p_1\geq 0,\cdots,p_n\geq 0,\sum_{t=1}^n p_t=1,\sum_{t=1}^n p_t\bar{\boldsymbol{Y}}_t(\mu^*,\gamma^*,\phi)=\mathbf{0}\right\},$$

and consider the profile empirical likelihood function $\bar{L}^P(\phi)=\max_{\mu^*,\gamma^*}\bar{L}(\mu^*,\gamma^*,\phi)$. Put $\bar{l}(\phi)=-2\log\bar{L}^P(\phi)$.

Theorem 2.7. *Suppose models (2.4) and (2.2) with $p\geq 1, q=0$ satisfy conditions **C1**–**C3**. Choose $m\geq p$. Then, under $H_0:\phi_0=1$, we have $\bar{l}(1)\xrightarrow{d}\chi_1^2$ as $n\to\infty$.*

Theorem 2.8. *Suppose conditions of Theorem 2.7 hold.*

*i) Assume $\gamma_0=0$ and condition **C5** holds with $\alpha=\delta/(2\delta-1)$. Then under $H_a:\phi_0=1-\frac{d_1}{n^\alpha}$ for some $d_1>0$, we have*

$$\bar{l}(1)=\frac{\{\frac{1}{\sqrt{n}}\sum_{t=1}^n W_t+d_1d_4\Delta_2\}^2}{\mathsf{E}\,W_1^2}+o_p(1),$$

where

$$\begin{aligned}d_4&=\{4\int_0^1 L^*(s)\,ds-6\int_0^1 sL^*(s)\,ds\}\int_0^1 L^*(s)|L^*(s)|^{-3/2}\,ds\\&\quad+\{-6\int_0^1 L^*(s)\,ds+12\int_0^1 sL^*(s)\,ds\}\int_0^1 sL^*(s)|L^*(s)|^{-3/2}\,ds-\int_0^1|L^*(s)|^{1/2}\,ds,\end{aligned}$$

Δ_2 and $L^*(s)$ are defined in Theorem 2.2 and Theorem 2.4, respectively.

*ii) Assume $\gamma_0\neq 0$ and condition **C4** holds. Then under $H_a:\phi_0=1-\frac{d_1}{n}$ for some $d_1\in\mathbb{R}$, we have $\bar{l}(1)\xrightarrow{d}\chi_1^2$ as $n\to\infty$.*

Remark 2.6. *The above Theorem 2.8 ii) shows that the proposed test has no power under the alternative $H_a:\phi_0=1-d_1/n^\alpha$ when $\gamma_0\neq 0$, which is not surprising because $X_{[ns]}/n\xrightarrow{p}s$*

in this case, and distinguishing the terms t and X_{t-1} becomes impossible. In other words, one should prefer model (2.3) to model (2.4). Also note that models (2.3) and (2.4) are equivalent when $\phi_0 = 1$ and tests for both models have comparable power when $\gamma_0 = 0$.

2.2.3 Robust Estimation and Inference for Stationary AR Processes

When the above unit root hypothesis is rejected, an interesting question is to estimate parameters μ and ϕ consistently. We consider the more general stationary AR(r) model:

$$X_t = \mu + \sum_{j=1}^{r} \phi_j X_{t-j} + e_t. \tag{2.10}$$

Since Zhang and Ling (2015) showed that the LSE may be inconsistent depending on the moments of errors, we consider the following weighted least squares estimator (WLSE) $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\phi}_1, \cdots, \hat{\phi}_r)^T$, which minimizes

$$\sum_{t=1}^{n} \left\{ X_t - \mu - \sum_{j=1}^{r} \phi_j X_{t-j} \right\}^2 \frac{1}{1 + \sum_{k=1}^{m+1} X_{t-k}^2} \tag{2.11}$$

with respect to $\boldsymbol{\theta} = (\mu, \phi_1, \cdots, \phi_r)^T$.

Denote the true value of $\boldsymbol{\theta}$ as $\boldsymbol{\theta_0} = (\mu_0, \phi_{1,0}, \cdots, \phi_{r,0})^T$. The following theorem states the limit distribution of the proposed estimator $\hat{\boldsymbol{\theta}}$.

Theorem 2.9. *Suppose model (2.10) is strictly stationary, model (2.2) satisfies $p \geq 1, q = 0$ and conditions **C1**–**C3** hold. Choose $m \geq p + r$. Then*

$$\sqrt{n}\boldsymbol{B}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma})$$

as $n \to \infty$, where $\boldsymbol{B} = (b_{i,j})_{1 \leq i,j \leq r+1}$, $\boldsymbol{\Gamma} = (\gamma_{i,j})_{1 \leq i,j \leq r+1}$,

$$b_{1,1} = \mathsf{E}\frac{1}{1 + \sum_{k=1}^{m+1} X_{t-k}^2}, \; b_{1,i+1} = \mathsf{E}\frac{X_{t-i}}{1 + \sum_{k=1}^{m+1} X_{t-k}^2}, \; b_{i+1,j+1} = \mathsf{E}\frac{X_{t-i}X_{t-j}}{(1 + \sum_{k=1}^{m+1} X_{t-k}^2)^2},$$

$$\gamma_{1,1} = \mathsf{E}\frac{\sigma_t^2}{(1 + \sum_{k=1}^{m+1} X_{t-k}^2)^2}, \; \gamma_{1,i+1} = \mathsf{E}\frac{X_{t-i}\sigma_t^2}{(1 + \sum_{k=1}^{m+1} X_{t-k}^2)^2}, \; \gamma_{i+1,j+1} = \mathsf{E}\frac{X_{t-i}X_{t-j}\sigma_t^2}{(1 + \sum_{k=1}^{m+1} X_{t-k}^2)^2}$$

for $i, j = 1, \cdots, r$.

For interval estimation of $\boldsymbol{\theta_0}$, one could simply apply the empirical likelihood method to the above score equations without estimating $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$. More specifically, define

$$\hat{Y}_{t,1}(\boldsymbol{\theta}) = \left\{X_t - \mu - \sum_{j=1}^{r} \phi_j X_{t-j}\right\} \frac{1}{1+\sum_{k=1}^{m+1} X_{t-k}^2},$$
$$\hat{Y}_{t,i+1}(\boldsymbol{\theta}) = \left\{X_t - \mu - \sum_{j=1}^{r} \phi_j X_{t-j}\right\} \frac{X_{t-i}}{1+\sum_{k=1}^{m+1} X_{t-k}^2},$$

for $i = 1, \cdots, r$ and put $\hat{\mathbf{Y}}_t(\boldsymbol{\theta}) = (\hat{Y}_{t,1}(\boldsymbol{\theta}), \cdots, \hat{Y}_{t,r+1}(\boldsymbol{\theta}))^T$. Then the empirical likelihood function for $\boldsymbol{\theta}$ is

$$\hat{L}(\boldsymbol{\theta}) = \sup\left\{\prod_{t=1}^{n}(np_t) : p_1 \geq 0, \cdots, p_n \geq 0, \sum_{t=1}^{n} p_t = 1, \sum_{t=1}^{n} p_t \hat{\mathbf{Y}}_t(\boldsymbol{\theta}) = \mathbf{0}\right\}.$$

Theorem 2.10. *Suppose model (2.10) is strictly stationary, model (2.2) satisfies* $\mathrm{p} \geq 1, \mathrm{q} = 0$*, and conditions* ***C1**–**C3*** *hold. Choose* $m \geq \mathrm{p} + \mathrm{r}$*. Then* $-2\log\hat{L}(\boldsymbol{\theta_0}) \xrightarrow{d} \chi^2_{\mathrm{r}+1}$ *as* $n \to \infty$*.*

Remark 2.7. *Confidence region at the confidence level* $1-\tau$ *for* $\boldsymbol{\theta_0}$ *is constructed as*

$\{\boldsymbol{\theta} : -2\log\hat{L}(\boldsymbol{\theta}) \leq \chi^2_{\mathrm{r}+1,1-\tau}\}$.

Confidence region (interval) for a subset of $\boldsymbol{\theta_0}$ *can be obtained by using the profile empirical likelihood method. However, it remains unknown to us how to deal with a stationary ARMA process instead of the AR*(r) *model.*

Remark 2.8. *When* ϵ_t *in model (2.2) has zero median instead of zero mean, Zhu and Ling (2015) proposed a SLADE for a stationary ARMA process without restriction on the moments of GARCH errors, and a random weighting approach for inference. This is different from the proposed new inference procedures in Theorems 2.9 and 2.10. Since the asymptotic variance matrices for both estimators are complicated, we compare these two methods numerically in Section 2.3 instead of theoretically when the innovation in the GARCH errors has simuatlaneously zero mean and zero median, which shows that the proposed WLSE performs better in finite samples than the SLADE in Zhu and Ling (2015). It is also worth* *mentioning that the proposed estimator has an explicit formula unlike the estimator in Zhu and Ling (2015), which requires a proper optimization procedure with some initial value.*

2.2.4 Extension to GARCH Errors

In this subsection, we shall generalize the results in the preceding two subsections for ARCH errors to GARCH errors. Note that the key idea in the above methods is to bound the conditional standard deviation σ_t in model (2.2) by some known weight function. When e_t follows model (2.2) with $p \geq 1$ and $q \geq 1$, it is hard to find a simple weight function to bound σ_t almost surely. However, proofs for Theorems 2.1–2.10 in Section 2.6 show that we only need a finite $(2+\delta)$-th moment for a weighted σ_t with some positive δ. To better understand the extensions, we look at the question of testing $H_0 : \phi_0 = 1$ for model (2.1) with errors satisfying model (2.2).

Write

$$\frac{\sigma_t^2}{1+\sum_{k=1}^{\max(p,q)} e_{t-k}^2} \leq (p+1)\max(\alpha_0,\alpha_1,\cdots,\alpha_p) + \frac{\sum_{j=1}^q \beta_j \sigma_{t-j}^2}{1+\sum_{k=1}^{\max(p,q)} e_{t-k}^2}$$

and for $j = 1, \cdots, q$,

$$\begin{aligned}\left(\frac{\sigma_{t-j}^2}{1+\sum_{k=1}^{\max(p,q)} e_{t-k}^2}\right)^{1+(\delta_2-\delta_1)/2} &= \sigma_{t-j}^{1+\delta_2}|\epsilon_{t-j}|^{-(1-\delta_1)} \frac{|\sigma_{t-j}\epsilon_{t-j}|^{1-\delta_1}}{(1+\sum_{k=1}^{\max(p,q)} e_{t-k}^2)^{1+(\delta_2-\delta_1)/2}} \\ &\leq \sigma_{t-j}^{1+\delta_2}|\epsilon_{t-j}|^{-(1-\delta_1)}.\end{aligned}$$

Hence, when $\mathsf{E}\,\sigma_t^{1+\delta_2} < \infty$ and $\mathsf{E}\,|\epsilon_t|^{-(1-\delta_1)} < \infty$ for some $0 < \delta_1 < \delta_2 < 1$, there exists $\delta > 0$ such that

$$\mathsf{E}\left(\frac{\sigma_t^2}{1+\sum_{k=1}^{\max(p,q)}(X_{t-k}-X_{t-k-1})^2}\right)^{1+\delta} < \infty. \tag{2.12}$$

Therefore, the following theorem can be established by using the above arguments and similar proofs for Theorems 2.1–2.10.

Theorem 2.11. *Assume $\mathsf{E}\,\sigma_t^{1+\delta_2} < \infty$ and $\mathsf{E}\,|\epsilon_t|^{-(1-\delta_1)} < \infty$ for some $0 < \delta_1 < \delta_2 < 1$. Choose $m \geq \max(p+\mathrm{r}, q+\mathrm{r})$. Then Theorems 2.1–2.10 hold when e_t follows model (2.2) with $p \geq 1$ and $q \geq 1$.*

Remark 2.9. *If the density of ϵ_t is finite at zero like normal distribution and t-distribution (i.e., condition **C3** holds), then $\mathsf{E}\,|\epsilon_t|^{-(1-\delta_1)} < \infty$ for any $\delta_1 \in (0,1)$.*

2.3 Simulation Study

In this section, we examine the finite sample performance of the proposed unit root test, robust estimation and inference. Comparison will be made with analogous tests, estimators and inference methods. The R package `emplik` is used to compute the empirical likelihood function, and the R function `nlm` is employed to calculate the profile empirical likelihood function. To facilitate exposition of this section, we denote the empirical likelihood unit root tests in Theorem 2.1 and Theorem 2.3 as well as their respective extension in Theorem 2.11 as ELT type I and ELT type II, respectively. In all simulations, we draw $10,000$ random samples.

2.3.1 Unit Root Test

It is known that the commonly employed augmented Dickey-Fuller (ADF) test for a unit root assumes uncorrelated and finite variance errors, and the Phillips-Perron (PP) test works for stationary and finite variance errors. As reviewed in the introduction, the two tests with a wild sieve bootstrap implementation proposed by Cavaliere et al. (2018), denoted by Q_T and R_T with tuning parameter κ, require that $\{e_t\}$ is a linear process with symmetrically distributed errors, i.e., the method may be invalid for model (2.1) and model (2.3) with GARCH errors.

To demonstrate that these existing tests fail in the presence of infinite variance GARCH errors, we draw samples from models (2.1) and (2.2) with $p = 1, q = 0, \phi_0 = 1, \alpha_0 = 1, \alpha_1 = 2.5$, ϵ_t following a standardized skew normal $(0, 1, 10)$ distribution such that $\mathsf{E}\,\epsilon_t = 0$ and $\mathsf{E}\,\epsilon_t^2 = 1$, and sample size $n = 1000$, 2000 and 5000. To study ELT Type II under model (2.3) with nonzero intercept, we draw separate samples from models (2.3) and (2.2) with the same setup as above except $\mu_0 = 0.01$. Note that the above setup implies that $\mathsf{E}\,e_t^2 = \infty$.

As suggested in Cavaliere et al. (2018), we choose the tuning parameter $\kappa = 4$ and 12 for computing the lag length used in the ADF regression, and the bootstrap sample size $B = 1999$ for implementing the sieve wild bootstrap method, and we use 'NA' to the denote the case where the bootstrap method fails due to the algorithm of using a linear process to approximate e_t in model (2.1). To implement the ADF test, we employ the R package `fUnitRoots`, where ADF of type I, type II, type III correspond to the type of unit root regression with no drift nor linear trend, with drift but no linear trend, and with both drift and linear trend, respectively. P-values for both the PP test and ADT test are based on interpolating the asymptotic critical values from Table 10.A.2 in Fuller (1996). To implement ELT type II for model (2.3), the added pseudo sample $\{W_t : t = 1, \dots, n\}$ are computed by $W_t = 1/\sqrt{10000} \sum_{i=1}^{10000} W_{t,i}$, where $W_{t,i}$'s are independent and identically distributed random variables from $N(0, \bar{\sigma}^2)$ and

$$\bar{\sigma} = 1.5 \sqrt{\frac{1}{n} \sum_{t=1}^{n} \tilde{Y}_{t1}^2(\tilde{\mu}, 1)},$$

where $\tilde{\mu}$ is the solution to Equation (2.7).

In Table 2.1, we report the empirical size for our proposed empirical likelihood tests based on (2.1) and (2.3) as well as for the abovementioned PP test, ADF tests, and Q_T and R_T tests at levels $\tau = 0.05, 0.10, 0.25$. Results in this table clearly show that the proposed empirical likelihood tests are correctly sized for infinite variance ARCH errors, while the other tests have an incorrect size with PP and ADF tests significantly over-sized and the sieve wild bootstrap tests under-sized. This is in line with the theoretical arguments. Additionally, the merit of ELT type II encompassing the cases of zero and nonzero intercept is well exhibited by its accurate size for model (2.3) with both zero and nonzero μ_0.

We further investigate the size and power properties of these tests by drawing samples from models (2.1) and (2.2) with local alternatives $\phi_0 = 1 - d/n$, where we set $d = 0, 5, 10$, and

$$\alpha_0 = 4.7170\text{e}{-07}, (\alpha_1, \beta_1) = (0.1216, 0.8329), (0.1216, 0.8784), (0.1266, 0.8784),$$

ϵ_t being a standardized skew normal random variable with location parameter 0, shape parameter 1, slant parameters 0 and 10. To render the simulation studies of unit root tests more informative and empirically relevant, we tailor the features of simulated data such that the parameters $(\alpha_0, \alpha_1, \beta_1) = (4.7170\text{e}{-07}, 0.1216, 0.8329)$ are extracted from fitting models (2.1) and (2.2) to the monthly long-term yield in the period 1976/01–2016/12 in Section 2.4. Note that $\alpha_1 + \beta_1 \geq 1$ implies $\mathsf{E}\, e_t^2 = \infty$, and results for $d = 0$ and $d \neq 0$ represent the empirical size and power of the tests, respectively. Since Theorem 2.4 i) shows that ELT type II for model (2.3) with zero intercept has no power under the local alternative $\phi_0 = 1 - d/n$, we instead choose $\phi_0 = 1 - d/(5\sqrt{n})$ to study this test for models (2.3) and (2.2). With the same configuration on the GARCH errors as above, we consider $\mu_0 = 0.01$ and $\phi_0 = 1 - d/(5n)$ for ELT type II under model (2.3) with nonzero intercept as suggested by Theorem 2.4 ii). To implement Q_T and R_T tests in Cavaliere et al. (2018), we again choose $\kappa = 4$ and 12. As α_0 is very small, most values of e_t's and σ_t's will be quite small. Consequently, the constant one in the weight $(1 + X_{t-1}^2)^{1/2}(1 + \sum_{k=1}^{m}(X_{t-k} - X_{t-k-1})^2)^{1/2}$ could be significantly larger than $\sum_{k=1}^{m}(X_{t-k} - X_{t-k-1})^2$ for a finite sample size, which is used to bound σ_t. That is, this simple weight function, albeit plausible theoretically, overweights σ_t for a finite n. Due to this reason, we replace the constant one by

$$\Delta_n = \min\left\{1, \left(\frac{1}{n-1}\sum_{s=2}^{n}\left|X_s - X_{s-1} - \frac{X_n - X_1}{n-1}\right|\right)^2\right\},$$

where $(X_n - X_1)/(n-1)$, the average of $\{X_s - X_{s-1}\}_{s=2}^{n}$, estimates μ under the unit root hypothesis. Hence, we construct the following weight function

$$\left\{\Delta_n + X_{t-1}^2\right\}^{1/2}\left\{\Delta_n + \sum_{k=1}^{m}\left(X_{t-k} - X_{t-k-1} - \frac{X_n - X_1}{n-1}\right)^2\right\}^{1/2}$$

for the study of ELT type II here and for its application to predictive variables in Section 2.4. Note that the phenomenon of very small α_0 in model (2.2) is common in the literature of financial time series modeling. For example, Zhu and Ling (2015) reported a very small estimate of α_0 for exchange rate returns; McElroy and Jach (2019) also reported very small α_0 estimates in fitting GARCH(1,1) models to the CAC 400 returns and FTSE 100 returns; Li et al. (2018) proposed a first-order zero-drift GARCH model via omitting α_0 in the classical GARCH model in Equation (2.2).

Results with respect to $d = 0$ in Table 2.2 show that the proposed ELT type I, along with ELT type II when $\mu_0 \neq 0$, has an accurate size regardless of finite or infinite variance. ELT type II when $\mu_0 = 0$ tends to be over-sized when the sample size is small, but over-sizing for this test significantly improves as the sample size increases. These observations are well supported by theoretical results as inference for the model with a nonzero intercept is usually more powerful than that with zero intercept, and ELT type II with $\mu_0 = 0$ sacrifices the efficiency in unifying the cases of zero and nonzero intercept. PP test and ADF test, as expected, are infeasible for the case of infinite variance. Strikingly, the wild bootstrap Q_T and R_T ADF tests, developed for a linear process driven by infinite unconditional variance innovations in Cavaliere et al. (2018), have quite robust performance against the infinite variance GARCH errors as seen from their good finite sample size, especially when $\kappa = 4$. Nonetheless, the presence of 'NA's for the Q_T and R_T tests when $\kappa = 12$ indicates the practical infeasibility in implementing the wild bootstrap algorithm with a large lag length, at least under the GARCH setting. Turning to empirical power, results with respect to $d = 5$ and 10 show that the proposed empirical likelihood tests have a nontrivial power, although ELT type I is less powerful than the wild bootstrap tests. Note that comparing the power between the two empirical likelihood tests is not meaningful since we set $\phi_0 = 1 - d/n$ for ELT type I, $\phi_0 = 1 - d/(5\sqrt{n})$ for ELT type II when $\mu_0 = 0$, and $\phi_0 = 1 - d/(5n)$ for ELT type II when $\mu_0 \neq 0$. We do not report the power for ELT type II under $\mu_0 = 0$ and the local alternative $\phi_0 = 1 - d/n$ here, which indeed shows it is much less powerful than ELT type I.

To provide further evidence that all the above unit root tests except our proposed empirical likelihood test are problematic for infinite variance GARCH errors, although these extant tests have already been demonstrated to have an inaccurate size in this case, we study the size performance under the setting $(\alpha_0, \alpha_1, \beta_1) = (4.7170\text{e} - 07, 0.1266, 0.8784)$, $\epsilon_t \sim$ standardized skew normal$(0, 1, 10)$, and sample size $n = 10,000$. Using such a large sample size is to ensure that some extreme values from the model will be generated. Figure 2.1 plots the histograms of P-values of these tests under the unit root hypothesis from 10,000 replications. The distributions of P-values of the proposed empirical likelihood tests are largely uniform, while the P-values from other tests fail to have a correct asymptotic uniform distribution. In particular, the wild bootstrap tests from Cavaliere et al. (2018) have 5 'NA's when $\kappa = 12$, and are asymptotically invalid under infinite variance GARCH errors with skewed innovations.

2.3.2 Estimation and Inference for Stationary AR Processes

We examine the finite sample performance of the WLSE in Theorem 2.9 and compare it with the LSE and the SLADE in Zhu and Ling (2015) by generating data from the AR(1)-ARCH(3) model with

$$(\mu, \phi, \alpha_0, \alpha_1, \alpha_2, \alpha_3) = (1.9037\text{e}{-03}, -0.1954, 8.4511\text{e}{-05}, 0.6228, 0.4040, 0.2898).$$

We take sample size $n = 500, 2000, 5000$, and choose ϵ_t to follow a standardized t_ν distribution with $\nu = 2.8, 5, 10$ such that $\mathsf{E}\,\epsilon_t = 0$ and $\mathsf{E}\,\epsilon_t^2 = 1$, or a standard normal distribution. The AR(1)-ARCH(3) parameter setting with $\epsilon_t \sim t_{2.8}$ is obtained from modeling the log-returns of daily HKD/USD exchange rate, to which we will apply our methods in Section 2.4, with the R package `fGarch`. In implementing WLSE here and in analyzing the exchange rate in Section 2.4, due to the very small value of α_0, we change the weight function in Equation (2.11) to

$$\min\left\{1, \left(\frac{1}{n}\sum_{s=1}^{n}|X_s^*|\right)^2\right\} + \sum_{k=1}^{m+1} X_{t-k}^{*2},$$

where $X_t^* = X_t - \frac{1}{n}\sum_{s=1}^{n} X_s$. The rationale behind using this new weight function is the same as that in the above unit root tests, i.e., the new weight is comparable to σ_t^2 as X_t^* is comparable to e_t. For computing the SLADE, we use the weight in equation (2.5) in Zhu and Ling (2015) with the tuning parameter C selected as the 95% quantile of the observations $\{X_1, \ldots, X_n\}$ as suggested therein. Moreover, we set the WLSE of (μ, ϕ) as the initial value for minimizing $\tilde{L}_{sn}$ in equation (2.1) in Zhu and Ling (2015) and employ the R function `optim` since the function $\tilde{L}_{sn}$ is not differentiable. Table 2.3 reports the mean, standard deviation (SD) and root-mean-square error (RMSE) of WLSE, SLADE and LSE. It is apparent that LSE is uniformly inferior in terms of SD and RMSE. Comparing WLSE with SLADE, WLSE has smaller biases than and comparable SD and RMSE to SLADE for μ, and outperforms SALDE for ϕ in terms of bias, SD and RMSE. In summary, WLSE performs better than SLADE under our considered AR-ARCH setup.

We further assess the performance of the empirical likelihood inference method in Theorem 2.10 by computing the coverage probabilities with the above settings, which are reported in Table 2.4. The coverage probabilities remarkably close to the nominal levels even when $n = 500$ demonstrate the good finite sample performance of the proposed empirical likelihood inference procedure.

2.4 Applications

2.4.1 Unit Root Testing in Predictive Variables

We test for unit roots for the monthly dividend-price ratio (d/p), dividend yield (d/y), book-to-market value ratio (b/m), long-term yield (lty) and term spread (tms), which are some commonly employed predictive variables for testing the predictability of stock returns in the predictive regression literature; see Kostakis et al. (2015) for details.

Before applying the proposed unified unit root tests to these variables, we fit model (2.1) for each variable and plot the residuals and Hill estimates for estimating the tail index of the residuals in Figures 2.2–2.5 for the time periods 1953/01–2016/12 (post-war) and 1976/01–2016/12 (after the oil shock recession), respectively. Figures 2.2 and 2.4 suggest that e_t's in model (2.1) for each variable exhibit a similar pattern to a GARCH sequence. Figures 2.3 and 2.5 suggest that errors for the 1976–2016 period may have infinite variance and a heavier tail than errors for the 1953–2016 period. Therefore, it is interesting to see how results from the tests in the above simulation study change with regard to these two periods. Theoretically, PP test, ADF test, and the sieve wild bootstrap implementation of the ADF tests fail for the case of infinite variance GARCH errors. It is known that these unit root tests need $X_0/n \xrightarrow{p} 0$ to derive the asymptotic limit. To eliminate the effect of initial value X_0 in model (2.1) for a medium sample size, we apply these unit root tests to $\{X_t - X_1\}_{t=2}^{n}$ instead of $\{X_t\}_{t=1}^{n}$. We choose $m = 1$ and 2 in implementing the proposed unified empirical likelihood tests and report P-values in Tables 2.5 and 2.6 for the periods 1953/01–2016/12 and 1976/01–2016/12, respectively. Results in Tables 2.5 and 2.6 show that at the 5% level, none of the tests rejects the unit root hypothesis for the dividend-price ratio, dividend yield, book-to-market value ratio and long-term yield, but all tests except the empirical likelihood test with drift reject the unit root hypothesis for the term spread for the period 1953/01–2016/12. When the period 1976/01–2016/12 is concerned, the empirical likelihood tests do not reject the unit root hypothesis for any of the variables whereas all other tests reject the unit root hypothesis for term spread at the 5% level, and the PP test and ADF test (type III) reject the unit root hypothesis for long-term yield at the 10% level. When the dividend-price ratio, book-to-market value ratio and long-term yield become more heavy-tailed for the period 1976/01–2016/12, P-values for the empirical likelihood test without drift are considerably larger than those for Q_T test and R_T test. Due to their robustness to heavy tails and ease of computation, the proposed unified unit root tests thus provide practitioners a powerful tool for pretesting the time series properties of predictive regressors without the need to distinguish whether the errors have finite or infinite variance.

2.4.2 Inference for HKD/USD Exchange Rate

We re-investigate the daily HKD/USD exchange rate from January 21, 1998 to July 6, 2000 studied in Zhu and Ling (2015). We note that the log-returns of the series have 621 observations, which is different from the 600 observations mentioned therein. Denote the log-returns (×100) of the data by $\{y_t\}_{t=1}^{621}$.

As a benchmark result, we first obtain the estimates and confidence intervals for μ

and ϕ by using the `garchFit` function in the R package `fGarch` to fit an AR(1)-ARCH(3) model. The obtained intervals require finite fourth moment of GARCH errors since they are constructed based on a normal limit of the maximum likelihood estimator (MLE). However, the Hill estimates reported in Zhu and Ling (2015) indicate y_t has infinite variance, which implies that the intervals obtained from the MLE are inaccurate. We then obtain the WLSE of (μ, ϕ) in Theorem 2.9 with $m = 3$ and employ the empirical likelihood method in Theorem 2.10 and the profile empirical likelihood procedures in Remark 2.7 to construct their confidence intervals. We re-estimate model (2.3) using the SLADE method in and implement the citeZhuLing2015JASA random weighting procedure with $J = 500$ developed in the same paper for estimating the standard errors and constructing the confidence intervals.

The estimates and corresponding confidence intervals at levels 90% and 95% for the three estimators are reported in Table 2.7. Estimates for μ are close for all estimators, but the SLADE estimate is quite different from the other two estimates for ϕ. Examining the confidence intervals, it is noteworthy that the confidence intervals for μ at both levels 90% and 95% based on WLSE contain zero whereas the confidence intervals based on MLE and WLSE do not.

2.5 Conclusion

The contribution of this paper is two-fold. For an AR(1) model with GARCH errors, the limiting null distributions of existing unit root tests depend on the tail index of the errors, i.e., whether the errors have finite or infinite variance, thus requiring bootstrap or simulation methods to approximate the asymptotic null distributions for inference. This paper proposes unit root tests for an AR(1) model with GARCH errors, which are asymptotically valid without prior on the moments of ARCH errors and with a little more than finite mean of GARCH errors. Unlike the tests in Cavaliere et al. (2018), the proposed tests permit asymmetric innovations, but involve choosing some tuning parameter based on the order in the GARCH errors and requires simulating a pseudo sample when the model has deterministic terms. The tests are generalized to an AR(r) model with GARCH errors in the so-called augmented Dickey-Fuller form. Since the proposed unit root tests do not estimate the unknown parameters in the GARCH errors and always have a chi-squared limit, they are robust and computationally fast in implementation.

When the unit root hypothesis is rejected by the above unit root inference, statistical inference for a stationary AR process with GARCH errors is of interest. Least squares estimator for the AR parameters without estimating the GARCH errors may be inconsistent when the sequence has infinite variance, and has a non-normal limit when the sequence has infinite fourth moment. The asymptotic normality of the quasi-maximum likelihood estimator requires the finite fourth moment of both the sequence itself and the errors in the GARCH model. Some existing estimation procedures valid in the presence of infinite variance innovations require that the innovations in the GARCH model have a symmetric distribution or zero median. Complementing existing methods, this paper develops an inference procedure regardless of the tail heaviness of the GARCH errors via minimizing a weighted least squares distance. The proposed estimator has an explicit formula without estimating the unknown parameters in the GARCH errors and the empirical likelihood inference attains a chi-squared limit.

2.6 Proofs

Let $\mathcal{F}_t = \sigma\{\epsilon_t, \epsilon_{t-1}, \ldots\}$ be the σ-field generated by the sequence $\{\epsilon_t, \epsilon_{t-1}, \ldots\}$.

Lemma 2.1. *Under conditions of Theorem 2.1, we have as $n \to \infty$,*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_t(1) \xrightarrow{d} N(0, \sigma^2), \tag{2.13}$$

$$\frac{1}{n} \sum_{t=1}^{n} Y_t^2(1) \xrightarrow{p} \sigma^2, \tag{2.14}$$

$$\max_{1 \le t \le n} |Y_t(1)| = o_p(n^{1/2}), \tag{2.15}$$

where $\sigma^2 = \mathsf{E}\, \frac{\sigma_t^2}{1+\sum_{k=1}^m e_{t-k}^2}$.

Proof of Lemma 2.1. Note that

$$Y_t(1) = \frac{e_t X_{t-1}}{\sqrt{(1+X_{t-1}^2)(1+\sum_{k=1}^m e_{t-k}^2)}}.$$

When $\phi_0 = 1$, $|X_t| \xrightarrow{p} \infty$ as $t \to \infty$, i.e., $X_t^2/(1+X_t^2) \overset{p}{\to} 1$, hence it follows from the stationarity of $\{\sigma_t^2\}$ such that

$$\begin{aligned} V_n^2 &= \tfrac{1}{n}\textstyle\sum_{t=1}^n \mathsf{E}\,(Y_t^2(1)|\mathcal{F}_{t-1}) = \tfrac{1}{n}\sum_{t=1}^n \frac{\sigma_t^2 X_{t-1}^2}{(1+X_{t-1}^2)(1+\sum_{k=1}^m e_{t-k}^2)} \\ &= \tfrac{1}{n}\textstyle\sum_{t=1}^n \frac{\sigma_t^2}{1+\sum_{k=1}^m e_{t-k}^2} + o_p(1) \overset{p}{\to} \sigma^2. \end{aligned} \tag{2.16}$$

Put $\alpha_* = \max\{\alpha_0, \ldots, \alpha_p\}$, which is positive. We have for any $c > 0$,

$$\frac{1}{n}\sum_{t=1}^n \mathsf{E}\,[Y_t^2(1) I(|Y_t(1)| > c\sqrt{n})|\mathcal{F}_{t-1}] \le (c\sqrt{n})^{-d^*} \alpha_*^{1+\frac{d^*}{2}}\, \mathsf{E}\,|\epsilon_1|^{2+d^*} \xrightarrow{p} 0. \tag{2.17}$$

Hence, (2.13) follows from (2.16), (2.17) and Corollary 3.1 of Hall and Heyde (1980).

Since

$$\sup_n P(V_n^2 > \lambda) \le P(\alpha_* > \lambda) = 0 \quad \text{as} \quad \lambda \to \infty,$$

(2.14) follows from (2.17) and Theorem 2.23 of Hall and Heyde (1980).

Further, (2.15) follows from that $\max_{1\le t\le n} |Y_t(1)| \le \sqrt{\alpha_*} \max_{1\le t\le n} |\epsilon_t|$ and $E|\epsilon_t|^{2+d^*} < \infty$. Hence Lemma 2.1 follows. □

Proof of Theorem 2.1. It follows from Lemma 2.1 and standard arguments in Chapter 11 of Owen (2001) that

$$\lambda = \frac{\sum_{t=1}^n Y_t(1)}{\sum_{t=1}^n Y_t^2(1)} + o_p(1/\sqrt{n})$$

and

$$\begin{aligned} l(1) &= 2\textstyle\sum_{t=1}^n \{\lambda Y_t(1) - \frac{1}{2}\lambda^2 Y_t^2(1)\} + o_p(1) \\ &= \frac{(\sum_{t=1}^n Y_t(1))^2}{\sum_{t=1}^n Y_t^2(1)} + o_p(1) \\ &\overset{d}{\to} \chi^2(1) \quad \text{as} \quad n \to \infty, \end{aligned}$$

i.e., Theorem 2.1 follows. □

Proof of Theorem 2.2. Like the proof of Lemma 1 in Phillips (1987), it follows from condition **C4** that

$$\begin{aligned}
\frac{X_{[ns]}}{n^{1/\delta}L(n)} &= \tfrac{1}{n^{1/\delta}L(n)} \textstyle\sum_{t=1}^{[ns]} \phi_0^{[ns]-t} \int_{(t-1)/n}^{t/n} dS_n(r) + o_p(1) \\
&= \tfrac{1}{n^{1/\delta}L(n)} \textstyle\sum_{t=1}^{[ns]} \int_{(t-1)/n}^{t/n} \phi_0^{(s-r)n} \, dS_n(r) + o_p(1) \\
&= \tfrac{1}{n^{1/\delta}L(n)} \textstyle\int_0^s \phi_0^{(s-r)n} \, dS_n(r) + o_p(1) \\
&= \tfrac{1}{n^{1/\delta}L(n)} \{S_n(s) + \textstyle\int_0^s S_n(r)\phi_0^{(s-r)n} n\log(\phi_0)\, dr\} + o_p(1) \\
&= \tilde{W}_\delta(s) - d_1\bar{d}_1 \textstyle\int_0^s \tilde{W}_\delta(r) e^{-(s-r)d_1\bar{d}_1} \, dr + o_p(1) \\
&= \tilde{J}_\delta(s) + o_p(1),
\end{aligned} \tag{2.18}$$

where $S_n(r) = \sum_{t=1}^{[nr]} e_t$, and

$$X_{[ns]} - X_{[ns]-1} = e_{[ns]} + o_p(1) \quad \text{uniformly in} \quad s \in [0,1]. \tag{2.19}$$

Write

$$Y_t(1) = Y_t(\phi_0) - \frac{d_1 X_{t-1}^2}{n^{1/2+1/\delta} L(n) \sqrt{(1+X_{t-1}^2)(1+\sum_{k=1}^m (X_{t-k} - X_{t-k-1})^2)}}.$$

Put $S_n^* = \sum_{t=1}^n \frac{e_t}{\sqrt{1+\sum_{k=1}^m e_{t-k}^2}}$ and $S_0^* = 0$. Then it follows from (2.18) and (2.19) that

$$\begin{aligned}
& \tfrac{1}{\sqrt{n}} \textstyle\sum_{t=1}^n Y_t(\phi_0) \\
= \quad & \tfrac{1}{\sqrt{n}} \textstyle\sum_{t=1}^n (S_t^* - S_{t-1}^*) \frac{X_{t-1}}{\sqrt{1+X_{t-1}^2}} \\
= \quad & \tfrac{1}{\sqrt{n}} \textstyle\sum_{t=1}^n S_t^* \frac{X_{t-1}}{\sqrt{1+X_{t-1}^2}} - \tfrac{1}{\sqrt{n}} \textstyle\sum_{t=1}^{n-1} S_t^* \frac{X_t}{\sqrt{1+X_t^2}} \\
= \quad & \tfrac{1}{\sqrt{n}} S_n^* \frac{X_{n-1}}{\sqrt{1+X_{n-1}^2}} + \tfrac{1}{\sqrt{n}} \textstyle\sum_{t=1}^{n-1} S_t^* \frac{X_{t-1}-X_t}{(1+\xi_t^2)^{3/2}} \\
= \quad & sgn(\tilde{J}_\delta(1)) \tfrac{1}{\sqrt{n}} \textstyle\sum_{t=1}^n \frac{e_t}{\sqrt{1+\sum_{k=1}^m e_{t-k}^2}} + o_p(1),
\end{aligned} \tag{2.20}$$

where ξ_t lies between X_{t-1} and X_t. Like the proof of Lemma 2.1, under $H_a : \phi_0 = 1 - \frac{d_1}{n^{1/2+1/\delta}L(n)}$, using (2.18) and (2.19), we have

$$\frac{1}{n} \sum_{t=1}^n Y_t^2(\phi_0) = \mathsf{E}\,(W^2(1))\mathsf{E}\,(\frac{\sigma_t^2}{1+\sum_{k=1}^m e_{t-k}^2}) + o_p(1), \tag{2.21}$$

$$\begin{aligned}
& \tfrac{1}{\sqrt{n}} \textstyle\sum_{t=1}^n \frac{d_1 X_{t-1}^2}{n^{1/2+1/\delta}L(n)\sqrt{(1+X_{t-1}^2)(1+\sum_{k=1}^m (X_{t-k}-X_{t-k-1})^2)}} \\
= \quad & \tfrac{1}{n} \textstyle\sum_{t=1}^n \frac{d_1 X_{t-1}^2}{n^{1/\delta}L(n)\sqrt{1+X_{t-1}^2}\sqrt{1+\sum_{k=1}^m e_{t-k}^2}} + o_p(1) \\
= \quad & d_1 \mathsf{E}\,(\frac{1}{\sqrt{1+\sum_{k=1}^m e_{t-k}^2}}) \textstyle\int_0^1 |\tilde{J}_\delta(s)| \, ds + o_p(1)
\end{aligned} \tag{2.22}$$

and

$$\frac{1}{n}\sum_{t=1}^{n}\frac{d_1^2X_{t-1}^4}{n^{1+2/\delta}L^2(n)(1+X_{t-1}^2)\{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2\}}\xrightarrow{p}0. \tag{2.23}$$

Therefore, it follows from (2.20)–(2.23) and arguments in Chapter 11 of Owen (2001) that

$$\lambda=\frac{\sum_{t=1}^n Y_t(1)}{\sum_{t=1}^n Y_t^2(1)}+o_p(n^{-1/2})=O_p(n^{-1/2})$$

and

$$\begin{aligned} l(1) &= 2\sum_{t=1}^n \log(1+\lambda Y_t(1))\\ &= \frac{(\sum_{t=1}^n Y_t(1))^2}{\sum_{t=1}^n Y_t^2(1)}+o_p(1)\\ &= \frac{\left\{sgn(\tilde{J}_\delta(1))\frac{1}{\sqrt{n}}\sum_{t=1}^n\frac{e_t}{\sqrt{1+\sum_{k=1}^m e_{t-k}^2}}-d_1\Delta_2\int_0^1|\tilde{J}_\delta(s)|\,ds\right\}^2}{\Delta_1}+o_p(1), \end{aligned}$$

i.e., the theorem holds. □

Lemma 2.2. *Under conditions of Theorem 2.3, we have* $\frac{1}{\sqrt{n}}\sum_{t=1}^n\tilde{\mathbf{Y}}_t(\mu_0,1)\xrightarrow{d}N(\mathbf{0},\tilde{\mathbf{\Sigma}})$ *as* $n\to\infty$, *where* $\tilde{\mathbf{\Sigma}}=(\tilde{\sigma}_{ij})_{1\le i,j\le 2}$ *with* $\tilde{\sigma}_{11}=\mathsf{E}\,\frac{\sigma_t^2}{1+\sum_{k=1}^m(\mu_0+e_{t-k})^2}$, $\tilde{\sigma}_{22}=\bar{\sigma}^2$, *and* $\tilde{\sigma}_{12}=\tilde{\sigma}_{21}=0$.

Proof of Lemma 2.2. Note that

$$\begin{cases}\tilde{Y}_{t1}(\mu_0,1)=\frac{\sigma_t\epsilon_t}{\sqrt{1+\sum_{k=1}^m(\mu_0+e_{t-k})^2}},\\ \tilde{Y}_{t2}(\mu_0,1)=\frac{X_{t-1}\sigma_t\epsilon_t}{(1+X_{t-1}^2)^{0.75}\sqrt{1+\sum_{k=1}^m(\mu_0+e_{t-k})^2}}+W_t.\end{cases} \tag{2.24}$$

Similar to the proof of Lemma 2.1, we obtain $\frac{1}{\sqrt{n}}\sum_{t=1}^n\tilde{Y}_{t1}(\mu_0,1)\xrightarrow{d}N(0,\tilde{\sigma}_{11})$, by noting that

$$\begin{aligned}\frac{\sigma_t^2}{1+\sum_{k=1}^m(\ \mu_0+e_{t-k})^2} &\le \frac{\alpha_0+2\sum_{k=1}^p\alpha_k(\mu_0+e_{t-k})^2+2\mu_0^2\sum_{k=1}^p\alpha_k}{1+\sum_{k=1}^m(\mu_0+e_{t-k})^2}\\ &\le \max\left\{\alpha_0+2\mu_0^2\sum_{k=1}^p\alpha_k,\ 2\alpha_1,\ \cdots,\ 2\alpha_p\right\}.\end{aligned}$$

Next, when $\phi_0=1$, since $|X_t|\xrightarrow{p}\infty$ as $t\to\infty$, we have

$$\frac{1}{n}\sum_{t=1}^{n}\frac{\sigma_t^2X_{t-1}^2}{(1+X_{t-1}^2)^{1.5}\{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2\}}\xrightarrow{p}0, \tag{2.25}$$

which implies $\frac{1}{\sqrt{n}}\sum_{t=1}^n\tilde{Y}_{t2}(\mu_0,1)=\frac{1}{\sqrt{n}}\sum_{t=1}^n W_t+o_p(1)\xrightarrow{d}N(0,\tilde{\sigma}_{22})$.

Observing that $\frac{1}{n}\sum_{t=1}^{n} \mathsf{E}\left(\tilde{Y}_{t1}(\mu_0,1)\tilde{Y}_{t2}(\mu_0,1)|\mathcal{F}_{t-1}\right) = o_p(1)$, Lemma 2.2 follows from the Cramér-Wold device. □

Lemma 2.3. *Under conditions of Theorem 2.3, we have as $n \to \infty$,*

$$\frac{1}{n}\sum_{t=1}^{n} \tilde{\boldsymbol{Y}}_t(\mu_0,1)\tilde{\boldsymbol{Y}}_t^T(\mu_0,1) \xrightarrow{p} \tilde{\boldsymbol{\Sigma}} \quad \textit{and} \quad \max_{1\le t\le n} \|\tilde{\boldsymbol{Y}}_t(\mu_0,1)\| = o_p(n^{1/2}),$$

where $\tilde{\boldsymbol{\Sigma}}$ is defined in Lemma 2.2.

Proof of Lemma 2.3. The lemma follows from (2.24)–(2.25) and the weak law of large numbers for martingale differences (see Hall and Heyde, 1980). □

Lemma 2.4. *Under conditions of Theorem 2.3, as $n \to \infty$, with probability one, $\tilde{L}(\mu,1)$ attains its maximum value at $\bar{\mu}$ such that $|\bar{\mu}-\mu_0| < n^{-1/d_0}$ for some $d_0 \in (2, 2+d_1)$, and $\bar{\mu}$ and $\bar{\boldsymbol{\lambda}}$ satisfy $\boldsymbol{Q}_{1n}(\bar{\mu},\bar{\boldsymbol{\lambda}}) = 0$ and $\boldsymbol{Q}_{2n}(\bar{\mu},\bar{\boldsymbol{\lambda}}) = 0$, where*

$$\boldsymbol{Q}_{1n}(\mu,\boldsymbol{\lambda}) := \frac{1}{n}\sum_{t=1}^{n}\frac{\tilde{\boldsymbol{Y}}_t(\mu,1)}{1+\boldsymbol{\lambda}^T\tilde{\boldsymbol{Y}}_t(\mu,1)} \quad \text{and} \quad \boldsymbol{Q}_{2n}(\mu,\boldsymbol{\lambda}) := \frac{1}{n}\sum_{t=1}^{n}\frac{1}{1+\boldsymbol{\lambda}^T\tilde{\boldsymbol{Y}}_t(\mu,1)}\left(\frac{\partial\tilde{\boldsymbol{Y}}_t(\mu,1)}{\partial\mu}\right)^T\boldsymbol{\lambda}.$$

Proof of Lemma 2.4. This can be shown in the same way as Lemma 1 of Qin and Lawless (1994) by using Lemmas 2.2 and 2.3. □

Proof of Theorem 2.3. Using the same arguments as in the proof of Theorem 1 in Qin and Lawless (1994), it follows from Lemmas 2.2 and 2.3 that

$$\begin{pmatrix} \bar{\boldsymbol{\lambda}} \\ \bar{\mu}-\mu_0 \end{pmatrix} = \boldsymbol{S}_n^{-1}\begin{pmatrix} -\boldsymbol{Q}_{1n}(\mu_0,\boldsymbol{0}) + o_p(n^{-1/2}) \\ o_p(n^{-1/2}) \end{pmatrix},$$

where

$$\boldsymbol{S}_n = \begin{pmatrix} \frac{\partial \boldsymbol{Q}_{1n}(\mu_0,\boldsymbol{0})}{\partial\boldsymbol{\lambda}^T} & \frac{\partial \boldsymbol{Q}_{1n}(\mu_0,\boldsymbol{0})}{\partial\mu} \\ \frac{\partial \boldsymbol{Q}_{2n}(\mu_0,\boldsymbol{0})}{\partial\boldsymbol{\lambda}^T} & \boldsymbol{0} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{12}^T & \boldsymbol{0} \end{pmatrix}$$

with $\boldsymbol{S}_{11} = -\tilde{\boldsymbol{\Sigma}} = -\begin{pmatrix} \tilde{\sigma}_{11} & 0 \\ 0 & \tilde{\sigma}_{22} \end{pmatrix}$ and $\boldsymbol{S}_{12} = \left(-\mathsf{E}\frac{1}{\sqrt{1+\sum_{k=1}^{m}(\mu_0+e_{t-k})^2}}, 0\right)^T := (s_{12,1}, 0)^T$, where $\tilde{\sigma}_{11}$ and $\tilde{\sigma}_{22}$ are defined in Lemma 2.2.

By the standard arguments of empirical likelihood method (see the proof of Theorem 1 in Owen, 1990) and Lemmas 2.2 and 2.3, we have

$$\begin{aligned}
\tilde{l}(1) &= 2\textstyle\sum_{t=1}^{n}\log\{1+\bar{\boldsymbol{\lambda}}^T\tilde{\boldsymbol{Y}}_t(\bar{\mu},1)\}\\
&= 2n(\bar{\boldsymbol{\lambda}}^T,\ \bar{\mu}-\mu_0)\left(\boldsymbol{Q}_{1n}^T(\mu_0,\mathbf{0}),\ \mathbf{0}\right)^T\\
&\quad +n(\bar{\boldsymbol{\lambda}}^T,\ \bar{\mu}-\mu_0)\boldsymbol{S}_n(\bar{\boldsymbol{\lambda}},\ \bar{\mu}-\mu_0)^T+o_p(1)\\
&= -n(\boldsymbol{Q}_{1n}^T(\mu_0,\mathbf{0}),\ \mathbf{0})\boldsymbol{S}_n^{-1}\left(\boldsymbol{Q}_{1n}^T(\mu_0,\mathbf{0}),\ \mathbf{0}\right)^T+o_p(1)\\
&= -(\boldsymbol{Z}^T,\ \mathbf{0})\begin{pmatrix}\boldsymbol{S}_{11} & \boldsymbol{S}_{12}\\ \boldsymbol{S}_{12}^T & \mathbf{0}\end{pmatrix}^{-1}(\boldsymbol{Z}^T,\ \mathbf{0})^T+o_p(1)
\end{aligned}$$

as $n\to\infty$, where $\boldsymbol{Z}=(\boldsymbol{Z}_1,\boldsymbol{Z}_2)^T\sim N(\mathbf{0},\tilde{\boldsymbol{\Sigma}})$. Since

$$\begin{pmatrix}\boldsymbol{S}_{11} & \boldsymbol{S}_{12}\\ \boldsymbol{S}_{12}^T & \mathbf{0}\end{pmatrix}^{-1}=\begin{pmatrix}\boldsymbol{S}_{11}^{-1}-\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1} & \boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\\ \boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1} & -\boldsymbol{\Delta}^{-1}\end{pmatrix},$$

where $\boldsymbol{\Delta}=\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}=-s_{12,1}^2/\tilde{\sigma}_{11}$, we have $\boldsymbol{S}_{11}^{-1}=\begin{pmatrix}\tilde{\sigma}_{11}^{-1} & 0\\ 0 & \tilde{\sigma}_{22}^{-1}\end{pmatrix}$ and

$$\begin{aligned}
-(\boldsymbol{Z}^T,\mathbf{0})\begin{pmatrix}\boldsymbol{S}_{11} & \boldsymbol{S}_{12}\\ \boldsymbol{S}_{12}^T & \mathbf{0}\end{pmatrix}^{-1}(\boldsymbol{Z}^T,\mathbf{0})^T &= -\boldsymbol{Z}^T\left(\boldsymbol{S}_{11}^{-1}-\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1}\right)\boldsymbol{Z}\\
&= \boldsymbol{Z}^T\begin{pmatrix}0 & 0\\ 0 & \tilde{\sigma}_{22}^{-1}\end{pmatrix}\boldsymbol{Z}=(\boldsymbol{Z}_2/\sqrt{\tilde{\sigma}_{22}})^2\\
&\xrightarrow{d}\chi^2(1).
\end{aligned}$$

□

Proof of Theorem 2.4. Put

$$\tilde{\mu}=\mu+\frac{(\phi_0-1)\sum_{t=1}^{n}\frac{X_{t-1}}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}}{\sum_{t=1}^{n}\frac{1}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}}$$

and

$$\tilde{\mu}_0=\mu_0+\frac{(\phi_0-1)\sum_{t=1}^{n}\frac{X_{t-1}}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}}{\sum_{t=1}^{n}\frac{1}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}}.$$

Then, we have

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\tilde{Y}_{t1}(\tilde{\mu}_0,1)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{e_t}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}$$

and

$$\begin{aligned}
&\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\tilde{Y}_{t2}(\tilde{\mu}_0,1)\\
=\;&\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}W_t-\frac{(\phi_0-1)\sum_{t=1}^{n}\frac{X_{t-1}}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}}{\sum_{t=1}^{n}\frac{1}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}}\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\frac{X_{t-1}}{(1+X_{t-1}^2)^{0.75}\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}\\
&+(\phi_0-1)\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\frac{X_{t-1}^2}{(1+X_{t-1}^2)^{0.75}\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}+o_p(1).
\end{aligned}$$

i) Like the derivation of (2.18), we can show that

$$\begin{aligned}
\tfrac{X_{[ns]}}{n^{\alpha/\delta}}&=\textstyle\int_0^s\phi_0^{(s-r)n}\,d\{\frac{S_n(r)-S_n(s)}{n^{\alpha/\delta}}\}\\
&=-\textstyle\int_0^1\phi_0^{rsn}\,d\{\frac{S_n(s-sr)-S_n(s)}{n^{\alpha/\delta}}\}\\
&=-\textstyle\int_0^{n^{1-\alpha}}\phi_0^{rsn^{\alpha}}\,d\{\frac{S_n(s-srn^{\alpha-1})-S_n(s)}{n^{\alpha/\delta}}\}\\
&=-\textstyle\int_0^{\infty}e^{-d_1sr}\,d\Gamma_\alpha(s,r)+o_p(1)\quad\text{uniformly in}\quad s\in[0,1],
\end{aligned}\tag{2.26}$$

which is used to show that $X_{[ns]}-X_{[ns]-1}=e_{[ns]}+o_p(1)$ uniformly in $s\in[0,1]$,

$$\begin{aligned}
&\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\tilde{Y}_{t2}(\tilde{\mu}_0,1)\\
=\;&\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}W_t+d_1\{\int_0^1L^*(s)\,ds\}\{\Delta_2\int_0^1L^*(s)|L^*(s)|^{-3/2}\,ds\}-d_1\Delta_2\int_0^1|L^*(s)|^{1/2}\,ds+o_p(1)\\
=\;&\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}W_t+d_1d_2\Delta_2+o_p(1),
\end{aligned}$$

$$\frac{1}{n}\sum_{t=1}^{n}\tilde{Y}_{t1}^2(\tilde{\mu}_0,1)\xrightarrow{p}\Delta_1,\qquad\frac{1}{n}\sum_{t=1}^{n}\tilde{Y}_{t2}^2(\tilde{\mu}_0,1)\xrightarrow{p}\mathsf{E}\,W_1^2.$$

Hence it follows from the proof of Theorem 2.3 that

$$\tilde{l}(1)=\frac{\{\sum_{t=1}^{n}\tilde{Y}_{t2}(\tilde{\mu}_0,1)\}^2}{\sum_{t=1}^{n}\tilde{Y}_{t2}^2(\tilde{\mu}_0,1)}+o_p(1)=\frac{\{\frac{1}{\sqrt{n}}\sum_{t=1}^{n}W_t+d_1d_2\Delta_2\}^2}{\mathsf{E}\,W_1^2}+o_p(1),$$

i.e., Theorem 2.4 i) holds.

ii) Write

$$X_t=\mu_0\frac{1-\phi_0^t}{1-\phi_0}+\phi_0^tX_0+\sum_{j=1}^{t}\phi_0^{t-j}e_j.$$

Then, like the proof of (2.18), we can show that

$$X_{[ns]}/n=\mu_0\frac{1-e^{-d_1s}}{d_1}+o_p(1)\quad\text{and}\quad X_{[ns]}-X_{[ns]-1}=\mu_0e^{-d_1s}+e_{[ns]}$$

uniformly in $s \in [0,1]$, which are used to show that

$$\begin{aligned} & \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^n \tilde{Y}_{t2}(\tilde{\mu}_0,1) \\ = \;& \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^n W_t + |\mu_0|^{1/2} d_1 \frac{E\int_0^1 \frac{f(s;d_1)}{\sqrt{1+\sum_{k=1}^m (-\mu_0 e^{-d_1 s}+e_{t-k})^2}}\, ds}{E\int_0^1 \frac{1}{\sqrt{1+\sum_{k=1}^m (-\mu_0 e^{-d_1 s}+e_{t-k})^2}}\, ds} \mathsf{E}\int_0^1 \frac{f^{-1/2}(s;d_1)}{\sqrt{1+\sum_{k=1}^m (-\mu_0 e^{-d_1 s}+e_{t-k})^2}}\, ds \\ & -|\mu_0|^{1/2} d_1 \mathsf{E}\int_0^1 \frac{f^{1/2}(s;d_1)}{\sqrt{1+\sum_{k=1}^m (-\mu_0 e^{-d_1 s}+e_{t-k})^2}}\, ds + o_p(1) \\ = \;& \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^n W_t + |\mu_0|^{1/2} d_1 d_3 + o_p(1), \end{aligned}$$

$$\frac{1}{n}\sum_{t=1}^n \tilde{Y}_{t1}^2(\tilde{\mu}_0,1) \xrightarrow{p} \Delta_1, \quad \frac{1}{n}\sum_{t=1}^n \tilde{Y}_{t2}^2(\tilde{\mu}_0,1) \xrightarrow{p} \mathsf{E}\, W_1^2.$$

Hence, it follows from the proof of Theorem 2.3 that

$$\tilde{l}(1) = \frac{\{\sum_{t=1}^n \tilde{Y}_{t2}(\tilde{\mu}_0,1)\}^2}{\sum_{t=1}^n \tilde{Y}_{t2}^2(\tilde{\mu}_0,1)} + o_p(1) = \frac{\left\{\frac{1}{\sqrt{n}}\sum_{t=1}^n W_t + |\mu_0|^{1/2} d_1 d_3\right\}^2}{\mathsf{E}\, W_1^2} + o_p(1),$$

i.e., Theorem 2.4 ii) holds. □

Proof of Theorem 2.5. Under **C4**, there exists a stable process $W(s)$ such that

$$\frac{X_{[ns]}}{n^{1/\delta} L(n)} \xrightarrow{D} W(s) \quad \text{in} \quad D([0,1]).$$

Like the proofs of Lemma 2.1 and (2.20), we can show that

$$\frac{1}{n}\sum_{t=1}^n \boldsymbol{Y}_t^*(1,\boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0},\boldsymbol{\Sigma}^*), \quad \frac{1}{n}\sum_{t=1}^n \boldsymbol{Y}_t^*(1,\boldsymbol{\theta}_0)\{\boldsymbol{Y}_t^*(1,\boldsymbol{\theta}_0)\}^T \xrightarrow{p} \boldsymbol{\Sigma}^*,$$

$$\frac{1}{n}\sum_{t=1}^n \frac{\partial \boldsymbol{Y}_{t,1}^*(1,\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} = sgn(W(1))\boldsymbol{b}_1 + o_p(1), \quad \frac{1}{n}\sum_{t=1}^n \frac{\partial \boldsymbol{Y}_{t,i}^*(1,\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} = \boldsymbol{b}_i + o_p(1)$$

for $i = 2, \cdots, r+1$, where $\boldsymbol{\Sigma}^*$ is a positive definite matrix and $\boldsymbol{b}_i$'s are vectors. Put $\boldsymbol{S}_{11} = -\boldsymbol{\Sigma}^*$, $\boldsymbol{S}_{12} = (sgn(W(1))\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{r+1})$, $\boldsymbol{\Delta} = \boldsymbol{S}_{12}^T \boldsymbol{S}_{11}^{-1} \boldsymbol{S}_{12}$ and let $\boldsymbol{Z} \sim N(\boldsymbol{0},\boldsymbol{\Sigma}^*)$. Like the proof of Theorem 2.3, we have

$$\begin{aligned} l^*(1) &= -\boldsymbol{Z}^T(\boldsymbol{S}_{11}^{-1} - \boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1})\boldsymbol{Z} + o_p(1) \\ &= (\boldsymbol{\Sigma}^{*1/2}\boldsymbol{Z})^T(\boldsymbol{I}_{(r+1)\times(r+1)} - \boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1/2})(\boldsymbol{\Sigma}^{*1/2}\boldsymbol{Z}) + o_p(1), \end{aligned}$$

where $\boldsymbol{I}_{(r+1)\times(r+1)}$ denotes the $(r+1)\times(r+1)$ identity matrix. Since

$$trace(\boldsymbol{I}_{(r+1)\times(r+1)} - \boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1/2}) = r+1-trace(\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}) = 1$$

and the matrix $\boldsymbol{I}_{(r+1)\times(r+1)} - \boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1/2}$ is idempotent, we have $l^*(1) \xrightarrow{d} \chi_1^2$ as $n \to \infty$. □

Proof of Theorem 2.6. Like the proofs of Lemma 2.1 and (2.20), we can show that

$$\frac{1}{n}\sum_{t=1}^n \boldsymbol{Y}_t^*(1,\boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0},\boldsymbol{\Sigma}^*), \quad \frac{1}{n}\sum_{t=1}^n \boldsymbol{Y}_t^*(1,\boldsymbol{\theta}_0)\{\boldsymbol{Y}_t^*(1,\boldsymbol{\theta}_0)\}^T \xrightarrow{p} \boldsymbol{\Sigma}^*,$$

$$\frac{1}{n}\sum_{t=1}^n \frac{\partial \boldsymbol{Y}_{t,1}^*(1,\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} \xrightarrow{p} \boldsymbol{b}_1, \quad \frac{1}{n}\sum_{t=1}^n \frac{\partial \boldsymbol{Y}_{t,2}^*(1,\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} \xrightarrow{p} \boldsymbol{0}, \quad \frac{1}{n}\sum_{t=1}^n \frac{\partial \boldsymbol{Y}_{t,i}^*(1,\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} \xrightarrow{p} \boldsymbol{b}_i$$

for $i = 3, \cdots, r+2$, where $\boldsymbol{\Sigma}^*$ is a positive definite matrix and $\boldsymbol{b}_i$'s are vectors. Put $\boldsymbol{S}_{11} = -\boldsymbol{\Sigma}^*$, $\boldsymbol{S}_{12} = (\boldsymbol{b}_1, \boldsymbol{0}, \boldsymbol{b}_3 \cdots, \boldsymbol{b}_{r+2})$, $\boldsymbol{\Delta} = \boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}$ and let $\boldsymbol{Z} \sim N(\boldsymbol{0},\boldsymbol{\Sigma}^*)$. Like the proof of Theorem 2.3, we have

$$\begin{aligned}\bar{l}^*(1) &= -\boldsymbol{Z}^T(\boldsymbol{S}_{11}^{-1} - \boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1})\boldsymbol{Z} + o_p(1)\\ &= (\boldsymbol{\Sigma}^{*1/2}\boldsymbol{Z})^T(\boldsymbol{I}_{(r+2)\times(r+2)} - \boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1/2})(\boldsymbol{\Sigma}^{*1/2}\boldsymbol{Z}) + o_p(1).\end{aligned}$$

Since

$$trace(\boldsymbol{I}_{(r+2)\times(r+2)} - \boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1/2}) = r+2-trace(\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}) = 1$$

and the matrix $\boldsymbol{I}_{(r+2)\times(r+2)} - \boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{\Delta}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1/2}$ is idempotent, we have $\bar{l}^*(1) \xrightarrow{d} \chi_1^2$ as $n \to \infty$. □

Proof of Theorem 2.7. Under $H_0 : \phi_0 = 1$, we have $\mu_0^* = \gamma_0$, $\gamma_0^* = 0$,

$$\begin{aligned}\bar{Y}_{t1}(\mu_0^*,\gamma_0^*,1) &= \tfrac{\sigma_t\epsilon_t}{\sqrt{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2}},\\ \bar{Y}_{t2}(\mu_0^*,\gamma_0^*,1) &= \tfrac{t\sigma_t\epsilon_t}{n\sqrt{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2}},\\ \bar{Y}_{t3}(\mu_0^*,\gamma_0^*,1) &= \tfrac{\sigma_t\epsilon_t}{(1+X_{t-1}^2)^{0.75}\sqrt{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2}} + W_t.\end{aligned}$$

Like the proofs of Lemmas 2.2 and 2.3, we have

$$\begin{cases}\frac{1}{\sqrt{n}}\sum_{t=1}^n \bar{\boldsymbol{Y}}_t(\mu_0^*,\gamma_0^*,1) \xrightarrow{d} N(\boldsymbol{0},\bar{\boldsymbol{\Sigma}}),\\ \frac{1}{n}\sum_{t=1}^n \bar{\boldsymbol{Y}}_t(\mu_0^*,\gamma_0^*,1) \xrightarrow{p} \bar{\boldsymbol{\Sigma}},\\ \max_{1\le t\le n}||\bar{\boldsymbol{Y}}_t(\mu_0^*,\gamma_0^*,1)|| = o_p(n^{1/2}),\end{cases} \tag{2.27}$$

where

$$\bar{\boldsymbol{\Sigma}} = \begin{pmatrix} \mathsf{E}\,\frac{\sigma_t^2}{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2} & \frac{1}{2}\mathsf{E}\,\frac{\sigma_t^2}{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2} & 0 \\ \frac{1}{2}\mathsf{E}\,\frac{\sigma_t^2}{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2} & \frac{1}{3}\mathsf{E}\,\frac{\sigma_t^2}{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2} & 0 \\ 0 & 0 & \bar{\sigma}^2 \end{pmatrix}.$$

Put $\bar{\boldsymbol{Z}} = (\bar{\boldsymbol{Z}}_1, \bar{\boldsymbol{Z}}_2, \bar{\boldsymbol{Z}}_3)^T \sim N(\mathbf{0}, \bar{\boldsymbol{\Sigma}})$, $\bar{\boldsymbol{S}}_{11} = -\bar{\boldsymbol{\Sigma}}$,

$$\bar{\boldsymbol{S}}_{12} = \begin{pmatrix} -\mathsf{E}\,\frac{1}{\sqrt{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2}} & -\frac{1}{2}\mathsf{E}\,\frac{1}{\sqrt{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2}} \\ -\frac{1}{2}\mathsf{E}\,\frac{1}{\sqrt{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2}} & -\frac{1}{3}\mathsf{E}\,\frac{1}{\sqrt{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2}} \\ 0 & 0 \end{pmatrix},$$

$$\bar{\boldsymbol{\Delta}} = \bar{\boldsymbol{S}}_{12}^T\bar{\boldsymbol{S}}_{11}^{-1}\bar{\boldsymbol{S}}_{12}, \quad a = \mathsf{E}\,\frac{\sigma_t^2}{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2} \quad \text{and} \quad b = E\frac{1}{\sqrt{1+\sum_{k=1}^m(\gamma_0+e_{t-k})^2}}.$$

Then

$$\bar{\boldsymbol{\Sigma}}^{-1} = \begin{pmatrix} 4/a & -6/a & 0 \\ -6/a & 12/a & 0 \\ 0 & 0 & \bar{\sigma}^2 \end{pmatrix} \quad \text{and} \quad \bar{\boldsymbol{\Delta}} = \begin{pmatrix} b^2/a & b^2/(2a) \\ b^2/(2a) & b^2/(3a) \end{pmatrix}.$$

Like the proof of Theorem 2.3, we have

$$\begin{aligned} \bar{l}(1) &= -\bar{\boldsymbol{Z}}^T(\bar{\boldsymbol{S}}_{11}^{-1} - \bar{\boldsymbol{S}}_{11}^{-1}\bar{\boldsymbol{S}}_{12}\bar{\boldsymbol{\Delta}}^{-1}\bar{\boldsymbol{S}}_{12}^T\bar{\boldsymbol{S}}_{11}^{-1})\bar{\boldsymbol{Z}} + o_p(1) \\ &= \bar{\boldsymbol{Z}}^T \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \bar{\sigma}^{-2} \end{pmatrix} \bar{\boldsymbol{Z}} + o_p(1) \qquad (2.28) \\ &= (\bar{\boldsymbol{Z}}_3/\bar{\sigma})^2 + o_p(1) \xrightarrow{d} \chi^2(1). \end{aligned}$$

□

Proof of Theorem 2.8. Put

$$S_{n1} = \sum_{t=1}^n \frac{1}{\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}}, \quad S_{n2} = \sum_{t=1}^n \frac{t}{\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}},$$

$$S_{n3} = \sum_{t=1}^n \frac{X_{t-1}}{\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}}, \quad S_{n4} = \sum_{t=1}^n \frac{t}{n\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}},$$

$$S_{n5} = \sum_{t=1}^n \frac{t^2}{n\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}}, \quad S_{n6} = \sum_{t=1}^n \frac{tX_{t-1}}{n\sqrt{1+\sum_{k=1}^m(X_{t-k}-X_{t-k-1})^2}},$$

$$\bar{\mu}* = \mu^* + (\phi_0 - 1)\frac{S_{n3}S_{n5} - S_{n2}S_{n6}}{S_{n1}S_{n5} - S_{n2}S_{n4}}, \quad \bar{\mu}_0^* = \mu_0^* + (\phi_0 - 1)\frac{S_{n3}S_{n5} - S_{n2}S_{n6}}{S_{n1}S_{n5} - S_{n2}S_{n4}},$$

$$\bar{\gamma}^* = \gamma^* + (\phi_0 - 1)\frac{S_{n3}S_{n4} - S_{n1}S_{n6}}{S_{n2}S_{n4} - S_{n1}S_{n5}}, \quad \bar{\gamma}_0^* = \gamma_0^* + (\phi_0 - 1)\frac{S_{n3}S_{n4} - S_{n1}S_{n6}}{S_{n2}S_{n4} - S_{n1}S_{n5}}.$$

Then, we have

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\bar{Y}_{t1}(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{e_t}{\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}},$$

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\bar{Y}_{t2}(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{e_t t}{n\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}}$$

and

$$\begin{aligned}
& \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\bar{Y}_{t3}(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) \\
= \; & \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}W_t - (\phi_0-1)\tfrac{S_{n3}S_{n5}-S_{n2}S_{n6}}{S_{n1}S_{n5}-S_{n2}S_{n4}}\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\tfrac{X_{t-1}}{(1+X_{t-1}^2)^{0.75}\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}} \\
& -(\phi_0-1)\tfrac{S_{n3}S_{n4}-S_{n1}S_{n6}}{S_{n2}S_{n4}-S_{n1}S_{n5}}\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\tfrac{tX_{t-1}}{(1+X_{t-1}^2)^{0.75}\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}} \\
& +(\phi_0-1)\tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\tfrac{X_{t-1}^2}{(1+X_{t-1}^2)^{0.75}\sqrt{1+\sum_{k=1}^{m}(X_{t-k}-X_{t-k-1})^2}} + o_p(1).
\end{aligned}$$

i) Since $\mu_0^* = \mu_0(1-\phi_0)$, $\gamma_0^* = 0$ and $X_t = \mu_0(1-\phi_0^t) + \phi_0^t X_0 + \sum_{j=1}^{t}\phi_0^{t-j}e_j$, like the proof of (2.26), we can show that

$$\frac{X_{[ns]}}{n^{\alpha/\delta}} = -\int_0^\infty e^{-d_1 sr}\, d\Gamma_\alpha(s,r) + o_p(1) \quad \text{uniformly in} \quad s \in [0,1],$$

which is used to show that $X_{[ns]} - X_{[ns]-1} = e_{[ns]} + o_p(1)$ uniformly in $s \in [0,1]$,

$$\frac{S_{n1}}{n} \xrightarrow{p} \Delta_2, \quad \frac{S_{n2}}{n^2} \xrightarrow{p} \frac{\Delta_2}{2}, \quad \frac{S_{n3}}{n^{1+\alpha/\delta}} \xrightarrow{d} \Delta_2\int_0^1 L^*(s)\, ds,$$

$$\frac{S_{n4}}{n} \xrightarrow{p} \frac{\Delta_2}{2}, \quad \frac{S_{n5}}{n^2} \xrightarrow{p} \frac{\Delta_2}{3}, \quad \frac{S_{n6}}{n^{1+\alpha/\delta}} \xrightarrow{d} \Delta_2\int_0^1 sL^*(s)\, ds,$$

$$\begin{aligned}
& \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\bar{Y}_{t3}(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) \\
= \; & \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}W_t + d_1\{4\int_0^1 L^*(s)\, ds - 6\int_0^1 sL^*(s)\, ds\}\Delta_2\int_0^1 L^*(s)|L^*(s)|^{-3/2}\, ds \\
& +d_1\{-6\int_0^1 L^*(s)\, ds + 12\int_0^1 sL^*(s)\, ds\}\Delta_2\int_0^1 sL^*(s)|L^*(s)|^{-3/2}\, ds \\
& -d_1\Delta_2\int_0^1 |L^*(s)|^{1/2}\, ds + o_p(1) \\
= \; & \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}W_t + d_1 d_4\Delta_2 + o_p(1),
\end{aligned}$$

$$\frac{1}{n}\sum_{t=1}^{t}\bar{Y}_{t1}^2(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) \xrightarrow{p} \Delta_1, \quad \frac{1}{n}\sum_{t=1}^{n}\bar{Y}_{t2}^2(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) \xrightarrow{p} \frac{\Delta_1}{3}, \quad \frac{1}{n}\sum_{t=1}^{n}\bar{Y}_{t3}^2(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) \xrightarrow{p} \mathsf{E}\, W_1^2.$$

Hence it follows from the proof of Theorem 2.7 that

$$\bar{l}(1) = \frac{\{\sum_{t=1}^n \bar{Y}_{t3}(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1)\}^2}{\sum_{t=1}^n \bar{Y}_{t3}^2(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1)} + o_p(1) = \frac{\{\frac{1}{\sqrt{n}}\sum_{t=1}^n W_t + d_1 d_4 \Delta_2\}^2}{\mathsf{E}\, W_1^2} + o_p(1),$$

i.e., Theorem 2.8 i) holds.

ii) Since $\mu_0^* = \mu_0(1-\phi_0) + \phi_0\gamma_0$, $\gamma_0^* = \gamma_0(1-\phi_0)$ and

$$X_t = \mu_0^* \frac{1-\phi_0^t}{1-\phi_0} + \gamma_0^* \sum_{j=1}^t \phi_0^{t-j} j + \sum_{j=1}^t \phi_0^{t-j} e_j + \phi_0^t X_0,$$

like the proof of (2.18), we have

$$\begin{aligned}
\tfrac{X_{[ns]}}{n} &= \mu_0^* \tfrac{1-(1-d_1/n)^{[ns]}}{d_1} + \gamma_0^* \tfrac{1}{n} \textstyle\sum_{j=1}^{[ns]} (1 - \tfrac{d_1}{n})^{[ns]-j} j + \tfrac{1}{n} \textstyle\sum_{j=1}^{[ns]} (1-\tfrac{d_1}{n})^{[ns]-j} e_j + o_p(1) \\
&= \mu_0^* \tfrac{1-e^{-d_1 s}}{d_1} + \gamma_0^* n \textstyle\int_0^s t e^{-(s-t)d_1}\, dt + o_p(1) \\
&= \tfrac{\gamma_0(1-e^{-d_1 s})}{d_1} + \gamma_0 d_1 \textstyle\int_0^s t e^{-(s-t)d_1}\, dt + o_p(1) \\
&= \gamma_0 s + o_p(1) \quad \text{uniformly in} \quad s \in [0,1],
\end{aligned}$$

which is used to show that

$$X_{[ns]} - X_{[ns]-1} = \mu_0^* + \gamma_0^*[ns] + (\phi_0 - 1)X_{[ns]-1} + e_{[ns]} = \gamma_0 + e_{[ns]} + o_p(1) \quad \text{uniformly in} \quad s \in [0,1],$$

$$\frac{S_{n1}}{n} \xrightarrow{p} \Delta_3, \quad \frac{S_{n2}}{n^2} \xrightarrow{p} \frac{\Delta_3}{2}, \quad \frac{S_{n3}}{n^2} \xrightarrow{p} \frac{\gamma_0\Delta_3}{2}, \quad \frac{S_{n4}}{n} \xrightarrow{p} \frac{\Delta_3}{2}, \quad \frac{S_{n5}}{n^2} \xrightarrow{p} \frac{\Delta_3}{3}, \quad \frac{S_{n6}}{n^2} \xrightarrow{p} \frac{\gamma_0\Delta_3}{3}$$

with $\Delta_3 = \mathsf{E}\, \frac{1}{\sqrt{1+\sum_{k=1}^m (\gamma_0 + e_{t-k})^2}}$, and

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{t=1}^n \bar{Y}_{t3}(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) = \frac{1}{\sqrt{n}} \sum_{t=1}^n W_t + o_p(1), \\
&\frac{1}{n} \sum_{t=1}^n \bar{Y}_{t1}^2(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) \xrightarrow{p} \Delta_1, \quad \frac{1}{n} \sum_{t=1}^n \bar{Y}_{t2}^2(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) \xrightarrow{d} \frac{\Delta_1}{3}, \quad \frac{1}{n} \sum_{t=1}^n \bar{Y}_{t3}^2(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1) \xrightarrow{p} \mathsf{E}\, W_1^2.
\end{aligned}$$

Hence it follows from the proof of Theorem 2.7 that

$$\bar{l}(1) = \frac{\{\sum_{t=1}^n \bar{Y}_{t3}(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1)\}^2}{\sum_{t=1}^n \bar{Y}_{t3}^2(\bar{\mu}_0^*, \bar{\gamma}_0^*, 1)} + o_p(1) \xrightarrow{d} \chi_1^2,$$

i.e., Theorem 2.8 ii) holds.

□

Proof of Theorem 2.9. Write

$$\begin{aligned}
&\textstyle\sum_{t=1}^n (X_t - \mu - \sum_{j=1}^r \phi_j X_{t-j}) \frac{1}{1+\sum_{k=1}^{m+1} X_{t-k}^2} \\
= \;&\textstyle\sum_{t=1}^n \frac{\sigma_t \epsilon_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2} - (\mu - \mu_0) \sum_{t=1}^n \frac{1}{1+\sum_{k=1}^{m+1} X_{t-k}^2} - \sum_{j=1}^r (\phi_j - \phi_{j,0}) \sum_{t=1}^n \frac{X_{t-j}}{1+\sum_{k=1}^{m+1} X_{t-k}^2}
\end{aligned}$$

and

$$\begin{aligned}
&\textstyle\sum_{t=1}^n (X_t - \mu - \sum_{j=1}^r \phi_j X_{t-j}) \frac{X_{t-i}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} \\
= \;&\textstyle\sum_{t=1}^n \frac{X_{t-i}\sigma_t \epsilon_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2} - (\mu - \mu_0) \sum_{t=1}^n \frac{X_{t-i}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} - \sum_{j=1}^r (\phi_j - \phi_{j,0}) \sum_{t=1}^n \frac{X_{t-i}X_{t-j}}{1+\sum_{k=1}^{m+1} X_{t-k}^2}
\end{aligned}$$

for $i = 1, \cdots, r$. Since $\{X_t\}$ is strictly stationary and both $\frac{\sigma_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2}$ and $\frac{X_{t-1}\sigma_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2}$ are bounded by a constant, we can show that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n \left(\frac{\sigma_t \epsilon_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2}, \frac{X_{t-1}\sigma_t \epsilon_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2}, \cdots, \frac{X_{t-r}\sigma_t \epsilon_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2} \right)^T \xrightarrow{d} N(\mathbf{0}, \mathbf{\Gamma})$$

and

$$\frac{1}{n} \sum_{t=1}^n \begin{pmatrix} \frac{1}{1+\sum_{k=1}^{m+1} X_{t-k}^2} & \frac{X_{t-1}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} & \cdots & \frac{X_{t-r}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} \\ \frac{X_{t-1}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} & \frac{X_{t-1}X_{t-1}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} & \cdots & \frac{X_{t-1}X_{t-r}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} \\ \cdot & \cdot & \cdots & \cdot \\ \frac{X_{t-r}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} & \frac{X_{t-r}X_{t-1}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} & \cdots & \frac{X_{t-r}X_{t-r}}{1+\sum_{k=1}^{m+1} X_{t-k}^2} \end{pmatrix} \xrightarrow{p} \boldsymbol{B},$$

i.e., Theorem 2.9 holds. □

Lemma 2.5. *Under conditions of Theorem 2.10, we have*

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{t=1}^n \hat{\mathbf{Y}}_t(\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Gamma}), \\
&\frac{1}{n} \sum_{t=1}^n \hat{\mathbf{Y}}_t(\boldsymbol{\theta}_0) \hat{\mathbf{Y}}_t^T(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{\Gamma}, \\
&\max_{1 \le t \le n} \left\| \hat{\boldsymbol{Y}}_t(\boldsymbol{\theta}_0) \right\| = o_p(n^{1/2}),
\end{aligned}$$

as $n \to \infty$, where $\mathbf{\Gamma}$ is defined in Theorem 2.9.

Proof of Lemma 2.5. This can be shown by using the central limit theorem for martingale differences and by noting that $\{X_t\}$ is strictly stationary and both $\frac{\sigma_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2}$ and $\frac{X_{t-1}\sigma_t}{1+\sum_{k=1}^{m+1} X_{t-k}^2}$ are bounded by a constant. □

Proof of Theorem 2.10. Theorem 2.10 can be proved by using Lemma 2.5 and standard arguments in empirical likelihood method (see Chapter 11 of Owen, 2001). □

Proof of Theorem 2.11. When GARCH(p, 0) errors in Theorems 2.1–2.10 are replaced by GARCH(p, q) errors, the arguments of $\sigma_t^2/\{1+\sum_{k=1}^m (X_{t-k}-X_{t-k-1})^2\}$ being bounded by a constant are no longer valid. However, all proofs for the above lemmas and theorems are still valid if

$$\mathsf{E}\left\{\frac{\sigma_t^2}{1+\sum_{k=1}^m (X_{t-k}-X_{t-k-1})^2}\right\}^{1+\delta} < \infty \quad \text{for some} \quad \delta > 0. \tag{2.29}$$

Now equation (2.29) easily follows from the arguments in proving (2.12). Hence, Theorem 2.11 follows. □

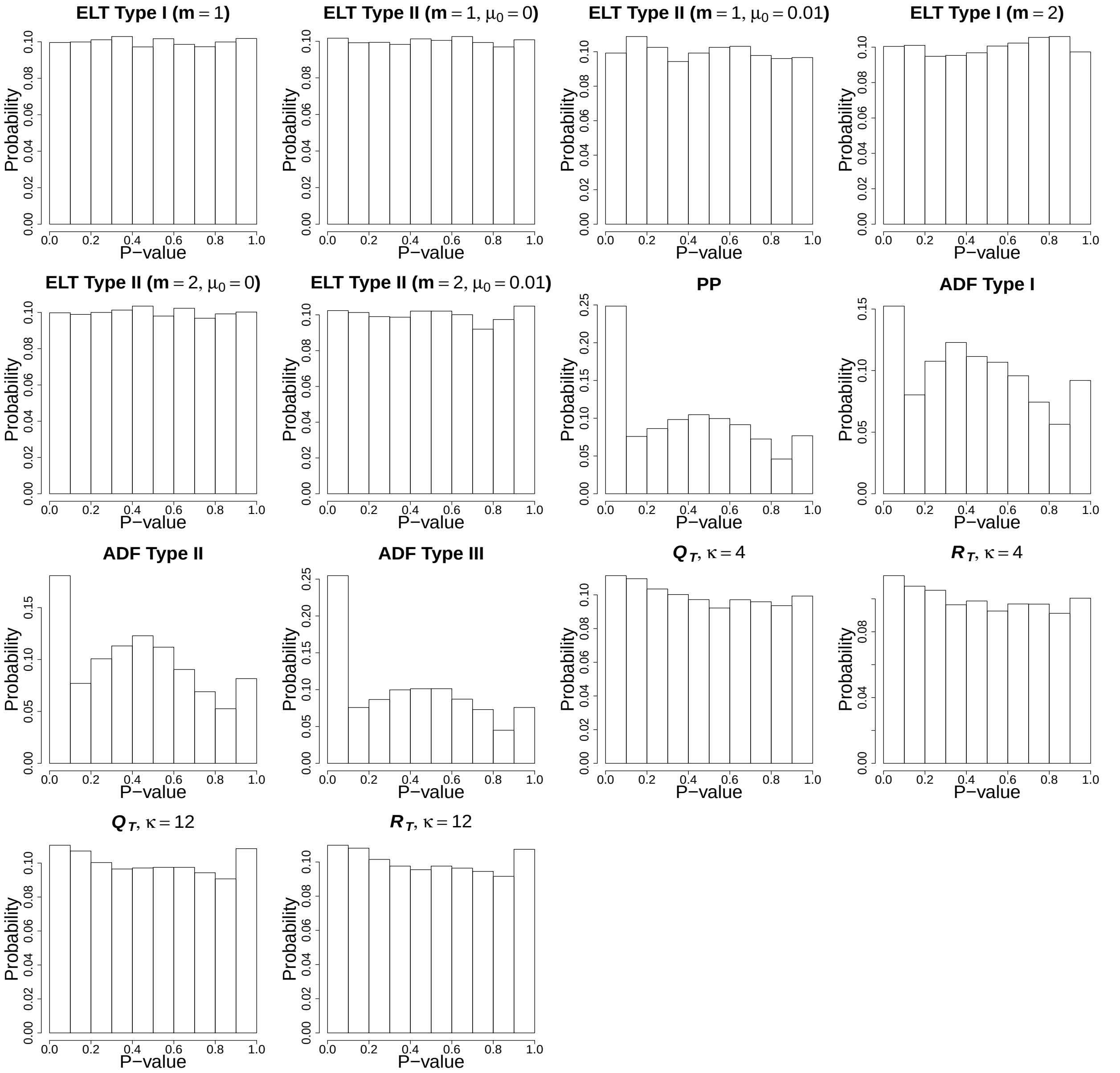


Figure 2.1. A large-sample comparison of the asymptotic validity of several unit root tests for model (2.1) with infinite variance GARCH errors.

This figure plots the histograms of P-values under the unit root null hypothesis from the empirical likelihood unit root tests based on the extension of Theorem 2.1 (ELT Type I) and Theorem 2.3 (ELT Type II) in Theorem 2.11, the Phillips-Perron (PP) test, the augmented Dickey-Fuller (ADF) test (Type I: no drift, no linear trend; Type II: with drift, no linear trend; Type III: with drift and linear trend), and the sieve wild bootstrap ADF tests (Q_T and R_T, $\kappa = 4, 12$). AR(1)-GARCH(1, 1) model is generated with $(\phi_0, \alpha_0, \alpha_1, \beta_1) = (1, 4.717\text{e} - 07, 0.1266, 0.8784)$, $\epsilon_t \sim$ standardized skew normal$(0, 1, 10)$, and $n = 10000$. (There are 5 'NA's for the wild bootstrap tests when $\kappa = 12$.)

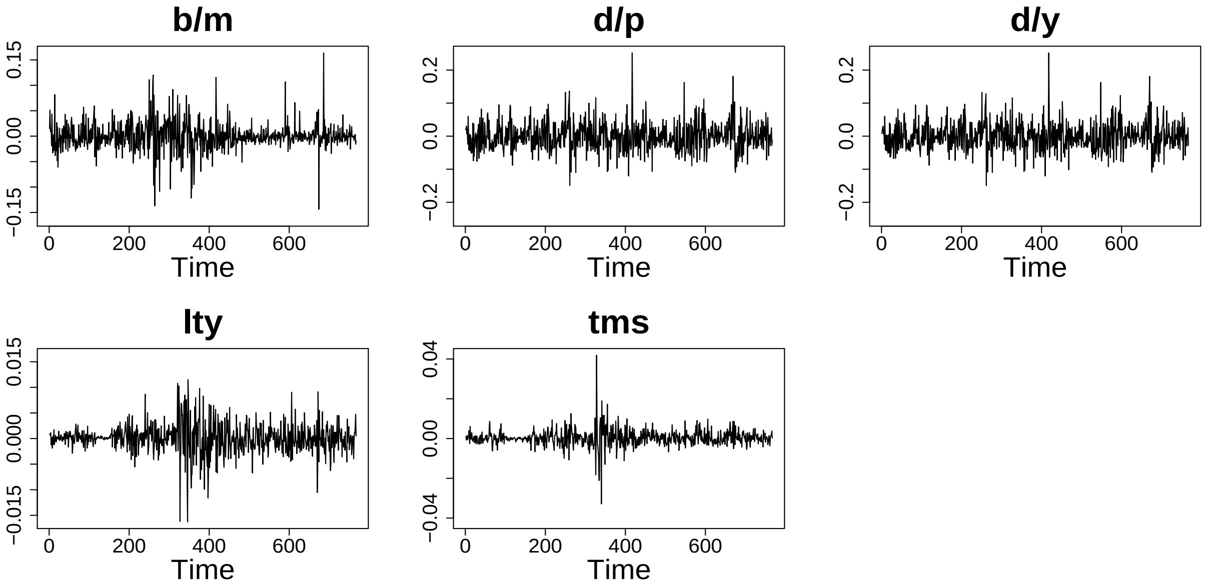


Figure 2.2. Residuals of financial ratios in the post-war period.

This figure plots the residuals in model (2.1) for monthly dividend-price ratio (d/p), dividend yield (d/y), book-to-market value ratio (b/m), long-term yield (lty), term pread (tms) in the post-war period (1953/01–2016/12).

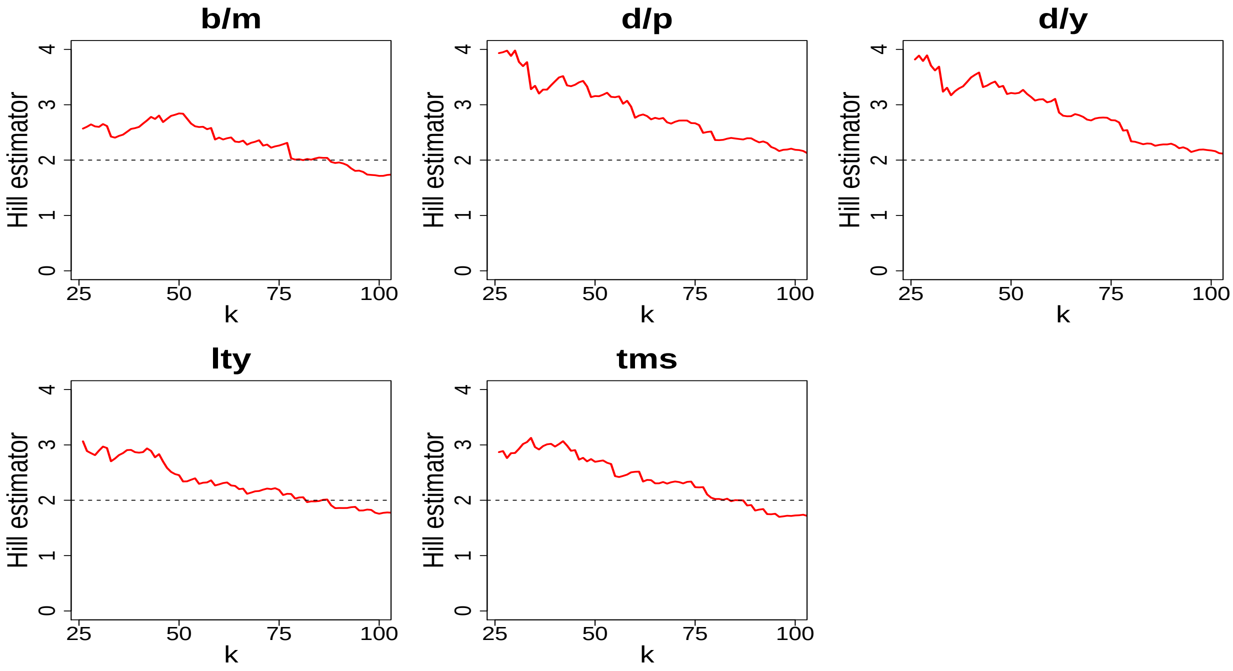


Figure 2.3. Hill estimates for residuals of financial ratios in the post-war period.

This figure plots Hill estimates against the sample fraction k for the residuals in model (2.1) for monthly dividend-price ratio (d/p), dividend yield (d/y), book-to-market value ratio (b/m), long-term yield (lty), term spread (tms) in the post-war period (1953/01–2016/12).

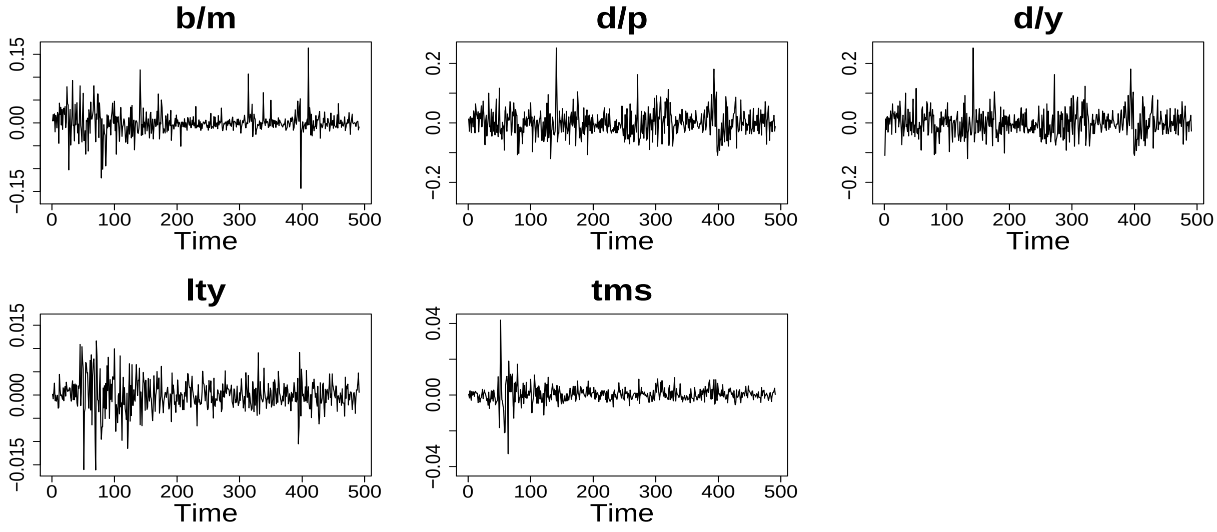


Figure 2.4. Residuals of financial ratios in the period after the oil shock recession.

This figure plots the residuals in model (2.1) for monthly dividend-price ratio (d/p), dividend yield (d/y), book-to-market value ratio (b/m), long-term yield (lty), term spread (tms) in the period after the oil shock recession (1976/01–2016/12).

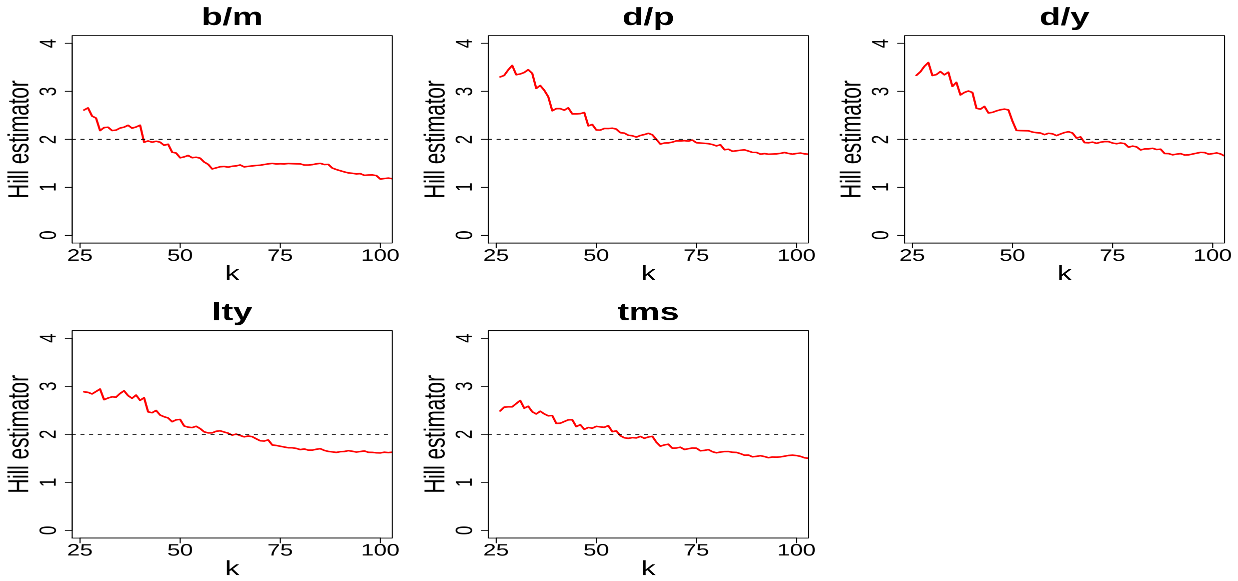


Figure 2.5. Hill estimates for residuals of financial ratios in the period after the oil shock recession.

This figure plots Hill estimates against the sample fraction k for the residuals in model (2.1) for monthly dividend-price ratio (d/p), dividend yield (d/y), book-to-market value ratio (b/m), long-term yield (lty), term spread (tms) in the period after the oil shock recession (1976/01–2016/12).

Table 2.1. Empirical size of several unit root tests for AR processes with ARCH errors

This table reports the empirical size of empirical likelihood unit root tests (ELT Type I in Theorem 2.1, ELT Type II in Theorem 2.3), Phillips-Perron (PP) test, augmented Dickey-Fuller (ADF) test (Type I: no drift nor linear trend; Type II: with drift, no linear trend; Type III: with drift and linear trend) and sieve wild bootstrap ADF tests (Q_T and R_T). Except for ELT Type II with $\mu_0 = 0.01$, AR(1)-ARCH(1) model is generated with $\phi_0 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2.5$, $\epsilon_t \sim$ standardized skew normal $(0, 1, 10)$ and $n = 1000, 2000, 5000$. Empirical size (multiplied by 100) is computed at levels $\tau = 0.05, 0.10, 0.25$. (NA counts the number of cases where the sieve wild bootstrap ADF test fails to compute the P-value.)

		ELT			PP	ADF			Wild Bootstrap ADF					
									$\kappa = 4$			$\kappa = 12$		
		Type I	Type II			Type I	Type II	Type III	Q_T	R_T	NA	Q_T	R_T	NA
n	τ		$\mu_0 = 0$	$\mu_0 = 0.01$										
1000	0.05	5.01	4.86	4.91	29.16	20.42	30.27	34.34	4.35	4.34	**113**	3.53	3.37	**254**
	0.10	9.88	9.62	9.56	32.68	24.40	34.78	38.74	8.33	8.16		7.49	7.18	
	0.25	25.63	25.08	25.06	38.46	31.56	42.48	45.18	20.39	20.25		19.30	18.74	
2000	0.05	4.97	4.61	4.61	29.52	21.68	32.74	36.96	4.15	4.11	**71**	3.68	3.43	**172**
	0.10	10.45	9.38	9.40	32.97	25.54	37.23	41.01	7.95	7.83		6.90	6.83	
	0.25	25.57	25.01	24.99	38.09	32.44	44.04	46.56	20.25	20.12		18.80	18.38	
5000	0.05	4.96	5.31	5.32	27.68	20.74	32.31	36.78	3.70	3.77	**43**	3.07	3.06	**87**
	0.10	9.96	10.29	10.29	30.87	24.33	36.53	40.83	6.98	7.06		6.61	6.31	
	0.25	25.12	25.66	25.68	36.14	31.14	43.67	46.17	19.51	19.55		17.79	17.83	

Table 2.2. Empirical size and power of several unit root tests for AR processes with GARCH errors

This table reports the empirical size and power of empirical likelihood unit root tests based on the extension of Theorem 2.1 (ELT Type I) and Theorem 2.3 (ELT Type II) in Theorem 2.11, Phillips-Perron (PP) test, augmented Dickey-Fuller (ADF) test (Type I: no drift nor linear trend; Type II: with drift, no linear trend; Type III: with drift and linear trend) and sieve wild bootstrap ADF tests (Q_T and R_T). AR(1)-GARCH(1,1) model is generated with $\alpha_0 = 4.717\text{e}{-07}$, $\alpha_1 + \beta_1 = 0.9545, 1, 1.005$, $\epsilon_t \sim$ standardized skew normal$(0, 1, \xi)$, $\xi = 0, 10$, and $n = 200, 500, 2000$. Empirical sizes and powers (multiplied by 100) are computed at the level $\tau = 0.05$. Local alternatives for ELT Type II when $\mu_0 = 0$ and $\mu_0 = 0.01$ are $\phi_0 = 1 - d/(5\sqrt{n})$ and $\phi_0 = 1 - d/(5n)$, respectively; all other tests are based on $\phi_0 = 1 - d/n$, with $d = 0, 5, 10$. (NA counts the number of cases where the wild bootstrap ADF test fails to compute the P-value.)

			ELT			PP	ADF			Wild Bootstrap ADF				
										$\kappa = 4$		$\kappa = 12$		
			Type I	Type II			Type I	Type II	Type III	Q_T	R_T	Q_T	R_T	NA
n	ξ	d		$\mu_0 = 0$	$\mu_0 = 0.01$									
	$\alpha_1 = 0.1216,\ \beta_1 = 0.8329$													
200	0	0	5.63	7.56	4.81	5.87	5.58	4.92	5.46	4.67	4.42	5.15	3.81	
		5	17.26	49.92	71.17	10.44	33.59	12.88	8.82	28.50	27.07	25.09	19.12	
		10	43.95	89.85	80.59	22.68	75.44	32.31	19.51	64.37	62.29	48.75	41.43	
	10	0	5.64	8.01	5.09	6.48	6.22	5.21	5.72	5.30	5.14	5.31	3.97	
		5	17.16	46.90	73.54	11.44	35.15	14.02	9.73	29.48	27.85	25.88	20.02	**1**
		10	42.75	86.10	81.85	24.61	77.61	35.17	21.33	65.53	63.40	50.11	42.89	**1**
500	0	0	5.21	5.96	5.08	6.72	5.30	6.36	6.45	4.90	4.81	5.25	4.77	
		5	17.19	68.38	81.37	11.02	32.98	13.46	10.11	30.01	28.93	28.63	25.51	
		10	42.32	98.18	90.65	23.07	75.57	32.99	21.28	69.12	67.70	61.81	58.19	
	10	0	5.33	6.79	4.96	6.94	5.45	5.97	6.55	4.81	4.78	5.07	4.46	
		5	16.83	64.18	83.57	11.74	34.69	14.41	10.59	31.19	30.23	29.21	26.20	
		10	41.33	96.44	91.50	24.09	76.91	34.86	21.95	69.39	68.23	61.77	58.47	
2000	0	0	5.45	5.72	5.43	5.73	4.84	5.56	5.67	4.71	4.70	4.46	4.31	
		5	16.38	93.09	89.80	9.52	32.49	12.36	9.41	31.34	30.61	30.46	29.54	
		10	40.40	99.99	96.15	19.96	74.88	31.62	19.49	73.04	72.27	70.14	68.59	
	10	0	4.91	5.25	5.22	6.17	5.20	5.34	6.23	4.93	4.99	5.11	4.99	
		5	16.91	90.69	91.45	9.95	34.41	13.88	9.91	32.72	32.22	31.29	29.92	
		10	41.52	99.95	96.70	21.60	76.81	34.55	21.69	73.71	72.95	71.06	69.53	
	$\alpha_1 = 0.1216,\ \beta_1 = 0.8784$													
200	0	0	5.59	6.55	4.70	5.26	7.12	4.07	4.73	4.89	4.83	4.73	3.89	
		5	15.75	45.67	37.74	10.51	34.54	13.39	8.96	25.53	24.73	21.37	17.13	**5**
		10	39.58	85.54	50.54	22.57	73.70	32.38	20.23	57.68	56.51	42.00	35.76	**8**
	10	0	5.39	6.94	5.01	5.73	6.96	3.79	4.92	5.34	5.34	5.21	4.11	
		5	15.87	43.06	40.42	11.11	36.70	14.16	9.60	26.75	25.73	22.93	18.54	**2**
		10	39.90	82.66	53.33	24.41	75.95	35.55	21.38	59.14	57.96	43.79	37.46	**9**
500	0	0	5.29	5.60	4.90	6.82	7.28	6.08	6.29	4.84	4.87	5.42	5.00	**2**
		5	16.39	61.13	32.66	12.40	37.03	15.53	11.17	27.63	26.97	26.55	24.48	
		10	37.55	95.16	47.56	24.74	74.54	36.22	23.09	60.68	60.05	54.16	51.84	**1**
	10	0	5.23	6.14	4.78	7.18	7.33	5.14	6.56	4.89	4.82	5.31	4.75	**3**
		5	16.16	56.69	36.82	12.72	39.10	16.48	11.31	28.83	27.88	27.43	25.19	**2**
		10	37.84	92.39	52.77	25.93	78.42	38.84	23.99	62.65	62.10	56.48	54.09	**2**
2000	0	0	5.16	5.51	5.51	10.21	8.03	9.26	10.18	4.92	4.77	4.93	4.86	
		5	17.37	86.91	29.43	15.35	38.89	18.80	14.96	28.12	27.51	28.38	26.97	
		10	38.35	99.41	44.14	26.92	74.46	38.76	26.81	61.64	60.94	60.77	59.36	
	10	0	4.85	5.30	5.15	11.89	8.44	8.83	11.80	5.37	5.36	5.61	5.60	**1**
		5	18.10	79.88	34.15	18.46	44.21	24.18	18.63	31.45	30.83	30.97	29.78	**1**
		10	39.96	98.48	51.11	33.68	81.27	47.45	32.97	67.99	67.24	65.15	63.79	**1**

Table 2.2 (cont'd). Empirical size and power of several unit root tests for AR processes with GARCH errors

			ELT			PP	ADF			Wild Bootstrap ADF				
										$\kappa = 4$			$\kappa = 12$	
			Type I	Type II			Type I	Type II	Type III	Q_T	R_T	Q_T	R_T	NA
n	ξ	d		$\mu_0 = 0$	$\mu_0 = 0.01$									
	$\alpha_1 = 0.1266,\ \beta_1 = 0.8784$													
200	0	0	5.64	6.52	4.71	5.43	7.61	4.12	4.79	4.96	4.88	4.73	3.87	
		5	15.49	44.28	33.48	11.00	35.13	13.67	9.40	24.99	24.18	20.88	16.90	**8**
		10	38.41	83.96	45.45	23.29	73.19	32.77	20.60	56.01	54.93	40.43	34.66	**15**
	10	0	5.31	6.70	4.93	5.87	7.14	3.69	4.89	5.27	5.25	5.16	4.30	
		5	15.68	42.19	36.20	11.41	36.93	14.46	9.83	26.20	25.24	22.28	18.09	**5**
		10	39.15	81.38	48.11	24.70	75.33	35.81	21.74	57.65	56.55	42.36	35.99	**13**
500	0	0	5.33	5.33	4.92	6.83	8.30	5.93	6.39	4.87	4.95	5.54	4.98	**4**
		5	15.70	58.27	25.52	13.03	37.82	15.91	11.82	26.76	25.98	25.32	23.46	**4**
		10	36.07	93.26	37.46	25.33	74.02	36.83	23.66	57.87	57.20	51.74	49.62	**5**
	10	0	5.21	5.79	4.78	7.31	7.92	5.08	6.57	4.83	4.79	5.17	4.84	**6**
		5	15.68	54.60	29.57	13.40	39.77	17.27	11.92	27.48	26.70	26.10	24.32	**3**
		10	37.02	90.49	43.43	26.50	77.83	39.37	24.64	60.34	59.49	53.91	51.70	**4**
2000	0	0	5.52	5.59	5.50	11.60	9.97	10.55	11.55	5.09	5.08	5.18	4.97	**5**
		5	17.24	79.93	18.32	17.19	41.13	21.42	17.15	26.35	25.71	26.25	25.02	**5**
		10	36.38	97.15	28.01	25.33	74.02	36.83	23.66	57.87	57.20	51.74	49.62	**5**
	10	0	5.00	4.83	5.15	7.31	7.92	5.08	6.57	4.83	4.79	5.17	4.84	**6**
		5	17.95	73.80	22.74	13.40	39.77	17.27	11.92	27.48	26.70	26.10	24.32	**3**
		10	39.03	96.06	35.32	26.50	77.83	39.37	24.64	60.34	59.49	53.91	51.70	**4**

Table 2.3. Empirical comparison of several estimators for AR parameters in AR processes with ARCH errors

This table reports the mean, standard deviation (SD) and root-mean-square error (RMSE) of the point estimate $(\hat{\mu}, \hat{\phi})$ of (μ_0, ϕ_0) in model 2.3 by weighted least squares estimator (WLSE), self-weighted least absolute deviation estimator (SLADE) in Zhu and Ling (2015) and least squares estimator (LSE). AR(1)-ARCH(3) model is generated with $(\mu_0, \phi_0, \alpha_0, \alpha_1, \alpha_2, \alpha_3) = (1.9037\text{e}{-03}, -0.1954, 8.4511\text{e}{-05}, 0.6228, 0.4040, 0.2898)$, $\epsilon_t \sim t_\nu/\sqrt{\nu/(\nu-2)}$, $\nu = 2.8, 5, 10$ or N(0,1), and $n = 500, 2000, 5000$.

n	$\epsilon_t \sim$		$\hat{\mu}$			$\hat{\phi}$		
			WLSE	SLADE	LSE	WLSE	SLADE	LSE
500	$t_{2.8}$	Mean	1.8934e-03	1.4876e-03	1.8882e-03	−0.1953	−0.1638	−0.1875
		SD	6.7086e-04	4.8192e-04	3.3900e-03	0.0787	0.0763	0.1466
		RMSE	6.7094e-04	6.3672e-04	3.3900e-03	0.0787	0.0826	0.1468
	t_5	Mean	1.9014e-03	1.0890e-03	1.8074e-03	−0.1957	−0.1526	−0.1851
		SD	1.0101e-03	9.2938e-04	2.1031e-02	0.0727	0.1036	0.1596
		RMSE	1.0101e-03	1.2359e-03	2.1031e-02	0.0727	0.1121	0.1600
	t_{10}	Mean	1.8985e-03	7.6660e-03	3.8710e-03	−0.1965	−0.1476	−0.1860
		SD	3.6954e-03	1.3059e-03	1.6313e-01	0.0742	0.1174	0.1647
		RMSE	3.6954e-03	1.7316e-03	1.6314e-01	0.0742	0.1267	0.1649
	N(0,1)	Mean	1.8780e-03	4.2936e-04	-4.1111e-03	−0.1956	−0.1425	−0.1842
		SD	1.0567e-02	1.7662e-03	7.6673e-01	0.0739	0.1220	0.1673
		RMSE	1.0567e-02	2.3007e-03	7.6676e-01	0.0739	0.1330	0.1677
2000	$t_{2.8}$	Mean	1.8960e-03	1.4826e-03	1.8918e-03	−0.1958	−0.1636	−0.1873
		SD	3.2271e-04	2.4212e-04	1.5048e-03	0.0410	0.0468	0.1336
		RMSE	3.2280e-04	4.8579e-04	1.5048e-03	0.0410	0.0566	0.1339
	t_5	Mean	1.8978e-03	1.0739e-03	1.8010e-03	−0.1955	−0.1529	−0.1874
		SD	4.5444e-04	4.7637e-04	2.7802e-02	0.0365	0.0796	0.1497
		RMSE	4.5448e-04	9.5685e-04	2.7802e-02	0.0365	0.0903	0.1499
	t_{10}	Mean	1.8944e-03	7.6295e-04	2.8758e-03	−0.1957	−0.1498	−0.1854
		SD	6.9074e-04	6.4926e-04	6.2189e-02	0.0395	0.0994	0.1592
		RMSE	6.9081e-04	1.3126e-03	6.2196e-02	0.0395	0.1093	0.1595
	N(0,1)	Mean	1.9166e-03	4.2149e-04	3.8944e-04	−0.1955	−0.1486	−0.1844
		SD	2.3490e-03	8.4741e-04	4.7601e-01	0.0442	0.1199	0.1616
		RMSE	2.3490e-03	1.7074e-03	4.7601e-01	0.0442	0.1287	0.1620
5000	$t_{2.8}$	Mean	1.8999e-03	1.4859e-03	1.8818e-03	−0.1956	−0.1641	−0.1904
		SD	1.9917e-04	1.5500e-04	9.5334e-04	0.0243	0.0331	0.1262
		RMSE	1.9920e-04	4.4564e-04	9.5359e-04	0.0243	0.0455	0.1263
	t_5	Mean	1.9019e-03	1.0744e-03	1.8007e-03	−0.1955	−0.1530	−0.1848
		SD	2.8350e-04	3.0964e-04	1.7578e-02	0.0234	0.0692	0.1468
		RMSE	2.8350e-04	8.8521e-04	1.7578e-02	0.0234	0.0812	0.1472
	t_{10}	Mean	1.9390e-03	7.6721e-04	7.8363e-03	−0.1956	−0.1514	−0.1887
		SD	5.3188e-03	4.2073e-04	9.3850e-01	0.0277	0.0919	0.1558
		RMSE	5.3190e-03	1.2119e-03	9.3852e-01	0.0277	0.1018	0.1559
	N(0,1)	Mean	1.9143e-03	4.2191e-04	2.7365e-03	−0.1957	−0.1518	−0.1854
		SD	9.0584e-04	5.5958e-04	3.1451e-01	0.0296	0.1156	0.1595
		RMSE	9.0590e-04	1.5839e-03	3.1451e-01	0.0296	0.1235	0.1598

Table 2.4. Empirical coverage probabilities of the empirical likelihood confidence region for AR parameters in AR processes with ARCH errors

This table reports the empirical coverage probabilities of the empirical likelihood confidence region for (μ_0, ϕ_0) in model (2.3) based on Theorem 2.10 at levels $1 - \tau = 0.90, 0.95$. AR(1)-ARCH(3) model is generated with $(\mu_0, \phi_0, \alpha_0, \alpha_1, \alpha_2, \alpha_3) = (1.9037\text{e}{-}03, -0.1954, 8.4511\text{e}{-}05, 0.6228, 0.4040, 0.2898)$, $\epsilon_t \sim t_\nu / \sqrt{\nu/(\nu-2)}$, $\nu = 2.8, 5, 10$ or N(0,1), and $n = 500, 2000, 5000$.

$\epsilon_t \sim$	$n = 500$		$n = 2000$		$n = 5000$	
	$1 - \tau = 0.90$	$1 - \tau = 0.95$	$1 - \tau = 0.90$	$1 - \tau = 0.95$	$1 - \tau = 0.90$	$1 - \tau = 0.95$
$t_{2.8}$	0.8800	0.9347	0.8898	0.9436	0.8933	0.9445
t_5	0.8927	0.9454	0.9051	0.9541	0.8974	0.9501
t_{10}	0.8927	0.9448	0.8953	0.9466	0.9014	0.9508
N(0,1)	0.8950	0.9492	0.8962	0.9505	0.8994	0.9492

Table 2.5. Unit root tests for monthly financial ratios of stock return predictability during the period 1953/01–2016/12

This table reports P-values of the proposed unified empirical likelihood unit root tests based on the extension of Theorem 2.1 (ELT Type I) and Theorem 2.3 (ELT Type II) in Theorem 2.11, Phillips-Perron (PP) test, augmented Dickey-Fuller (ADF) test (Type I: no drift nor linear trend; Type II: with drift, no linear trend; Type III: with drift and linear trend) and sieve wild bootstrap ADF tests (Q_T and R_T).

	ELT				PP	ADF			Wild Bootstrap ADF			
Variable	Type I		Type II						$\kappa = 4$		$\kappa = 12$	
	$m = 1$	$m = 2$	$m = 1$	$m = 2$		Type I	Type II	Type III	Q_T	R_T	Q_T	R_T
b/m	0.5842	0.7817	0.9272	0.8417	0.5665	0.3362	0.4953	0.6222	0.2591	0.2351	0.4465	0.4248
d/p	0.9188	0.7262	0.8077	0.9848	0.4609	0.5311	0.3915	0.5575	0.5208	0.5048	0.5879	0.5809
d/y	0.7599	0.5648	0.5759	0.7565	0.4770	0.5226	0.4027	0.5686	0.5002	0.4932	0.5812	0.5805
lty	0.6415	0.7725	0.7015	0.6092	0.8287	0.3272	0.5288	0.8100	0.3845	0.4365	0.3871	0.4381
tms	0.0381	0.0492	0.2765	0.3280	0.0100	0.0100	0.0100	0.0100	0.0083	0.0090	0.0010	0.0013

Table 2.6. Unit root tests for monthly financial ratios of stock return predictability during the period 1976/01–2016/12

This table reports P-values of the proposed unified empirical likelihood unit root tests based on the extension of Theorem 2.1 (ELT Type I) and Theorem 2.3 (ELT Type II) in Theorem 2.11, Phillips-Perron (PP) test, augmented Dickey-Fuller (ADF) test (Type I: no drift nor linear trend; Type II: with drift, no linear trend; Type III: with drift and linear trend) and sieve wild bootstrap ADF tests (Q_T and R_T).

	ELT				PP	ADF			Wild Bootstrap ADF			
Variable	Type I		Type II						$\kappa = 4$		$\kappa = 12$	
	$m = 1$	$m = 2$	$m = 1$	$m = 2$		Type I	Type II	Type III	Q_T	R_T	Q_T	R_T
b/m	0.9458	0.9009	0.8617	0.8663	0.7833	0.5459	0.5907	0.8296	0.4985	0.4932	0.5915	0.5962
d/p	0.8512	0.8537	0.9124	0.7798	0.6878	0.4416	0.6182	0.7450	0.3951	0.3938	0.4428	0.4491
d/y	0.9028	0.9624	0.5727	0.6385	0.6958	0.4898	0.6227	0.7468	0.4715	0.4755	0.5335	0.5365
lty	0.5552	0.4094	0.6793	0.4794	0.0700	0.4347	0.7298	0.0685	0.4818	0.4295	0.5552	0.5052
tms	0.1057	0.1771	0.0511	0.1177	0.0116	0.0100	0.0100	0.0190	0.0113	0.0117	0.0050	0.0050

Table 2.7. Estimation and inference for log-returns of the daily HKD/USD exchange rate

This table reports the point estimates and estimated confidence intervals (CI) at confidence levels 90% and 95% of (μ, ϕ) in model (2.3) for log-returns of the daily HKD/USD exchange rate (January 21, 1998–July 6, 2000). We estimate (μ, ϕ) by the maximum likelihood estimator (MLE) in the `fGarch` package, the weighted least squares estimator (WLSE) in the extension of Theorems 2.9 in Theorem 2.11 with $m = 3$, and the self-weighted least absolute deviation estimator (SLADE) in Zhu and Ling (2015). Confidence intervals for WLSE and SLADE are constructed by profile empirical likelihood in Remark 2.7 and a random weighting approach in Zhu and Ling (2015), respectively.

Estimator	$(\hat{\mu}, \hat{\phi})$	90% CI		95% CI	
		$\hat{\mu}$	$\hat{\phi}$	$\hat{\mu}$	$\hat{\phi}$
MLE	(0.0019, -0.1954)	(1.3386e-03, 2.4688e-03)	(-0.2619, -0.1289)	(1.2304e-03, 2.5770e-03)	(-0.2746, -0.1161)
WLSE	(0.0015, -0.1660)	(-1.7572e-04, 2.4853e-03)	(-0.2560, -0.0700)	(-2.0329e-04, 2.6449e-03)	(-0.2729, -0.0503)
SLADE	(0.0012, -0.0918)	(3.5403e-04, 2.0153e-03)	(-0.1665, -0.0170)	(1.9490e-04, 2.1744e-03)	(-0.1808, -0.0027)

CHAPTER 3

Comovements and Asymmetric Tail Dependence in State Housing Prices in the US[1]

Abstract

We re-examine the methods used in estimating comovements among U.S. regional home prices and find that there are insufficient moments to ensure a normal limit necessary for employing the quasi-maximum likelihood estimator. Hence, we propose applying the self-weighted quasi-maximum exponential likelihood estimator and a bootstrap method to test and account for the asymmetry of comovements as well as different magnitudes across state pairs. Our results reveal interstate asymmetric tail dependence based on observed house price indices rather than residuals from fitting AR-GARCH models.

[1] This chapter is based on the joint work: Huang, H., Peng, L., & Yao, V. W. (2019). Comovements and asymmetric tail dependence in state housing prices in the USA: A nonparametric approach. *Journal of Applied Econometrics*, 34(5), 843-849.

3.1 Introduction

The existence of comovements and contagion in the housing market, especially during extreme market upswings and downswings, is an important stylized fact in the literature (Del Negro and Otrok, 2007; Glaeser and Gyourko, 2006; Shiller, 2007; Kuethe and Pede, 2011; Kallberg et al., 2014). The observation has broad policy implications: Leamer (2007) argued that housing is an important precursor of national business cycles; thus, regional housing markets must be studied to understand the transmission of business cycles and develop better policy insights (Hamilton and Owyang, 2012). In addition, since the recent Great Recession was largely characterized by housing and housing-induced financial crises, the study of comovements in regional housing prices is particularly useful in assessing the risk of structured securities, such as collateralized debt obligations (CDOs; Zimmer, 2012, 2015), as well as managing portfolios. As Coval et al. (2009) stated, correlated default probabilities are amplified when CDOs are sold in tranches. The authors argued that the housing crisis was exacerbated by (belief in) misspecified statistical distributions.

The evaluation of CDO ratings has traditionally relied on the Gaussian copula, which is popular because of its simplicity. However, its assumption that extreme events such as steep housing declines are unrelated can lead to significant underestimation of the magnitude of comovements (Zimmer, 2012) and mislead the valuation of housing and mortgage-related securities. Accurate modeling and understanding of comovements among regional home prices are of broad importance for better risk management (Kole et al., 2007; Siburg et al., 2015).

In this paper, we propose a new econometric framework and apply it to re-examine house price comovements among four so-called Sand States that were severely hit by the housing crisis, namely, California (CA), Florida (FL), Nevada (NV), and Arizona (AZ).[2] Our work is built on the work of Zimmer (2012) and Ho et al. (2016, 2019), who advocated

[2] We use the quarterly house price index (HPI) reports estimated and published by the Federal Housing Finance Agency (FHFA).

both parametric and nonparametric copulas in modeling comovements in housing as well as energy markets.

Recent research has found that US housing prices exhibit persistent volatility[3], which motivates the application of AR-GARCH models. Let $X_{i,t}$ denote quarterly percentage changes in the HPI for the ith state at time t. We can fit an AR(r)-GARCH(1,1) model to $X_{i,t}$ with the following equation:

$$X_{i,t} = \mu_i + \sum_{j=1}^{r} \phi_{i,j} X_{i,t-j} + \varepsilon_{i,t}, \quad \varepsilon_{i,t} = \sigma_{i,t} e_{i,t}, \quad \sigma_{i,t}^2 = \alpha_{i,0} + \alpha_{i,1}\varepsilon_{i,t-1}^2 + \beta_{i,1}\sigma_{i,t-1}^2, \tag{3.1}$$

where $\mu_i, \phi_{i,1}, \cdots, \phi_{i,r} \in \mathbb{R}$, $\alpha_{i,0} > 0, \alpha_{i,1} > 0, \beta_{i,1} > 0$ for $i = 1, \cdots, m$, $\{\boldsymbol{e}_t = (e_{1,t}, \cdots, e_{m,t})^T\}_{t=1}^n$ is a sequence of independent and identically distributed random vectors with zero means and unit variances. We have $m = 4$ in our data analysis.

To study housing price comovements, Zimmer (2012) proposed estimating

$$P(e_{j,t} < -k | e_{i,t} < -k) = \frac{C_{ij}(G_i(-k), G_j(-k))}{G_i(-k)}, \; P(e_{j,t} > k | e_{i,t} > k) = \frac{\bar{C}_{ij}(G_i(k), G_j(k))}{1 - G_i(k)} \tag{3.2}$$

via fitting a parametric copula to $C_{ij}(u_1, u_2) = P(G_i(e_{i,t}) \leq u_1, G_j(e_{j,t}) \leq u_2)$, $u_1, u_2 \in [0, 1]$, and a parametric family to each marginal distribution $G_i(x) = P(e_{i,t} \leq x)$ and $G_j(y) = P(e_{j,t} \leq y)$, where $\bar{C}_{ij}(u_1, u_2) = P(G_i(e_{i,t}) > u_1, G_j(e_{j,t}) > u_2)$. The two conditional probabilities in Eq. (3.2) could explain the lower and upper comovements of the residuals from fitting the AR(r)-GARCH(1,1) models to the HPI, respectively. However, the validity of this approach, as cautioned by Zimmer (2012), relies heavily on correct specifications of both the copula and marginal distributions.

To overcome the robustness issue of fitting restrictive parametric families to the copula and marginals, Ho et al. (2016) re-estimated these two quantities by using a nonparametric copula estimator and nonparametric smoothing distribution estimators for the marginals. Nonetheless, as shown in Figures 2–4 of their paper, the bootstrap confidence intervals are so wide, especially for the upper tail dependence, that it is difficult to justify the asymmetric tail dependence and significant changes in comovements due to FHFA data revision during the sample period.

[3] See, for example, Miao et al. (2011), Miles (2011), and Zhu et al. (2013).

Generally, comovement in the lower (upper) tail, between $X_{i,t}$ and $X_{j,t}$, is defined as the conditional probability of $X_{i,t}$ below a low threshold (above a high threshold) given $X_{j,t}$ below a low threshold (above a high threshold). The thresholds can be the same or different, such that the probability of being below (above) the threshold is the same for both variables. When the $X_{i,t}$ values are dependent and exhibit persistent volatility, the calculations of comovements between $e_{i,t}$ and $e_{j,t}$ are often studied after filtering $X_{i,t}$ and $X_{j,t}$ by time series models, as in Eq. (3.2). Note that $P(e_{i,t} \leq -k) \neq P(e_{i,t} > k)$ holds in most cases when the distribution of $e_{i,t}$ is asymmetric. Without knowing whether the distribution of $e_{i,t}$ is symmetric or not, one can draw spurious conclusions about the upper and lower comovements for the same k due to the fact that the conditional probabilities are intrinsically different. On the other hand, since financial products are directly related to $X_{i,t}$ rather than the unobserved $e_{i,t}$, the comovements between $X_{i,t}$ and $X_{j,t}$ could be more important to understand than those between $e_{i,t}$ and $e_{j,t}$.

These considerations motivate us to depart from the copula framework of Zimmer (2012) and to model the contemporaneous lower and upper comovements as follows:

$$\gamma^{-}_{j|i,e}(p) = P(G_i(e_{j,t}) < p|G_i(e_{i,t}) < p),\ \gamma^{+}_{j|i,e}(p) = P(G_i(e_{j,t}) > 1-p|G_i(e_{i,t}) > 1-p), \tag{3.3}$$

$$\gamma^{-}_{j|i,X}(p) = P(F_i(X_{j,t}) < p|F_i(X_{i,t}) < p),\ \gamma^{+}_{j|i,X}(p) = P(F_i(X_{j,t}) > 1-p|F_i(X_{i,t}) > 1-p), \tag{3.4}$$

$$\gamma^{*-}_{j|i,e}(p) = P(G_j(e_{j,t}) < p|G_i(e_{i,t}) < p),\ \gamma^{*+}_{j|i,e}(p) = P(G_j(e_{j,t}) > 1-p|G_i(e_{i,t}) > 1-p), \tag{3.5}$$

$$\gamma^{*-}_{j|i,X}(p) = P(F_j(X_{j,t}) < p|F_i(X_{i,t}) < p),\ \gamma^{*+}_{j|i,X}(p) = P(F_j(X_{j,t}) > 1-p|F_i(X_{i,t}) > 1-p), \tag{3.6}$$

where $p \in (0,1)$ and F_i and G_i denote the distribution functions of $X_{i,t}$ and $e_{i,t}$, respectively.

Let $G_i^{\leftarrow}$ denote the generalized inverse of G_i, and we rewrite Eq. (3.3) as

$$\gamma^{-}_{j|i,e}(p) = P(e_{j,t} < G_i^{\leftarrow}(p)|e_{i,t} < G_i^{\leftarrow}(p)),\ \gamma^{+}_{j|i,e}(p) = P(e_{j,t} > G_i^{\leftarrow}(1-p)|e_{i,t} > G_i^{\leftarrow}(1-p)),$$

which uses the same threshold for $e_{i,t}$ and $e_{j,t}$ as in Eq. (3.2) based on Zimmer (2012). Unlike Zimmer (2012), the new definition assumes the same probability for the conditional events with respect to the upper and lower tail dependence, i.e., $P(e_{i,t} < G_i^{\leftarrow}(p)) = P(e_{i,t} > G_i^{\leftarrow}(1-p)) = p$. In addition, comovements in Eq. (3.3) depend on the magnitude of $e_{j,t}$, while the quantities in Eq. (3.5) are invariant to the marginals and use different thresholds but with the same probability of being below or above the chosen threshold. On the other hand, comovement definitions in Eqs. (3.4) and (3.6) are more useful to analyze portfolios since a portfolio may not follow an AR-GARCH model even when its individual assets do. By design, the comovement measures in Eqs. (3.5) and (3.6) are invariant in terms of the ordering of the states, meaning that the effects of housing price changes in one state on the housing prices of the other state are the same for a particular state pair, whereas the comovement measures in Eqs. (3.3) and (3.4) vary with the order.

In this paper, we intend to provide rigorous procedures to answer the following questions: i) Are the wide intervals of Ho et al. (2016) due to the infeasible quasi-maximum likelihood estimation and bootstrap method without a guaranteed normal limit? ii) How does one test asymmetric tail dependence, that is, the difference between upper and lower comovements? iii) Do the magnitudes of lower (or upper) comovements differ across state pairs? iv) How different would the answers be for the above three questions if one considered the HPI rather than the residuals in the fitted time series models? Specifically, we first investigate the moment condition and then propose a correct AR-GARCH estimation procedure to ensure the validity of bootstrapping the residuals. Using the new estimation procedure along with novel comovement measures and the bootstrap method, we further propose formal statistical tests for asymmetric tail dependence and differences in the magnitudes of lower (or upper) tail dependence.

Based on the new framework, we find no evidence of asymmetric tail dependence or differences in the magnitudes of lower (or upper) tail dependence based on the residuals of the HPI, consistent with the wide confidence intervals of the conditional probabilities of Ho et al. (2016). However, asymmetric dependence is supported for some of the state pairs when we use the comovement measures based on the HPI change series.

Our study contributes to the literature in several dimensions. First, housing prices, like other economic and financial data, can have heavy tails. Therefore, in the first-stage AR-GARCH estimation, it is imperative that the moment conditions be tested and tail heaviness be accounted for. We propose a new procedure so that subsequent analyses are not biased by inconsistent AR-GARCH estimates. Second, we complement the copula framework in modeling housing price comovements by proposing new measures of comovements, as well as new nonparametric estimators. Because we do not use parametric copula selection as does Zimmer (2012) or kernel smoothing as do Ho et al. (2016), our approach is computationally advantageous. Third, the simplicity and flexibility of the comovement measures and associated estimators allow us to formulate hypothesis tests to directly test asymmetric tail dependence. Given that financial risks are largely determined by tail dependence, the proposed tests have important implications in gauging risk measures and choosing a parametric copula family with symmetric or asymmetric tail dependence (see Siburg et al., 2015; White et al., 2015). Unlike Zimmer (2012) and Ho et al. (2016), our proposed comovement defintions based on observations instead of residuals are applicable to the study of portfolios, and an application to computing the upper and lower Value-at-Risk (VaR) of portfolios illustrates the usefulness of the proposed test for asymmetric tail dependence.

The paper is organized as follows. Section 3.2 provides details on the methodologies. Section 3.3 reports the empirical results. Section 3.4 concludes the paper with discussions. The Appendix contains further explanation of methodologies and additional figures and tables.

3.2 Methodologies

3.2.1 AR-GARCH Estimation

To fit model (3.1) to the quarterly changes of a state's HPI, Zimmer (2012) and Ho et al. (2016) employed the well-known quasi-maximum likelihood estimation, for which asymptotic normality requires both $Ee_{i,t}^4 < \infty$ and $EX_{i,t}^4 < \infty$ (Francq and Zakoian, 2004). Since the validity of a bootstrap confidence interval requires that the asymptotic distribution of the involved parameter estimation be normal, we first check whether $EX_{i,t}^4 < \infty$ and

$Ee_{i,t}^4 < \infty$ by using the well-known Hill (1975) estimator.[4] We also check the autocorrelation functions of the residuals by using the quasi-maximum likelihood estimator, the self-weighted quasi-maximum likelihood estimator of Ling (2007), and the self-weighted quasi-maximum exponential likelihood estimator (SWQMELE) of Zhu and Ling (2011).

We find that both $EX_{i,t}^4$ and $Ee_{i,t}^4$ could be infinite and that application of the SWQMELE to fit model (3.1) with $r = 3$ is justifiable. The final model fitting is reported in Table 3.1 and details pertinent to tail index estimation and model diagnostics are plotted in Figures 3.A.1–3.A.8 in the Appendix.

3.2.2 Comovement Estimation and Hypothesis Tests

To estimate the quantities in Eqs. (3.3)–(3.6), we employ the fitted models from above and denote the resultant estimator by $\hat{\boldsymbol{\theta}}_i$ for $\boldsymbol{\theta}_i = (\mu_i, \phi_{i,1}, \cdots, \phi_{i,r}, \alpha_{i,0}, \alpha_{i,1}, \beta_{i,1})^T$.

We can write $\hat{\varepsilon}_{i,t} = X_{i,t} - \hat{\mu}_i - \sum_{j=1}^r \hat{\phi}_{i,j} X_{i,t-j}$, $\hat{\sigma}_{i,t}^2 = \hat{\alpha}_{i,0} + \hat{\alpha}_{i,1}\hat{\varepsilon}_{i,t-1}^2 + \hat{\beta}_{i,1}\hat{\sigma}_{i,t-1}^2$, $\hat{e}_{i,t} = \hat{\varepsilon}_{i,t}/\hat{\sigma}_{i,t}$. A nonparametric estimator for $\gamma_{j|i,e}^-(p)$ is $\hat{\gamma}_{j|i,e}^-(p) = \frac{1}{np}\sum_{t=1}^n I(G_{ni}(\hat{e}_{i,t}) \leq p, G_{ni}(\hat{e}_{j,t}) \leq p)$, where $I(\cdot)$ denotes the indicator function and $G_{ni}(x) = \frac{1}{n+1}\sum_{t=1}^n I(\hat{e}_{i,t} \leq x)$. The estimators for the remaining quantities in Eqs. (3.3)–(3.6) can be defined in the same fashion.

To assess the asymmetry of tail dependence for a specific state pair and compare the magnitudes of comovements across different state pairs, we define the distance-based test statistics based on residuals as follows: $D_{j|i,e}(p) = \int_0^p |\hat{\gamma}_{j|i,e}^-(s) - \hat{\gamma}_{j|i,e}^+(s)|^2\, ds$, $\bar{D}_{j,k|i,e}^+(p) = \int_0^p |\hat{\gamma}_{j|i,e}^+(s) - \hat{\gamma}_{k|i,e}^+(s)|^2\, ds$, and $\bar{D}_{j,k|i,e}^-(p) = \int_0^p |\hat{\gamma}_{j|i,e}^-(s) - \hat{\gamma}_{k|i,e}^-(s)|^2\, ds$. In the Appendix, we also define analogous test statistics $D_{j|i,X}(p), \bar{D}_{j,k|i,X}^+(p), \bar{D}_{j,k|i,X}^-(p)$ based on changes in the HPI rather than the residuals. These test statistics are in line with the well-known Crámer-von Mises statistic for testing the goodness-of-fit of distribution functions. Hence, when the defined distance is too large, one will reject the null hypothesis of no difference.

To obtain critical values for the above test statistics, we adopt a bootstrap method via resampling from the residuals in model (3.1). Specifically, we draw samples with replacement

[4] The asymptotic properties of the Hill estimator for dependent data are available from Drees (2003).

from $\{\hat{e}_{i,t}\}_{t=1}^{n}$, say, $\{\hat{e}_{i,t}^{*}\}_{t=1}^{n}$, and then refit the model (3.1) with $\boldsymbol{\theta}_i$ and $e_{i,t}$ replaced by $\hat{\boldsymbol{\theta}}_i$ and $\hat{e}_{i,t}^{*}$, respectively, yielding the bootstrap samples $\{X_{i,t}^{*}\}_{t=1}^{n}$. The bootstrap test statistics are obtained using these bootstrap samples. Critical values are computed by repeating this procedure $1,000$ times in our analysis.

Using the same procedure, we also test asymmetry in tail dependence based on the comovements defined in Eqs. (3.5) and (3.6) and examine the effects of data revision on the comovements defined in Eqs. (3.3)–(3.6). The relevant test statistics are included in the Appendix. The importance of testing for asymmetric tail dependence is also illustrated by comparing upper and lower VaR of portfolios.

3.3 Results

In Figures 3.A.9 and 3.A.9 in the Appendix, we plot the cross-state comovement estimates of the measures in Eqs. (3.3) and (3.4) against $p = 0.25$, 0.24, ..., 0.01 based on the residual series of fitted AR(3)-GARCH(1,1) models as well as the quarterly change series of the state HPI from 1975:Q2 to 2017:Q1, respectively.

One interesting observation from these two figures is that the lower comovements are weaker than the upper comovements for all states conditional on CA when the comovements are defined on the residuals, while the relation is reversed when the comovements are defined on the original series. The former relation appears in line with the findings of Zimmer (2012), that upper tail dependence is stronger and more prevalent when residuals from AR-GARCH models are used, and the latter relation corroborates the evidence of Kuethe and Pede (2011) that economic shocks in CA have a great impact on the housing prices in the neighboring states of AZ and NV. Another observation is that lower comovements are larger in magnitude for most state pairs when the comovements are defined based on the original HPI. Similar plots for the measures in Eqs. (3.5) and (3.6) can be found in Figures 3.A.11 and 3.A.12 in the Appendix.

Table 3.2 reports the test results of asymmetric tail dependence based on the comovement measures defined in Eqs. (3.3) and (3.4). The p-values of the test statistic $D_{j|i,e}(p)$ fail to reject the null hypothesis of symmetry based on the residuals for all the state pairs, whereas the p-values of $D_{j|i,X}(p)$ reject the null hypothesis of symmetry for three state pairs AZ–CA, NV–AZ, and AZ–NV based on the original series.

The comovement measures in Eqs. (3.3) and (3.4) have the appealing order-varying property for each given state pair; that is, the comovements of state j conditional on i could differ from those of state i conditional on state j. For example, our results show asymmetric dependence for AZ conditional on CA but not vice versa. This means that the housing prices in AZ respond differently from the housing price upswings and downswings in CA but not vice versa.

In Table 3.3, we report the results for testing asymmetric tail dependence based on Eqs. (3.5) and (3.6). The results from the test statistics defined on observed percentage changes in the HPI show strong evidence of asymmetric dependence for the AZ–NV pair and moderate evidence for the NV–CA pair. This result is consistent with the evidence in Table 3.2 of asymmetric dependence between AZ and NV.

To the extent that the proposed comovements based on observations are applicable to portfolios, we examine the impact of asymmetric tail dependence on the upper and lower VaR of portfolios $X_{i,t} + X_{j,t}$ by comparing the measure DV(p) defined as the difference of VaR at levels p and $1 - p$ multiplied by the skewness of the portfolio. The large values of DV(0.90) and DV(0.95) for portfolios NV+CA, AZ+CA and AZ+NV in Table 3.A.7 are consistent with the small p-values of $D_{j|i,X}(0.10)$ and $D_{j|i,X}(0.05)$ for NV/CA, AZ/CA and AZ/NV in Table 3.3, i.e., higher asymmetric tail dependence implies larger difference in upper and lower VaR.

More results are reported in the Appendix. For example, results for testing the difference in magnitudes of the upper (or lower) comovements are in Tables 3.A.1-3.A.2; those for examining whether the revised HPI data published by the FHFA lead to significant changes in comovements are reported in Tables 3.A.3-3.A.6; those for diversification effect are in Table 3.A.8.

3.4 Conclusions

We re-examine the methods used in modeling the comovements of state HPIs. We find that, due to the heavy tails of the HPI data, the previously adopted quasi-maximum likelihood estimator in fitting an AR-GARCH model has a non-normal limit. We thus propose employing the self-weighted quasi-maximum exponential likelihood estimator. Based on the new estimation procedure and a bootstrap method based on residuals, we propose hypothesis tests of asymmetry between lower and upper comovements and differences in the magnitudes of comovements across state pairs. Our test results support the asymmetric dependence of housing prices between certain states, using measures defined based on original HPI change series rather than on the residuals from the fitted AR-GARCH models. We also find that data revision has little impact on comovements. The proposed methods based on observations are applicable to the study of portfolios.

Table 3.1. Summary statistics of estimates from AR(3)-GARCH(1,1) models

We fit AR(3)-GARCH(1,1) models using the SWQMELE to quarterly changes in the HPI for four states from 1975:Q2 to 2017:Q1: CA, FL, NV, and AZ. In this table, we report the median and mean of the standardized residuals, the mean of the absolute value of standardized residuals, and the parameter estimates, with bootstrap standard deviations in parentheses.

State	Median $\hat{e}_{i,t}$	Mean $\hat{e}_{i,t}$	Mean $\lvert\hat{e}_{i,t}\rvert$	AR(3)-GARCH(1,1) Parameter Estimates						
				$\hat{\mu}_i$	$\hat{\phi}_{i,1}$	$\hat{\phi}_{i,2}$	$\hat{\phi}_{i,3}$	$\hat{\alpha}_{i,0}$	$\hat{\alpha}_{i,1}$	$\hat{\beta}_{i,1}$
CA	-0.0020	-0.0827	1.0032	0.2925	0.7133	0.0142	0.1200	0.2703	0.2727	0.2853
	(0.0321)	(0.0846)	(0.0258)	(0.1524)	(0.1146)	(0.1286)	(0.1023)	(0.1359)	(0.1324)	(0.2096)
FL	-0.0283	-0.0381	0.9692	0.2190	0.5306	0.0199	0.3253	0.1084	0.1870	0.5738
	(0.0249)	(0.0831)	(0.0263)	(0.1175)	(0.0764)	(0.0815)	(0.0714)	(0.1024)	(0.1051)	(0.2117)
NV	-0.0639	0.0414	0.9804	0.2650	0.3681	0.0497	0.3280	0.0462	0.2115	0.6843
	(0.0438)	(0.0877)	(0.0229)	(0.1863)	(0.1117)	(0.1180)	(0.1070)	(0.1016)	(0.1051)	(0.1711)
AZ	-0.0692	0.0092	0.9934	0.3759	0.2642	0.1568	0.2873	0.0335	0.3340	0.5346
	(0.0486)	(0.0831)	(0.0218)	(0.1569)	(0.0962)	(0.0931)	(0.0862)	(0.0934)	(0.1303)	(0.1344)

Table 3.2. Test results for asymmetric tail dependence based on Eqs. (3.3) and (3.4)

This table reports the p-values of tests based on $D_{j|i,e}(p)$ and $D_{j|i,X}(p)$ using the quarterly changes in the state HPI from 1975:Q2 to 2017:Q1. The superscripts ** and * denote significance at the 0.05 and 0.1 levels, respectively.

State pair	$D_{j\|i,e}(p)$					$D_{j\|i,X}(p)$				
	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
FL/CA	0.954	0.947	0.946	0.890	0.847	0.402	0.354	0.283	0.199	0.202
NV/CA	0.466	0.483	0.520	0.475	0.392	0.446	0.364	0.267	0.219	0.208
AZ/CA	0.637	0.639	0.656	0.608	0.646	0.209	0.155	0.114	0.077^{*}	0.097^{*}
CA/FL	0.837	0.854	0.817	0.729	0.738	0.859	0.856	0.854	0.818	0.816
NV/FL	0.597	0.531	0.513	0.472	0.418	0.713	0.645	0.543	0.458	0.425
AZ/FL	0.568	0.522	0.687	0.632	0.489	0.866	0.809	0.760	0.708	0.629
CA/NV	0.905	0.857	0.796	0.957	0.790	0.802	0.772	0.701	0.618	0.473
FL/NV	0.841	0.804	0.750	0.660	0.761	0.396	0.323	0.236	0.201	0.172
AZ/NV	0.907	0.873	0.828	0.742	0.734	0.165	0.124	0.084^{*}	0.071^{*}	0.187
CA/AZ	0.709	0.677	0.637	0.582	0.645	0.644	0.634	0.638	0.515	0.403
FL/AZ	0.948	0.934	0.906	0.909	0.877	0.871	0.826	0.768	0.743	0.611
NV/AZ	0.556	0.507	0.421	0.334	0.400	0.188	0.138	0.086^{*}	0.047^{**}	0.021^{**}

Table 3.3. Test results for asymmetric tail dependence based on Eqs. (3.5) and (3.6)

Note: In this table, we report the p-values of tests based on $D_{j|i,e}(p)$ and $D_{j|i,X}(p)$ using the quarterly changes of the state HPI from 1975:Q2 to 2017:Q1. The superscripts ** and * denote significance at the 0.05 and 0.1 levels, respectively.

State pair	$D_{j\|i,e}(p)$					$D_{j\|i,X}(p)$				
	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
FL/CA	0.845	0.801	0.786	0.741	0.759	0.504	0.467	0.446	0.421	0.630
NV/CA	0.413	0.415	0.485	0.533	0.518	0.288	0.250	0.198	0.150	0.082*
AZ/CA	0.775	0.761	0.787	0.785	0.654	0.271	0.237	0.193	0.146	0.141
NV/FL	0.713	0.692	0.656	0.612	0.737	0.275	0.240	0.187	0.152	0.328
AZ/FL	0.516	0.488	0.532	0.547	0.450	0.811	0.801	0.758	0.754	0.647
AZ/NV	0.787	0.781	0.752	0.703	0.754	0.042**	0.036**	0.023**	0.014**	0.048**

Appendix

Appendix 3.A Measuring Comovements

Let $X_{i,t}$ denote the quarterly percentage changes of state-level HPI for the i-th state at time t. For each state, we fit an AR(r)-GARCH(1,1) model to the HPI series:

$$X_{i,t} = \mu_i + \sum_{j=1}^{r} \phi_{i,j} X_{i,t-j} + \varepsilon_{i,t}, \quad \varepsilon_{i,t} = \sigma_{i,t} e_{i,t}, \quad \sigma_{i,t}^2 = \alpha_{i,0} + \alpha_{i,1}\varepsilon_{i,t-1}^2 + \beta_{i,1}\sigma_{i,t-1}^2, \tag{3.A.1}$$

where $\mu_i, \phi_{i,1}, \cdots, \phi_{i,r} \in \mathbb{R}$, $\alpha_{i,0} > 0, \alpha_{i,1} > 0, \beta_{i,1} > 0$ for $i = 1, \cdots, m$, $\{\boldsymbol{e}_t = (e_{1,t}, \cdots, e_{m,t})^T\}_{t=1}^n$ is a sequence of independent and identically distributed random vectors with zero means and unit variances. Here A^T denotes the transpose of matrix or vector A.

We propose to quantify the contemporaneous upper and lower comovements by

$$\gamma_{j|i,e}^{-}(p) = P(G_i(e_{j,t}) < p | G_i(e_{i,t}) < p), \ \gamma_{j|i,e}^{+}(p) = P(G_i(e_{j,t}) > 1-p | G_i(e_{i,t}) > 1-p), \tag{3.A.2}$$

$$\gamma_{j|i,X}^{-}(p) = P(F_i(X_{j,t}) < p | F_i(X_{i,t}) < p), \ \gamma_{j|i,X}^{+}(p) = P(F_i(X_{j,t}) > 1-p | F_i(X_{i,t}) > 1-p), \tag{3.A.3}$$

$$\gamma_{j|i,e}^{*-}(p) = P(G_j(e_{j,t}) < p | G_i(e_{i,t}) < p), \ \gamma_{j|i,e}^{*+}(p) = P(G_j(e_{j,t}) > 1-p | G_i(e_{i,t}) > 1-p), \tag{3.A.4}$$

$$\gamma_{j|i,X}^{*-}(p) = P(F_j(X_{j,t}) < p | F_i(X_{i,t}) < p), \ \gamma_{j|i,X}^{*+}(p) = P(F_j(X_{j,t}) > 1-p | F_i(X_{i,t}) > 1-p), \tag{3.A.5}$$

where $p \in (0,1)$, F_i and G_i denote the distribution functions of $X_{i,t}$ and $e_{i,t}$, respectively.

Appendix 3.B AR-GARCH Estimation

To fit model (3.A.1) to the state HPI series, Zimmer (2012) and Zimmer (2012) employed the well-known quasi-maximum likelihood estimation, of which the asymptotic normality requires both $Ee_{i,t}^4 < \infty$ and $EX_{i,t}^4 < \infty$ (Francq and Zakoian, 2004).

Since the validity of a bootstrap confidence interval based on the residuals of a time series model requires that the asymptotic distribution of the involved parameter estimation is normal, we first check whether $EX_{i,t}^4 < \infty$ by assuming that $P(|X_{i,t}| > x)$ has a heavy tail, i.e., $\lim_{s\to\infty} P(|X_{i,t}| > sx)/P(|X_{i,t}| > s) = x^{-\alpha_i}$ for all $x > 0$, where $\alpha_i > 0$ is called the tail index. When $\alpha_i > 4$, we have $E|X_{i,t}|^4 < \infty$. To estimate the tail index α_i, we adopt the well-known Hill estimator (Hill, 1975) defined as $\hat{\alpha}_i(k) = \left\{\frac{1}{k}\sum_{j=1}^{k} \log \frac{X_{i,n,n-j+1}}{X_{i,n,n-k}}\right\}^{-1}$, where $X_{i,n,1} \leq \cdots \leq X_{i,n,n}$ denote the order statistics of $X_{i,1}, \cdots, X_{i,n}$, and $k = k(n) \to \infty$ and $k/n \to 0$ as $n \to \infty$. Asymptotic behavior of the Hill estimator for a dependent sequence is studied in Drees (2003) and references therein.

In Figure 3.A.1, we use the state HPI data from 1975:Q2 to 2017:Q1 to plot $\hat{\alpha}_i(k)$ against $k = 5, 6, \cdots, 70$. The figure shows that the tail index for each state is between 2 and 4, i.e., $X_{i,t}$ has a finite variance, but a possible infinite fourth moment. Moreover, as shown in Figure 3.A.2, the autocorrelation functions (ACFs) of the estimated $\{e_{i,t}\}$ via the quasi-maximum likelihood estimator (QMLE) from fitting AR(1)-GARCH(1,1) models to the same data, point to the inadequacy of such model specifications. Hence, confidence intervals constructed from the bootstrap method based on QMLE are inaccurate due to the non–normal limit in fitting model (3.A.1).

To relax the moment condition of $X_{i,t}$, Ling (2007) proposed a so-called self-weighted quasi-maximum likelihood estimator (SWQMLE) to fit model (3.A.1), which only requires $Ee_{i,t}^4 < \infty$ to ensure a normal limit. We fit model (3.A.1) by using SWQMLE in Ling (2007) with weight $w_{i,t} = \left\{\max\left(1, \frac{1}{C_i}\sum_{j=1}^{t-1} \frac{|X_{i,t-j}|I(|X_{i,t-j}|>C_i)}{j^9}\right)\right\}^{-4}$, where C_i is taken as the 95% percentile of $X_{i,1}, \cdots, X_{i,n}$. Note that, in the optimization process, we delete the first ten terms in the summation of the log-likelihood function (approximately 10% of the length of the original data) to remove the effect of the initial values in computing $\sigma_{i,t}$ by setting $\varepsilon_{i,t} = X_{i,t} = \sigma_{i,t} = 0$ for $t \leq 0$[5]. With this fitting procedure, we obtain estimates for $e_{i,1}, \cdots, e_{i,n}$, say $\hat{e}_{i,t}^{SWQMLE}$, for $t = 1, \cdots, n$. Using these residual estimates without the first ten, we compute and plot the ACFs of $\{e_{i,t}\}$ and $\{e_{i,t}^2\}$ in Figures 3.A.3 and 3.A.4, which show that the fitting of AR(3)-GARCH(1,1) models is quite reasonable. As the asymptotic normality of this estimator needs a finite fourth moment of $e_{i,t}$, we plot the Hill estimators based on $\{\hat{e}_{i,t}^{SWQMLE}\}$ in Figure 3.A.5, which well indicates that $Ee_{i,t}^4 = \infty$. That is, the estimator in Ling (2007) also can not ensure the validity of the bootstrap method due to the non–normal limit.

We further relax the moment conditions of both $X_{i,t}$ and $e_{i,t}$ by using the self-weighted quasi-maximum exponential likelihood estimator (SWQMELE) in Zhu and Ling (2011) with the above weight $w_{i,t}$ to fit model (3.A.1). It requires $E|e_{i,t}| = 1$ and $e_{i,t}$ to have median zero instead of mean zero and variance one. Again, we delete the first ten terms in the sum of log likelihood function for computing this estimator and estimating $e_{i,t}$'s. Denote these residuals by $\{\hat{e}_{i,t}^{SWQMELE}\}_{t=11}^{n}$. The ACFs of $\{e_{i,t}\}$ and $\{e_{i,t}^2\}$ and the Hill estimators based on $\{\hat{e}_{i,t}^{SWQMELE}\}_{t=11}^{n}$ are plotted in Figures 3.A.6–3.A.8. These figures suggest the model fitting and estimation procedure are adequate. Results for the median of residuals and the mean of the absolute values of residuals reported in the paper show that the assumptions on $\{e_{i,t}\}$ in using the SWQMELE are satisfied. In summary, it is sound to apply the SWQMELE with the above weight $w_{i,t}$ to fit model (3.A.1) with $r = 3$.

Appendix 3.C Comovement Estimation and Hypothesis Tests

In order to estimate the quantities in Eq. (3.A.2)–(3.A.5), we first infer model (3.A.1) with $r = 3$. We employ the SWQMELE for $\boldsymbol{\theta}_i = (\mu_i, \phi_{i,1}, \cdots, \phi_{i,r}, \alpha_{i,0}, \alpha_{i,1}, \beta_{i,1})^T$ in model (3.A.1)

[5] In separate and unreported simulation studies, this initialization approach works reasonably well for AR-GARCH estimation.

to ensure a normal limit, where we assume model (3.A.1) holds with $e_{i,t}$ having median zero and $E|e_{i,t}| = 1$. Denote the resultant estimator by $\hat{\boldsymbol{\theta}}_i$. Write $\hat{\varepsilon}_{i,t} = X_{i,t} - \hat{\mu}_i - \sum_{j=1}^{r} \hat{\phi}_{i,j} X_{i,t-j}$, $\hat{\sigma}_{i,t}^2 = \hat{\alpha}_{i,0} + \hat{\alpha}_{i,1}\hat{\varepsilon}_{i,t-1}^2 + \hat{\beta}_{i,1}\hat{\sigma}_{i,t-1}^2$, $\hat{e}_{i,t} = \hat{\varepsilon}_{i,t}/\hat{\sigma}_{i,t}$. Then the quantities in Eq. (3.A.2) through (3.A.5) can be estimated by

$$\hat{\gamma}_{j|i,e}^{-}(p) = \frac{1}{np}\sum_{t=1}^{n} I(G_{ni}(\hat{e}_{i,t}) \le p, G_{ni}(\hat{e}_{j,t}) \le p),$$

$$\hat{\gamma}_{j|i,e}^{+}(p) = \frac{1}{np}\sum_{t=1}^{n} I(G_{ni}(\hat{e}_{i,t}) > 1-p, G_{ni}(\hat{e}_{j,t}) > 1-p),$$

$$\hat{\gamma}_{j|i,X}^{-}(p) = \frac{1}{np}\sum_{t=1}^{n} I(F_{ni}(X_{i,t}) \le p, F_{ni}(X_{j,t}) \le p),$$

$$\hat{\gamma}_{j|i,X}^{+}(p) = \frac{1}{np}\sum_{t=1}^{n} I(F_{ni}(X_{i,t}) > 1-p, F_{ni}(X_{j,t}) > 1-p),$$

$$\hat{\gamma}_{j|i,e}^{*-}(p) = \frac{1}{np}\sum_{t=1}^{n} I(G_{ni}(\hat{e}_{i,t}) \le p, G_{nj}(\hat{e}_{j,t}) \le p),$$

$$\hat{\gamma}_{j|i,e}^{*+}(p) = \frac{1}{np}\sum_{t=1}^{n} I(G_{ni}(\hat{e}_{i,t}) > 1-p, G_{nj}(\hat{e}_{j,t}) > 1-p),$$

$$\hat{\gamma}_{j|i,X}^{*-}(p) = \frac{1}{np}\sum_{t=1}^{n} I(F_{ni}(X_{i,t}) \le p, F_{nj}(X_{j,t}) \le p),$$

$$\hat{\gamma}_{j|i,X}^{*+}(p) = \frac{1}{np}\sum_{t=1}^{n} I(F_{ni}(X_{i,t}) > 1-p, F_{nj}(X_{j,t}) > 1-p),$$

where $I(\cdot)$ denotes the indicator function,

$$G_{ni}(x) = \frac{1}{n+1}\sum_{t=1}^{n} I(\hat{e}_{i,t} \le x) \quad \text{and} \quad F_{ni}(x) = \frac{1}{n+1}\sum_{t=1}^{n} I(X_{i,t} \le x).$$

We assess the asymmetry of tail dependence of a given state pair and the differences in the upper (lower) comovements of different state pairs using the following distance-based test statistics:

$$
\begin{aligned}
D_{j|i,e}(p) &= \int_0^p |\hat{\gamma}^-_{j|i,e}(s) - \hat{\gamma}^+_{j|i,e}(s)|^2\, ds,\\
D_{j|i,X}(p) &= \int_0^p |\hat{\gamma}^-_{j|i,X}(s) - \hat{\gamma}^+_{j|i,X}(s)|^2\, ds,\\
\bar{D}^+_{j,k|i,e}(p) &= \int_0^p |\hat{\gamma}^+_{j|i,e}(s) - \hat{\gamma}^+_{k|i,e}(s)|^2\, ds,\\
\bar{D}^-_{j,k|i,e}(p) &= \int_0^p |\hat{\gamma}^-_{j|i,e}(s) - \hat{\gamma}^-_{k|i,e}(s)|^2\, ds,\\
\bar{D}^+_{j,k|i,X}(p) &= \int_0^p |\hat{\gamma}^+_{j|i,X}(s) - \hat{\gamma}^+_{k|i,X}(s)|^2\, ds,\\
\bar{D}^-_{j,k|i,X}(p) &= \int_0^p |\hat{\gamma}^-_{j|i,X}(s) - \hat{\gamma}^-_{k|i,X}(s)|^2\, ds.
\end{aligned}
$$

Appendix 3.D Effects of Data Revision on Comovements

Let $\tilde{X}_{i,t}$ denote the revised data and we fit the data using AR(3)-GARCH(1,1) models with residuals $\tilde{e}_{i,t}$ via SWQMELE. We use the following statistics and a similar bootstrap method to test the difference in the magnitude of comovements between the HPI used in Zimmer (2012) and the FHFA revised HPI:

$$
\begin{aligned}
\tilde{D}^+_{j|i,e,\tilde{e}}(p) &= \int_0^p |\hat{\gamma}^+_{j|i,e}(s) - \hat{\gamma}^+_{j|i,\tilde{e}}(s)|^2\, ds,\\
\tilde{D}^-_{j|i,e,\tilde{e}}(p) &= \int_0^p |\hat{\gamma}^-_{j|i,e}(s) - \hat{\gamma}^-_{j|i,\tilde{e}}(s)|^2\, ds,\\
\tilde{D}^+_{j|i,X,\tilde{X}}(p) &= \int_0^p |\hat{\gamma}^+_{j|i,X}(s) - \hat{\gamma}^+_{j|i,\tilde{X}}(s)|^2\, ds,\\
\tilde{D}^-_{j|i,X,\tilde{X}}(p) &= \int_0^p |\hat{\gamma}^-_{j|i,X}(s) - \hat{\gamma}^-_{j|i,\tilde{X}}(s)|^2\, ds.
\end{aligned}
$$

Appendix 3.E Value-at-Risk of Portfolios

As the proposed comovements based on observations is arguably applicable to portfolios, we examine the impact of asymmetric tail dependence on the upper and lower VaR. To take into account the effects of marginal distributions in different portfolios, we define a new measure $\mathrm{DV}(p)$ as the sum of VaR at levels p and $1-p$ of a portfolio scaled by its skewness. To examine how asymmetric tail dependence could affect the diversification benefits for an equally weighted portfolio of state house price changes, we study the ratio $\mathrm{DB}(p) = (\mathrm{VaR}_{X_{i,t}}(p) + \mathrm{VaR}_{X_{j,t}}(p))/\mathrm{VaR}_{X_{i,t}+X_{j,t}}(p)$. A larger value of $\mathrm{DB}(p)$ for a given level p means greater diversification benefits of investing in the portfolio.

Appendix 3.F Additional Tables and Figures

Results on fitting model (3.A.1) are reported in Figures 3.A.1–3.A.8, including tail index estimation and autocorrelation function plots. Estimates of comovements are plotted in Figures 3.A.9–3.A.12. P-values for the proposed tests for different magnitudes of comovements between state pairs and effects of data revision are reported in Tables 3.A.1–3.A.6. Calculations of VaR of portfolios and the new measure $\mathrm{DV}(p)$ for examining the impact of asymmetric tail dependence on risk measures are reported in Table 3.A.7. The diversification benefits at the lower and upper tails are reported in Table 3.A.8. The comovements between portfolios of equally weighted HPI change series in two states are plotted in Figure 3.A.13 using the comovement measure in Eq. (3.A.5).

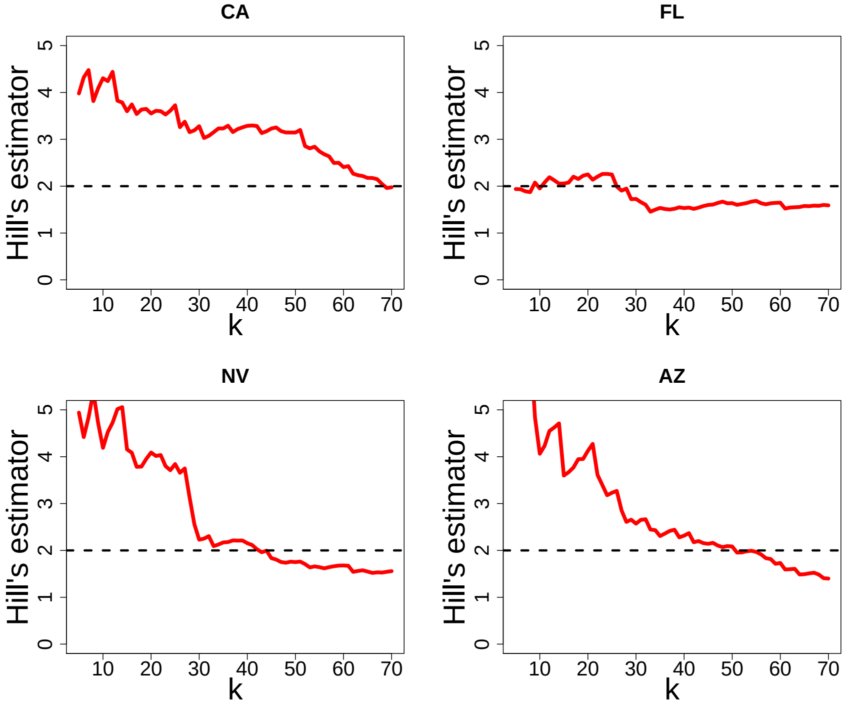


Figure 3.A.1. Hill's estimators for quarterly changes of HPI (1975:Q2 - 2017:Q1)

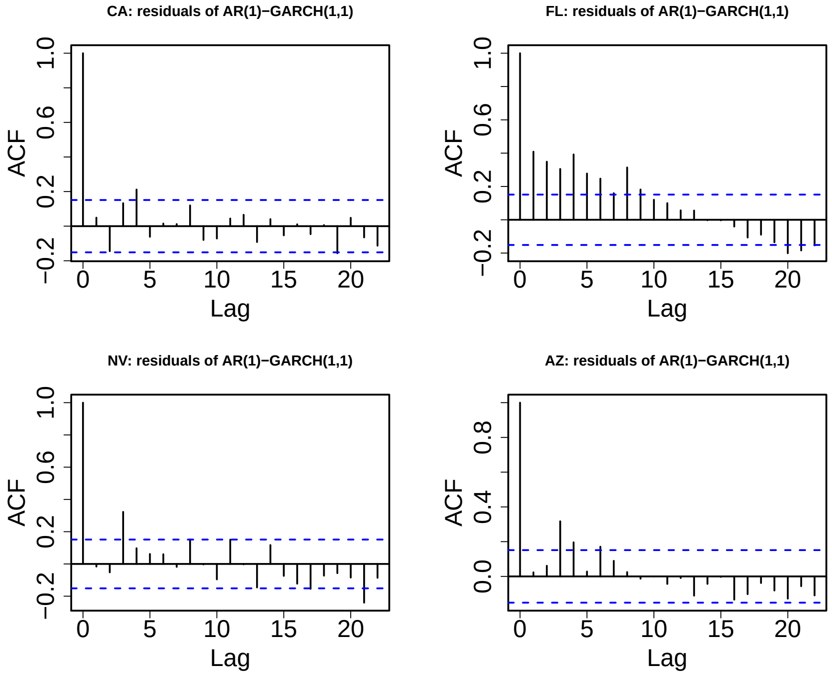


Figure 3.A.2. Autocorrelation functions of AR(1)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1)

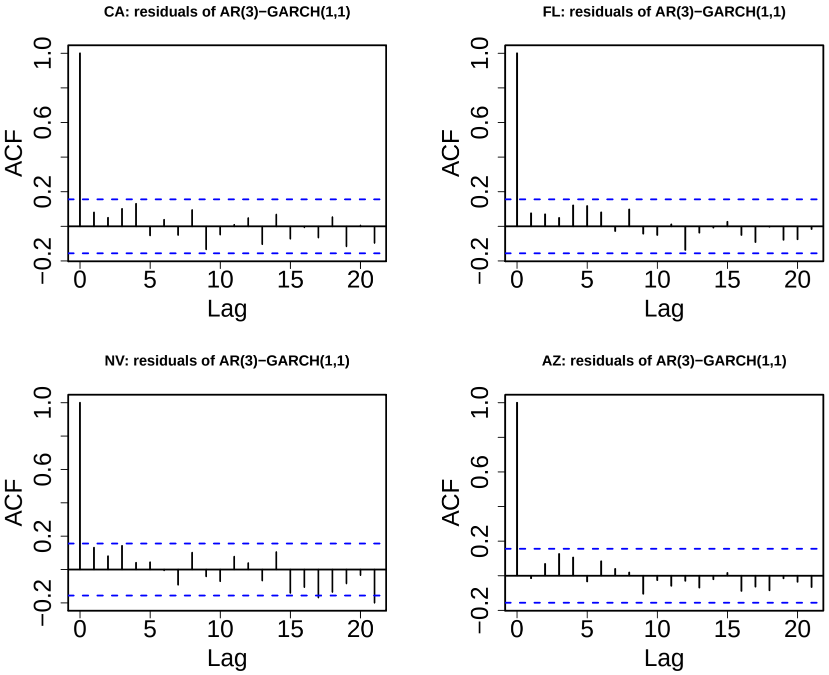


Figure 3.A.3. Autocorrelation functions of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1)

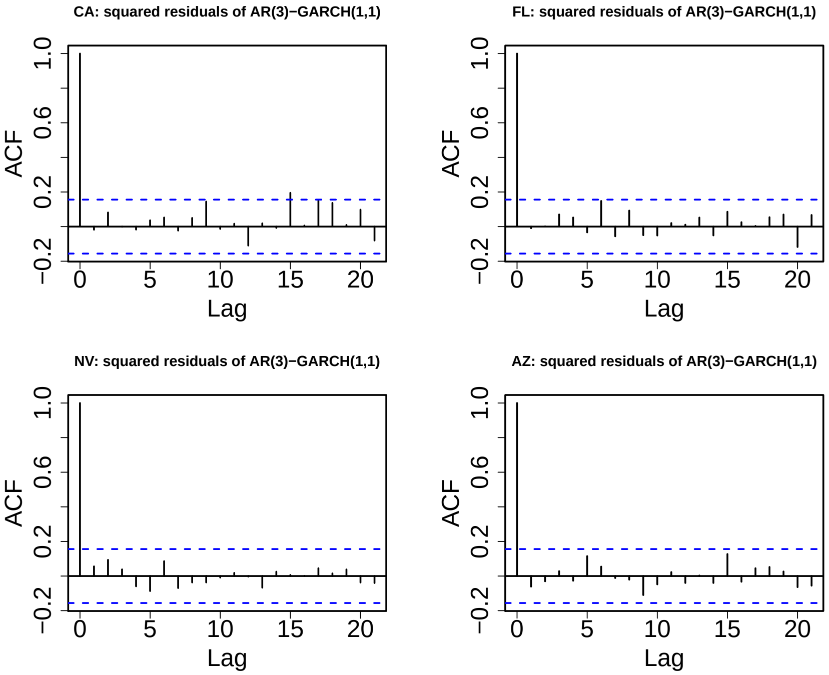


Figure 3.A.4. Autocorrelation functions of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1)

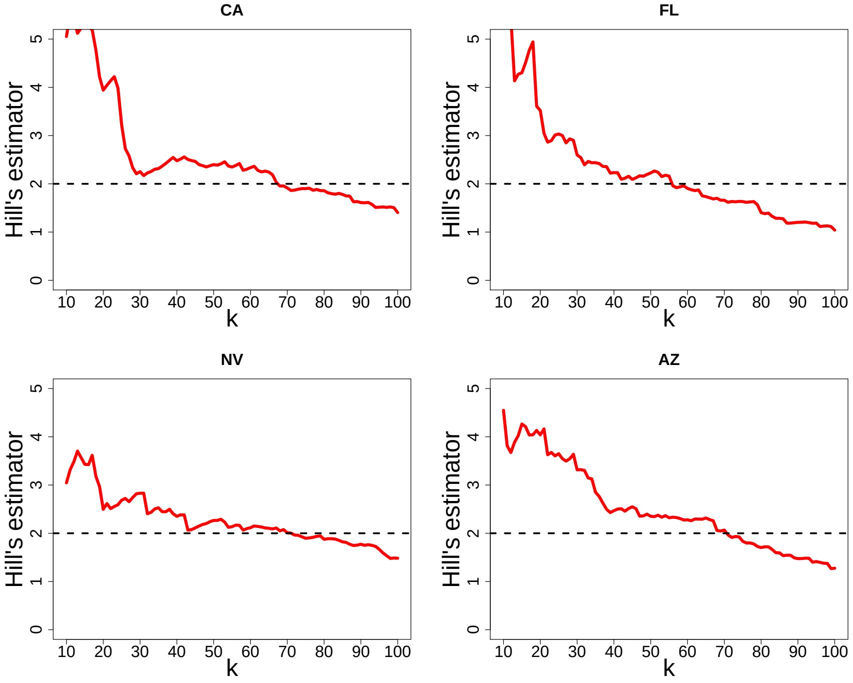


Figure 3.A.5. Hill's estimators of AR(3)-GARCH(1,1) model residuals for the quarterly changes of HPI (1975:Q2 - 2017:Q1)

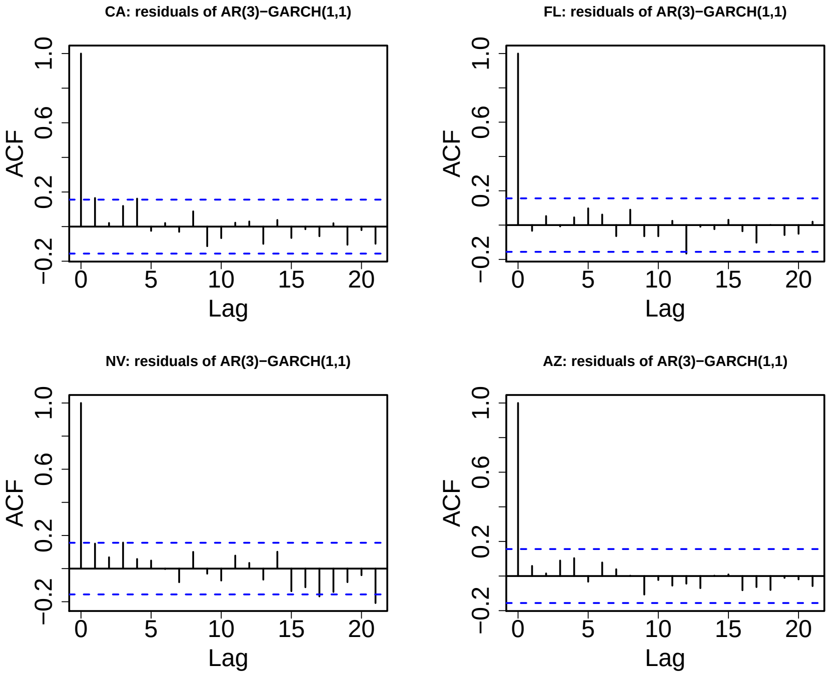


Figure 3.A.6. Autocorrelation functions of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1)

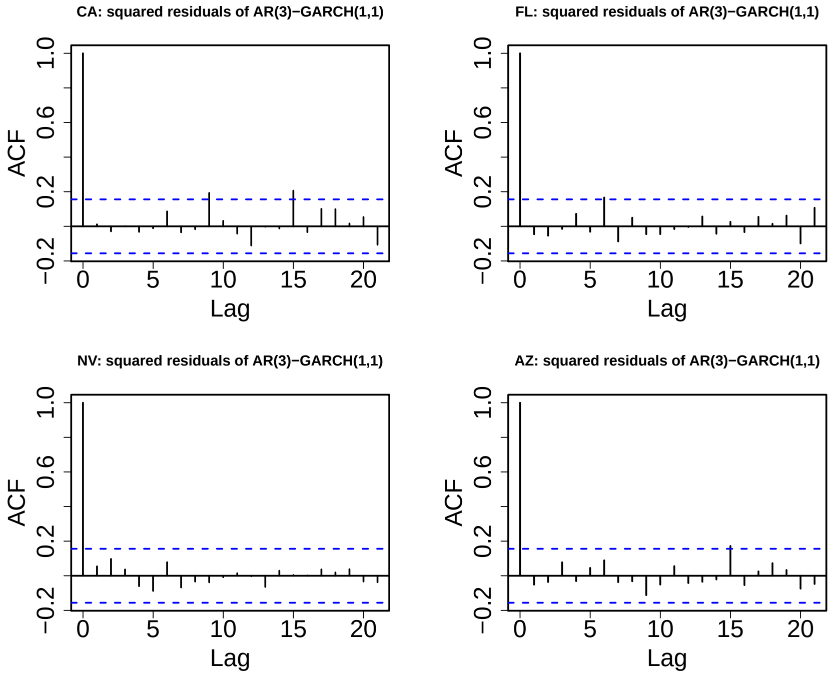


Figure 3.A.7. Autocorrelation functions of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1)

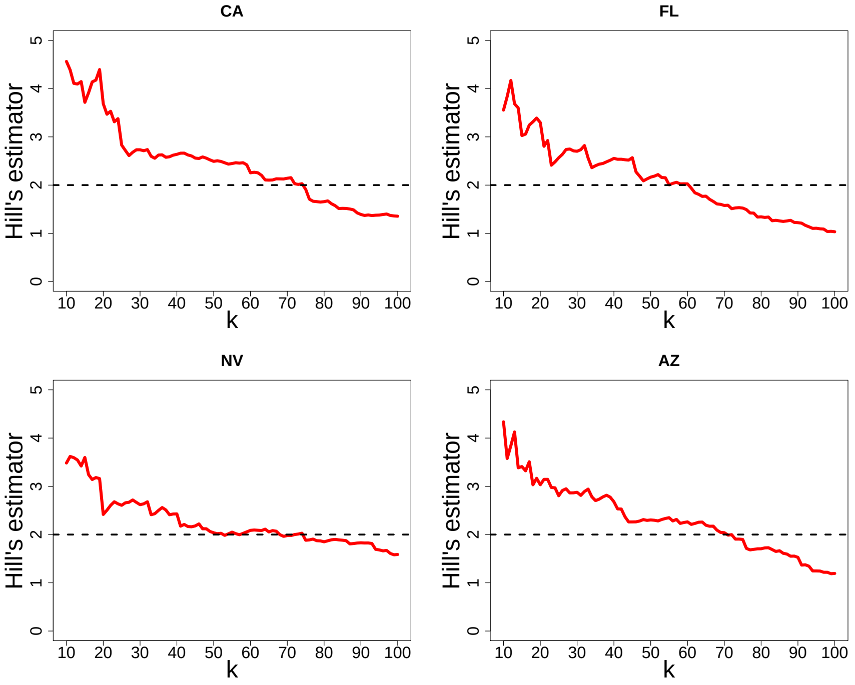


Figure 3.A.8. Hill's estimators of AR(3)-GARCH(1,1) model residuals for quarterly changes of HPI (1975:Q2 - 2017:Q1)

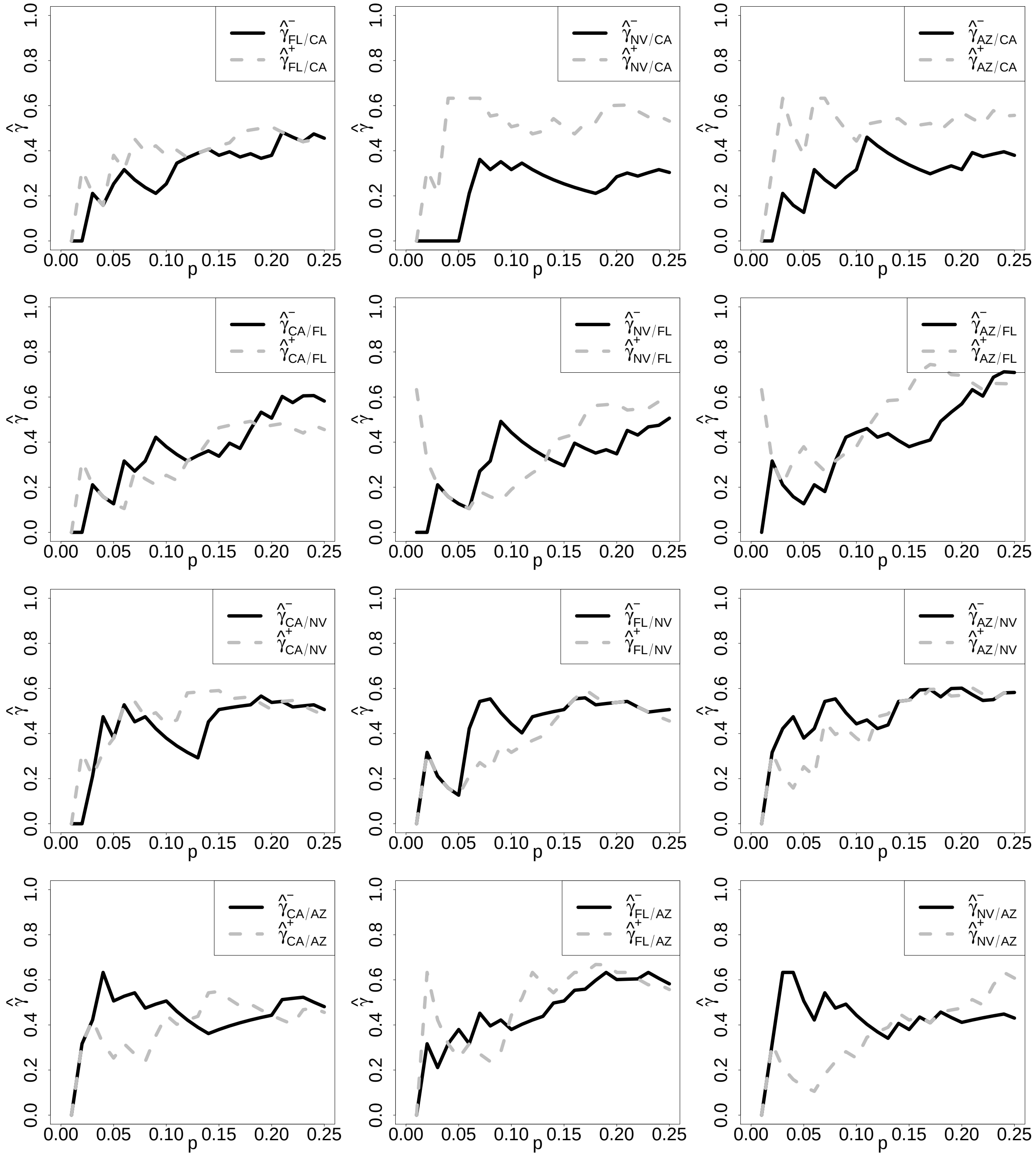


Figure 3.A.9. Upper tail dependence ($\hat{\gamma}^+_{i|j,e}$) and lower tail dependence ($\hat{\gamma}^-_{i|j,e}$) for AR(3)-GARCH(1,1) model residuals by SWQMELE

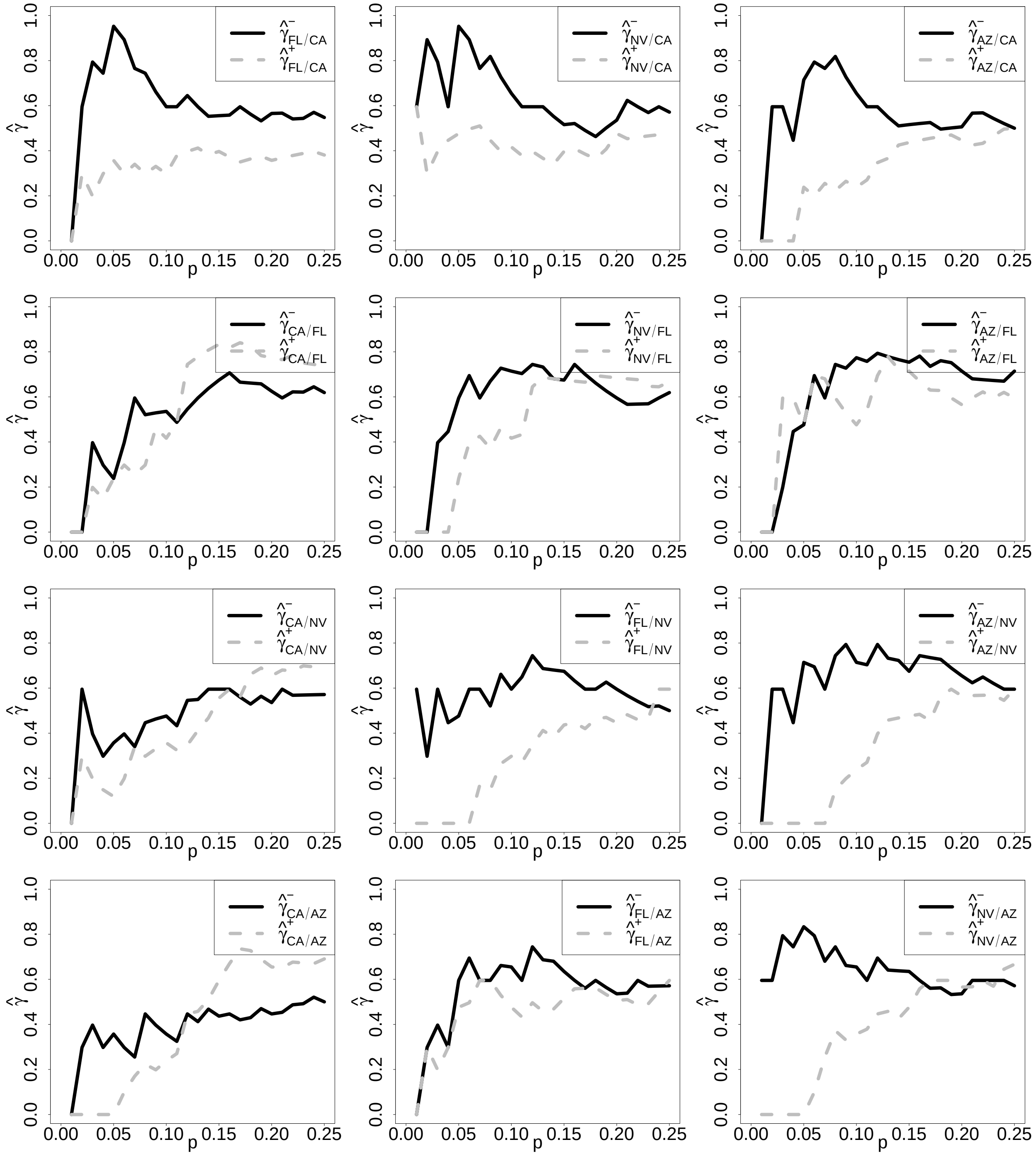


Figure 3.A.10. Upper tail dependence ($\hat{\gamma}^+_{i|j,X}$) and lower tail dependence ($\hat{\gamma}^-_{i|j,X}$) for the original HPI series

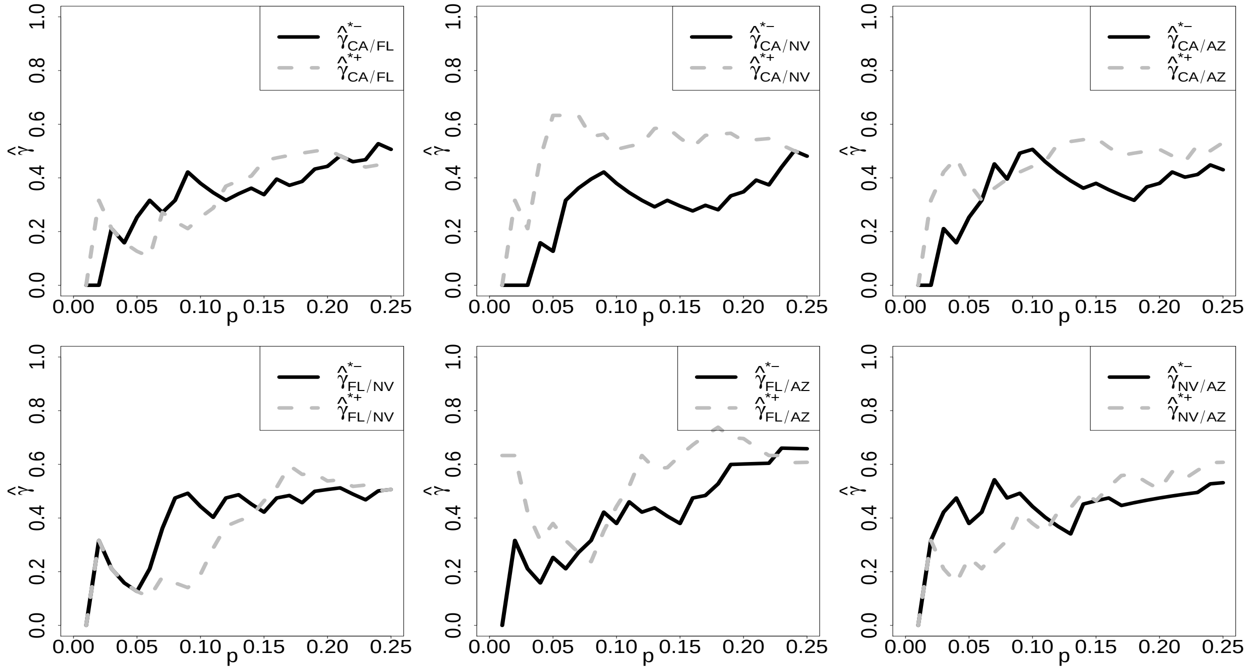


Figure 3.A.11. Upper tail dependence ($\hat{\gamma}^{*+}_{i|j,e}$) and lower tail dependence ($\hat{\gamma}^{*-}_{i|j,e}$) for AR(3)-GARCH(1,1) model residuals by SWQMELE

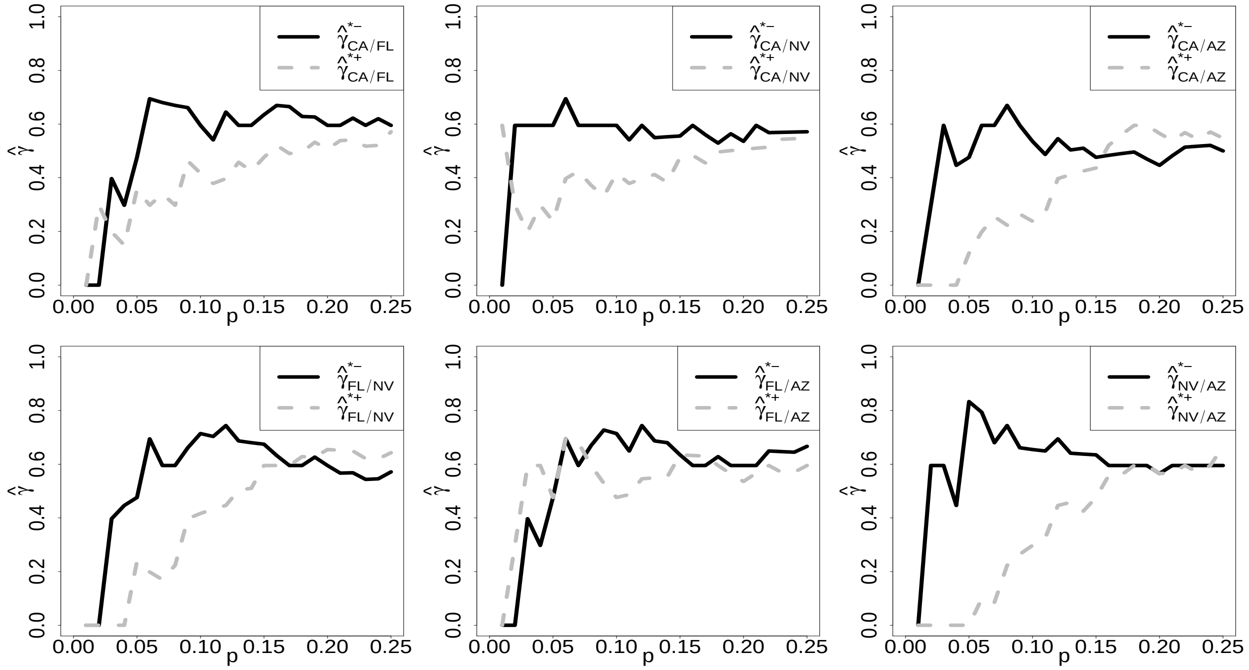


Figure 3.A.12. Upper tail dependence ($\hat{\gamma}^{*+}_{i|j,X}$) and lower tail dependence ($\hat{\gamma}^{*-}_{i|j,X}$) for the original HPI series

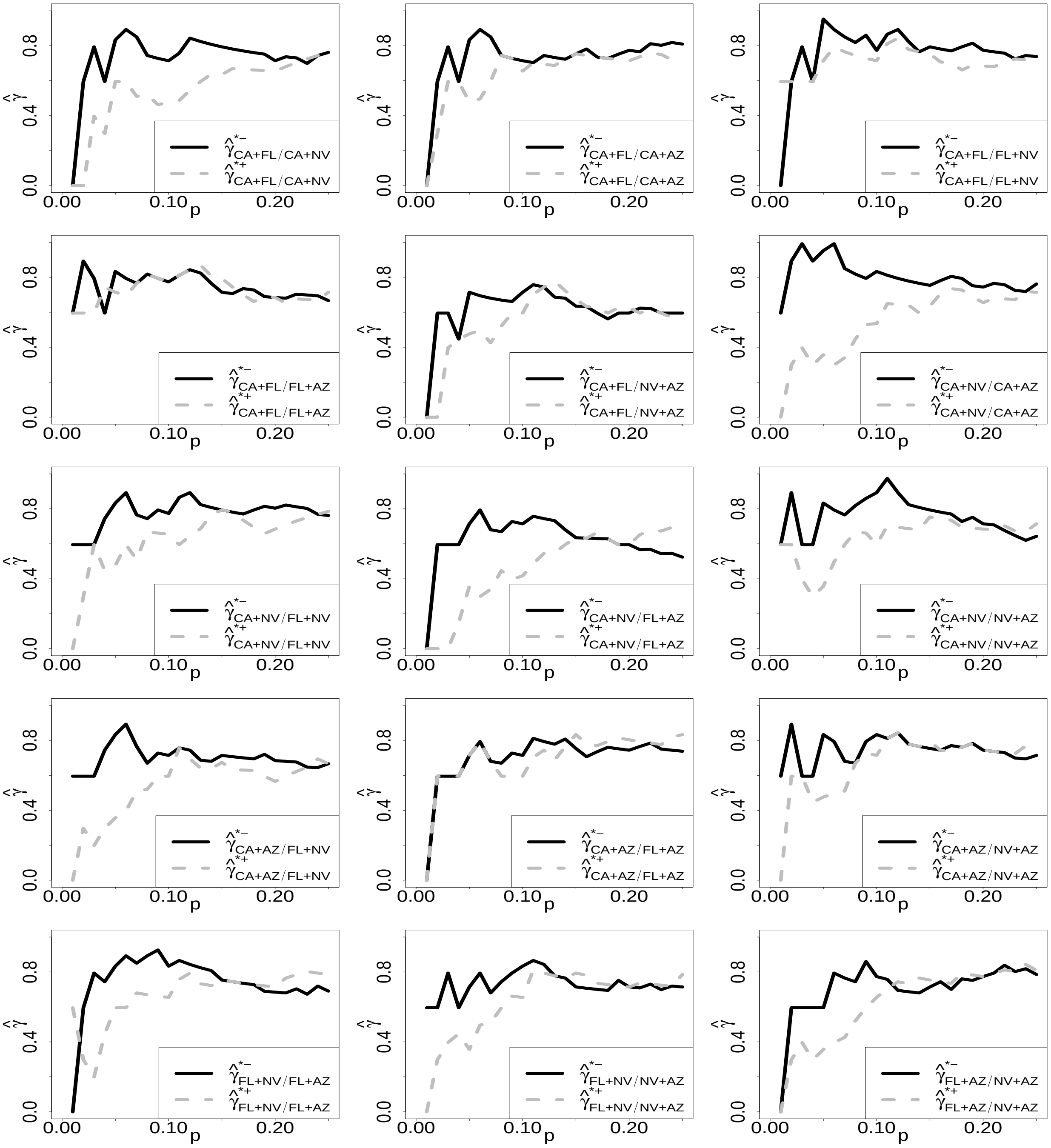


Figure 3.A.13. Upper tail dependence ($\hat{\gamma}^{*+}_{i|j,X}$) and lower tail dependence ($\hat{\gamma}^{*-}_{i|j,X}$) for portfolios of equally weighted original HPI series in two states

Table 3.A.1. Test results for the difference in interstate housing price comovements between two state pairs based on residuals

In this table, we report the p-values of tests based on $\bar{D}^{+}_{j,k|i,e}(p)$ and $\bar{D}^{-}_{j,k|i,e}(p)$ using the quarterly changes in state HPI (1975:Q2 - 2017:Q1).

State	$\bar{D}^{+}_{j,k\|i,e}(p)$					$\bar{D}^{-}_{j,k\|i,e}(p)$				
	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
FL, NV/CA	0.490	0.486	0.430	0.389	0.337	0.533	0.566	0.603	0.521	0.382
FL, AZ/CA	0.507	0.481	0.436	0.443	0.363	0.967	0.978	0.973	0.959	0.769
NV, AZ/CA	0.812	0.791	0.743	0.640	0.473	0.559	0.515	0.477	0.432	0.280
CA, NV/FL	0.482	0.431	0.364	0.318	0.288	0.771	0.850	0.864	0.748	0.856
CA, AZ/FL	0.494	0.474	0.507	0.464	0.376	0.898	0.892	0.813	0.772	0.637
NV, AZ/FL	0.527	0.502	0.578	0.656	0.594	0.592	0.766	0.781	0.667	0.459
CA, FL/NV	0.570	0.529	0.480	0.493	0.671	0.619	0.577	0.489	0.411	0.284
CA, AZ/NV	0.747	0.717	0.606	0.535	0.554	0.726	0.672	0.578	0.526	0.430
FL, AZ/NV	0.724	0.704	0.603	0.548	0.380	0.723	0.681	0.624	0.520	0.376
CA, FL/AZ	0.852	0.898	0.943	0.965	0.878	0.721	0.692	0.704	0.622	0.597
CA, NV/AZ	0.794	0.787	0.742	0.691	0.691	0.980	0.981	0.958	0.936	0.882
FL, NV/AZ	0.559	0.541	0.612	0.627	0.543	0.579	0.633	0.643	0.556	0.472

Table 3.A.2. Test results for the difference in interstate housing price comovements between two state pairs based on the original HPI series

In this table, we report the p-values of tests based on $\bar{D}^{+}_{j,k|i,X}(p)$ and $\bar{D}^{-}_{j,k|i,X}(p)$ using the quarterly percentage changes of state HPI (1975:Q2 - 2017:Q1).

State	$\bar{D}^{+}_{j,k\|i,X}(p)$					$\bar{D}^{-}_{j,k\|i,X}(p)$				
	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
FL, NV/CA	0.664	0.618	0.537	0.433	0.332	0.646	0.575	0.489	0.378	0.227
FL, AZ/CA	0.699	0.644	0.581	0.479	0.328	0.805	0.741	0.662	0.532	0.387
NV, AZ/CA	0.360	0.322	0.282	0.226	0.183	0.583	0.511	0.412	0.318	0.203
CA, NV/FL	0.823	0.794	0.801	0.730	0.586	0.752	0.693	0.634	0.585	0.616
CA, AZ/FL	0.574	0.557	0.539	0.434	0.469	0.696	0.645	0.594	0.598	0.618
NV, AZ/FL	0.458	0.421	0.376	0.312	0.254	0.861	0.859	0.848	0.800	0.674
CA, FL/NV	0.573	0.595	0.585	0.459	0.352	0.467	0.422	0.336	0.277	0.219
CA, AZ/NV	0.689	0.665	0.602	0.475	0.432	0.433	0.389	0.334	0.300	0.322
FL, AZ/NV	0.916	0.894	0.870	0.802	0.691	0.588	0.533	0.486	0.404	0.309
CA, FL/AZ	0.393	0.371	0.334	0.258	0.407	0.466	0.425	0.395	0.443	0.550
CA, NV/AZ	0.881	0.858	0.837	0.782	0.758	0.271	0.229	0.190	0.163	0.157
FL, NV/AZ	0.661	0.600	0.496	0.377	0.421	0.596	0.528	0.444	0.343	0.230

Table 3.A.3. Test results for effects of data revision on interstate housing price comovements based on Eq. (3.A.2)

In this table, we report the p-values of tests based on $\tilde{D}^{+}_{j|i,e,\tilde{e}}(p)$ and $\tilde{D}^{-}_{j|i,e,\tilde{e}}(p)$ using quarterly changes of state HPI (1975:Q2 - 2009:Q1) in Zimmer (2012) and FHFA revised data in the same period.

State	$\tilde{D}^{+}_{j\|i,e,\tilde{e}}(p)$					$\tilde{D}^{-}_{j\|i,e,\tilde{e}}(p)$				
	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
FL/CA	0.210	0.204	0.192	0.160	0.191	0.975	0.953	0.920	0.852	0.656
NV/CA	0.326	0.332	0.320	0.289	0.247	0.616	0.686	0.614	0.545	0.487
AZ/CA	0.384	0.372	0.338	0.289	0.215	0.705	0.664	0.661	0.822	0.489
CA/FL	0.968	0.942	0.918	0.864	0.541	0.728	0.796	0.804	0.810	0.700
NV/FL	0.925	0.848	0.943	0.862	0.614	0.746	0.675	0.687	0.600	0.380
AZ/FL	0.905	0.875	0.906	0.877	0.767	0.669	0.621	0.612	0.488	0.388
CA/NV	0.889	0.932	0.877	0.796	0.645	0.957	0.940	0.923	0.884	0.684
FL/NV	0.908	0.889	0.857	0.742	0.478	0.621	0.756	0.720	0.683	0.610
AZ/NV	0.465	0.468	0.413	0.331	0.423	0.771	0.845	0.872	0.818	0.717
CA/AZ	0.361	0.502	0.457	0.374	0.302	0.269	0.242	0.220	0.192	0.156
FL/AZ	0.593	0.559	0.502	0.549	0.635	0.816	0.769	0.726	0.602	0.537
NV/AZ	0.347	0.328	0.329	0.314	0.280	0.895	0.866	0.834	0.768	0.687

Table 3.A.4. Test results for effects of data revision on interstate housing price comovements based on Eq. (3.A.3)

In this table, we report the p-values of tests based on $\tilde{D}^{+}_{j|i,X,\tilde{X}}(p)$ and $\tilde{D}^{-}_{j|i,X,\tilde{X}}(p)$ using quarterly changes of state HPI (1975:Q2 - 2009:Q1) in Zimmer (2012) and FHFA revised data in the same period. * denotes significance at the 0.1 level.

State	$\tilde{D}^{+}_{j\|i,X,\tilde{X}}(p)$					$\tilde{D}^{-}_{j\|i,X,\tilde{X}}(p)$				
	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
FL/CA	0.761	0.704	0.620	0.564	0.408	0.503	0.444	0.360	0.276	0.197
NV/CA	0.554	0.527	0.490	0.409	0.393	0.262	0.212	0.171	0.136	0.088*
AZ/CA	0.863	0.834	0.807	0.744	0.600	0.501	0.464	0.415	0.335	0.267
CA/FL	0.831	0.787	0.713	0.761	0.676	0.927	0.907	0.871	0.859	0.703
NV/FL	0.783	0.745	0.712	0.726	0.676	0.809	0.807	0.779	0.732	0.573
AZ/FL	0.963	0.952	0.929	0.914	0.829	0.780	0.738	0.694	0.626	0.518
CA/NV	0.837	0.776	0.829	0.744	0.575	0.308	0.288	0.265	0.221	0.135
FL/NV	0.817	0.765	0.671	0.577	0.534	0.809	0.766	0.726	0.637	0.519
AZ/NV	0.878	0.851	0.838	0.740	0.668	0.355	0.326	0.274	0.234	0.179
CA/AZ	0.875	0.855	0.806	0.741	0.633	0.479	0.434	0.383	0.319	0.213
FL/AZ	0.519	0.466	0.394	0.302	0.194	0.538	0.496	0.449	0.366	0.242
NV/AZ	0.864	0.816	0.753	0.718	0.660	0.461	0.420	0.360	0.281	0.165

Table 3.A.5. Test results for effects of data revision on interstate housing price comovements based on Eq. (3.A.4)

In this table, we report the p-values of tests based on $\tilde{D}^{+}_{j|i,e,\tilde{e}}(p)$ and $\tilde{D}^{-}_{j|i,e,\tilde{e}}(p)$ using quarterly changes of state HPI (1975:Q2 - 2009:Q1) in Zimmer (2012) and FHFA revised data in the same period. * denotes significance at the 0.1 level.

State	$\tilde{D}^{+}_{j\|i,e,\tilde{e}}(p)$					$\tilde{D}^{-}_{j\|i,e,\tilde{e}}(p)$				
	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
FL/CA	0.184	0.185	0.183	0.173	0.172	0.919	0.919	0.880	0.854	0.681
NV/CA	0.945	0.918	0.893	0.859	0.764	0.865	0.846	0.865	0.796	0.674
AZ/CA	0.179	0.172	0.160	0.127	0.090 *	0.894	0.849	0.796	0.768	0.661
NV/FL	0.948	0.986	0.947	0.834	0.517	0.707	0.755	0.682	0.661	0.587
AZ/FL	0.634	0.589	0.587	0.548	0.546	0.610	0.564	0.502	0.430	0.317
AZ/NV	0.587	0.611	0.564	0.655	0.563	0.915	0.919	0.903	0.839	0.700

Table 3.A.6. Test results for effects of data revision on interstate housing price comovements based on Eq. (3.A.5)

In this table, we report the p-values of tests based on $\tilde{D}^{+}_{j|i,X,\tilde{X}}(p)$ and $\tilde{D}^{-}_{j|i,X,\tilde{X}}(p)$ using quarterly changes of state HPI (1975:Q2 - 2009:Q1) in Zimmer (2012) and FHFA revised data in the same period. ** and * denote significance at the 0.05 and 0.1 level, respectively.

State	$\tilde{D}^{+}_{j\|i,X,\tilde{X}}(p)$					$\tilde{D}^{-}_{j\|i,X,\tilde{X}}(p)$				
	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.25$	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
FL/CA	0.531	0.489	0.420	0.322	0.290	0.900	0.862	0.784	0.725	0.484
NV/CA	0.929	0.900	0.879	0.822	0.631	0.286	0.299	0.301	0.261	0.151
AZ/CA	0.974	0.957	0.902	0.815	0.717	0.416	0.385	0.307	0.223	0.242
NV/FL	0.843	0.810	0.776	0.789	0.728	0.905	0.919	0.863	0.790	0.579
AZ/FL	0.999	0.997	0.986	0.970	0.867	0.893	0.860	0.797	0.714	0.514
AZ/NV	0.928	0.924	0.900	0.815	0.781	0.152	0.141	0.109	0.079*	0.043**

Table 3.A.7. Value-at-Risk (VaR(p)) and DV(p) for equally weighted portfolios of house prices

In this table, DV(p) is defined as VaR(p) + VaR($1 - p$) multiplied by the skewness of a portfolio.

Portfolio	VaR(0.9)	VaR(0.1)	VaR(0.95)	VaR(0.05)	DV(0.9)	DV(0.95)
FL+CA	7.9674	-2.7599	9.9076	-7.0206	0.9620	0.5333
NV+CA	8.8854	-2.6335	11.0836	-6.8387	-2.7689	-1.8801
AZ+CA	9.2317	-2.6715	10.2663	-5.8529	-2.8624	-1.9257
NV+FL	8.5454	-3.8385	11.0665	-8.5886	1.4024	0.7383
AZ+FL	8.0527	-4.2648	11.5089	-7.6134	0.1543	0.1587
AZ+NV	9.6326	-3.2424	11.1258	-8.2008	-3.7741	-1.7275

Table 3.A.8. Diversification benefits of equally weighted portfolios of house prices

In this table, we calculate $\mathrm{DB}(p) = (\mathrm{VaR}_{X_{i,t}}(p) + \mathrm{VaR}_{X_{j,t}}(p))/\mathrm{VaR}_{X_{i,t}+X_{j,t}}(p)$ for each equally weighted portfolio.

Portfolio	DB(0.90)	DB(0.10)	DB(0.99)	DB(0.01)
FL+CA	1.1045	1.1988	1.0892	1.0044
NV+CA	1.1012	1.4544	1.1457	1.0340
AZ+CA	0.9894	1.5226	1.1410	1.0292
NV+FL	1.1066	1.1731	1.2354	1.2316
AZ+FL	1.0935	1.1115	1.0703	1.1635
AZ+NV	1.0163	1.6229	1.2498	1.0150

CHAPTER 4

Bootstrap Analysis of Mutual Fund Performance[1]

Abstract

We show that two prominent bootstrap tests for fund performance evaluation have distorted test sizes and lack test power to detect skilled funds when a substantial number of unskilled funds are present. We develop the theory for a valid bootstrap Hotelling's T-squared test allowing for serial correlations and cross-sectional dependence in fund residuals. Applying the new bootstrap test in a sequential testing procedure, our empirical analysis finds that skilled funds are more engaged in active management and hold stocks with higher expected anomalous returns.

4.1 Introduction

Are the funds with top-ranking alphas (or alpha t-statistics) skilled? To address this question, Kosowski, Timmermann, Wermers, and White (2006, KTWW) and Fama and French (2010) advocate the bootstrap tests for the joint zero-alpha null hypothesis. They analyze the

[1] This chapter is based on the following joint work: Huang, H., Jiang, L., Leng, X., & Peng, L. (2020). Bootstrap Analysis of Mutual Fund Performance. Working Paper.

significance of the alphas of extreme funds by comparing the cross-sectional distribution of estimated alphas with that of bootstrapped alphas at multiple percentiles. However, the test statistic constructed from bootstrap samples in these two studies is unconventional as a traditional bootstrap method is often used to approximate the distribution function of the test statistic for obtaining critical values. Moreover, neither a theoretical justification nor a numerical assessment of its performance is provided for the test statistic.[2] For example, if the null hypothesis is true such that all funds are zero-alpha, do the tests achieve an asymptotically correct size at a given significance level? When the null is not true such that some funds display skill, do they have the test power to reject the null? A growing finance literature has adopted the arguably convenient but unconventional bootstrap methods for various empirical investigations, for instance, performance evaluation for actively-managed mutual funds (Blake, Caulfield, Ioannidis, and Tonks, 2014, 2017), hedge funds (Kosowski, Naik, and Teo, 2007), and index funds (Crane and Crotty, 2018). Bootstrap methods have also been applied to a wide range of related studies in finance, such as Barras, Scaillet, and Wermers (2010), Ferson and Chen (2015), Chordia, Goyal, and Saretto (2017, 2020), Yan and Zheng (2017), and Harvey and Liu (2019). Strikingly, none of them has answered the fundamental and economically meaningful question of whether these bootstrap tests are statistically valid for these financial applications.

In this study, we first systematically analyze the size and power properties of the bootstrap tests in Kosowski et al. (2006) and Fama and French (2010). In a simplified framework of independent fund residuals, we show that these two tests suffer from size bias and low power after accounting for the following salient features of mutual fund data. First, the number of funds is much larger than the number of time-series observations of fund returns. Second, the return residuals from fund-by-fund regressions exhibit various levels of skewness.

[2] Kosowski et al. (2006) is related to the theoretical work in White (2000), although White (2000) requires that the time-series dimension goes to infinity and the cross-sectional dimension is fixed, which is different from the setting of mutual fund studies where the number of funds is much larger than the number of time-series observations. Furthermore, Bajgrowicz and Scaillet (2012) discuss the lack of power of the bootstrap reality check in White (2000) and this critique also applies to Kosowski et al. (2006). In contrast, Fama and French (2010) do not cite a theoretical origin for their bootstrap method.

Third, there are overwhelmingly more funds with negative alphas than those with positive alphas.

More specifically, the Kosowski et al. (2006) approach assumes that fund residuals are serially and cross-sectionally independent. It formulates an unconventional test statistic using bootstrapped t-statistics from the standard residual bootstrap method. Under the independent assumptions, this test has a correct asymptotic test size only when the sample size is large enough to the extent that $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} = 0$, where N is the number of funds, and T_i is the sample size (track record length) of the i-th fund. That is, its bootstrapped p-value at a given percentile converges in distribution to the desired uniform distribution on $[0, 1]$. However, when $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} > 0$, the bootstrap fails to correct the higher-order error terms in the test statistic and thereby produces inaccurate p-values due to the accumulation of estimation errors. On the other hand, the presence of a large proportion of negative alphas severely erodes the test power for positive alphas, making this test difficult to uncover skilled funds even if they do exist. Hansen (2005) similarly discusses how the inclusion of many poor and irrelevant alternatives adversely affects the power of the test in White (2000). Fan, Liao, and Yao (2015) underscore the problems of low power arising from the accumulation of estimation errors in high-dimensional cross-sectional testing. To contextualize the theoretical results, we observe in Figure 4.1 that in a population of 2650 funds, the majority have sample sizes less than a few hundred (Panel A), and the skewness of funds residuals in the cross-section is not negligible (Panel D).

[Figure 4.1 about here.]

The Fama and French (2010) approach is well motivated empirically to handle the possible cross-sectional dependence among fund returns, for which they suggest simultaneously resampling fund returns and factors. When fund residuals are independent in the cross-section, however, it is well expected that the Fama and French (2010) approach cannot correct the higher-order approximation error T_i^{-1} along the lines of the Kosowski et al. (2006) approach, i.e., its size is distorted when $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} > 0$. Because of the joint resampling of fund returns and factors, it is challenging to derive higher-order expansions for the test size and power as we cannot employ the same conditioning technique on factors as we do in developing theories for Kosowski et al. (2006). Nonetheless, our view is that it deserves further scrutiny. In fact, we conjecture that the joint resampling scheme in Fama and French (2010) cannot even correct the approximation error term $T_i^{-1/2}$ in the sense that the test size is biased when $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1/2}<\infty$ (while the Kosowski et al. (2006) method can correct this error term). In other words, the Fama and French (2010) approach is similarly challenged by small sample sizes vis-à-vis a large cross-section like Kosowski et al. (2006) and faces further threat from its unsubstantiated joint resampling scheme. Monte Carlo simulation studies confirm that the bootstrapped p-value obtained from the Fama and French (2010) approach is biased in the ideal scenario of normal and independent fund residuals with large sample sizes. Relative to the Kosowski et al. (2006) method, the test in Fama and French (2010) is heavily under-sized and consequently has little-to-no power to detect fund skill.

To overcome the caveats of the extant bootstrap tests, this paper proposes and theoretically justifies a zero-alpha test using Hotelling's T-squared statistic with bootstrap calibration. Although Pesaran and Yamagata (2017) study the Hotelling's T-squared test under the condition $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1}=0$, our theoretical result shows that the test has a biased size when $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1}>0$, which is the case for monthly mutual fund returns. Taking advantage of the residual-based bootstrap method, we propose to automatically correct this bias, so the Hotelling's T-squared test with bootstrap calibration has an asymptotically correct size whenever $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1}<\infty$. On the contrary, it is infeasible to make the unconventional bootstrap test in Kosowski et al. (2006) valid when $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1}\in(0,\infty)$. We initially develop the bootstrap Hotelling's T-squared test under the simplifying assumption of independent fund residuals. We further weaken the restrictive assumption and extend the test to the practical setting, where fund residuals are serially correlated and cross-sectionally dependent.

To separate skilled funds from zero-alpha funds, we provide a sequential testing procedure, which applies the bootstrap Hotelling's T-squared test sequentially to identify a data-driven p-value threshold and a maximum set of zero-alpha funds. The p-value threshold is also used to screen out a set of top-performing funds with large positive t-statistics. In the final step, the bootstrap Hotelling's T-squared test is conducted on the combined set of the top-performing funds and the predetermined zero-alpha funds. The top-performing funds are deemed skilled (relative to the zero-alpha funds) if the null hypothesis is rejected, that is, the top-performing fund set is significantly different from the zero-alpha fund set.

Our new test procedure improves the bootstrap tests in Kosowski et al. (2006) and Fama and French (2010) along several dimensions. First of all, our theory ensures that the bootstrap Hotelling's T-squared test has an asymptotically correct size even when the cross-sectional dimension is much larger than the time-series dimension, essentially circumventing the difficulties confronted with existing bootstrap approaches. Second, we offer a power enhancement procedure for implementation, where we exploit the set of zero-alpha funds as a reference and leave out those potentially unskilled funds when testing for skilled funds. This technique effectively draws on the information contained in the zero-alpha fund set, shrinks the number of funds to be tested, and enhances the test power for skilled funds. The idea of sequential testing shares the spirit of Hansen, Lunde, and Nason (2011) and Grønborg, Lunde, Timmermann, and Wermers (2021). The screening method of enhancing test power by eliminating inferior alternatives has been adopted in different testing problems, such as Hansen (2005) and Giglio, Liao, and Xiu (2020). Different from these studies, the choice of the screening threshold in our procedure is entirely data-driven. Last but not least, we validate the bootstrap Hotelling's T-squared test in a more general setting, where the fund residuals are both serially correlated and cross-sectionally dependent.

We illustrate the empirical relevance of the new test procedure in evaluating the performance of actively managed U.S. domestic equity mutual funds from January 1980 to December 2018. Applying the proposed bootstrap Hotelling's T-squared test, we find that there exist a minority of skilled funds after adjusting for several popular risk factors with the Carhart (1997) four-factor model. Most of the skilled funds are younger in a big fund family, have a lower turnover ratio and expense ratio, and attract more inflow from investors. Funds identified with skill are also more engaged in active management with a lower factor model R-squared (Amihud and Goyenko, 2013), higher active share (Cremers and Petajisto, 2009), and active weight (Doshi, Elkamhi, and Simutin, 2015). Recently, Li and Rossi (2021) propose to select skilled mutual funds using stock holding characteristics. We examine the stock holding difference between skilled and unskilled funds. We find that skilled funds hold stocks with higher bid-ask spread, dispersion in forecasted EPS, idiosyncratic volatility, Amihud ratio, return volatility, volatility of liquidity based on both dollar trading volume and share turnover, and stocks with a greater number of zero trading days than unskilled funds do. Also, skilled funds hold smaller stocks based on market capitalization and industrial adjusted market capitalization. Hou, Xue, and Zhang (2015) classify those anomalies into the category of “Trading Frictions”. Similarly, skilled fund managers also hold stocks with higher R&D expense to market capitalization and R&D expense to sales in the category of “Intangibles”. The results of stock holdings are mixed for anomalies in other categories, such as “Profitability”, “Investment”, “Value-Versus-Growth”, and “Momentum”. For example, skilled funds hold “Profitability” stocks in terms of a high return on asset. But at the same time, they also hold stocks with a low return on equity. Based on asset pricing literature such as Amihud and Mendelson (1989), Diether et al. (2002), Ali et al. (2003), Amihud (2002), Liu (2006), and Li (2011), illiquid stocks and stocks with higher R&D have a higher expected return, which can also be seen from the higher hypothetical excess return of the portfolios of skilled funds. It is possible that funds happen to hold those stocks and enjoy the premium from those characteristics and deliver outperformance. An alternative explanation is that skilled funds choose to hold illiquid assets on purpose for anomalous returns, because they have a lower turnover ratio and do not trade stocks often. A study that disentangles the two hypotheses constitutes a good topic for future research. Here, we only provide evidence for the association between stock holdings and fund skill.

Our study contributes to the literature in the following aspects. Firstly, the asymptotic theory we derive, along with simulation evidence, provides a cautionary note on the empirical application of two prevailing bootstrap methods in financial research. Blake et al. (2017) attempt to compare the two bootstrap methods in an empirical exercise by weighing them with the same mutual fund data. They suggest that the evaluation of fund performance depends crucially on the employed bootstrap methodology, but provide limited insight into the actual theoretical properties or performance of the methods themselves. Compared to theirs, our study takes a more fundamental approach: we systematically delineate the size and power properties of these two bootstrap methods and assess how they perform in simulations guided by theory and calibrated to real data. Other studies have also raised the issue of test power with the Fama and French (2010) approach. Harvey and Liu (2020a,b) find the low test power in this approach through simulations from real data and attribute it to the undersampling of funds with a relatively short sample period and low signal-to-noise ratio in mutual fund data. Although we do not establish formal theories for the size and power of the Fama and French (2010) approach, we argue that it cannot handle the challenge of small sample sizes compared to the large number of funds like Kosowski et al. (2006), which is supported by our simulation results.

The size and power deficiencies in these two methods have economic implications. They could help shed light on the opposing empirical conclusions of whether and to what extent skilled mutual funds exist in Kosowski et al. (2006) and Fama and French (2010). Kosowski et al. (2006) find evidence for a significant fraction of outperforming funds, i.e., they reject the zero-alpha null hypothesis. Conversely, Fama and French (2010) fail to reject this null. A plausible explanation based on our findings is that the Fama and French (2010) method is overly conservative in test size, limiting its power to detect outperformance. Even when a substantial number of funds are skilled with reasonably large alphas, the Fama and French (2010) test may still erroneously conclude that all funds are zero-alpha. Harvey and Liu (2020a) offer similar reconciliations over the conflicting findings in these two studies in terms of test power.

Secondly, our study is broadly related to a large body of econometric and finance literature analyzing the impact of estimation errors in large-scale testing problems, where the cross-sectional dimension N can be much larger than the sample size T_i. In particular, we argue that it is essential to allow $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1}\in(0,\infty)$ for evaluating mutual fund performance. One strand of this literature focuses on testing or estimating a high-dimensional structural parameter. Fan, Liao, and Yao (2015) and Pesaran and Yamagata (2017) highlight the challenges in designing such tests that can guard against the accumulation of high-dimensional estimation errors when $N >> T_i$. Gonçalves and Perron (2014, 2020) investigate the validity of bootstrap methods in estimating factor-augmented regression models of large dimensions, although our study is concerned with bootstrap testing in linear factor models. Another line of this literature is dedicated to multiple testing for evaluating fund performance, which is similarly challenged by estimation errors in large dimensions. Barras et al. (2010) apply the false discovery rate control approach to the field of mutual fund performance, which uses the residual bootstrap method as in Kosowski et al. (2006) to estimate the p-values. Through simulation studies, Andrikogiannopoulou and Papakonstantinou (2019) point out that this approach can be markedly biased in estimating the proportions of zero and non-zero alpha funds, particularly after accounting for the "large N small T" feature of mutual fund data.[3] Liu and Shao (2014) derive the accuracy of bootstrap calibration in estimating p-values in large-scale multiple testing and show that the accumulated estimation errors in large dimensions can invalidate the false discovery rate control. Recently, Giglio, Liao, and Xiu (2020) propose a high-dimensional multiple testing method that can validly control the false discovery rate in large cross-sections.

Finally, our theories for the bootstrap method in Kosowski et al. (2006) and the proposed new test are developed for the null hypothesis of zero alpha, i.e., $H_0 : \alpha_i = 0$ for $i = 1, \ldots, N$. Alternatively, the null hypothesis $H_0' : \max_{1\leq i\leq N}\alpha_i \leq 0$ can be used to test for the existence of at least one fund with a positive alpha. The advantage of using the second null hypothesis is that the test power will not be affected by the existence of unskilled funds. In general, such a test is quite conservative in that the size is asymptotically below the nominal level,

[3] Barras, Scaillet, and Wermers (2020) argue that the bias can be tremendously alleviated when choosing reasonable parameters. However, their response is limited to the assumption that the residuals are all normal.

leading to a less powerful test. In any event, how to develop a valid bootstrap test under H_0' for large cross-sections is an interesting problem that requires further research. It should be pointed out that the theory in White (2000) is not applicable as it requires that N is fixed and $T_i = T \to \infty$ for $i = 1, \ldots, N$, which is different from the setting in this study that N is much larger than T_i's.

We organize the rest of the paper as follows. Section 4.2 introduces the general framework and develops asymptotic theories to explain the pitfalls of the existing bootstrap methods in evaluating mutual fund performance in an ideal setting with independent fund residuals. Section 4.3 proposes a new bootstrap test for zero-alpha funds and a sequential testing approach to determine skilled funds. To better appreciate the methodology, we first develop the theory for the Hotelling's T-squared test with and without bootstrap calibration in an ideal setting of independent fund residuals. We then validate the bootstrap Hotelling's T-squared test in a general setting with serial correlations and cross-sectional dependence. Section 4.4 presents the empirical analysis comparing various characteristics of funds identified by our test as having or lacking skill. Section 4.5 concludes. To save space, we put all theoretical derivations, simulation studies for extant and proposed bootstrap methods, and additional results for empirical applications in a supplementary file.

4.2 Existing Bootstrap Methods

The bootstrap tests in Kosowski et al. (2006) and Fama and French (2010) are concerned with the overall null hypothesis of zero alpha for all mutual funds. Apart from different fund selection criteria and sample periods, there are two distinctions between these studies. First and foremost, they differ in bootstrap procedures. Kosowski et al. (2006) mainly use a standard residual bootstrap from a regression model and rely on the cross-sectionally bootstrapped p-value for formal inference, while the procedure in Fama and French (2010) jointly resamples fund returns and factor returns for all funds and use the “likelihood” for informal inference.[4] Although not theoretically grounded, the joint-resampling design has been widely acclaimed as an advantage of Fama and French (2010) over Kosowski et al. (2006) in capturing the cross-sectional dependence among funds, which may derive from mutual fund herding (Wermers, 1999), idea sharing (Cujean, 2019), and common information or liquidity shocks. Second, they reach different economic conclusions. Kosowski et al. (2006) conclude that a sizable minority of managers possess skills to deliver positive alpha, whereas Fama and French (2010) find that few funds can outperform.

There is mounting research to reconcile the distinct empirical findings, yet scant effort to question the statistical validity or assess the performance of the methods themselves. Fama and French (2010, p. 1940) claim that "whatever [fund] inclusion rules are used, failure to account for the joint distribution of fund returns, and of the fund and factor returns, biases the inferences of Kosowski et al. (2006) toward positive performance". Using the same fund inclusion criteria over the same sample period, Blake et al. (2017) directly compare the two alternative bootstrap methods in the context of U.K. mutual funds. They posit that different bootstrap resampling schemes lead to divergent findings from the two methods. In more recent studies, Harvey and Liu (2020a,b) propose a simulation-based double bootstrap method to evaluate the test size and power of the two bootstrap methods. Harvey and Liu (2020a) assert that the lack of power of the Fama and French approach to detect outperforming funds may help reconcile the difference between Kosowski et al. (2006) and Fama and French (2010). Harvey and Liu (2020b) compare several different bootstrap implementations and recommend the Fama and French approach with some modifications for future research. As they simulate fund returns by directly resampling the actual fund returns instead of from some known distribution, it is difficult to develop insights into the statistical properties of the bootstrap methods, in particular, whether the tests have an asymptotically correct size and sufficient power and whether the Fama and French (2010) test is indeed capable of dealing with cross-sectional dependence. These properties are of

[4] Although unreported, Kosowski et al. (2006) also randomize the factor returns in time series and find robust results.

fundamental importance in empirical finance research as they directly pertain to whether a hypothesis test based on these bootstrap procedures leads to correct and credible inferences.

Besides, there is ambiguity in the literature over how to conduct inference with the bootstrap methods. Kosowski et al. (2006) conduct formal hypothesis test by reporting the cross-sectionally bootstrapped p-values at different percentiles. Blake et al. (2017) suggest constructing a confidence interval of the bootstrap distribution of t-statistics at each percentile and testing whether the actual t-statistic at that percentile lies within the confidence interval to determine abnormal performance.[5] In contrast, Fama and French (2010) compare (qualitatively) percentiles of the cross-section of t-statistics with the corresponding average values from bootstrap simulations, and rely on the likelihood to gain perspective on the evidence of skill.[6] The informal inference in Fama and French (2010) evolves into incorrect inference in Crane and Crotty (2018), who apply the Fama and French methodology to evaluate the performance of index funds. They determine index funds as skilled if the percentile of actual t-statistics is larger than the corresponding average value from bootstrap simulations, or if the bootstrap-based likelihood is larger than 0.5, above the 50th percentile.[7]

In what follows, we encapsulate these statistical challenges into the foundational question: what is the asymptotic distribution of the bootstrap test under the null hypothesis of zero alpha for all funds? We tackle the challenges by formalizing the bootstrap-based inference in a rigorous theoretical framework.

[5] Blake et al. (2017, p. 1291) note: “If the actual $t(\hat{\alpha})$ lies to the right (left) of the CI at a given percentile point, this provides robust evidence of managerial outperformance (underperformance) at that percentile point.”

[6] Fama and French (2010, p. 1931) state: “we infer that some managers lack skill sufficient to cover costs if low fractions of the simulation runs produce left tail percentiles of $t(\alpha)$ below those from actual net fund returns, or equivalently if large fractions of the simulation runs beat the left tail t_α estimates from actual net fund returns.”

[7] Crane and Crotty (2018, p. 43) suggest: “In particular, if the bootstrapped and actual distributions are equal at a given percentile (i.e., zero alpha), the likelihood value should be 0.5.”

4.2.1 Measuring Fund Performance

We evaluate fund returns with a set of J benchmark factors and model fund excess returns as

$$r_{i,t} = \alpha_i + \sum_{j=1}^{J} \beta_{ij} f_{jt} + \varepsilon_{i,t},\ j = 1, \ldots, J,\ t = t_i + 1, \ldots, t_i + T_i,\ i = 1, \ldots, N, \tag{4.1}$$

where $r_{i,t}$ is the excess return (i.e., net return minus the one-month treasury bill rate) for fund i in period t, α_i is fund alpha, β_{ij} is fund i's risk loading on the j-th factor f_{jt}, and $\varepsilon_{i,t}$ is fund residual. Depending on the factor model, some popular sets of factors are the market factor in Jensen (1968), the market factor plus the SMB (small minus big) and HML (high minus low) in Fama and French (1996), and the Fama-French three factors plus the momentum factor in Carhart (1997). We allow the observation window $[t_i + 1, t_i + T_i]$ and sample size T_i for each fund to be different, as is the case in mutual funds.

To simplify exposition, let $Y_{i,t} = r_{i,t}$, $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iJ})'$ and $\boldsymbol{X}_t = (f_{1t}, f_{2t}, \ldots, f_{Jt})'$, where A' denotes the transpose of the vector or matrix A. Then we write equation (4.1) as

$$Y_{i,t} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{X}_t + \varepsilon_{i,t},\ t = t_i + 1, \ldots, t_i + T_i,\ i = 1, \ldots, N. \tag{4.2}$$

Let $\overline{Y_i} = T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} Y_{i,t}$, $\overline{\boldsymbol{X}}_i = T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} \boldsymbol{X}_t$, and $\sigma_i^2 = E(\varepsilon_{i,t}^2) < \infty$. For each i, the least-squares estimators of $\boldsymbol{\beta}_i$ and α_i based on (4.2) are

$$\begin{cases} \hat{\boldsymbol{\beta}}_i = \left\{ T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} (\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)' \right\}^{-1} \left\{ T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} (Y_{i,t} - \overline{Y_i})(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i) \right\}, \\ \hat{\alpha}_i = \overline{Y}_i - \hat{\boldsymbol{\beta}}_i' \overline{\boldsymbol{X}}_i. \end{cases} \tag{4.3}$$

Define

$$\hat{\varepsilon}_{i,t} = Y_{i,t} - \hat{\alpha}_i - \hat{\boldsymbol{\beta}}_i' \boldsymbol{X}_t, \hat{\sigma}_i^2 = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} \hat{\varepsilon}_{i,t}^2, \text{ and } \Sigma_i = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} (\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)'.$$

The t-statistic to test for $H_{0,i} : \alpha_i = 0$ is $\hat{t}_i(0)$ with

$$\hat{t}_i(\alpha_i) := \sqrt{T_i} \frac{\hat{\alpha}_i - \alpha_i}{\hat{\sigma}_i \sqrt{1 + \overline{\boldsymbol{X}}_i' \Sigma_i^{-1} \overline{\boldsymbol{X}}_i}} \text{ for } i = 1, \ldots, N.$$

The i-th fund is deemed skilled (unskilled) if $\hat{t}_i(0)$ is significantly larger (smaller) than zero. Otherwise, the i-th fund is declared as zero-alpha. In performance evaluation, much attention has been paid to answer the question of whether mutual funds with the largest (smallest) t-statistics are skilled (unskilled). Alternatively, if the goal is to determine the proportions of skilled and unskilled funds, a multiple hypothesis test is in place with the null hypotheses $H_{0,i} : \alpha_i = 0$ for $i = 1, \ldots, N$. The testing approach relies on the sequence $\{\hat{t}_1(0), \ldots, \hat{t}_N(0)\}$, which may entail the risk of false discoveries resulting from estimation errors of the sequence (Barras et al., 2010).

4.2.2 Assessing Fund Performance

Let $\hat{t}_{(1)}(0) \le \ldots \le \hat{t}_{(N)}(0)$ denote the order statistics of $\hat{t}_1(0), \ldots, \hat{t}_N(0)$. Kosowski et al. (2006) and Fama and French (2010) propose to compare $\hat{t}_{([pN])}(0)$ with its bootstrapped counterpart at some percentile level $p \in (0, 1)$, where $[pN]$ denotes the integer part of pN. More specifically, Kosowski et al. (2006) randomly resample the residuals from $\{\hat{\varepsilon}_{i,t}\}_{t=t_i+1}^{t_i+T_i}$ with replacement, say $\{\varepsilon_{i,t}^b\}_{t=t_i+1}^{t_i+T_i}$, for the b-th bootstrapped residuals with $b = 1, \ldots, B$, and compute pseudo excess returns by $Y_{i,t}^b = \hat{\boldsymbol{\beta}}_i' \boldsymbol{X}_t + \varepsilon_{i,t}^b$. Using the pseudo excess returns of each fund i, they rerun the regressions

$$Y_{i,t}^b = \alpha_i^* + \boldsymbol{\beta}_i{}' \boldsymbol{X}_t + \varepsilon_{i,t}^*, \ t = t_i + 1, \ldots, t_i + T_i.$$

Similar to (4.3), the above regression yields

$$\begin{cases} \hat{\boldsymbol{\beta}}_i^b = \left\{ T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} (\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)' \right\}^{-1} \left\{ T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} (Y_{i,t}^b - \overline{Y}_i^b)(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i) \right\}, \\ \hat{\alpha}_i^b = \overline{Y}_i^b - (\hat{\boldsymbol{\beta}}_i^b)' \overline{\boldsymbol{X}}_i, \end{cases}$$

where $\overline{Y}_i^b = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} Y_{i,t}^b$. The bootstrapped t-statistic is

$$\hat{t}_i^b(0) = \sqrt{T_i} \frac{\hat{\alpha}_i^b}{\hat{\sigma}_i^b \sqrt{1 + \overline{\boldsymbol{X}}_i' \Sigma_i^{-1} \overline{\boldsymbol{X}}_i}}, \tag{4.4}$$

where $\hat{\varepsilon}_{i,t}^b = Y_{i,t}^b - \hat{\alpha}_i^b - (\hat{\boldsymbol{\beta}}_i^b)' \boldsymbol{X}_t$ and $(\hat{\sigma}_i^b)^2 = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} (\hat{\varepsilon}_{i,t}^b)^2$ for $b = 1, \ldots, B$. Note that we do not use $\hat{\alpha}_i^b - \hat{\alpha}_i$ in (4.4) as $Y_{i,t}^b$ is equal to $\hat{\boldsymbol{\beta}}' \boldsymbol{X}_t + \varepsilon_{i,t}^b$ rather than $\hat{\alpha}_i + \hat{\boldsymbol{\beta}}' \boldsymbol{X}_t + \varepsilon_{i,t}^b$. To test for the existence of skilled (unskilled) funds, Kosowski et al. (2006) use the bootstrapped p-value

$$S(p) = \frac{1}{B}\sum_{b=1}^{B} I\big(\hat{t}^{b}_{([pN])}(0) \leq \hat{t}_{([pN])}(0)\big)$$

when $p \in (0, 0.5)$, and $1-S(p)$ when $p \in (0.5, 1)$, where $\hat{t}^{b}_{([pN])}(0)$ is the $[pN]$-th order statistic of $\hat{t}^{b}_{1}(0), \ldots, \hat{t}^{b}_{N}(0)$. The test is conducted based on the premise that $S(p)$ is asymptotically uniformly distributed under $H_0 : \alpha_i = 0, i = 1, \ldots, N$. Traditionally, a test statistic only depends on the sample, and the bootstrap method is employed to generate bootstrapped test statistics for accurately approximating the distribution function of the test statistic. In this sense, the test using $S(p)$ is unconventional.

The independent resampling scheme of Kosowski et al. (2006) discards the potential cross-sectional dependence among fund returns. In an attempt to account for the cross-sectional dependence, Fama and French (2010) propose to resample both the factor returns and fund returns according to a reindexed time sequence drawn from $\{1, \ldots, T\}$ with replacement, where $T = \max\{t_1+T_1, \ldots, t_N+T_N\}$, leading to the bootstrapped sample $\{Y^{b}_{i,t}, \boldsymbol{X}^{b}_{t}\}$ for $b = 1, \ldots, B$. By subtracting $\hat{\alpha}_i$ from the bootstrapped fund returns, they run the following regression

$$Y^{b}_{i,t} - \hat{\alpha}_i = \alpha^{*}_{i} + \boldsymbol{\beta}_i{}'\boldsymbol{X}^{b}_{t} + \varepsilon^{*}_{i,t}, \qquad t = t_i + 1, \ldots, t_i + T_i,$$

for each fund i. With the least-squares estimator $\hat{\alpha}^{b}_{i}$ of α^{*}_{i}, they obtain the t-statistic $\hat{t}^{b}_{i}(0)$ defined in (4.4) and $S(p)$ for $p \in (0, 1)$, which is called the "likelihood". In the resampling scheme, they implicitly assume a missing at random design as in Gagliardini et al. (2016).[8]

Recently, Blake et al. (2017) examine whether $\hat{t}_{([pN])}(0)$ is contained by the stochastic interval

$$\Big[\hat{t}^{([aB/2])}_{([pN])}(0), \quad \hat{t}^{(B-[aB/2])}_{([pN])}(0)\Big] \text{ for a given } a \in (0, 1),$$

[8] A notable practical problem with random sampling the same time sequence for all funds in the Fama and French approach is that in each simulation run, if a fund does not exist for the entire sample period, which is the case for most funds, the number of observations for the bootstrap sample may differ from that for the actual sample. In particular, some funds may end up with a bootstrap time series shorter than the minimum fund return requirement for fund selection.

where $\hat{t}^{(1)}_{([pN])}(0) \leq \ldots \leq \hat{t}^{(B)}_{([pN])}(0)$ denote the order statistics of $\hat{t}^{1}_{([pN])}(0), \ldots, \hat{t}^{b}_{([pN])}(0)$. Under $H_0 : \alpha_i = 0, i = 1, \ldots, N$, they implicitly conjecture that the confidence interval has the asymptotically correct coverage probability, i.e., $P\big(\hat{t}_{([pN])}(0) \in \big[\hat{t}^{([aB/2])}_{([pN])}(0), \hat{t}^{(B-[aB/2])}_{([pN])}(0)\big]\big) \to 1 - a$ as $N \to \infty$.

While the bootstrap methods have been widely applied to evaluate fund performance, no theoretical justification is given to guarantee the plausibility of the implicit conjectures and hence the statistical validity of bootstrap tests. In other words, no paper formally tests $H_0 : \alpha_i = 0, i = 1, \ldots, N$ by providing a critical region to reject H_0 for these bootstrap methods. This requires deriving and estimating the asymptotic distribution of $S(p)$ under H_0.

4.2.3 Size and Power Properties of the Kosowski et al. (2006) Bootstrap Test

We present the asymptotic theories of test size and power associated with the Kosowski et al. (2006) bootstrap method for fund performance evaluation. In a simplified setting of an unbalanced panel with independent fund residuals, we use Edgeworth expansion to characterize the approximation error in estimating α_i and unveil how small sample sizes of some funds relative to the number of funds and skewness in fund residuals complicate the asymptotic limits of the bootstrap tests. The theorem we derive shows that the Kosowski et al. (2006) bootstrap test does not have an asymptotically correct size when $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} \in (0, \infty)$. It is known that the accuracy of a two-sided test is $O(1/T_i)$ for the i-th fund, and the accuracy of bootstrap two-sided test is $o(1/T_i)$. This is why we argue that the bootstrap test using $S(p)$ is unconventional, which fails to achieve the accuracy of a traditional bootstrap method. Without doubt, the bootstrap test would have a more severely biased size when $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} \in (0, \infty)$, and when the ideal independent setting is replaced by a practical setting with serial correlations and cross-sectional dependence.[9]

Following the well-known Edgeworth expansions in Section 4.3.4 of Hall (1992) for his

[9] While we acknowledge that the Kosowski et al. (2006) test is not justified for such a practical setting, Table 4.A.5 in the Appendix shows that this test could be substantially oversized under cross-sectional dependence.

statistics T and T^* with $x_0 = 0$, corresponding to our $\hat{t}_i(\alpha_i)$ and $\hat{t}_i^b(0)$, we have

$$P(\hat{t}_i(\alpha_i) \le z | \{\boldsymbol{X}_t\}) = \Phi(z) + T_i^{-1/2}\phi(z) q_{i,1,\boldsymbol{x}}(z) + T_i^{-1}\phi(z) q_{i,2,\boldsymbol{x}}(z) + O_P(T_i^{-3/2}) \tag{4.5}$$

and

$$P(\hat{t}_i^b(0) \le z | \{Y_{i,t}, \boldsymbol{X}_t\}) = \Phi(z) + T_i^{-1/2}\phi(z) \hat{q}_{i,1,\boldsymbol{x}}(z) + T_i^{-1}\phi(z) \hat{q}_{i,2,\boldsymbol{x}}(z) + O_P(T_i^{-3/2}) \tag{4.6}$$

where

$$q_{i,1,\boldsymbol{x}}(z) = -\frac{1}{6}\gamma_i\left(\gamma_{i,\boldsymbol{x}} - \frac{3}{\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\right)z^2 - \frac{1}{6}\gamma_i\gamma_{i,\boldsymbol{x}},$$

$$q_{i,2,\boldsymbol{x}}(z) = -z\left\{2 + \frac{\gamma_i^2}{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i} + \frac{1}{24}\left[\kappa_i\kappa_{i,\boldsymbol{x}} + 6 - \frac{8\gamma_i^2}{\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\cdot\right.\right.$$

$$\left.\left.\left(\gamma_{i,\boldsymbol{x}} - \frac{3}{\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\right)(z^2-3)\right] + \frac{1}{72}\gamma_i^2\left(\gamma_{i,\boldsymbol{x}}^2 - \frac{3}{\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\right)^2(z^4 - 10z^2 + 15)\right\},$$

$$\gamma_i = \frac{E(\varepsilon_{i,t}^3)}{(E\varepsilon_{i,t}^2)^{3/2}},\ \hat{\gamma}_i = \frac{T_i^{-1}\sum_{t=t_i+1}^{t_i+T_i}\hat{\varepsilon}_{i,t}^3}{\left(T_i^{-1}\sum_{t=t_i+1}^{t_i+T_i}\hat{\varepsilon}_{i,t}^2\right)^{3/2}},\ \kappa_i = \frac{E(\varepsilon_{i,t}^4)}{(E\varepsilon_{i,t}^2)^2} - 3,\ \hat{\kappa}_i = \frac{T_i^{-1}\sum_{t=t_i+1}^{t_i+T_i}\hat{\varepsilon}_{i,t}^4}{\left(T_i^{-1}\sum_{t=t_i+1}^{t_i+T_i}\hat{\varepsilon}_{i,t}^2\right)^2} - 3,$$

$$\gamma_{i,\boldsymbol{x}} = \frac{1}{T_i}\sum_{t=t_i+1}^{t_i+T_i}\left\{\frac{1-\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)}{\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\right\}^3,\ \kappa_{i,\boldsymbol{x}} = \frac{1}{T_i}\sum_{t=t_i+1}^{t_i+T_i}\left\{\frac{1-\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)}{\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\right\}^4 - 3,$$

$\hat{q}_{i,1,\boldsymbol{x}}$ and $\hat{q}_{i,2,\boldsymbol{x}}$ equal $q_{i,1,\boldsymbol{x}}$ and $q_{i,2,\boldsymbol{x}}$ with γ_i and κ_i replaced by $\hat{\gamma}_i$ and $\hat{\kappa}_i$, respectively, and $\Phi(x)$ and $\phi(x)$ denote the distribution function and density function of a standard normal random variable, respectively.

To develop theories for the test size and power of the bootstrap methods, we need the following regularity conditions:

(**C1**) For each $i = 1, \dots, N$, $\{\varepsilon_{i,t}, t = t_i + 1, \dots, t_i + T_i\}$ is a sequence of independent and identically distributed random variables with mean zero and finite variance $\sigma_i^2 > 0$. Further assume that these N sequences are independent of each other and $\sup_{i\ge 1} E(|\varepsilon_{i,t}|^{4+\delta}) < \infty$ for some $\delta > 0$.

(**C2**) $\{\boldsymbol{X}_t, t = 1, \ldots, T\}$ is stationary and ergodic, where $T = \max(t_1 + T_1, \ldots, t_N + T_N)$, $E(\|\boldsymbol{X}_t\|^{4+\delta}) < \infty$ for some $\delta > 0$, and $\|\cdot\|$ denotes the usual Euclidean norm of a vector. Further assume that $\{\boldsymbol{X}_t, t = 1, \ldots, T\}$ is independent of all sequences $\{\varepsilon_{i,t}, t = t_i + 1, \ldots, t_i + T_i\}$ for $i = 1, \ldots, N$.

(**C3**) The covariance matrix $\Sigma := T^{-1}\sum_{t=1}^{T}(\boldsymbol{X}_t - \overline{\boldsymbol{X}})(\boldsymbol{X}_t - \overline{\boldsymbol{X}})'$ is nonsingular where $\overline{\boldsymbol{X}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{X}_t$. Assume that $\max_{1\le i\le N}\|\Sigma_i - \Sigma\| \xrightarrow{p} 0$ and $\min_{1\le i\le N} T_i \to \infty$ as $N \to \infty$, where Σ_i defined in Section 4.2.2.

(**C4**) $\lim\limits_{N\to\infty}\dfrac{1}{\sqrt{N}}\sum\limits_{i=1}^{N}T_i^{-1} < \infty.$

The following theorem explains the test size of the residual-based bootstrap method in Kosowski et al. (2006), revealing that small sample sizes of funds complicate the applicability of the residual-based bootstrap method of Kosowski et al. (2006) in a large cross-sectional dimension.

Theorem 4.1 (Test Size of the KTWW Approach)**.** *Suppose conditions (C1)–(C4) hold. Consider the residual-based bootstrap method in Kosowski et al. (2006). Under $H_0 : \alpha_i = 0, i = 1, \ldots, N$, for any fixed $p \in (0,1)$ and a given significance level $a \in (0,1)$, we have*

$$P\big(S(p) \le a|\{\boldsymbol{X}_t\}\big) = P\left(\Phi\left(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}(A_{N1} + A_{N2} + A_{N3})\right) \le a|\{\boldsymbol{X}_t\}\right) + o_P(1) \tag{4.7}$$

and

$$\begin{aligned}&P\left(\hat{t}_{([pN])}(0) \in \big[\hat{t}^{([aB/2])}_{([pN])}(0),\ \hat{t}^{(B-[aB/2])}_{([pN])}(0)\big]|\{\boldsymbol{X}_t\}\right)\\ =&P\left(a/2 \le \Phi\left(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}(A_{N1} + A_{N2} + A_{N3})\right) \le 1 - a/2|\{\boldsymbol{X}_t\}\right) + o_P(1)\end{aligned} \tag{4.8}$$

as $N \to \infty$ and $B \to \infty$, where

$$\begin{aligned} A_{N1} &= \sqrt{N}\big(\hat{t}_{([pN])}(0) - Q_p\big) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1} q_{i,2,\boldsymbol{x}}(Q_p), \\ A_{N2} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1}\sqrt{T_i}\big(\hat{q}_{i,1,\boldsymbol{x}}(Q_p) - q_{i,1,\boldsymbol{x}}(Q_p)\big), \\ A_{N3} &= \frac{1}{2} Q_p \sqrt{N}\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\}^2 \\ &\qquad - \sqrt{N}\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2} q'_{i,1,\boldsymbol{x}}(Q_p)\}\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\}, \end{aligned}$$

$Q_p = \Phi^{\leftarrow}(p)$ with $\Phi^{\leftarrow}(x)$ denoting the inverse function of $\Phi(x)$, $q'_{i,1,\boldsymbol{x}}(z) = dq_{i,1,\boldsymbol{x}}(z)/dz$, and $\frac{\phi(Q_p)}{\sqrt{p(1-p)}} A_{N1}$ has the standard normal limit.

Furthermore, when

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} = 0 \quad \text{and} \quad \sup_{i\geq 1} E(\varepsilon_{i,t}^6) < \infty, \tag{4.9}$$

we have

$$P\big(S(p) \leq a | \{\boldsymbol{X}_t\}\big) = a + o_P(1) \tag{4.10}$$

and

$$P\left(\hat{t}_{([pN])}(0) \in \big[\hat{t}_{([pN])}^{([aB/2])}(0),\ \hat{t}_{([pN])}^{(B-[aB/2])}(0)\big] | \{\boldsymbol{X}_t\}\right) = 1 - a + o_P(1). \tag{4.11}$$

We can replace the first equation in (4.9) by $\lim_{N\to\infty} \sqrt{N}\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2}\}^2 = 0$, but these two equations are equivalent when

$$\liminf \frac{1}{N} \sum_{i=1}^{N} \left(\frac{T_i}{\min_{1\leq j\leq N} T_j}\right)^{-1/2} > 0. \tag{4.12}$$

This is because

$$\sqrt{N}\left\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2}\right\}^2 \leq \sqrt{N}\frac{1}{N}\sum_{i=1}^{N} T_i^{-1} \leq \sqrt{N}\left\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2}\right\}^2 \left\{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{T_i}{\min_{1\leq j\leq N} T_j}\right)^{-1/2}\right\}^{-1}. \tag{4.13}$$

Nevertheless, we use (4.9) for ease of comparison with the Hotelling's T-squared test in Section 4.3.

The result (4.10) above shows that $S(p)$ follows a uniform distribution when T_i's are not small. In this case, the bootstrap test in Kosowski et al. (2006) is statistically valid with an asymptotically correct size. When a significant fraction of funds have a small sample size such that the first equation in (4.9) does not hold, but (4.12) holds, A_{N3} invalidates (4.10). When some fund residuals don't have finite 6-th moments, A_{N2} may destroy (4.10). In either case, the bias due to A_{N2} and/or A_{N3} depends on the skewness of fund residuals and complicates the limit of $S(p)$.

The impact of the accumulated estimation errors on the test size is well supported by the simulation evidence in Tables 4.A.1 – 4.A.3 in the Appendix. For a finite N, larger residual skewness will enlarge the influence of the bias term A_{N2} on test size. The simulation results in Table 4.A.1 suggest that $T = 200$ is large enough to ensure an accurate size for the case of zero skewness, while the results in Table 4.A.3 indicate that $T = 200$ produces a distorted size when fund residuals are heavily skewed. To contextualize the statistical biases, the average number of monthly return observations of actively managed U.S. equity funds in our empirical study is 204 (with a median of 186 and a standard deviation of 103), and the average absolute residual skewness estimated from the four-factor model is 0.385 (with a median of 0.246 and a standard deviation of 0.673). Furthermore, among 2650 mutual funds in our data sample, 1318 funds have a sample period shorter than 186 months, and 1325 funds have absolute residual skewness greater than 0.246. Table 4.1 in Section 4.4 summarizes the statistics for these mutual fund data characteristics.

Next, we study the test power of Kosowski et al. (2006) under condition (4.9), i.e., $S(p)$ has the uniform distribution on $[0, 1]$ under the null hypothesis. From the result (4.10) of Theorem 4.1, the rejection region at the level a is $S(p) \leq a$ for some $p < 0.5$ in favor of the existence of unskilled funds, and $S(p) \geq 1 - a$ for some $p > 0.5$ in favor of the existence of skilled funds. Denote

$$\delta_i = \frac{\sqrt{T_i}\alpha_i}{\sigma_i\sqrt{1 + \overline{\mathbf{X}}_i'\Sigma_i^{-1}\overline{\mathbf{X}}_i}}, \quad \hat{\delta}_i = \frac{\sqrt{T_i}\alpha_i}{\hat{\sigma}_i\sqrt{1 + \overline{\mathbf{X}}_i'\Sigma_i^{-1}\overline{\mathbf{X}}_i}}, \quad \text{and} \quad \Delta_N = \frac{1}{N}\sum_{i=1}^{N}\{\Phi(Q_p) - \Phi(Q_p - \delta_i)\}.$$

Note that δ_i measures the individual departure of fund i from the null hypothesis $\alpha_i = 0$.

First consider the case of $\lim_{N\to\infty} \Delta_N \neq 0$. As $\Phi(x)$ is an increasing function, this case suggests that the overall departure contributed by positive-alpha funds is not comparable to that by negative-alpha funds. We need condition (C5) below, which ensures that the difference between $\hat{\hat{\delta}}_i$ and δ_i is asymptotically negligible over i.

$$(\mathbf{C5})\ \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \frac{|\delta_i|}{\sqrt{T_i}} = 0.$$

Theorem 4.2 (Test Power of the KTWW Approach, Case 1)**.** *Under conditions (C1)–(C3), (C5), and* (4.9)*, if* $\lim\limits_{N\to\infty} \Delta_N = \delta_0 \neq 0$*, then for any fixed* $p \in (0,1)$ *and all* $a \in (0,1)$

$P(S(p) \leq a|\{\boldsymbol{X}_t\}) \xrightarrow{p} I(\delta_0 < 0)$ *as* $N \to \infty$ *and* $B \to \infty$*.*

Theorem 4.2 states that, when only positive-alpha funds exist such that $\delta_0 > 0$, the KTWW test has power approaching one for some percentile $p > 0.5$, i.e., the test is powerful in the detection of skilled funds; likewise when only negative-alpha funds exist. This would make the KTWW approach appealing provided that the fund sample contains no unskilled funds. The simulation results in Table 4.A.4 (where the proportion of negative-alpha funds $\pi^- = 0$) in the Appendix confirm this theorem.

However, Theorem 4.2 warrants an important empirical consideration for the KTWW method regarding its test power. The presence of a large number of negative-alpha funds (*poor alternatives* in Hansen, 2005) could unfavorably affect the power in detecting skilled funds. As implied by Theorem 4.2, when negative and positive alphas are both present, the bootstrap method shall only have the power to detect either skilled or unskilled funds, which is determined by the sign of δ_0. In particular, the vast majority of mutual funds typically have zero or negative alphas in the actual data, which makes $\delta_0 < 0$ possible, and the KTWW bootstrap approach could suffer from low power to identify outperforming funds. Panels A and B of Table 4.A.4 also confirm this observation when negative-alpha funds are prevalent in the fund population.

Consider further the case of $\lim_{N\to\infty} \Delta_N = 0$, which is true when all δ_i's are small. In contrast to the case of Theorem 4.2, this case suggests that the overall departure from the null hypothesis in positive-alpha funds is comparable to that in negative-alpha funds.

$$\textbf{(C6)}\ \lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{|\delta_i|}{\sqrt{T_i}}=0,\text{ and }\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\{\phi(Q_p)-\phi(Q_p-\delta_i)\}=0.$$

The first condition of (C6) assures that the difference between $\hat{\delta}_i$ and δ_i is asymptotically negligible over i when $\lim_{N\to\infty}\Delta_N = 0$. The second one measures the difference in the departure from the null between positive-alpha and negative-alpha funds, which is quite small. It holds when $\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}|\delta_i|=0$.

Theorem 4.3 (Test Power of the KTWW Approach, Case 2)**.** *Under conditions (C1)–(C3), (C6) and* (4.9)*, if* $\lim_{N\to\infty}\Delta_N=0$*, then for any fixed* $p\in(0,1)$ *and all* $a\in(0,1)$ *we have*

$$P\big(S(p)\le a|\{\boldsymbol{X}_t\}\big)=P\big(\Phi(B_{N1}+B_{N2})\le a|\{\boldsymbol{X}_t\}\big)+o_P(1)$$

as $N\to\infty$ *and* $B\to\infty$*, where*

$$B_{N1}:=\frac{\phi(Q_p)}{\sqrt{p(1-p)}}\left(\sqrt{N}(\hat{t}_{([pN])}(0)-Q_p)+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1/2}q_{i,1,\boldsymbol{x}}(Q_p)+\frac{\sqrt{N}\Delta_N}{\phi(Q_p)}\right)$$

has a standard normal limit given $\{\boldsymbol{X}_t\}$ *and* $B_{N2}:=-\frac{\sqrt{N}\Delta_N}{\sqrt{p(1-p)}}$.

Theorem 4.3 shows that in the case where the signal-to-noise ratio is low (e.g., alphas are small), the test power is determined by B_{N2}, which depends on both negative and positive alphas. For example, when both skilled and unskilled funds exist and negative δ_i's dominate positive δ_i's in the sense that Δ_N is quite negative for some $p<0.5$ (i.e., B_{N2} is quite positive), it is unlikely that the null hypothesis is rejected at the level a (and thereby leads to low test power). Panel C of Table 4.A.4 in the Appendix shows that the presence of unskilled funds significantly lowers the test power for skilled funds.

4.2.4 Theoretical Difficulty of the Fama and French (2010) Bootstrap Test

Fama and French (2010) propose to resample both the factor returns and fund returns based on the same time sequence drawn from $\{1,\dots,T\}$ with replacement, where $T=\max\{t_1+T_1,\dots,t_N+T_N\}$, leading to the bootstrapped sample $\{Y_{i,t}^b,\boldsymbol{X}_t^b\}$ for $b=1,\dots,B$.

By subtracting $\hat{\alpha}_i$ given in (4.3) from the bootstrapped fund returns, they run the following regression

$$Y_{i,t}^b - \hat{\alpha}_i = \alpha_i^* + \boldsymbol{\beta}_i' \boldsymbol{X}_t^b + \varepsilon_{i,t}^*, \qquad t = t_i + 1, \ldots, t_i + T_i,$$

for each fund i. Let $\hat{\alpha}_i^b$ and $\hat{\boldsymbol{\beta}}_i^b$ denote the least-squares estimators for α_i^* and $\boldsymbol{\beta}_i$.

Define

$$\varepsilon_{i,t}^b = Y_{i,t}^b - \boldsymbol{\beta}_i' \boldsymbol{X}_t^b, \ \overline{\boldsymbol{X}}_i^b = T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} \boldsymbol{X}_t^b, \ \Sigma_i^b = T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} (\boldsymbol{X}_t^b - \overline{\boldsymbol{X}}_i^b)(\boldsymbol{X}_t^b - \overline{\boldsymbol{X}}_i^b)',$$

and

$$(\hat{\sigma}_i^b)^2 = T_i^{-1} \sum_{t=t_i+1}^{t_i+T_i} \{Y_{i,t}^b - \hat{\alpha}_i - \hat{\alpha}_i^b - (\hat{\boldsymbol{\beta}}_i^b)' \boldsymbol{X}_t^b\}^2.$$

Then,

$$\begin{aligned}
\hat{t}_i^b(0) &= \sqrt{T_i} \frac{\hat{\alpha}_i^b}{\hat{\sigma}_i^b \sqrt{1 + (\overline{\boldsymbol{X}}_i^b)'(\Sigma_i^b)^{-1} \overline{\boldsymbol{X}}_i^b}} \\
&= T_i^{-1/2} \sum_{t=t_i+1}^{t_i+T_i} \frac{\varepsilon_{i,t}^b}{\hat{\sigma}_i^b} \frac{1 - (\boldsymbol{X}_t^b - \overline{\boldsymbol{X}}_i^b)'(\Sigma_i^b)^{-1} \overline{\boldsymbol{X}}_i^b}{\sqrt{1 + (\overline{\boldsymbol{X}}_i^b)'(\Sigma_i^b)^{-1} \overline{\boldsymbol{X}}_i^b}} - \frac{\sqrt{T_i} \hat{\alpha}_i}{\hat{\sigma}_i^b \sqrt{1 + (\overline{\boldsymbol{X}}_i^b)'(\Sigma_i^b)^{-1} \overline{\boldsymbol{X}}_i^b}}.
\end{aligned}$$

Hence, the likelihood in Fama and French (2010) (i.e., the bootstrapped p-value in Kosowski et al. (2006)) is

$$S(p) = \frac{1}{B} \sum_{b=1}^{B} I\big(\hat{t}_{([pN])}^b(0) \le \hat{t}_{([pN])}(0)\big) \text{ for } p \in (0, 1).$$

Reminiscent of the theorems above for Kosowski et al. (2006), to develop theories, it is necessary to derive Edgeworth expansions for $\hat{t}_i^b(0)$ till the order T_i^{-1}. Because of the joint resampling scheme, $\boldsymbol{X}_t^b$ is tied to $\varepsilon_{i,t}^b$ by the same bootstrapped time index, which complicates the derivation of higher-order Edgeworth expansion. In contrast, $\varepsilon_{i,t}^b$ is resampled independently of $\boldsymbol{X}_t$ in Kosowski et al. (2006), which makes the derivation of higher-order Edgeworth expansion easier conditional on $\boldsymbol{X}_t$. Although we can not derive theorems for the size and power of the Fama and French (2010) approach, it is well expected that it has a biased size in the same vein as Kosowski et al. (2006) when (4.9) does not hold because of the unusual way of formulating $S(p)$ from all bootstrapped t-statistics. Furthermore, we conjecture that the joint resampling scheme brings about additional bias in that it cannot even correct the approximation error $T_i^{-1/2}$ in using t-statistics, that is, the test has a biased size even when $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1/2}<\infty$.

Simulation results in the Internet Appendix indicate that the Fama and French (2010) method has serious issues with both size and power. For example, Table 4.A.1 shows that when all fund residuals are normal and independent with $T = 468$ (the largest sample size an individual fund can have in the data), the Fama and French (2010) test has a distorted size when $N = 500$ (Panel A), and the size distortion is substantial when $N = 2650$ (Panel B). Consistent with its conservativeness, it tends to have very low power as evidenced in Table 4.A.4. To further address the concern that the bias may occur only for cross-sectionally independent funds residuals, we conduct a simulation exercise for cross-sectionally dependent fund residuals. Table 4.A.5 in the Appendix demonstrates that the size bias is not mitigated for the dependent case.

Collectively, an important message of the above findings for the empirical finance literature is that both bootstrap tests are inadequate for fund performance evaluation because of the stylized fact that sample sizes are much smaller than the number of funds in monthly data. The Kosowski et al. (2006) test is challenged by its size distortion in large dimensions and unsatisfactory power properties in the presence of a large number of negative-alpha funds. The conservativeness of the Fama and French (2010) test is so high that it can mask the evidence of skilled funds. Our paper is closely related to but different from Harvey and Liu (2020a,b), who rely on simulation studies and resample fund returns from mutual fund data instead of simulating them from known distributions. Our analyses regarding the impact of accumulated estimation errors when $N >> T$ can also speak to the debate over the applicability of the false discovery rate control method, another popular approach in evaluating fund performance, such as Barras et al. (2010, 2020) and Andrikogiannopoulou and Papakonstantinou (2019). This literature is similarly challenged by large cross-sections and small sample sizes and too many inferior alternatives, as studied by Giglio, Liao, and

Xiu (2020).

4.3 Methodology

Given the above discoveries on the perils and pitfalls of the bootstrap methods in Kosowski et al. (2006) and Fama and French (2010), we propose an alternative test for $H_0 : \alpha_i = 0, i = 1, \ldots, N$ with an accurate size even when the sample sizes of many funds in the population are relatively small. More specifically, our theories show that the bootstrap Hotelling's T-squared test attains an asymptotically correct size when $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} \in (0, \infty)$ by automatically correcting the bias in the Hotelling's T-squared test statistic. We further provide a flexible procedure combining sequential testing and screening to identify skilled funds by first locating a set of zero-alpha funds from the bootstrap Hotelling's T-squared test.

4.3.1 Bootstrap Hotelling's T-squared Test In the Ideal Setting

For each fixed i, $\hat{t}_i(0)$ has an asymptotic normal distribution under the zero-alpha null hypothesis. This motivates us to study the Hotelling's T-squared test statistic

$$HT = \frac{1}{\sqrt{2N}} \sum_{i=1}^{N} \{\hat{t}_i^2(0) - 1\}.$$

Note that HT should not have a standard normal limit because small sample sizes of some funds make the approximation error between $\hat{t}_i(0)$ and $N(0,1)$ non-negligible. In the following theorem, we show the asymptotic limit of HT using (4.5) under the zero-alpha null hypothesis.

Theorem 4.4 (Test Size of Hotelling's T-squared Test)**.** *Under conditions (C1)–(C4) and* $H_0 : \alpha_i = 0, i = 1, \ldots, N$, *we have, as* $N \to \infty$,

$$P(HT \leq z | \{\boldsymbol{X}_t\}) = \Phi\left(z - \frac{1}{\sqrt{2N}} \sum_{i=1}^{N} T_i^{-1} \int_{-\infty}^{\infty} s^2 \, d\{\phi(s) q_{i,2,\boldsymbol{x}}(s)\} \right) + o_P(1) \text{ for } z \in \mathbb{R}. \quad (4.14)$$

The Theorem above shows that the bias in (4.14) caused by the skewness is asymptotically negligible when the average fund sample size is large in the sense that $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} = 0$. However, when some funds have small sample sizes ($\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} > 0$), we have to correct the bias. Here, we propose to employ the residual-based bootstrap method.

Using notations in Section 4.2.2 of the residual-based bootstrap method, we draw the b-th bootstrap sample and compute the bootstrap t-statistic $\hat{t}_i^b(0)$ for $i = 1, \dots, N$, which results in the bootstrap Hotelling's T-squared test statistics

$$HT^b = \frac{1}{\sqrt{2N}} \sum_{i=1}^{N} \left\{ (\hat{t}_i^b(0))^2 - 1 \right\} \text{ for } b = 1, \dots, B.$$

Theorem 4.5 (Test Size of Bootstrap Hotelling's T-squared Test)**.** *Under conditions (C1)–(C4), we have, as* $N \to \infty$,

$$P\big(HT^b \le z | \{Y_{i,t}, \boldsymbol{X}_t\}\big) = \Phi\left(z - \frac{1}{\sqrt{2N}} \sum_{i=1}^{N} T_i^{-1} \int_{-\infty}^{\infty} s^2 \, d\{\phi(s)\hat{q}_{i,2,\boldsymbol{x}}(s)\} \right) + o_P(1) \textit{ for } z \in \mathbb{R}. \quad (4.15)$$

The Theorem above shows that the residual-based bootstrap method automatically corrects the bias in the test statistic HT. Let $HT^{(1)} \le \ldots \le HT^{(B)}$ denote the order statistics of $HT^1, \dots, HT^B$. Using Theorems 4.4 and 4.5, we reject the zero-alpha null hypothesis at level a whenever $HT < HT^{([aB/2])}$ or $HT > HT^{(B-[aB/2])}$. Simulation results in Table 4.A.6 in the Appendix show that the proposed test achieves remarkable size control. In contrast, because of the unconventional statistic $S(p)$, it is infeasible to make the bootstrap test using $S(p)$ valid when $\limsup_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} > 0$.

Theorem 4.6 (Test Power of Bootstrap Hotelling's T-squared Test)**.** *Under conditions (C1)–(C4), if* $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} |\delta_i|/\sqrt{T_i} < \infty$ *and* $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \delta_i^2 < \infty$, *we have, as*

$N \to \infty$ *and* $B \to \infty$,

$$
\begin{aligned}
&P\big(HT < HT^{([aB/2])} \textit{ or } HT > HT^{(B-[aB/2])} | \{\boldsymbol{X}_t\}\big) \\
=&\Phi\Bigg(\Phi^{\leftarrow}(a/2) - \frac{1}{\sqrt{2N}} \sum_{i=1}^{n} \Bigg\{ \frac{\delta_i}{\sqrt{T_i}} \Bigg(2 \int_{-\infty}^{\infty} s \, d\{\phi(s) q_{i,1,\boldsymbol{x}}(s)\} - \frac{E\varepsilon_{i,t}^3}{\sigma_i^3 \sqrt{1 + \overline{\boldsymbol{X}}_i' \Sigma_i^{-1} \overline{\boldsymbol{X}}_i}} \Bigg) + \delta_i^2 \Bigg\} \Bigg) \\
&+ \Phi\Bigg(\Phi^{\leftarrow}(a/2) + \frac{1}{\sqrt{2N}} \sum_{i=1}^{n} \Bigg\{ \frac{\delta_i}{\sqrt{T_i}} \Bigg(2 \int_{-\infty}^{\infty} s \, d\{\phi(s) q_{i,1,\boldsymbol{x}}(s)\} - \frac{E\varepsilon_{i,t}^3}{\sigma_i^3 \sqrt{1 + \overline{\boldsymbol{X}}_i' \Sigma_i^{-1} \overline{\boldsymbol{X}}_i}} \Bigg) + \delta_i^2 \Bigg\} \Bigg) \\
&+ o_P(1). \qquad (4.16)
\end{aligned}
$$

Theorem 4.6 shows that the skewness of residuals plays an important role in the test power. The simulation results in Table 4.A.6 show that the test is very powerful when fund returns display realistic levels of skewness and when a small fraction of skilled funds exist.

4.3.2 Bootstrap Hotelling's T-squared Test In A Practical Setting

Although the study focuses on monthly returns, one may still concern with serial dependence within fund residuals as well as cross-sectional dependence across fund residuals. In this section, we generalize the bootstrap Hotelling's T-squared test to the case where the errors in model (4.2) follow from an AR model with cross-sectional dependence. Further extension to ARMA-GARCH model is straightforward, but the method may not be efficient due to small sample sizes of monthly returns. More specifically, we consider

$$Y_{i,t} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{X}_t + \varepsilon_{i,t}, \ t = t_i + 1, \ldots, t_i + T_i, \ i = 1, \ldots, N,$$

and

$$\varepsilon_{i,t} = \sum_{j=1}^{p_i} \phi_{i,j} \varepsilon_{i,t-j} + \eta_{i,t}, \ \eta_{i,t} = \zeta_i U_t + e_{i,t}, \qquad (4.17)$$

where U_t denotes a common latent factor and ζ_i is referred to as the latent factor loading. Model (4.17) introduces serial correlations and cross-sectional dependence into mutual fund residuals.[10] Note that in the mutual fund data, many funds have different inception and exiting times, and the returns of two funds may be non-overlapping or only spuriously overlapping. As a result, it becomes challenging to test for the existence and strength of cross-sectional dependence as such a method, if any, will require paired (or balanced) data.[11] Nevertheless, we use ζ_i's to limit the impact of the cross-sectional dependence.

Throughout this section, we assume that

(C7) For each i, $\{\varepsilon_{i,t}\}$ is a strictly stationary sequence.

(C8) For each i, $\{U_t\}$ and $\{e_{i,t}\}$ are sequences of independent and identically distributed random variables with mean zero and all sequences over i are independent. Further assume $E(|U_t|^{4+\kappa}) < \infty$, and $\sup_i E(|e_{i,t}|^{4+\kappa}) < \infty$ for some $\kappa > 0$.

(C9) The deterministic constants ζ_i's satisfy that $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} |\zeta_i| = 0$.

Under condition (C4), an example for (C9) is that $\zeta_i = T_i^{-1-\kappa_i}$ for all $\kappa_i > 0$. Another example is that the number of funds with $|\zeta_i| > 0$ for $i = 1, \ldots, N$ is $o(\sqrt{N})$.

If we employ the Newey and West (1987) t-test with serial-correlation correction for each fund, it is challenging to figure out and correct the bias in the Hotelling's T-squared test, which is essential when $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} \in (0, \infty)$. For example, if we use the blockwise bootstrap method, the number of blocks is $o(T_i)$, which can not achieve the accuracy of $o(1/T_i)$ for the t-statistic constructed for the i-th fund. Hence, it seems necessary to utilize the error structure for inference to allow that $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} \in (0, \infty)$.

Denote $\boldsymbol{\theta}_i := (\theta_{i,1}, \ldots, \theta_{i,K_i})' = (\alpha_i, \boldsymbol{\beta}_i', \phi_{i,1}, \ldots, \phi_{i,p_i})'$ with K_i denoting the dimension of $\boldsymbol{\theta}_i$ and write

$$\eta_{i,t}(\boldsymbol{\theta}_i) = Y_{i,t} - \alpha_i - \boldsymbol{\beta}_i'\boldsymbol{X}_t - \sum_{j=1}^{p_i} \phi_{i,j}(Y_{i,t-j} - \alpha_i - \boldsymbol{\beta}_i'\boldsymbol{X}_{t-j}).$$

[10] Kosowski et al. (2006, p. 2582) implement a stationary bootstrap as a robustness check for serial correlations. Fama and French (2010, p. 1925) claim that autocorrelation is a minor issue for their test quoting the literature on autocorrelations of stock returns. Several empirical finance papers have adopted the second part of Equation (4.17) to model cross-sectional dependence; see, for example, Jones and Shanken (2005), Barras et al. (2010), Andrikogiannopoulou and Papakonstantinou (2019), and Harvey, Liu, and Saretto (2020).

[11] For all fund pairs with overlapping returns of at least 60 months, the average pairwise correlation is around 0.08 for the four-factor model residuals, indicating a minor degree of cross-sectional dependence.

Then, the least-squares estimator of $\boldsymbol{\theta}_i$ is

$$\widetilde{\boldsymbol{\theta}}_i = (\widetilde{\theta}_{i,1}, \ldots, \widetilde{\theta}_{i,K_i})' = \arg\min_{\boldsymbol{\theta}_i} \sum_{t=t_i+1}^{t_i+T_i} \eta_{i,t}^2(\boldsymbol{\theta}_i),$$

which solves the score equations $\sum_{t=t_i+1}^{t_i+T_i} \Delta_{i,t,k}(\boldsymbol{\theta}_i) = 0$ for $k = 1, \ldots, K_i$ with

$$\Delta_{i,t,k}(\boldsymbol{\theta}_i) = \eta_{i,t}(\boldsymbol{\theta}_i) \frac{\partial}{\partial \theta_{i,k}} \eta_{i,t}(\boldsymbol{\theta}_i).$$

To estimate the asymptotic variance of $\widetilde{\theta}_{i,1} = \widetilde{\alpha}_i$, construct t-statistic, and formulate the Hotelling's T-squared test, we need the following notations:

$$\Delta_{i,t,k}^{(j)}(\boldsymbol{\theta}_i) = \frac{\partial}{\partial \theta_{i,j}} \Delta_{i,t,k}(\boldsymbol{\theta}_i) \text{ for } j = 1, \ldots, K_i,$$

$$\boldsymbol{\Gamma}_{i,k} = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} (\Delta_{i,t,k}^{(1)}(\boldsymbol{\theta}_i), \ldots, \Delta_{i,t,k}^{(K_i)}(\boldsymbol{\theta}_i))', \ \boldsymbol{\Gamma}_i = (\boldsymbol{\Gamma}_{i,1}, \ldots, \boldsymbol{\Gamma}_{i,K_i}),$$

$$\widetilde{\boldsymbol{\Gamma}}_{i,k} = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} (\Delta_{i,t,k}^{(1)}(\widetilde{\boldsymbol{\theta}}_i), \ldots, \Delta_{i,t,k}^{(K_i)}(\widetilde{\boldsymbol{\theta}}_i))', \ \widetilde{\boldsymbol{\Gamma}}_i = (\widetilde{\boldsymbol{\Gamma}}_{i,1}, \ldots, \widetilde{\boldsymbol{\Gamma}}_{i,K_i}),$$

$$M_{i,k} = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} \Delta_{i,t,k}^2(\boldsymbol{\theta}_i), \ \boldsymbol{M}_i = diag(M_{i,1}, \ldots, M_{i,K_i}),$$

$$\widetilde{M}_{i,k} = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} \Delta_{i,t,k}^2(\widetilde{\boldsymbol{\theta}}_i), \ \widetilde{\boldsymbol{M}}_i = diag(\widetilde{M}_{i,1}, \ldots, \widetilde{M}_{i,K_i}).$$

Let $\boldsymbol{e}_i = (1, 0, \ldots, 0)'$ be the K_i dimensional vector. We consider the squared t-statistics

$$\widetilde{t}_i^2 = T_i \frac{\boldsymbol{e}_i' \widetilde{\boldsymbol{\theta}}_i \widetilde{\boldsymbol{\theta}}_i' \boldsymbol{e}_i}{\boldsymbol{e}_i' \widetilde{\boldsymbol{\Gamma}}_i^{-1} \widetilde{\boldsymbol{M}}_i (\widetilde{\boldsymbol{\Gamma}}_i^{-1})' \boldsymbol{e}_i}, \quad i = 1, \ldots, N,$$

where $\boldsymbol{e}_i' \widetilde{\boldsymbol{\theta}}_i \widetilde{\boldsymbol{\theta}}_i' \boldsymbol{e}_i = \widetilde{\alpha}_i^2$. Define the Hotelling's T-squared test statistic as

$$\widetilde{HT} = \frac{1}{\sqrt{2N}} \sum_{i=1}^{N} \{\widetilde{t}_i^2 - 1\}.$$

Because it is challenging to derive the Edgeworth expansions for the above t-statistics, we will calculate the bias of $E(\widetilde{t}_i^2)$ directly to prove the following theorem for analyzing the test size of this Hotelling's T-squared test.

Theorem 4.7 (Test Size of Hotelling's T-squared Test for AR Errors)**.** *Suppose models (4.2) and (4.17) satisfy (C2)–(C4), and (C7)–(C9). Under $H_0 : \alpha_i = 0, i = 1, \dots, N$, we have, as $N \to \infty$,*

$$P\big(\widetilde{HT} \le z\big) = \Phi\bigg(z - \frac{1}{\sqrt{2N}}\sum_{i=1}^{N}\mu_i\bigg) + o_P(1) \quad \textit{for } z \in \mathbb{R}, \tag{4.18}$$

where μ_i's are given in Equation (4.A.28) in the Appendix, and $\lim_{N\to\infty}|\frac{1}{\sqrt{2N}}\sum_{i=1}^{N}\mu_i| \in [0, \infty)$.

When $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}T_i^{-1} \in (0, \infty)$, the bias term above may not be zero asymptotically. Hence, it is necessary to correct this bias term. As it is nontrivial to estimate the bias, we employ the residual-based bootstrap method again. That is, for each i, draw a random sample with sample size T_i from $\{\eta_{i,t}(\widetilde{\boldsymbol{\theta}}_i)\}_{t=t_i+1}^{t_i+T_i}$ with replacement, say, $\{\eta_{i,t}^*\}_{t=t_i+1}^{t_i+T_i}$. After generating $\{\varepsilon_{i,t}^*\}_{t=t_i+1}^{t_i+T_i}$ from $\varepsilon_{i,t}^* = \sum_{j=1}^{p_i}\widetilde{\phi}_{i,j}\varepsilon_{i,t-j}^* + \eta_{i,t}^*$, we generate $\{Y_{i,t}^*\}_{t=t_i+1}^{t_i+T_i}$ by using

$$Y_{i,t}^* = \widetilde{\alpha}_i + \widetilde{\boldsymbol{\theta}}_i'\boldsymbol{X}_t + \varepsilon_{i,t}^*.$$

Therefore, we obtain $\widetilde{\boldsymbol{\theta}}_i^*$ by minimizing

$$\frac{1}{T_i}\sum_{t=t_i+1}^{t_i+T_i}\left\{Y_{i,t}^* - \alpha_i - \boldsymbol{\beta}_i'\boldsymbol{X}_t - \sum_{j=1}^{p_i}\phi_{i,j}(Y_{i,t-j}^* - \alpha_i - \boldsymbol{\beta}_i'\boldsymbol{X}_{t-j})\right\}^2.$$

Compute $\widetilde{\boldsymbol{\Gamma}}_i$ and $\widetilde{\boldsymbol{M}}_i$ using $Y_{i,t}^*$ and $\widetilde{\boldsymbol{\theta}}_i^*$ to get $\widetilde{\boldsymbol{\Gamma}}_i^*$ and $\widetilde{\boldsymbol{M}}_i^*$, which gives

$$\widetilde{t}_i^{2*} = T_i\frac{\boldsymbol{e}_i'(\widetilde{\boldsymbol{\theta}}_i^* - \widetilde{\boldsymbol{\theta}}_i)(\widetilde{\boldsymbol{\theta}}_i^* - \widetilde{\boldsymbol{\theta}}_i)'\boldsymbol{e}_i}{\boldsymbol{e}_i'(\widetilde{\boldsymbol{\Gamma}}_i^*)^{-1}\widetilde{\boldsymbol{M}}_i^*((\widetilde{\boldsymbol{\Gamma}}_i^*)^{-1})'\boldsymbol{e}_i}$$

and

$$\widetilde{HT}^* = \frac{1}{\sqrt{2N}}\sum_{i=1}^{N}\{\widetilde{t}_i^{2*} - 1\}.$$

Theorem 4.8 (Test Size of Bootstrap Hotelling's T-squared Test for AR Errors)**.** *Under conditions of Theorem 4.7 and $H_0 : \alpha_i = 0, i = 1, \dots, N$, we have, as $N \to \infty$,*

$$\sup_z\Big|P\big(\widetilde{HT}^* \le z|\{Y_{i,t}, \boldsymbol{X}_t\}\big) - P\big(\widetilde{HT} \le z\big)\Big| = o_P(1). \tag{4.19}$$

Using the above theorems, we can test H_0 by repeating the above bootstrap procedure B times to get $\widetilde{HT}^*_1, \ldots, \widetilde{HT}^*_B$ and then calculating the p-value as

$$2 \cdot \min\left(\frac{1}{B}\sum_{b=1}^{B} I\{\widetilde{HT} < \widetilde{HT}^*_b\}, \frac{1}{B}\sum_{b=1}^{B} I\{\widetilde{HT} > \widetilde{HT}^*_b\}\right).$$

We skip the tedious calculations for analyzing the test power. Simulations in Table 4.A.6 confirm Theorem 4.8 and the high power of the bootstrap test using $\widetilde{HT}$. It is straightforward to generalize model (4.17) to an ARMA process. However, when we model $\{\varepsilon_{i,t}\}$ by an ARMA-GARCH model, we have to infer both ARMA and GARCH models and employ the residual-based bootstrap method. If we only infer the ARMA model and employ the wild bootstrap method, then the resulted bootstrap test does not have an asymptotically correct size when $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1} \in (0, \infty)$. This is because the wild bootstrap method fails to catch the bias term in the Hotelling's T-squared test, which is a high-order term caused by the heteroscedasticity.

4.3.3 Sequential Testing and Thresholding

In this subsection, we provide a flexible procedure to select skilled mutual funds based on the bootstrap Hotelling's T-squared test. It proceeds in two steps: applying the new bootstrap test to identify a zero-alpha fund group and assessing whether the top- (bottom-) ranking funds are skilled (unskilled) compared to the zero-alpha group. We adopt a similar sequential testing approach to that in Hansen, Lunde, and Nason (2011) to determine a p-value threshold, and reduce the fund population into subsets of zero-alpha funds and non-zero alpha funds.

The zero-alpha funds are chosen by sequentially performing the bootstrap Hotelling's T-squared test on the sets of funds with p-values thresholded above incrementally decreasing levels. Similar to Hansen, Lunde, and Nason (2011), let $p_{N,1} \leq \ldots \leq p_{N,N}$ be the order statistics of $p_1, \ldots, p_N$ with each p_i computed from the standard t-test for $\alpha_i = 0$. Starting from a relatively large k, for those funds in the set $\mathcal{S}_k^0 = \{i = 1, \ldots, N : p_i \geq p_{N,k}\}$, we conduct the above Hotelling's T-squared test. The initial value of k can be chosen so that

$p_{N,k}$ is large. Then by decreasing k and thereby reducing $p_{N,k}$, we expand the set $\mathcal{S}_k^0$ to include more funds with smaller p-values (therefore more likely to be non-zero alpha) and repeat the bootstrap test iteratively on the expanded test set. If the p-value of the proposed bootstrap Hotelling's T-squared test is larger than the significance level, we determine the funds in $\mathcal{S}_k^0$ as zero-alpha. Obviously, as $p_{N,k}$ approaches 0, $\mathcal{S}_k^0$ will get closer to being the full sample of funds. The iteration is stopped when $k = 1$, that is, the original data sample is reached. We choose the smallest k, denoted as k^*, for which the new test does not reject the null to maximize the number of zero-alpha funds included in $\mathcal{S}_k^0$, and fix $p^* = p_{N,k^*}$ as the threshold of our test. The collection of funds in $\mathcal{S}_{k^*}^0$ is the zero-alpha fund group identified by our bootstrap test, which we denote as $\mathcal{S}^0$.

The second step of our procedure seeks to detect the true outperforming and underperforming funds. It operates on a reduced set of funds after having screened out the zero-alpha funds. Based on the threshold p^*, we further split the remaining funds not in $\mathcal{S}^0$ into two subsets: $\mathcal{S}^+ = \{i = 1, \ldots, N : p_i < p^*,\ \hat{t}_i(0) > 0\}$ and $\mathcal{S}^- = \{i = 1, \ldots, N : p_i < p^*,\ \hat{t}_i(0) < 0\}$. To investigate whether or not the funds in $\mathcal{S}^+$ are skilled compared to the zero-alpha funds in $\mathcal{S}^0$, we perform the bootstrap Hotelling's T-squared test again on the combined set $\mathcal{S}^0 \cup \mathcal{S}^+$.[12] If the zero-alpha null hypothesis is rejected on this combined set of funds, we claim that the funds in $\mathcal{S}^+$ are skilled. From an economic perspective, the skilled funds identified in this way are able to produce significantly positive alphas as they are benchmarked against the zero-alpha funds. Similarly, we could also confirm whether those funds in $\mathcal{S}^-$ are unskilled by repeating the test for funds in $\mathcal{S}^0 \cup \mathcal{S}^-$.

In the Appendix, we assess via simulation studies the accuracy of this sequential testing procedure in selecting the truly skilled funds with and without cross-sectional dependence (using HT and $\widetilde{HT}$). Figure 4.A.1 shows that the procedure is quite accurate for realistic levels of alphas when the true proportion of skilled funds is 2%, but it is downwardly biased when the true proportion of skilled funds is 5%, assuming that 20% of funds have a negative

[12] Alternatively, we can conduct the test on $\overline{\mathcal{S}}^0 \cup \mathcal{S}^+$, where $\overline{\mathcal{S}}^0$ is the subset of $\mathcal{S}^0$ excluding the $k^* - 1$ funds with the smallest p-values from $\mathcal{S}^0$, so the number of funds in $\overline{\mathcal{S}}^0 \cup \mathcal{S}^+$ is kept the same as that in $\mathcal{S}^0$. In empirical applications, the results remain very similar.

alpha of -0.30. In general, the procedure may conservatively estimate the proportion of skilled funds as it applies the most stringent (i.e., the smallest) threshold. This is not overly concerning if the goal is to test for the existence of skilled funds or to select the most profiting funds for investors.

What attributes make our methodology advantageous over that of Kosowski et al. (2006)? First, our bootstrap test is developed to overcome the statistical challenges plaguing the KTWW method. Our theories demonstrate that the bootstrap approach can correct the bias in the Hotelling's T-squared test statistic in a large cross-section with small sample sizes. The bootstrap Hotelling's T-squared test is further generalized to allow for serially correlated errors and is shown to be robust to cross-sectional dependence. Our simulation studies confirm good size and power properties of the proposed bootstrap test in settings calibrated to the characteristics of actual mutual fund data. Moreover, our approach has a high power of uncovering superior fund performance through a screening technique. The KTWW test could have a considerably reduced power of identifying skilled funds by taking the entire sample of funds altogether as the test set, which contains a large number of zero- and negative-alpha funds. This makes it difficult for the truly superior funds to stand out among a much larger crowd of mediocre and low-performing funds. In contrast, our data-driven thresholding approach screens out zero-alpha funds and excludes the large set of negative-alpha funds in the search for positive-alpha funds, thus reducing the dimension of the test set substantially and giving our test procedure increased power to discover skilled funds in a large fund population. In a study related to ours, Grønborg et al. (2021) eliminate funds with predicted inferior performance, but they do this through pairwise fund comparisons. In contexts different from ours, Hansen (2005) and Giglio, Liao, and Xiu (2020) similarly propose to increase the test power by removing poor and irrelevant alternatives.

4.4 Empirical Applications

In this empirical section, applying our proposed method, we first evaluate mutual fund performance by examining whether and how many skilled funds exist, and then examine the characteristics of funds with different levels of performance. We provide descriptive evidence on what kind of funds are likely to be skilled and unskilled. Our empirical analysis is based on monthly returns from January 1980 to December 2018 for all U.S. actively-managed equity funds. The results are obtained using the baseline bootstrap Hotelling's T-squared test. Section 4.C.1 in Appendix provides further details about the data we use. Section 4.C.2 presents the empirical results based on the generalized version of the bootstrap Hotelling's T-squared test, which accounts for serial correlations in fund returns.

4.4.1 Mutual Fund Performance

We use the Carhart (1997) four-factor model as the benchmark model to estimate fund alphas based on fund net returns as follows:[13]

$$r_{i,t} - r_{f,t} = \alpha_i + \sum_{j=1}^{4} \beta_{ij} X_{j,t} + \varepsilon_{i,t}.$$

The factors include CRSP value-weighted excess market return (Mktrf), size (SMB), book-to-market (HML), and momentum (UMD) factors. We require a minimum of 60 monthly observations in our estimation similar to Barras et al. (2010). The factors are obtained from Ken French's website.[14]

We first identify the funds with zero alpha. We obtain from fund-by-fund regressions the individual alpha, alpha t-statistic, and p-value of classical t-test for the i-th fund. For each p-value threshold between 0 and 0.1, we apply the bootstrap Hotelling's T-squared test to the set of funds with p-values larger than or equal to the threshold, i.e., the set $\mathcal{S}_k^0$ defined

[13] Results for the Jensen (1968) one-factor model and Fama and French (1996) three-factor model are very similar, and they are available upon request from the authors.

[14] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

in Section 4.3.3. In Panel A of Figure 4.2 , we plot the p-values of bootstrap Hotelling's T-squared test against the range of thresholds. This figure illustrates that there exist a range of thresholds for which the null is not rejected at the 10% level.

Having identified the zero-alpha funds, we implement the thresholding procedure by detecting skilled funds on a reduced set of funds without the extreme negative-alpha funds. Figure 4.2 Panel B presents the p-values of the bootstrap Hotelling's T-squared test on the set $\mathcal{S}^0 \cup \mathcal{S}^+$ for the range of thresholds with which the null is not rejected on $\mathcal{S}^0$. The bootstrap test rejects the null on the combined set for all values of thresholds. This provides strong evidence of the existence of superior funds in the data sample. Applying the smallest p-value threshold 0.0405, our sequential testing procedure uncovers 1.36% (36 out of 2650) of funds in the sample as skilled funds (with significantly positive alphas).

[Figure 4.2 about here.]

[Table 4.1 about here.]

While our test procedure concludes that some managers stand out with superior performance, extant bootstrap methods lead us to an opposite conclusion. In Panel B of Table 4.1, based on the Carhart model, the bootstrap tests in Kosowski et al. (2006) and Fama and French (2010) strongly reject the null at all upper percentile points and find no evidence for skilled funds in the population. From our prior analysis, the lack of evidence of skilled funds from these two tests may very well be an artifact of their inadequate power properties. The superior funds identified by our test, a non-negligible minority, exhibit skills to more than overcome their costs: they generate an average four-factor alpha of 0.336% per month (with an average alpha t-statistic of 2.598). This illustrates the power of our procedure and the importance of developing a valid bootstrap test that accounts for the empirical properties of mutual fund returns. For comparison, we further implement another influential approach for evaluating mutual fund performance, i.e., the false discovery rate (FDR) control method in Barras et al. (2010), which has been found to be biased in estimating the proportions of funds, especially when funds have small sample sizes (Andrikogiannopoulou and Papakonstantinou, 2019; Barras et al., 2020). Panel C shows that the FDR approach fails to identify any skilled funds.

As a robustness check, we also repeat the analyses using the generalized test statistic $\widetilde{HT}$, which takes into account the serial correlations in fund returns. Figure 4.A.2 in the Appendix shows that the result is very similar, with 1.06% of funds found to be skilled.

4.4.2 Characteristics of Different Fund Groups

For the zero-alpha, skilled, and unskilled funds identified by the sequential testing procedure, we examine and report the characteristics associated with each group in Table 4.2. Table 4.A.8 in the Appendix reports qualitatively similar results based on the $\widetilde{HT}$ statistic. The characteristics examined in the table include holding characteristics, fund characteristics, and performance/active management measures. The 93 stock holding characteristics and active weight are from March 1980 to February 2018, because portfolio holdings data from Thomson Reuters begin at the end of the first quarter in 1980. Furthermore, as “fdate” and “rdate” from MFLINK2 (as of April 2020) end in December 2017, we assume constant holding for three months after December 2017 (inclusive). We form fund portfolios with significant positive alpha, zero alpha, and significant negative alpha, respectively, based on our test results. We then report the time-series averages of the monthly cross-sectional means in each portfolio and the difference in means between the two extreme portfolios. Since mutual funds are required to disclose their holding every six months before May 2004 and every three months afterward, we compute t-statistics of the differences in means with Newey and West (1987) correction for time-series correlation with six lags.

[Table 4.2 about here.]

To begin with, we observe from this table that the stocks held by skilled funds and unskilled funds are dramatically different. This observation indicates that equity funds may achieve alpha and exhibit skills through holding stocks with certain characteristics, even though the Carhart benchmark model has explicitly taken them into account. Based on the *t*-statistics, we could see that for the 93 holding characteristics in Green et al. (2017), only 20 differences are insignificant at the 5% level, and the rest of the stock characteristics are significantly different for skilled funds and unskilled funds. For example, funds who exhibit significant positive alphas hold stocks with a higher bid-ask spread (*baspread*), higher standard deviation of earnings per share forecast (*disp*), higher idiosyncratic volatility (*idiovol*), higher Amihud ratio (*ill*), higher return volatility (*retvol*), higher volatility of liquidity based on both dollar trading volume (*std_dolvol*) and share turnover (*std_turn*), and a larger number of zero trading days (*zerotrade*).[15] Also, skilled funds hold smaller stocks based on market capitalization (*mve*) and industrial adjusted market capitalization(*mve_ia*).

Hou et al. (2015) classify all anomalies into six categories, including "Momentum", "Value-Versus-Growth", "Investment", "Profitability", "Intangibles", and "Trading Frictions". The above-mentioned anomalies are all in the category of "Trading Frictions". Furthermore, the asset pricing literature, such as Amihud and Mendelson (1989), Diether et al. (2002), Ali et al. (2003), Amihud (2002), Ang et al. (2006), Chordia et al. (2001), and Liu (2006), finds that stocks with those characteristics have higher anomalous returns. Since we also find that skilled funds have a lower fund turnover ratio compared to unskilled funds (see the variable *turn_ratio*), skilled managers may hold illiquid stocks purposely to extract illiquidity premium and at the same time prevent incurring trading costs. Besides, skilled funds also hold stocks with higher R&D expense to market capitalization (*rd_mve*) and R&D expense to sales (*rd_sale*). Guo et al. (2006) and Li (2011) suggest a higher expected stock return for stocks with those characteristics, which implies that part of the performance of skilled funds may be from the premium of holding intangible assets. For anomalies in the other categories, the results are either inconsistent or less significant.

From the perspective of fund characteristics, positive-alpha funds are younger in a big

[15] In Tables 4.2 and 4.A.8, we use the same variable abbreviations as in Table 1 of Green et al. (2017) for fund-level stock holding characteristics.

fund family, tend to have a lower expense ratio (about 1% a year), attract more inflow (1.7% a month), and as mentioned above have a lower turnover ratio (about 74% a year). Funds with positive alphas are also more engaged in active management since they have lower R-squared statistic (rsq), higher active share ($active\_share$) and active weight (aw). These findings are in line with the previous literature. For example, Amihud and Goyenko (2013), Cremers and Petajisto (2009), and Doshi et al. (2015) document that the performance of mutual funds can benefit from active management. In these studies, the measures of active management are "$1-R$-square", active share, and active weight. Kacperczyk et al. (2008) note that funds good at interim trading have better performance. Consistent with this study, we find that the funds in the skilled portfolio on average have a higher return gap ($retgap$) than funds in the unskilled portfolio. Finally, the hypothetical excess return based on stock holdings in skilled funds (1.13% per month) is much larger than that of unskilled funds (0.66% per month). It again suggests that part of fund skills may come from the higher expected return of stocks they hold. It is important to note that we are not trying to establish a causal relationship between fund characteristics and fund performance. Instead, we provide descriptive statistics for holding characteristics, fund characteristics, and performance measures based on the skilled and unskilled fund groups.

4.5 Conclusion

Finance researchers routinely evaluate the performance of thousands of mutual funds with short sample periods of monthly returns. Moreover, the vast majority of funds exhibit either zero or negative alphas. Our study shows that these stylized facts pose great challenges to the statistical validity of some existing performance evaluation methods. Originally proposed to separate luck from skill in mutual funds, the unconventional test statistic in Kosowski et al. (2006) fails to eliminate the higher order of approximation errors in using t-statistics. Although a well-intentioned proposal to deal with cross-sectional dependence, the joint resampling in Fama and French (2010) is not immune to the bias stemming from accumulated approximation errors under cross-sectional independence. Consequently, the Kosowski et al. (2006) test features size distortions in large cross-sections and may lack the power to detect skilled funds amongst a large number of funds with negative alphas. To the extent that fund sample sizes are relatively large, the Fama and French (2010) test still tends to have very low power, consistent with its heavy undersizing.

We develop the theory for adopting Hotelling's T-squared statistic for the zero-alpha test both in the ideal setting of independent errors and in the practical setting of serially correlated and cross-sectionally dependent fund residuals. To account for small sample sizes in large cross-sections, a residual-based bootstrap method is used to correct the bias in Hotelling's T-squared test and ensure accurate size control. When the zero-alpha null hypothesis is rejected, we provide an entirely data-driven sequential testing and thresholding procedure to evaluate mutual fund performance with enhanced power. Empirical analysis reveals modest evidence of superior performance. Skilled funds tend to be engaged in active management in big fund families and attract more inflows. They hold illiquid stocks and stocks with higher R&D expenses, both of which indicate higher expected stock returns and may contribute to the superior performance of those funds.

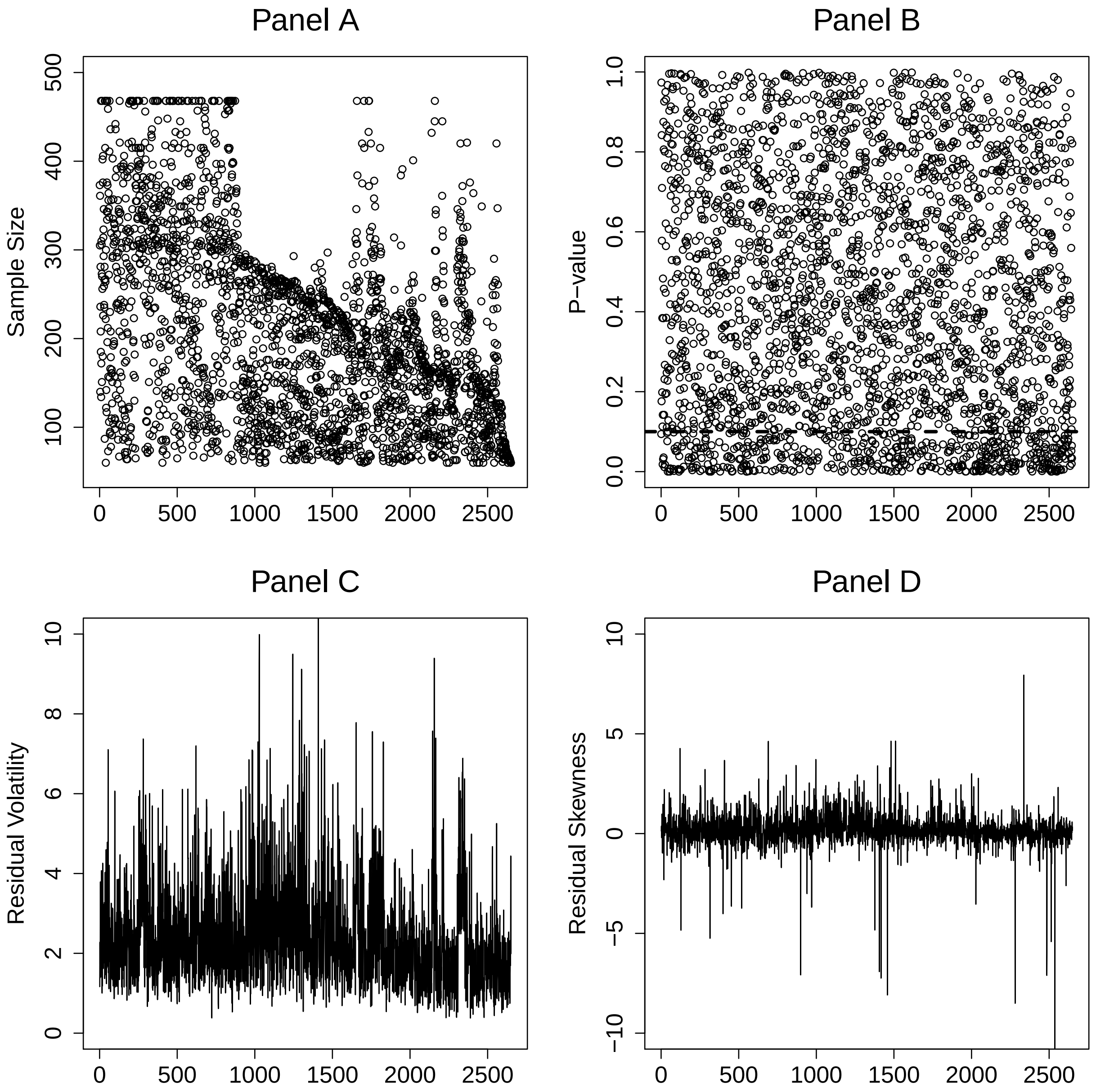


Figure 4.1. The cross section of monthly mutual fund returns.
The figure presents several statistical features of the cross section of monthly mutual fund returns from January 1980 to December 2018 for all U.S. actively-managed equity funds with at least 60 valid observations. Panel A shows the sample size of each mutual fund; Using the Carhart (1997) four-factor model, Panel B shows the p-value from standard t-test of zero alpha of each fund, and Panels C and D display the residual volatility and residual skewness for each fund, respectively.

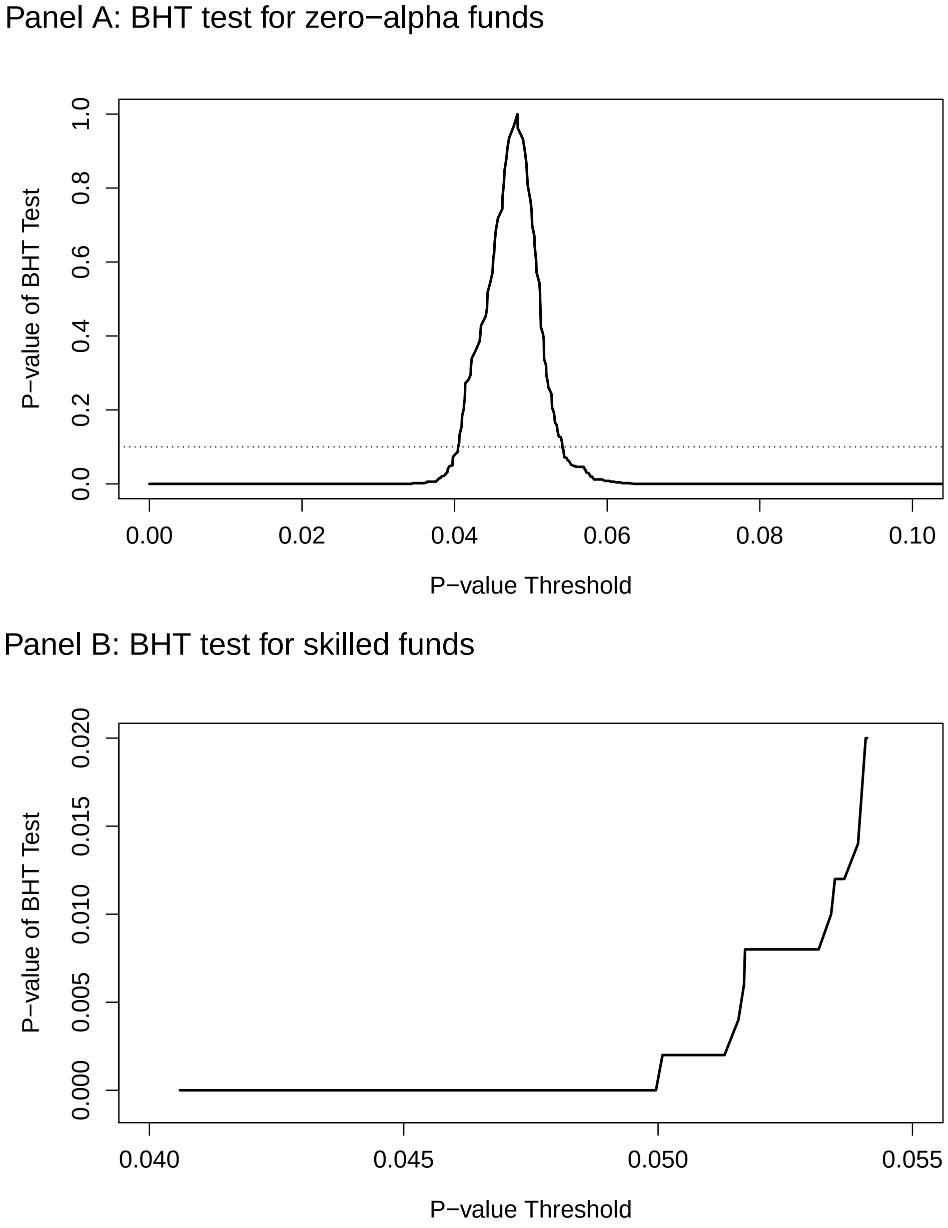


Figure 4.2. Bootstrap Hotelling's T-squared test for fund selection – *HT* test. This figure plots the p-values for the bootstrap Hotelling's T-squared (BHT) test for a range of p-value thresholds in the sequential fund selection procedure. All funds residuals are assumed to be serially uncorrelated. Panel A shows the BHT test for a zero-alpha fund set, and Panel B shows the test for confirming a skilled fund set relative to the zero-alpha fund set. The data sample is monthly returns from January 1980 to December 2018 for all U.S. actively-managed equity funds with at least 60 return observations.

Table 4.1. Summary statistics of the cross section of mutual fund returns and alternative performance evaluation tests

From January 1980 to December 2018, for all U.S. actively-managed equity funds, this table presents the statistical characteristics of mutual fund returns as well as alternative performance evaluation tests based on fund-by-fund regressions with the Carhart (1997) model. Panel A summarizes several key statistics for the mutual fund data. Panel B reports the cross-sectionally bootstrapped *p*-values at several percentile points from the bootstrap tests in Kosowski et al. (2006) (KTWW) and Fama and French (2010) (FF) based on the Newey and West (1987) heteroskedasticity- and autocorrelation-consistent standard errors. Note that the results are very similar without the Newey and West (1987) adjustment. Panel C reports the estimated fund proportions based on the FDR approach in Barras et al. (2010). We follow the internet appendix of Barras et al. (2010) to implement their procedure.

	Percentile										Median	Mean	SD
	0.01	0.03	0.05	0.10	0.15	0.85	0.90	0.95	0.97	0.99			
Panel A: Summary Statistics													
Sample size	62	66	70	82	92	314	355	414	454	468	186	204	103
Alpha	-0.690	-0.517	-0.426	-0.316	-0.254	0.081	0.122	0.196	0.244	0.387	-0.798	-0.916	0.205
Alpha *t*-statistic	-3.502	-2.989	-2.689	-2.252	-1.946	0.640	0.962	1.367	1.650	2.236	-0.700	-0.667	1.248
Residual volatility	0.477	0.647	0.760	0.918	1.019	2.668	3.082	3.811	4.465	5.479	1.613	1.851	1.003
Residual skewness	-2.167	-0.897	-0.662	-0.449	-0.345	0.444	0.566	0.779	0.948	1.442	0.051	0.018	0.775
Panel B: Bootstrap Test *p*-Values													
KTWW test	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	–	–	–
FF test	0.000	0.000	0.000	0.000	0.000	0.956	0.937	0.918	0.915	0.829	–	–	–
Panel C: Estimated Fund Proportions with the FDR Approach													
			Unskilled			Zero-Alpha			Skilled				
			30.34%			69.66%			0.00%				

Table 4.2. Mutual fund characteristics based on *HT* test

From January 1980 to December 2018, for all U.S. actively-managed equity funds with at least 60 valid observations, we compute alphas using the four-factor model and generate mutual fund portfolios with significantly positive alphas, zero alpha, and negative alphas, respectively, based on the *HT* test statistic. We report the time-series averages of the monthly cross-sectional means in each portfolio and the difference in means between the two extreme portfolios. We compute *t*-statistics of the differences with Newey and West (1987) correction for time-series correlation with 6 lags. The variables include fund level stock holding characteristics (using the same variable abbreviations as in table one of Green et al. (2017)), fund characteristics, and fund performance/active management measures. For ease of reading, *ear* and *sue* are scaled by 100, Amihud ratio by 1000000, and *mve_ia* by 1/1000. We take log for the total net asset ($ million), for the age of the fund's oldest share class (in years), and for the family total net asset ($ million). Annual turnover and expense ratio (both in percentage point) are the value weighted averages across all fund share classes. Fund flow (%) is the average monthly net growth in fund assets beyond reinvested dividends and portfolio returns. Return gap is in percentage point. Active weight is scaled by 2. The hypothetical excess returns are in percentage. The variables are defined in Section 4.C.1 of the Appendix. Statistical significance of 1, 5, and 10 percent are indicated by ***, **, and *, respectively.

Variable	Neg. Alpha	Zero Alpha	Pos. Alpha	Pos.−Neg.	*t*-stat
(Number of funds)	(352)	(2,262)	(36)		
Holding characteristics					
absacc	0.0660	0.0694	0.0754	0.009***	(11.24)
acc	-0.0213	-0.0214	-0.0209	0.000	(0.27)
aeavol	0.6368	0.6794	0.7685	0.132***	(12.84)
age	17.0742	16.0295	14.9119	-2.162***	(-22.34)
agr	0.1757	0.1957	0.2255	0.050***	(8.44)
baspread	0.0266	0.0277	0.0293	0.003***	(8.04)
beta	1.0291	1.0660	1.1521	0.123***	(7.82)
bm	0.5000	0.4811	0.3925	-0.107***	(-8.01)
bm_ia	32.3681	32.278	50.9588	18.591*	(1.91)
cash	0.1257	0.1352	0.1741	0.048***	(14.67)
cashdebt	0.2537	0.2667	0.3282	0.075***	(10.47)
cashpr	5.3980	6.7956	13.7102	8.312***	(9.99)
cfp	0.0752	0.0731	0.0648	-0.010***	(-3.47)
cfp_ia	15.6601	15.6715	23.1901	7.530**	(2.00)
chatoia	-0.0072	-0.0084	-0.0125	-0.005***	(-3.04)
chcsho	0.1639	0.1689	0.1953	0.031***	(4.33)
chempia	-0.1095	-0.1054	-0.1021	0.007	(1.18)
chfeps	0.0208	0.0202	0.0176	-0.003	(-1.41)
chinv	0.0111	0.0126	0.0153	0.004***	(5.29)
chmom	-0.0106	-0.0049	0.0047	0.015***	(2.84)
chpmia	0.2425	0.2374	0.2480	0.006	(0.04)
chtx	0.0019	0.0022	0.0030	0.001***	(5.00)
cinvest	-0.0020	-0.0027	-0.0065	-0.004***	(-2.61)
convind	0.1782	0.1814	0.1870	0.009	(1.43)
currat	2.4325	2.5461	2.7654	0.333***	(13.49)
depr	0.2125	0.221	0.2456	0.033***	(14.58)
disp	0.0642	0.0737	0.0774	0.013***	(7.60)
divi	0.0163	0.0195	0.0251	0.009***	(7.76)
divo	0.0115	0.0147	0.0177	0.006***	(8.22)

Table 4.2 (cont'd): Mutual fund characteristics based on *HT* test

Variable (Number of funds)	Neg. Alpha (352)	Zero Alpha (2,262)	Pos. Alpha (36)	Pos.−Neg.	*t*-stat
Holding characteristics					
dy	0.0222	0.0197	0.0120	-0.010***	(-10.50)
ear	0.7334	0.7985	0.8632	0.130**	(2.36)
egr	0.1852	0.2060	0.2365	0.051***	(6.96)
ep	0.0556	0.0507	0.0423	-0.013***	(-11.52)
fgr5yr	14.2460	15.2047	16.4715	2.226***	(9.94)
gma	0.3892	0.4122	0.5000	0.111***	(11.99)
grcapx	0.6035	0.6851	0.7825	0.179***	(9.11)
grltnoa	0.0971	0.1033	0.1104	0.013***	(10.26)
herf	0.0770	0.0782	0.0783	0.001	(1.08)
hire	0.0874	0.1045	0.1299	0.042***	(9.41)
idiovol	0.0406	0.0425	0.0456	0.005***	(11.27)
ill	0.0279	0.0439	0.0473	0.019***	(4.13)
indmom	0.1641	0.1639	0.1662	0.002	(0.44)
invest	0.0794	0.0879	0.0979	0.018***	(6.51)
ipo	0.0174	0.0247	0.0307	0.013***	(7.91)
lev	2.0203	1.7073	0.9081	-1.112***	(-16.27)
mom12m	0.2392	0.2564	0.2770	0.038***	(3.50)
mom1m	0.0175	0.0197	0.0235	0.006***	(7.72)
mom36m	0.5482	0.5746	0.6222	0.074***	(3.89)
ms	4.7309	4.7313	5.1134	0.382***	(13.64)
mve	15.3905	15.1364	14.9424	-0.448***	(-22.19)
mve_ia	12.8047	10.2769	8.6132	-4.192***	(-9.56)
nanalyst	18.4106	16.9041	17.0367	-1.374***	(-11.35)
nincr	1.2067	1.2390	1.3644	0.158***	(4.76)
operprof	0.9051	0.9189	0.9948	0.090***	(6.05)
orgcap	0.0082	0.0084	0.0095	0.001***	(9.51)
pchcapx_ia	7.7086	7.1562	10.1049	2.396	(1.34)
pchcurrat	0.0329	0.0369	0.0350	0.002	(0.70)
pchdepr	0.0416	0.0447	0.0519	0.010***	(4.51)
pchgm_pchsale	0.0083	0.0023	-0.0073	-0.016***	(-4.59)
pchsale_pchinvt	-0.0338	-0.0316	-0.0159	0.018***	(2.68)
pchsale_pchrect	-0.0310	-0.0333	-0.0314	-0.000	(-0.13)
pchsale_pchxsga	0.0057	0.0087	0.0073	0.002	(0.70)
pchsaleinv	0.0963	0.0913	0.0817	-0.015*	(-1.67)
pctacc	-0.8311	-0.9211	-0.9101	-0.079*	(-1.87)

Table 4.2 (cont'd): Mutual fund characteristics based on *HT* test

Variable	Neg. Alpha	Zero Alpha	Pos. Alpha	Pos.−Neg.	*t*-stat
(Number of funds)	(352)	(2,262)	(36)		
Holding characteristics					
pricedelay	0.1030	0.0953	0.0906	-0.012***	(-3.57)
ps	4.8307	4.8154	4.8241	-0.007	(-0.39)
rd	0.0955	0.0952	0.0999	0.004	(1.41)
rd_mve	0.0342	0.0344	0.0395	0.005***	(6.23)
rd_sale	0.1148	0.1197	0.1914	0.077***	(7.33)
realestate	0.3059	0.3009	0.3020	-0.004**	(-2.14)
retvol	0.0202	0.0210	0.0225	0.002***	(9.24)
roaq	0.0172	0.0174	0.0205	0.003***	(5.81)
roavol	0.0128	0.0141	0.0163	0.003***	(14.97)
roeq	0.0419	0.0403	0.0409	-0.001	(-0.96)
roic	0.1157	0.1173	0.1326	0.017***	(4.23)
rsup	0.0235	0.0248	0.0235	-0.000	(-0.02)
salecash	42.7258	43.8062	35.4452	-7.281***	(-6.05)
saleinv	28.4203	28.2467	27.2056	-1.215**	(-2.44)
salerec	10.8051	11.5833	10.9149	0.110	(0.75)
secured	0.2129	0.2552	0.2965	0.083***	(9.45)
securedind	0.3814	0.3997	0.4200	0.039***	(6.22)
sfe	0.0556	0.0495	0.0418	-0.014***	(-9.54)
sgr	0.1568	0.1759	0.2077	0.051***	(8.76)
sin	0.0146	0.0145	0.0100	-0.005***	(-8.71)
sp	1.2442	1.2233	0.9085	-0.336***	(-7.17)
std_dolvol	0.4971	0.5242	0.5492	0.052***	(13.41)
std_turn	3.4764	3.8437	4.3445	0.868***	(10.26)
stdcf	1.6649	2.0204	3.1414	1.477***	(4.70)
sue	0.0287	0.0318	0.0287	0.000	(0.00)
tang	0.4885	0.4931	0.5035	0.015***	(6.59)
tb	0.1218	0.1267	0.1681	0.046***	(2.71)
turn	1.3246	1.4065	1.5439	0.219***	(7.70)
zerotrade	0.0210	0.0405	0.0373	0.016***	(4.48)
Fund characteristics					
logtna	5.0724	5.5085	6.7391	1.667***	(13.57)
logage	2.4772	2.4586	2.3415	-0.136**	(-2.02)
logtna_family	7.8417	8.5416	10.4485	2.607***	(19.02)
turn_ratio	79.1506	78.9171	74.1319	-5.019	(-1.31)
flow_pct	0.4425	0.8094	1.7110	1.269***	(5.87)
exp_ratio	1.1863	1.1255	1.0088	-0.178***	(-11.15)

Table 4.2 (cont'd): Mutual fund characteristics based on *HT* test

Variable	Neg. Alpha	Zero Alpha	Pos. Alpha	Pos.−Neg.	*t*-stat
(Number of funds)	(352)	(2,262)	(36)		
Fund characteristics					
logtna	5.0724	5.5085	6.7391	1.667***	(13.57)
logage	2.4772	2.4586	2.3415	-0.136**	(-2.02)
logtna_family	7.8417	8.5416	10.4485	2.607***	(19.02)
turn_ratio	79.1506	78.9171	74.1319	-5.019	(-1.31)
flow_pct	0.4425	0.8094	1.7110	1.269***	(5.87)
exp_ratio	1.1863	1.1255	1.0088	-0.178***	(-11.15)
Performance/Active management measures					
rsq	0.9156	0.8752	0.8697	-0.046***	(-9.63)
idiovolm	0.0130	0.0167	0.0185	0.005***	(13.89)
retgap	-0.0902	-0.0331	0.0090	0.099***	(4.43)
active_share	0.7875	0.8471	0.8982	0.111***	(26.24)
aw	0.8387	0.8460	0.9171	0.078***	(11.90)
hrex	0.6550	0.8107	1.1255	0.471***	(6.40)

Appendix

This Appendix contains theoretical derivations, simulation results, and empirical results that supplement those in the paper. Section 4.A collects the proofs of all theorems. Section 4.B provides a simulation study to confirm the problems of the extant bootstrap methods. It also demonstrates the good finite-sample size and power properties of the new test and the accuracy of the sequential fund selection procedure. Section 4.C provides a description of the data and presents additional results for the empirical applications in the paper.

Appendix 4.A Theoretical Derivations

Throughout, Theorem 4.* and equation (4.*) refer to the corresponding ones in the main paper.

4.A.1 Proofs of Theorems 4.1–4.3

Proof of Theorem 4.1. Define

$$\mu_{i,0} = P(\hat{t}_i^b(0) \le \hat{t}_{([pN])}(0) | \{Y_{i,t}, \boldsymbol{X}_t\}), \quad \mu_0 = \frac{1}{N}\sum_{i=1}^{N}\mu_{i,0}, \quad \sigma_0^2 = \frac{1}{N}\sum_{i=1}^{N}\mu_{i,0}(1-\mu_{i,0}),$$

and write

$$\begin{aligned} & P\left(\hat{t}^b_{([pN])}(0) \le \hat{t}_{([pN])}(0) | \{Y_{i,t}, \boldsymbol{X}_t\}\right) \\ &= P\left(\sum_{i=1}^{N} I\left(\hat{t}_i^b(0) \le \hat{t}_{([pN])}(0)\right) \ge [pN] | \{Y_{i,t}, \boldsymbol{X}_t\}\right) \\ &= P\left(\sqrt{N}\frac{N^{-1}\sum_{i=1}^{N}\left(I(\hat{t}_i^b(0) \le \hat{t}_{([pN])}(0)) - \mu_{i,0}\right)}{\sigma_0} \ge \sqrt{N}\frac{[pN]/N - \mu_0}{\sigma_0} | \{Y_{i,t}, \boldsymbol{X}_t\}\right). \end{aligned}$$

Throughout, $\{Y_{i,t}, \boldsymbol{X}_t\}$ denotes the set $\{Y_{i,t}, \boldsymbol{X}_t : t = t_i + 1, \dots, t_i + T_i, i = 1, \dots, N\}$. Conditional on $\{Y_{i,t}, \boldsymbol{X}_t\}$, $\{I(\hat{t}_i^b(0) \le \hat{t}_{([pN])}(0)) - \mu_{i,0}\}_{i=1}^N$ is a sequence of independent but nonidentically distributed bounded random variables with zero means. Hence, it follows from the Central Limit Theorem for martingale differences (see Theorem 3.2 of Hall and Heyde, 1980) that, as $N \to \infty$,

$$P\left(\hat{t}^b_{([pN])}(0) \le \hat{t}_{([pN])}(0) | \{Y_{i,t}, \boldsymbol{X}_t\}\right) = 1 - \Phi\left(\sqrt{N}\frac{[pN]/N - \mu_0}{\sigma_0}\right) + o_P(1). \tag{4.A.1}$$

Similarly, conditional on $\{\boldsymbol{X}_t\}$,

$$\begin{aligned}
&P\big(\sqrt{N}(\hat{t}_{([pN])}(0) - Q_p) \le x|\{\boldsymbol{X}_t\}\big)\\
=&P\Bigg(\sum_{i=1}^{N} I\big(\hat{t}_i(0) \le Q_p + \frac{x}{\sqrt{N}}\big) \ge [pN]|\{\boldsymbol{X}_t\}\Bigg)\\
=&P\Bigg(\sqrt{N}\frac{N^{-1}\sum_{i=1}^{N}\big(I(\hat{t}_i(0) \le Q_p + \frac{x}{\sqrt{N}}) - P(\hat{t}_i(0) \le Q_p + \frac{x}{\sqrt{N}}|\{\boldsymbol{X}_t\})\big)}{\sqrt{p(1-p)}}\\
&\qquad \ge \sqrt{N}\frac{N^{-1}([pN] - \sum_{i=1}^{N} P(\hat{t}_i(0) \le Q_p + \frac{x}{\sqrt{N}}|\{\boldsymbol{X}_t\}))}{\sqrt{p(1-p)}}|\{\boldsymbol{X}_t\}\Bigg).
\end{aligned}$$

Using (4.5) and the Central Limit Theorem for martingale differences as before, we have

$$\begin{aligned}
&P\big(\sqrt{N}(\hat{t}_{([pN])}(0) - Q_p) \le x|\{\boldsymbol{X}_t\}\big)\\
=&P\Bigg(\sqrt{N}\frac{N^{-1}\sum_{i=1}^{N}\big(I(\hat{t}_i(0) \le Q_p + \frac{x}{\sqrt{N}}) - P(\hat{t}_i(0) \le Q_p + \frac{x}{\sqrt{N}}|\{\boldsymbol{X}_t\})\big)}{\sqrt{p(1-p)}}\\
&\ge \frac{1}{\sqrt{N}}\frac{[pN] - N\Phi(Q_p + \frac{x}{\sqrt{N}}) - \phi(Q_p)\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p) - \phi(Q_p)\sum_{i=1}^{N} T_i^{-1} q_{i,2,\boldsymbol{x}}(Q_p)}{\sqrt{p(1-p)}}|\{\boldsymbol{X}_t\}\Bigg)\\
=&1 - \Phi\Bigg(\frac{-\phi(Q_p)x - \phi(Q_p)\frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p) - \phi(Q_p)\frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1} q_{i,2,\boldsymbol{x}}(Q_p)}{\sqrt{p(1-p)}}\Bigg) + o_P(1)\\
=&\Phi\Bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}\{x + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1} q_{i,2,\boldsymbol{x}}(Q_p)\}\Bigg) + o_P(1),
\end{aligned}$$

implying that

$$P\Bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}} A_{N1} \le x|\{\boldsymbol{X}_t\}\Bigg) = \Phi(x) + o_P(1) \tag{4.A.2}$$

and

$$\hat{t}_{([pN])}(0) - Q_p = -\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\{1 + o_p(1)\} + O_P(N^{-1/2}). \tag{4.A.3}$$

By (4.6), (4.A.3), and Taylor expansion, we have

$$\begin{aligned}
\mu_{i,0} &= P\big(\hat{t}_i^b(0) \leq \hat{t}_{([pN])}(0) | \{Y_{i,t}, \boldsymbol{X}_t\}\big) \\
&= \Phi(\hat{t}_{([pN])}(0)) + T_i^{-1/2}\phi(\hat{t}_{([pN])}(0))\hat{q}_{i,1,\boldsymbol{x}}(\hat{t}_{([pN])}(0)) + T_i^{-1}\phi(\hat{t}_{([pN])}(0))\hat{q}_{i,2,\boldsymbol{x}}(\hat{t}_{([pN])}(0)) + O_P(T_i^{-3/2}) \\
&= p + \phi(Q_p)(\hat{t}_{([pN])}(0) - Q_p) - \frac{1}{2}Q_p\phi(Q_p)(\hat{t}_{([pN])}(0) - Q_p)^2 + o_P\big(\{\frac{1}{N}\sum_{i=1}^N T_i^{-1/2}\}^2\big) + o_P(N^{-1}) \\
&\quad + T_i^{-1/2}\phi(Q_p)\hat{q}_{i,1,\boldsymbol{x}}(Q_p) + T_i^{-1/2}\phi(Q_p)(\hat{q}'_{i,1,\boldsymbol{x}}(Q_p) - Q_p\hat{q}_{i,1,\boldsymbol{x}}(Q_p))(\hat{t}_{([pN])}(0) - Q_p) \\
&\quad + T_i^{-1}\phi(Q_p)\hat{q}_{i,2,\boldsymbol{x}}(Q_p) + o_P(T_i^{-1}) \\
&= p + \phi(Q_p)(\hat{t}_{([pN])}(0) - Q_p) - \frac{1}{2}Q_p\phi(Q_p)\{\frac{1}{N}\sum_{i=1}^N T_i^{-1/2}q_{i,1,\boldsymbol{x}}(Q_p)\}^2 + \phi(Q_p)T_i^{-1/2}q_{i,1,\boldsymbol{x}}(Q_p) \\
&\quad + \phi(Q_p)T_i^{-1}\sqrt{T_i}(\hat{q}_{i,1,\boldsymbol{x}}(Q_p) - q_{i,1,\boldsymbol{x}}(Q_p)) + Q_p\phi(Q_p)T_i^{-1/2}q_{i,1,\boldsymbol{x}}(Q_p)\{\frac{1}{N}\sum_{i=1}^N T_i^{-1/2}q_{i,1,\boldsymbol{x}}(Q_p)\} \\
&\quad - \phi(Q_p)T_i^{-1/2}q'_{i,1,\boldsymbol{x}}(Q_p)\{\frac{1}{N}\sum_{i=1}^N T_i^{-1/2}q_{i,1,\boldsymbol{x}}(Q_p)\} + \phi(Q_p)T_i^{-1}q_{i,2,\boldsymbol{x}}(Q_p) \\
&\quad + o_P(T_i^{-1/2}\frac{1}{N}\sum_{i=1}^N T_i^{-1/2}) + o_P(T_i^{-1}) + o_P\big(\{\frac{1}{N}\sum_{i=1}^N T_i^{-1/2}\}^2\big) + o_P(N^{-1}), \qquad (4.A.4)
\end{aligned}$$

which implies that

$$\sigma_0^2 = p(1-p) + o_P(1). \qquad (4.A.5)$$

Plugging (4.A.4) and (4.A.5) into (4.A.1), we have

$$
\begin{aligned}
&P\big(\hat{t}^{b}_{([pN])}(0) \le \hat{t}_{([pN])}(0) | \{Y_{i,t}, \boldsymbol{X}_t\}\big) \\
=&\Phi\left(\sqrt{N}\frac{\mu_0 - [pN]/N}{\sigma_0}\right) + o_P(1) \\
=&\Phi\bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}\big[\sqrt{N}(\hat{t}_{([pN])}(0) - Q_p) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1} q_{i,2,\boldsymbol{x}}(Q_p) \\
&+ \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1}\sqrt{T_i}(\hat{q}_{i,1,\boldsymbol{x}}(Q_p) - q_{i,1,\boldsymbol{x}}(Q_p)) \\
&+ \frac{1}{2} Q_p \sqrt{N}\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\}^2 - \sqrt{N}\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2} q'_{i,1,\boldsymbol{x}}(Q_p)\}\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\}\big]\bigg) \\
&+ o_P(1) \\
=&\Phi\bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}(A_{N1} + A_{N2} + A_{N3})\bigg) + o_P(1).
\end{aligned}
\tag{4.A.6}
$$

Conditional on $\{Y_{i,t}, \boldsymbol{X}_t\}$, because $\{I(\hat{t}^{b}_{([pN])}(0) \le \hat{t}_{([pN])}(0))\}_{b=1}^{B}$ is a sequence of independent variables, it follows from the law of large numbers that

$$
\frac{1}{B}\sum_{b=1}^{B} I\{\hat{t}^{b}_{([pN])}(0) \le \hat{t}_{([pN])}(0)\} = P\big(\hat{t}^{b}_{([pN])}(0) \le \hat{t}_{([pN])}(0) | \{Y_{i,t}, \boldsymbol{X}_t\}\big) + o_P(1), \tag{4.A.7}
$$

as $B \to \infty$. Thus, it follows from (4.A.6) and (4.A.7) that

$$
\begin{aligned}
&P\bigg(\frac{1}{B}\sum_{b=1}^{B} I\{\hat{t}^{b}_{([pN])}(0) \le \hat{t}_{([pN])}(0)\} \le a | \{\boldsymbol{X}_t\}\bigg) \\
=&E\bigg[P\bigg(\frac{1}{B}\sum_{b=1}^{B} I\{\hat{t}^{b}_{([pN])}(0) \le \hat{t}_{([pN])}(0)\} \le a | \{Y_{i,t}, \boldsymbol{X}_t\}\bigg) | \{\boldsymbol{X}_t\}\bigg] \\
=&P\bigg(\Phi\bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}(A_{N1} + A_{N2} + A_{N3})\bigg) \le a | \{\boldsymbol{X}_t\}\bigg) + o_P(1),
\end{aligned}
\tag{4.A.8}
$$

as $N \to \infty$ and $B \to \infty$. This shows (4.7).

(4.8) follows from the fact that

$$
\begin{aligned}
&P\left(\hat{t}_{([pN])}(0) \in \left[\hat{t}^{([aB/2])}_{([pN])}(0),\ \hat{t}^{(B-[aB/2])}_{([pN])}(0)\right] | \{\boldsymbol{X}_t\}\right)\\
=&P\bigg([aB/2]/B \leq \frac{1}{B}\sum_{b=1}^{B} I\{\hat{t}^{b}_{([pN])}(0) \leq \hat{t}_{([pN])}(0)\} \leq 1-[aB/2]/B | \{\boldsymbol{X}_t\}\bigg)\\
=&P\bigg(a/2 \leq \frac{1}{B}\sum_{b=1}^{B} I\{\hat{t}^{b}_{([pN])}(0) \leq \hat{t}_{([pN])}(0)\} \leq 1-a/2 | \{\boldsymbol{X}_t\}\bigg)\\
=&P\bigg(a/2 \leq \Phi\bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}(A_{N1}+A_{N2}+A_{N3})\bigg) \leq 1-a/2 | \{\boldsymbol{X}_t\}\bigg) + o_P(1).
\end{aligned}
$$

When $\lim_{N\to\infty} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1} = 0$ by (4.9), we have

$$
\lim_{N\to\infty} \sqrt{N}\left\{\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2}\right\}^2 \leq \lim_{N\to\infty} \sqrt{N}\frac{1}{N}\sum_{i=1}^{N} T_i^{-1} = 0.
$$

This implies $A_{N3} = o_P(1)$ and $A_{N2} = O_P(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1}) = o_P(1)$ conditional on $\{\boldsymbol{X}_t\}$ provided that $\sup_{i\geq 1} E(\varepsilon_{i,t}^6) < \infty$. Hence, (4.10) and (4.11) hold from (4.A.2). The proof of Theorem 4.1 is complete. □

Proof of Theorem 4.2. Because $\min_{1\leq i\leq N} T_i \to \infty$ in condition (C3), we have

$$
\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2} = 0.
$$

Denote $A_N := \min\{|\log(\frac{1}{N}\sum_{i=1}^{N} T_i^{-1/2})|, N^{1/4}\}$. Then,

$$
A_N \to \infty, \quad \frac{A_N}{\sqrt{N}} \to 0, \quad \text{and} \quad \frac{A_N}{N}\sum_{i=1}^{N} T_i^{-1/2} \to 0 \quad \text{as } N\to\infty. \tag{4.A.9}
$$

Noting that $\hat{\delta}_i/\delta_i - 1 = O_P(T_i^{-1/2})$ and $\frac{1}{N}\sum_{i=1}^N |\delta_i|/\sqrt{T_i} \to 0$, we have

$$\begin{aligned}
&P\big(A_N(\hat{t}_{([pN])}(0) - Q_p) \leq x|\{\boldsymbol{X}_t\}\big)\\
=&P\big(\sum_{i=1}^N I\{\hat{t}_i(\alpha_i) \leq -\hat{\delta}_i + Q_p + x/A_N\} \geq [pN]|\{\boldsymbol{X}_t\}\big)\\
=&P\bigg(\frac{1}{N}\sum_{i=1}^N I\{\hat{t}_i(\alpha_i) \leq Q_p - \delta_i\} \geq p|\{\boldsymbol{X}_t\}\bigg) + o_P(1).
\end{aligned}$$

Conditional on $\{\boldsymbol{X}_t\}$, because $\{I(\hat{t}_i(\alpha_i) \leq Q_p - \delta_i)\}_{i=1}^N$ is a sequence of independent but non-identically distributed bounded random variables, it follows from the law of large numbers that

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N I\{\hat{t}_i(\alpha_i) \leq Q_p - \delta_i\}\\
=&\frac{1}{N}\sum_{i=1}^N\{\Phi(Q_p - \delta_i) - \Phi(Q_p)\} + \Phi(Q_p) + o_P(1)\\
=&-\delta_0 + p + o_P(1),
\end{aligned}$$

as $N \to \infty$. Hence, for any fixed $x \in \Re$,

$P\big(A_N(\hat{t}_{([pN])}(0) - Q_p) \leq x|\{\boldsymbol{X}_t\}\big) = I\{\delta_0 \leq 0\} + o_P(1)$,

i.e.,

$$A_N(\hat{t}_{([pN]}(0) - Q_p) \xrightarrow{p} \begin{cases} \infty, & \delta_0 > 0 \\ -\infty, & \delta_0 < 0 \end{cases} \tag{4.A.10}$$

as $N \to \infty$. Following from the proof of (4.7) of Theorem 4.1, using (4.A.9) and (4.A.10), we have

$$\begin{aligned}
&P\big(S(p) \le a|\{\boldsymbol{X}_t\}\big) \\
&= P\bigg(\Phi\bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}\big(\sqrt{N}(\hat{t}_{([pN])}(0) - Q_p) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\big)\bigg) \le a|\{\boldsymbol{X}_t\}\bigg) + o_P(1) \\
&= P\bigg(\Phi\bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}\frac{\sqrt{N}}{A_N}\big(A_N(\hat{t}_{([pN])}(0) - Q_p) + \frac{A_N}{N}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\big)\bigg) \le a|\{\boldsymbol{X}_t\}\bigg) + o_P(1) \\
&= I\big(\delta_0 < 0\big) + o_P(1).
\end{aligned}$$

Hence, Theorem 4.2 follows. □

Proof of Theorem 4.3. Using (4.10), condition (C6), and similar arguments in the proof of Theorem 4.1, we have

$$\begin{aligned}
&P\big(\hat{t}^b_{([pN])}(0) \le \hat{t}_{([pN])}(0)|\{Y_{i,t}, \boldsymbol{X}_t\}\big) \\
= \ &\Phi\bigg(\tfrac{\phi(Q_p)}{\sqrt{p(1-p)}}\big(\sqrt{N}(\hat{t}_{([pN])}(0) - Q_p) + \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\big)\bigg) + o_P(1)
\end{aligned} \tag{4.A.11}$$

and

$$\begin{aligned}
&P\big(\sum_{i=1}^{N} I(\hat{t}_i(\alpha_i) \le Q_p - \delta_i + \frac{x}{\sqrt{N}}) \ge [pN]|\{\boldsymbol{X}_t\}\big) \\
=&P\bigg(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\big\{I(\hat{t}_i(\alpha_i) \le Q_p - \delta_i + \frac{x}{\sqrt{N}}) - \Phi(Q_p - \delta_i + \frac{x}{\sqrt{N}}) - T_i^{-1/2}\phi(Q_p)q_{i,1,\boldsymbol{x}}(Q_p)\big\} \\
&\qquad \ge \frac{1}{\sqrt{N}}\big\{[pN] - \sum_{i=1}^{N}\Phi(Q_p - \delta_i + \frac{x}{\sqrt{N}}) - \phi(Q_p)\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\big\}|\{\boldsymbol{X}_t\}\bigg) \\
=&1 - \Phi\bigg(\frac{1}{\sqrt{p(1-p)N}}\big\{[pN] - \sum_{i=1}^{N}\Phi(Q_p - \delta_i + \frac{x}{\sqrt{N}}) - \phi(Q_p)\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\big\}\bigg) + o_P(1) \\
=&1 - \Phi\bigg(\frac{1}{\sqrt{p(1-p)}}\big\{-x\phi(Q_p) - \sqrt{N}\Delta_N(Q_p) - \phi(Q_p)\frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p)\big\}\bigg) + o_P(1) \\
=&\Phi\bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}\big\{x + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p) + \frac{\sqrt{N}\Delta_N(Q_p)}{\phi(Q_p)}\big\}\bigg) + o_P(1).
\end{aligned} \tag{4.A.12}$$

Because $\hat{\delta}_i/\delta_i - 1 = O_P(T_i^{-1/2})$ and $\frac{1}{\sqrt{N}}\sum_{i=1}^{n}|\delta_i|/\sqrt{T_i} \to 0$, we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\big\{I(\hat{t}_i(\alpha_i) \le Q_p - \hat{\delta}_i + x/\sqrt{N}) - I(\hat{t}_i(\alpha_i) \le Q_p - \delta_i + x/\sqrt{N})\big\} = o_P(1). \qquad (4.A.13)$$

It follows from (4.A.12) and (4.A.13) that

$$\begin{aligned}
&P\big(\sqrt{N}(\hat{t}_{([pN])}(0) - Q_p) \le x|\{\boldsymbol{X}_t\}\big)\\
=&P\big(\sum_{i=1}^{N} I(\hat{t}_i(\alpha_i) \le Q_p - \hat{\delta}_i + \frac{x}{\sqrt{N}}) \ge [pN]|\{\boldsymbol{X}_t\}\big)\\
=&\Phi\bigg(\frac{\phi(Q_p)}{\sqrt{p(1-p)}}\{x + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} T_i^{-1/2} q_{i,1,\boldsymbol{x}}(Q_p) + \frac{\sqrt{N}\Delta_N(Q_p)}{\phi(Q_p)}\}\bigg) + o_P(1), \qquad (4.A.14)
\end{aligned}$$

i.e., conditional on $\{\boldsymbol{X}_t\}$, B_{N1} converges in distribution to a standard normal random variable.

By (4.A.11) and (4.A.14), we have

$$\begin{aligned}
P(S(p) \le z|\{\boldsymbol{X}_t\}) &= E\big(P(S(p) \le z|\{Y_{i,t}, \boldsymbol{X}_t\})|\{\boldsymbol{X}_t\}\big)\\
&= P\big(\Phi(B_{N1} + B_{N2}) \le z|\{\boldsymbol{X}_t\}\big) + o_P(1).
\end{aligned}$$

Hence the theorem follows. □

4.A.2 Proofs of Theorems 4.4–4.6

Proof of Theorem 4.4. Under H_0, it follows from (4.5) that

$$E\big(\hat{t}_i^2(0)|\{\boldsymbol{X}_t\}\big) = \int_{-\infty}^{\infty} z^2\, dP(\hat{t}_i(0) \le z|\{\boldsymbol{X}_i\}) = t_i + 1^{-1}\int_{-\infty}^{\infty} s^2\, d\{\phi(s) q_{i,2,\boldsymbol{x}}(s)\} + O_P(T_i^{-3/2})$$

and

$$E\big((\hat{t}_i^2(0) - 1)^2|\{\boldsymbol{X}_t\}\big) = \int_{-\infty}^{\infty}(z^2 - 1)^2\, dP(\hat{t}_i(0) \le z|\{\boldsymbol{X}_t\}) = 2 + o_P(1).$$

Conditional on $\{\boldsymbol{X}_t\}$, $\{\hat{t}_i^2(0) - E(\hat{t}_i^2(0)|\{\boldsymbol{X}_t\})\}_{i=1}^N$ is a sequence of independent but nonidentically distributed random variables. Hence, the theorem follows from the Central Limit Theorem for martingale differences and the above expansions. □

Proof of Theorem 4.5. We can prove the theorem in the same way as Theorem 4.4 using (4.6). □

Proof of Theorem 4.6. Because $\hat{t}_i(0) = \hat{t}_i(\alpha_i) + \hat{\delta}_i$ and $\hat{\delta}_i/\delta_i - 1 = O_P(T_i^{-1/2})$, we have

$$\begin{aligned}
&E\big(\hat{t}_i^2(0)|\{\boldsymbol{X}_t\}\big)\\
=&E\big(\hat{t}_i^2(\alpha_i)|\{\boldsymbol{X}_t\}\big) + \delta_i^2 E\big(\hat{\delta}_i^2/\delta_i^2|\{\boldsymbol{X}_t\}\big) + 2E\big(\hat{t}_i(\alpha_i)(\hat{\delta}_i - \delta_i + \delta_i)|\{\boldsymbol{X}_t\}\big)\\
=&E\big(\hat{t}_i^2(\alpha_i)|\{\boldsymbol{X}_t\}\big) + \delta_i^2\big(1 + O_P(T_i^{-1/2})\big) + 2E\big(\hat{t}_i(\alpha_i)(\hat{\delta}_i - \delta_i)|\{\boldsymbol{X}_t\}\big) + 2E\big(\hat{t}_i(\alpha_i)\delta_i|\{\boldsymbol{X}_t\}\big)
\end{aligned} \tag{4.A.15}$$

By (4.5), we obtain that

$$\begin{cases}
E\big(\hat{t}_i^2(\alpha_i)|\{\boldsymbol{X}_t\}\big) = t_i + 1^{-1}\int_{-\infty}^{\infty} s^2\, d\{\phi(s)q_{i,2,\boldsymbol{x}}(s)\} + O_P(T_i^{-3/2}),\\
E\big(\hat{t}_i(\alpha_i)\delta_i|\{\boldsymbol{X}_t\}\big) = T_i^{-1/2}\delta_i \int_{-\infty}^{\infty} s\, d\{\phi(s)q_{i,1,\boldsymbol{x}}(s)\} + O_P(T_i^{-3/2}).
\end{cases} \tag{4.A.16}$$

Because

$$\hat{t}_i(\alpha_i) = \frac{1}{\sqrt{T_i}} \sum_{t=t_i+1}^{t_i+T_i} \frac{\varepsilon_{i,t}}{\hat{\sigma}_i} \frac{1 - (\boldsymbol{X}_t - \overline{\boldsymbol{X}}_i)'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}{\sqrt{1 + \overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}$$

and

$$\hat{\sigma}_i^2 - \sigma_i^2 = \frac{1}{T_i} \sum_{t=t_i+1}^{t_i+T_i} \{\varepsilon_{i,t}^2 - E\varepsilon_{i,t}^2\} + O_P(T_i^{-1}),$$

we have

$$\begin{aligned}
E\big(\hat{t}_i(\alpha_i)(\delta_i-\hat{\delta}_i)|\{\boldsymbol{X}_t\}\big) =& E\big(\hat{t}_i(\alpha_i)\frac{\delta_i}{\hat{\sigma}_i}(\hat{\sigma}_i-\sigma_i)|\{\boldsymbol{X}_t\}\big)\\
=&\frac{\delta_i}{2\sigma_i^2}\big(1+O_P(T_i^{-1/2})\big)E\big(\hat{t}_i(\alpha_i)(\hat{\sigma}_i^2-\sigma_i^2)|\{\boldsymbol{X}_t\}\big)\\
=&\frac{\delta_i}{2\sigma_i^2}\big(1+O_P(T_i^{-1/2})\big)\left\{\frac{1}{\sqrt{T_i}}\frac{E\varepsilon_{i,t}^3}{\sigma_i\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}+O_P(T_i^{-1})\right\}\\
=&\frac{\delta_i}{\sqrt{T_i}}\frac{E\varepsilon_{i,t}^3}{2\sigma_i^3\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\big(1+O_P(T_i^{-1/2})\big). \qquad (4.A.17)
\end{aligned}$$

Hence, it follows from (4.A.15)–(4.A.17) that

$$\begin{aligned}
E\big(\hat{t}_i^2(0)|\{\boldsymbol{X}_t\}\big) =\ & t_i+1^{-1}\int_{-\infty}^{\infty}s^2\,d\{\phi(s)q_{i,2,\boldsymbol{x}}(s)\}\\
&+\frac{\delta_i}{\sqrt{T_i}}\left(2\int_{-\infty}^{\infty}s\,d\{\phi(s)q_{i,1,\boldsymbol{x}}(s)\}-\frac{E\varepsilon_{i,t}^3}{\sigma_i^3\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\right)+\delta_i^2\\
&+o_P\big(T_i^{-1}+|\delta_i|/\sqrt{T_i}+\delta_i^2\big) \qquad (4.A.18)
\end{aligned}$$

and

$$E\big((\hat{t}_i^2(0)-1)^2|\{\boldsymbol{X}_t\}\big)=2+o_P(1). \qquad (4.A.19)$$

Like the proof of Theorem 4.4, we have

$$P\big(HT\le z)|\{\boldsymbol{X}_t\}\big)=\Phi\left(z-\frac{1}{\sqrt{2N}}\sum_{i=1}^{N}T_i^{-1}\int_{-\infty}^{\infty}s^2\,d\{\phi(s)q_{i,2,\boldsymbol{x}}(s)\}-\Delta\right)+o_P(1), \qquad (4.A.20)$$

where

$$\Delta=\frac{1}{\sqrt{2N}}\sum_{i=1}^{N}\left\{\frac{\delta_i}{\sqrt{T_i}}\left(2\int_{-\infty}^{\infty}s\,d\{\phi(s)q_{i,1,\boldsymbol{x}}(s)\}-\frac{E\varepsilon_{i,t}^3}{\sigma_i^3\sqrt{1+\overline{\boldsymbol{X}}_i'\Sigma_i^{-1}\overline{\boldsymbol{X}}_i}}\right)+\delta_i^2\right\}.$$

Because $\hat{\gamma}_i\xrightarrow{p}\gamma_i$ and $\hat{\kappa}_i\xrightarrow{p}\kappa_i$ as $T_i\to\infty$, and Theorem 4.5 holds for any α_i's, we have

$$\Phi\left(HT^{([aB/2])}-\frac{1}{\sqrt{2N}}\sum_{i=1}^{N}T_i^{-1}\int_{-\infty}^{\infty}s^2\,d\{\phi(s)q_{i,2,\boldsymbol{x}}(s)\}\right)\xrightarrow{p}a/2, \qquad (4.A.21)$$

$$\Phi\left(HT^{(B-[aB/2])} - \frac{1}{\sqrt{2N}}\sum_{i=1}^{N} T_i^{-1}\int_{-\infty}^{\infty} s^2\, d\{\phi(s)q_{i,2,\boldsymbol{x}}(s)\}\right) \xrightarrow{p} 1 - a/2, \tag{4.A.22}$$

as $N \to \infty$ and $B \to \infty$. Therefore, the theorem follows from (4.A.20), (4.A.21) and (4.A.22). □

4.A.3 Proofs of Theorems 4.7 and 4.8

Proof of Theorem 4.7. The key idea in the proof is to expand $E(\widetilde{t}_i^2) - 1$ till the term with order $1/T_i$. Before computing this expectation, we introduce some notations.

$$\Delta_{i,t,k}^{(j,l)}(\boldsymbol{\theta}_i) = \frac{\partial^2}{\partial\theta_{i,j}\partial\theta_{i,l}}\Delta_{i,t,k}(\boldsymbol{\theta}_i),\ \boldsymbol{\Sigma}_{i,k} = \frac{1}{T_i}\sum_{t=t_i+1}^{t_i+T_i}(\Delta_{i,t,k}^{(j,l)}(\boldsymbol{\theta}_i))_{1\le j,l\le K_i},$$

$$\widetilde{\boldsymbol{\Sigma}}_{i,k} = \frac{1}{T_i}\sum_{t=t_i+1}^{t_i+T_i}(\Delta_{i,t,k}^{(j,l)}(\widetilde{\boldsymbol{\theta}}_i))_{1\le j,l\le K_i},$$

$$\mathbf{E\Gamma}_{i,k} = E(\boldsymbol{\Gamma}_{i,k}),\ \mathbf{E\Gamma}_i = E(\boldsymbol{\Gamma}_i),\ \mathbf{EM}_i = E(\boldsymbol{M}_i),$$

$$Z_{i,k} = \frac{1}{\sqrt{T}_i}\sum_{t=t_i+1}^{t_i+T_i}\Delta_{i,t,k}(\boldsymbol{\theta}_i),\ \boldsymbol{Z}_i = (Z_{i,1},\cdots,Z_{i,K_i})',$$

$$R_{i,k} = \frac{1}{2\sqrt{T_i}}\sqrt{T_i}(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)'\boldsymbol{\Sigma}_{i,k}\sqrt{T_i}(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) = O_P(1/\sqrt{T_i}),\ \boldsymbol{R}_i = (R_{i,1},\cdots,R_{i,K_i})',$$

$$\mathbf{D\Gamma}_i = \mathbf{E\Gamma}_i^{-1}\{\boldsymbol{\Gamma}_i - \mathbf{E\Gamma}_i\}\mathbf{E\Gamma}_i^{-1},\ \widetilde{\mathbf{D\Gamma}}_i = \mathbf{E\Gamma}_i^{-1}\{\widetilde{\boldsymbol{\Gamma}}_i - \mathbf{E\Gamma}_i\}\mathbf{E\Gamma}_i^{-1}.$$

It follows from Taylor expansion that

$$\begin{aligned} 0 &= \frac{1}{\sqrt{T}_i}\sum_{t=t_i+1}^{t_i+T_i}\Delta_{i,t,k}(\widetilde{\boldsymbol{\theta}}_i) \\ &= Z_{i,k} + \boldsymbol{\Gamma}_{i,k}'\sqrt{T_i}(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) + \frac{1}{2\sqrt{T_i}}\sqrt{T_i}(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)'\boldsymbol{\Sigma}_{i,k}\sqrt{T_i}(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) + O_P(1/T_i), \end{aligned}$$

for $k = 1, \ldots, K_i$, implying that

$$\sqrt{T_i}(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) = -\boldsymbol{\Gamma}_i^{-1}\boldsymbol{Z}_i + O_P(1/\sqrt{T_i}), \tag{4.A.23}$$

and

$$\begin{aligned} & \sqrt{T_i}(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \\ = \;& -\boldsymbol{\Gamma}_i^{-1}\boldsymbol{Z}_i - \boldsymbol{\Gamma}_i^{-1}\boldsymbol{R}_i + O_P(1/T_i) \qquad (4.A.24) \\ = \;& -\mathbf{E}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{Z}_i + \mathbf{D}\boldsymbol{\Gamma}_i\boldsymbol{Z}_i - \mathbf{E}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{R}_i + O_P(1/T_i). \end{aligned}$$

By (4.A.23), (4.A.24), and Taylor expansion, we expand the numerator and denominator of $\widetilde{t}_i^2$ as

$$T_i\boldsymbol{e}_i'(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)(\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)'\boldsymbol{e}_i = \sigma_i + B_i + C_i \qquad (4.A.25)$$

and

$$\{\boldsymbol{e}_i'\widetilde{\boldsymbol{\Gamma}}_i^{-1}\widetilde{\boldsymbol{M}}_i(\widetilde{\boldsymbol{\Gamma}}_i^{-1})'\boldsymbol{e}_i\}^{-1} = \{\sigma_i + A_i\}^{-1} = \sigma_i^{-1} - \sigma_i^{-2}A_i + 2\sigma_i^{-3}A_i^2 + O_P(1/T_i), \qquad (4.A.26)$$

where

$$\sigma_i = \boldsymbol{e}_i'\mathbf{E}\boldsymbol{\Gamma}_i^{-1}\mathbf{E}\mathbf{M}_i(\mathbf{E}\boldsymbol{\Gamma}_i^{-1})'\boldsymbol{e}_i,$$

$$A_i = -\boldsymbol{e}_i'\widetilde{\mathbf{D}\boldsymbol{\Gamma}}_i\widetilde{\boldsymbol{M}}_i(\widetilde{\boldsymbol{\Gamma}}_i^{-1})'\boldsymbol{e}_i + \boldsymbol{e}_i'\mathbf{E}\boldsymbol{\Gamma}_i^{-1}\widetilde{\mathbf{D}\mathbf{M}}_i(\widetilde{\boldsymbol{\Gamma}}_i^{-1})'\boldsymbol{e}_i - \boldsymbol{e}_i'\mathbf{E}\boldsymbol{\Gamma}_i^{-1}\mathbf{E}\mathbf{M}_i\widetilde{\mathbf{D}\boldsymbol{\Gamma}}_i'\boldsymbol{e}_i.$$

$$B_i = \boldsymbol{e}_i'\mathbf{E}\boldsymbol{\Gamma}_i^{-1}\{\boldsymbol{Z}_i\boldsymbol{Z}_i' - \mathbf{E}\mathbf{M}_i\}(\mathbf{E}\boldsymbol{\Gamma}_i^{-1})'\boldsymbol{e}_i$$

$$C_i = -2\boldsymbol{e}_i'\mathbf{E}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{Z}_i(\mathbf{D}\boldsymbol{\Gamma}_i\boldsymbol{Z}_i - \mathbf{E}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{R}_i)'\boldsymbol{e}_i + \boldsymbol{e}_i'(\mathbf{D}\boldsymbol{\Gamma}_i\boldsymbol{Z}_i - \mathbf{E}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{R}_i)(\mathbf{D}\boldsymbol{\Gamma}_i\boldsymbol{Z}_i - \mathbf{E}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{R}_i)'\boldsymbol{e}_i.$$

It follows from (4.A.25) and (4.A.26) that

$$\widetilde{t}_i^2 - 1 = \sigma_i^{-1}B_i + \mu_i + o_P(1/T_i), \qquad (4.A.27)$$

where

$$\mu_i = E\{\sigma_i^{-1}C_i - \sigma_i^{-1}A_i - \sigma_i^{-2}A_iB_i - \sigma_i^{-2}A_iC_i + 2\sigma_i^{-2}A_i^2 + 2\sigma_i^{-3}A_i^2B_i + 2\sigma_i^{-3}A_i^2C_i\}. \quad (4.A.28)$$

Some tedious calculations show

$$\mu_i = O(1/T_i). \qquad (4.A.29)$$

Next, we write

$$\boldsymbol{Z}_i = \boldsymbol{W}_i^{(1)} + \boldsymbol{W}_i^{(2)},$$

where

$$\boldsymbol{W}_i^{(1)} = \frac{1}{\sqrt{T_i}} \sum_{t=t_i+1}^{t_i+T_i} \zeta_i U_t \frac{\partial}{\partial \boldsymbol{\theta}_i} \eta_{i,t}(\boldsymbol{\theta}_i) \text{ and } \boldsymbol{W}_i^{(2)} = \frac{1}{\sqrt{T_i}} \sum_{t=t_i+1}^{t_i+T_i} e_{i,t} \frac{\partial}{\partial \boldsymbol{\theta}_i} \eta_{i,t}(\boldsymbol{\theta}_i).$$

Put

$$\mathbf{EM}_i^{(1)} = E\left\{ U_t^2 \frac{\partial}{\partial \boldsymbol{\theta}_i} \eta_{i,t}(\boldsymbol{\theta}_i) \frac{\partial}{\partial \boldsymbol{\theta}_i'} \eta_{i,t}(\boldsymbol{\theta}_i) \right\},\ \mathbf{EM}_i^{(2)} = E\left\{ e_{i,t}^2 \frac{\partial}{\partial \boldsymbol{\theta}_i} \eta_{i,t}(\boldsymbol{\theta}_i) \frac{\partial}{\partial \boldsymbol{\theta}_i'} \eta_{i,t}(\boldsymbol{\theta}_i) \right\},$$

$$B_i^{(1)} = \boldsymbol{e}_i' \mathbf{E}\boldsymbol{\Gamma}_i^{-1} \{ \boldsymbol{W}_i^{(1)} \boldsymbol{W}_i^{(1)'} - \mathbf{EM}_i^{(1)} \} (\mathbf{E}\boldsymbol{\Gamma}_i^{-1})' \boldsymbol{e}_i,$$

$$B_i^{(2)} = \boldsymbol{e}_i' \mathbf{E}\boldsymbol{\Gamma}_i^{-1} \{ \boldsymbol{W}_i^{(2)} \boldsymbol{W}_i^{(2)'} - \mathbf{EM}_i^{(2)} \} (\mathbf{E}\boldsymbol{\Gamma}_i^{-1})' \boldsymbol{e}_i$$

and

$$B_i^{(3)} = \boldsymbol{e}_i' \mathbf{E}\boldsymbol{\Gamma}_i^{-1} \{ \boldsymbol{W}_i^{(1)} \boldsymbol{W}_i^{(2)'} + \boldsymbol{W}_i^{(2)} \boldsymbol{W}_i^{(1)'} \} (\mathbf{E}\boldsymbol{\Gamma}_i^{-1})' \boldsymbol{e}_i.$$

Using condition (C9), we can show that

$$\begin{cases} \frac{1}{\sqrt{2N}} \sum_{i=1}^N \sigma_i^{-1} B_i^{(1)} = O_P(\frac{1}{\sqrt{N}} \sum_{i=1}^N \zeta_i^2) = o_P(1), \\ \frac{1}{\sqrt{2N}} \sum_{i=1}^N \sigma_i^{-1} B_i^{(3)} = O_P(\frac{1}{\sqrt{N}} \sum_{i=1}^N |\zeta_i|) = o_P(1). \end{cases} \tag{4.A.30}$$

As $\{\sigma_i^{-1} B_i^{(2)}\}_{i=1}^N$ is a sequence of independent but nonidentically distributed random variables with zero mean, it follows from the Central Limit Theorem for martingale differences that

$$\frac{1}{\sqrt{2N}} \sum_{i=1}^N \sigma_i^{-1} B_i^{(2)} \xrightarrow{d} N(0,1) \text{ as } N \to \infty. \tag{4.A.31}$$

Hence, the theorem follows from (4.A.27)–(4.A.31), and the fact that $B_i = B_i^{(1)} + B_i^{(2)} + B_i^{(3)}$. □

Proof of Theorem 4.8. Follow the same way in the proof of Theorem 4.7, we can show that

$$P(\widetilde{HT}_b^* \le z | \{Y_{i,t}, \boldsymbol{X}_t\}) = \Phi\left(z - \frac{1}{\sqrt{2N}} \sum_{i=1}^N \widetilde{\mu}_i \right) + o_P(1),$$

where $\widetilde{\mu}_i$ is an estimate of μ_i and $\sum_{i=1}^N \widetilde{\mu}_i / \sum_{i=1}^N \mu_i \xrightarrow{p} 1$. Hence, the theorem follows. □

Appendix 4.B Simulation Studies

In this section, we first provide simulation evidence to support our theoretical results of the Kosowski et al. (2006) method and our arguments for the Fama and French (2010) method. We then investigate the finite sample performance of the proposed bootstrap Hotelling's T-squared test. We calibrate simulation parameters to match the empirical quantities in the data set of actively-managed U.S. domestic equity mutual funds from January 1980 to December 2018; see Section 4.C of this Appendix for a detailed description of the data. All simulations are based on the Carhart (1997) four-factor model. Factor returns and alphas are in percentage per month.

4.B.1 Empirical Size and Power of Existing Bootstrap Tests

Empirical size

Simulation setup. We draw random samples of zero-alpha fund returns from the following four-factor model:

$$Y_{i,t} = \beta_{i1} X_{1,t} + \ldots + \beta_{i4} X_{4,t} + \varepsilon_{i,t},\ t = 1, \ldots, T_i,\ i = 1, \ldots, N.$$

Fund betas $\{\beta_{i,1}, \ldots, \beta_{i,4}\}_{i=1}^{N}$ are the regression estimates from real data. Factor returns $\{X_{1,t}\}, \{X_{2,t}\}, \{X_{3,t}\}$, and $\{X_{4,t}\}$ are mutually independent sequences with normal distributions $\mathcal{N}(0.642, 4.406^2)$, $\mathcal{N}(0.094, 3.000^2)$, $\mathcal{N}(0.255, 2.927^2)$, $\mathcal{N}(0.583, 4.460^2)$, respectively. The means and standard deviations of the simulated factors are matched to the corresponding empirical quantities of the market, size, book-to-market, and momentum factors. To better appreciate the effect of small sample sizes, we simulate three different panels of mutual fund data: (i) balanced panel with $N = 500$ and $T_1 = \ldots = T_N = T = 60,\ 200,\ 468;$[1] (ii) balanced panel with $N = 2650$, the number of funds in the data, and $T_1 = \ldots = T_N = T = 60,\ 200,\ 468$; (iii) unbalanced panel with $N = 2650$ and T_i matched to that in the real data. To examine the effect of skewness of fund residuals, we draw $\varepsilon_{i,t}$ independently and identically

[1] The fund betas in this case are a random draw of 500 estimated betas and kept fixed throughout.

from either $\mathcal{N}(0, 1.613^2)$, or standardized $\log \mathcal{N}(0, 0.100^2)$, or standardized $\log \mathcal{N}(0, 0.500^2)$. The skewness of the three residual distributions is 0, 0.302, 1.750, representing zero, moderate, and heavy skewness, respectively. Both log-normal distributions are standardized to have a mean of 0 and a standard deviation of 1.613. Note that 1.613 is the empirical median of residual volatility across funds. Furthermore, $\{\boldsymbol{X}_t\}$ and $\{\varepsilon_{i,t}\}$ are independently simulated. Because we are more interested in the performance of the tests for skilled funds, we focus on the percentiles above the median: $p_0 = 0.60, 0.70, 0.80, 0.85, 0.90, 0.95, 0.97, 0.99$. With 1000 repetitions and $B = 1000$ bootstrap samples, we report the empirical size of the KTWW and FF bootstrap tests in Tables 4.A.1–4.A.3 at the 10% significance level.

[Table 4.A.1 about here.]

[Table 4.A.2 about here.]

[Table 4.A.3 about here.]

Skewness. The size of the KTWW test is close to the nominal level for normally distributed residuals (Table 4.A.1) but is increasingly distorted when residual skewness grows (Tables 4.A.2 and 4.A.3), especially when T is small, i.e., the effect of A_{N2} and A_{N3} in Theorem 4.1 is significant. In general, size biases are more pronounced at more extreme percentiles.

In contrast, results in Table 4.A.1 demonstrate that the FF bootstrap test has marked size bias even when all residuals are normal and cross-sectionally independent, and all funds have long track records. Panel A shows that when $N = 500$ is comparable to $T = 468$, the size bias is appreciable, supporting our conjecture that the FF test has a biased size even when $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i^{-1/2} < \infty$, or in this balanced case, $\lim_{N\to\infty} N/T < \infty$. Panel B shows that when $N = 2650$ is much larger than $T = 468$, the size bias becomes substantial. The size distortion persists in unbalanced panels (Panel C). This reinforces our argument that the FF approach is also subject to the bias due to small sample sizes and a large number of funds. Note that in balanced panels, the practical issue of sample size mismatches between the bootstrap and actual samples is not present for the FF approach. This confirms that the determinant of size distortion for this approach is not simply due to undersampling of funds with a relatively short sample period, as advocated by Harvey and Liu (2020a,b). This argument also applies when we study its test power.[2]

Cross-sectional and time-series dimensions. Comparing Panel A to Panel B in Tables 4.A.2 and 4.A.3, as the cross-sectional dimension N is increased from 500 to 2650, the KTWW approach is more biased in size. Within either panel, the size bias is larger when the time-series dimension T is smaller. Panel C illustrates that in unbalanced panels, the KTWW test cannot achieve size control when fund residuals are heavily skewed. In nearly all cases, the FF approach produces worse sizes than the KTWW approach.

In summary, the simulation evidence is well in line with our theoretical conclusion in Theorem 4.1: the KTWW test is statistically invalid when many funds have short fund lengths and could have large size bias when fund residuals are skewed. While initially proposed to deal with cross-sectional dependence, the FF approach fails to deliver a test with a correct asymptotic size under cross-sectional independence. The FF test is heavily undersized even when all fund residuals are normal, cross-sectionally independent and have large sample sizes.

Empirical power

Simulation setup. We draw random samples of fund returns from the following model:

$$Y_{i,t} = \alpha_i + \beta_{i1} X_{1,t} + \ldots + \beta_{i4} X_{4,t} + \varepsilon_{i,t},\ t = 1, \ldots, T_i,\ i = 1, \ldots, N.$$

We assess test power under two scenarios to illustrate the effects of negative-alpha funds on test power: i) only funds with zero and positive alphas are present; and ii) funds with negative, zero, and positive alphas are all present. Let π^+ and π^- denote the proportions of skilled and unskilled funds, and let α^+ and α^- denote the alphas of skilled and unskilled funds, respectively. In the first scenario, we simulate skilled funds with the settings $(\pi^+, \alpha^+) = (1\%, 0.35), (2\%, 0.30), (20\%, 0.05)$. In the second

[2] Nevertheless, fund sample size mismatches may further exacerbate the size distortion or lack of power in its empirical application to unbalanced panels; see the discussion in Harvey and Liu (2020a,b).

scenario, in addition to skilled funds, we simulate unskilled funds with the settings $(\pi^-, \alpha^-) = (20\%, -0.30), (20\%, -0.30), (20\%, -0.10)$. In both cases, all other funds are zero-alpha. The alphas in the first two settings are close to the empirical averages in the data. The rest of the settings are the same as those for test size except that all fund residuals are drawn from $\mathcal{N}(0, 1.613^2)$ independently. To save space, we only report results for the unbalanced panel with $N = 2650$.

[Table 4.A.4 about here.]

The presence of negative-alpha funds. Table 4.A.4 illustrates how negative-alpha funds can erode the test power. Theorem 4.2 indicates that when negative-alpha funds overwhelm positive-alpha funds, the KTWW test could have insufficient power to detect skilled funds. This is confirmed by Panels A and B. In each panel, when there are positive alphas but no negative alphas ($\pi^- = 0$), the KTWW method has high power to detect skilled funds at almost all percentiles; when negative alphas are present ($\pi^- = 20\%$), the KTWW test maintains power at extreme percentiles but has significantly reduced power at smaller percentiles. Panel C further validates Theorem 4.3 that the KTWW test power is impacted by negative alphas when signal-to-noise ratio is low. In Panel C, the KTWW test is very powerful if only positive-alphas funds are present with very low signal-to-noise ratio. However, as comparable proportions of unskilled funds are introduced with slightly dominating negative alphas, the KTWW test cannot achieve high power at any percentile. In summary, these results support our prior theory that the presence of a substantial number of unskilled funds reduces the test power of the KTWW test for skilled funds.

In contrast, the FF test could have low power insofar as there are no negative-alpha funds. Additionally, the FF test power is also affected by negative alphas. The low test power of the FF approach has also been documented by Harvey and Liu (2020a,b). Note that the simulations in these studies do not include negative-alpha funds.

The alpha (signal-to-noise ratio) and proportion of skilled funds. Focusing on the case without negative-alpha funds ($\pi^- = 0$), in Table 4.A.4, the KTWW test is powerful either with small π^+ and large α^+ (Panel A), or small α^+ and large π^+ (Panel C). That is, the power is affected by both the alphas and proportions of truly skilled funds. Similar to Andrikogiannopoulou and Papakonstantinou (2019) and Harvey and Liu (2020a,b), a full array of simulation studies can be done by varying the alpha or the proportions of both unskilled and skilled funds while fixing the other. As the settings in Panels A and B are representative of the mutual fund data, we omit additional simulations to avoid clutter.

Overall, the simulation evidence presented above corroborates our theories of the KTWW test power, which is affected in particular by the presence of negative-alpha funds. The low power together with large size bias of the FF test should serve as a caution for applying this approach to performance evaluation of mutual funds.

Impact of cross-sectional dependence on empirical size

Given the empirical importance of cross-sectional dependence in fund returns, we briefly illustrate how it affects the size of the two bootstrap tests. This would also alleviate concerns regarding the size distortion of the FF test as all preceding simulations assume cross-sectional independence.

We simulate cross-sectionally dependent fund residuals based on Equation (4.17) without serial correlation: $\varepsilon_{i,t} = \zeta_i U_t + e_{i,t}$. We consider two specifications: i) all fund residuals are cross-sectionally weakly dependent: $\zeta_i = 0.1$ for $i = 1, \ldots, N$; ii) a small fraction of funds are cross-sectionally relatively strongly dependent: $\zeta_i = 0.4$ for 5% of funds and $\zeta_i = 0$ for the rest. These two specifications are also used in the simulations for bootstrap Hotelling's T-squared test in Section 4.B.2. $e_{i,t}$ is drawn independently from a normal distribution but when $\zeta_i \neq 0$, its standard deviation is adjusted such that the standard deviation of $\varepsilon_{i,t}$ remains 1.613. We report the results only for the unbalanced panel with $N = 2650$ and T_i the same as in the data.

[Table 4.A.5 about here.]

Table 4.A.5 shows that the size of the FF test is still considerably below the nominal level under the two different types of cross-sectional dependence. This is consistent with the results in Harvey and Liu (2020b) (e.g., Table A.2.1), although our simulations are different from theirs in that we explicitly control the level of cross-sectional dependence in the simulated fund returns whereas they simulate fund returns by random resampling directly from the data (after subtracting the estimated alphas). In conclusion, this simulation suggests that, quite opposite to the common belief in the empirical finance literature, the FF approach doesn't seem capable of capturing arbitrary degrees of cross-sectional dependence.[3] As expected, the KTWW test has a biased size under cross-sectional dependence. It can be substantially oversized even when the cross-sectional dependence is weak (Panel A). In contrast, as we will show in the following section, our proposed bootstrap Hotelling's T-squared test can achieve size control under these two specifications of cross-sectional dependence.

4.B.2 Empirical Size and Power of Bootstrap Hotelling's T-squared Test

Empirical size

To investigate the test size of the proposed bootstrap Hotelling's T-squared test for zero alpha, we simulate fund residuals $\varepsilon_{i,t}$ based on Equation (4.17): $\varepsilon_{i,t} = \sum_{j=1}^{p_i} \phi_{i,j}\varepsilon_{i,t-j} + \eta_{i,t}$, $\eta_{i,t} = \zeta_i U_t + e_{i,t}$. As in Section 4.B.1, we draw $e_{i,t}$ from three distributions to study the effects of skewness: normal, standardized $\log\mathcal{N}(0, 0.100^2)$ and standardized $\log\mathcal{N}(0, 0.500^2)$. We generate $\varepsilon_{i,t}$ without serial correlation ($\phi_{i,j} = 0$ for all funds) and with serial correlation ($\varepsilon_{i,t}$ is an AR(1) process and $\phi_{i,1} = 0.1$ for all funds). The test statistics HT and $\widetilde{HT}$ are used for these two cases, respectively. In addition, we specify three different cases of cross-sectional dependence: i) all residuals are cross-sectionally independent; ii) all residuals are cross-sectionally weakly dependent: $\zeta_i = 0.1$ for $i = 1, \ldots, N$; iii) a small fraction of funds are cross-sectionally relatively strongly dependent: $\zeta_i = 0.4$ for 5% of funds and $\zeta_i = 0$ for the rest. We draw $B = 1000$ samples for each fund residual series and compute the size based on the two-sided p-value and 1000 random samples. Each random sample is an unbalanced

[3] In further simulations unreported here, the undersizing of the FF test improves but does not vanish under stronger cross-sectional dependence.

panel of fund returns with $N = 2650$ and sample sizes T_i the same as in the data.

Panel A of Table 4.A.6 reports the empirical size at the significance level 10%. The proposed test delivers good size control in large cross-sections with small sample sizes, even when the fund residuals are skewed. Both test statistics are robust to weak cross-sectional dependence, and the test statistic $\widetilde{HT}$ can accommodate serial correlations.

[Table 4.A.6 about here.]

B.2.2 Empirical power

To examine the test power for skilled funds, we simulate funds endowed with zero and positive alphas.[4] The proportion and alpha of skilled funds are $(\pi^+, \alpha^+) = (1\%, 0.35), (2\%, 0.30)$. All other funds are zero-alpha. The remaining simulation settings are the same as those for test size. Panels B and C of Table 4.A.6 show that, across all cases, the new test is very powerful at detecting skilled funds even if they are scarce among the vast population of zero-alpha funds. Skewness and serial correlations can lead to power loss, but only to a small extent. Cross-sectional dependence doesn't have a noticeable impact on the test power.

In summary, contrary to existing bootstrap tests, the bootstrap Hotelling's T-squared test that we develop is well motivated in theory and supported by simulation evidence. The remarkable size and power properties of the proposed test makes our test procedure particularly viable for mutual fund performance evaluation.

Empirical accuracy of the sequential testing procedure

In this subsection, we assess the accuracy of applying the bootstrap Hotelling's T-squared test to select skilled funds in the sequential testing procedure. Recall that the sequential fund selection procedure searches for a p-value threshold to first isolate a zero-alpha fund

[4] Because the test is two-sided, we don't include negative-alpha funds as we did for studying existing bootstrap tests. Including negative-alpha funds would only increase test power.

set and then identify the skilled fund set. In practice, there exist a range of such thresholds, and the choice of the threshold directly affects the classification accuracy. As the empirical applications use the smallest threshold, we examine the accuracy associated with this threshold in the simulation.

The simulation settings are similar to those for studying the test power of existing bootstrap tests except that negative-alpha funds are now present. We simulate both cross-sectionally independent ($\delta_i = 0$ for $i = 1, \ldots, N$) and dependent ($\delta_i = 0.1$ for $i = 1, \ldots, N$) fund returns. The true proportion of skilled funds is set as $\pi^+ = 2\%$ or 5% with true alpha α^+ taken from $\{0.26, 0.28, 0.30, 0.32, 0.34\}$. The proportion and alpha of the unskilled funds are fixed at $(\pi^-, \alpha^-) = (20\%, -0.30)$. For reference, the top 2% and 5% of funds have an average four-factor alpha of 0.31 and 0.26, respectively, and the bottom 20% of funds have an average alpha of -0.30. All simulated funds have the same sample sizes and betas as those in the data ranked on their t-statistics. For example, the 2% skilled funds have the same T_i's and betas as the 2% top-ranked funds based on their t-statistics. All fund residuals follow $\mathcal{N}(0, 1.613^2)$. The simulation is repeated 500 times for each specification.

[Figure 4.A.1 about here.]

Figure 4.A.1 plots the average estimated proportions of skilled funds for varying alphas of skilled funds together with the standard deviations of the estimated proportions. Panel A shows the results for the case of cross-sectional independence, and Panel B for cross-sectional dependence. Panel A shows that when $\pi^+ = 2\%$, the average estimated proportion of skilled funds is very close to the true proportion with a slight upward bias, but the bias is mostly well within one standard deviation. When $\pi^+ = 5\%$, the sequential procedure underestimates the proportion of skilled funds with a downward bias, and it becomes increasingly accurate for larger alpha. In Panel B, same patterns emerge for the cross-sectional dependent case except that the standard deviation is inflated. Taken together, the bias associated with the smallest p-value threshold is small, and the sequential procedure can identify skilled funds with high accuracy.

Appendix 4.C Additional Results for Empirical Applications

4.C.1 Data

Similar to Blake et al. (2017) and Harvey and Liu (2020a), we focus on actively-managed U.S. equity mutual funds. From the Center for Research in Security Prices (CRSP) Survivorship Bias Free Mutual Fund Database, we take monthly returns, monthly total net assets (TNA), annual expense ratios, turnover ratios, and other fund characteristics for each share class uniquely identified by “crsp_fundno”. We aggregate multiple share classes based on the unique identifier, “wficn”, provided by MFLINK1. Following Elton et al. (2001), we exclude funds with less than $15 million in total net asset (TNA) and address the incubation bias issue following Evans (2010). We base our selection criteria on objective codes following Kacperczyk et al. (2008). We also exclude funds with an average percentage of common stocks lower than 80% of the total net asset. We identify index funds, ETF, and other passive funds using their names and the CRSP index fund identifier following Busse and Tong (2012) and Ferson and Lin (2014). Fund-level TNA is the sum of TNA across all share classes of the fund. The fund age is the years to the date of the oldest share class in the fund. Family TNA is the total TNA of each fund in a fund family (excluding the fund itself). The expense ratio and turnover ratio are the corresponding TNA-weighted averages of the expense ratios and turnover ratios across all fund share classes. We define fund flow as the average monthly net growth in fund assets beyond capital gains and dividends.

Since we focus on actively-managed equity funds, we expect that skilled funds and unskilled funds hold different stocks with different characteristics and different expected returns. We first generate 93 stock characteristics from CRSP, Compustat, and I/B/E/S based on Green et al. (2017).[5] They include all common stocks listed on NYSE, AMEX, and NASDAQ.[6] To calculate the fund level average of stock holding characteristics, we obtain the share volume of mutual fund portfolio holdings from the Thomson Reuters Mutual Fund Holdings database. We use the holding value as the weight and calculate the holding value-weighted stock characteristics. We filter out funds that hold less than ten stocks, and also exclude funds with the following Investment Objective Codes in the Thomson Reuters Mutual Fund Holdings database: International, Municipal Bonds, Bond and Preferred, Balanced, and Metals. Finally, we merge the CRSP Mutual Fund database and the Thomson Reuters Mutual Fund Holdings database using the MFLINKS tables provided by WRDS. Our sample is from January 1980 to December 2018.

[Table 4.A.7 about here.]

We also report the following performance measures and active management measures of mutual funds.

> *R-Squared Statistic* (*rsq*) in Amihud and Goyenko (2013), and Fund Idiosyncratic Volatility (*idiovolm*) in Jordan and Riley (2015). We regress fund excess net return on the Carhart (1997) four-factor model over a 36-month estimation period (with at least 12 valid observations) and obtain the R-squared statistic from this regression. We calculate the idiosyncratic volatility (*idiovolm*) based on residuals from the regression.
>
> *Return Gap* (*retgap*) in Kacperczyk et al. (2008) is the difference between fund gross return and holdings-based returns. We calculate the holdings-based gross portfolio return each month as the return of the disclosed portfolio by assuming constant fund portfolio holdings from the fund's most recent disclosure.
>
> *Active Share* (*active_share*) in Cremers and Petajisto (2009) captures the percentage of a manager's portfolio that differs from its benchmark index.[7] It is calculated by

[5] We appreciate Jeremiah Green for sharing SAS code at https://sites.google.com/site/jeremiahrgreenacctg/home.

[6] See Table 4.A.7 for a detailed definition of these 93 stock characteristics.

[7] https://activeshare.nd.edu/data/ and http://www.petajisto.net/data.html

aggregating the absolute differences between the weight of a portfolio's actual holdings and the weight of its closest matching index.

Active Weight (*aw*) in Doshi et al. (2015) is the absolute difference between the value weights and actual weights held by a fund, summed across its holdings.

Hypothetical Excess Return (*hrex*) is the hypothetical excess return for only common stocks in CRSP.

4.C.2 Additional Results Based on $\widetilde{HT}$ Test

In this subsection, we report the empirical results using the bootstrap Hotelling's T-squared test statistic $\widetilde{HT}$, which accounts for serial correlations in fund residuals.

For the four-factor residuals of each fund, we fit an AR model without intercept and pre-select the AR order using the *auto.arima* function from the R package *forecast* with a maximum order of 5. Around 53% of funds do not exhibit serial correlation, 24% and 14% of funds have an AR order of 1 and 2, respectively, and less than 10% of funds have an order above 3. For those funds having serially correlated residuals, we estimate the four-factor model together with the AR parameters as in Section 4.3.2 and obtain the t-statistics. The bootstrap Hotelling's T-squared test based on the test statistic $\widetilde{HT}$ is then applied in the sequential fund selection procedure to select skilled funds. Figure 4.A.2 shows the sequential test results for the zero-alpha fund set (Panel A) and the skilled fund set (Panel B). Compared to the test results based on HT, as shown in Figure 4.2, the p-value thresholds for zero-alpha fund sets (where the $\widetilde{HT}$ test p-value is above 0.1) become smaller. However, Panel B shows that for p-value threholds between 0.0323 and 0.0434, the $\widetilde{HT}$ test identifies a subset of positive-alpha funds as skilled against the corresponding zero-alpha fund sets. Applying the smallest threshold, 1.06% of funds are declared as skilled. Therefore, the empirical conclusion that a small subset of funds are skilled remains unchanged after taking serial correlations in fund returns into account.

Table 4.A.8 presents the stock holding characteristics, fund characteristics and alternative performance/active management measures for the fund portfolios formed by the sequential testing procedure with $\widetilde{HT}$. The results are very similar to those in Table 4.2.

[Figure 4.A.2 about here.]

[Table 4.A.8 about here.]

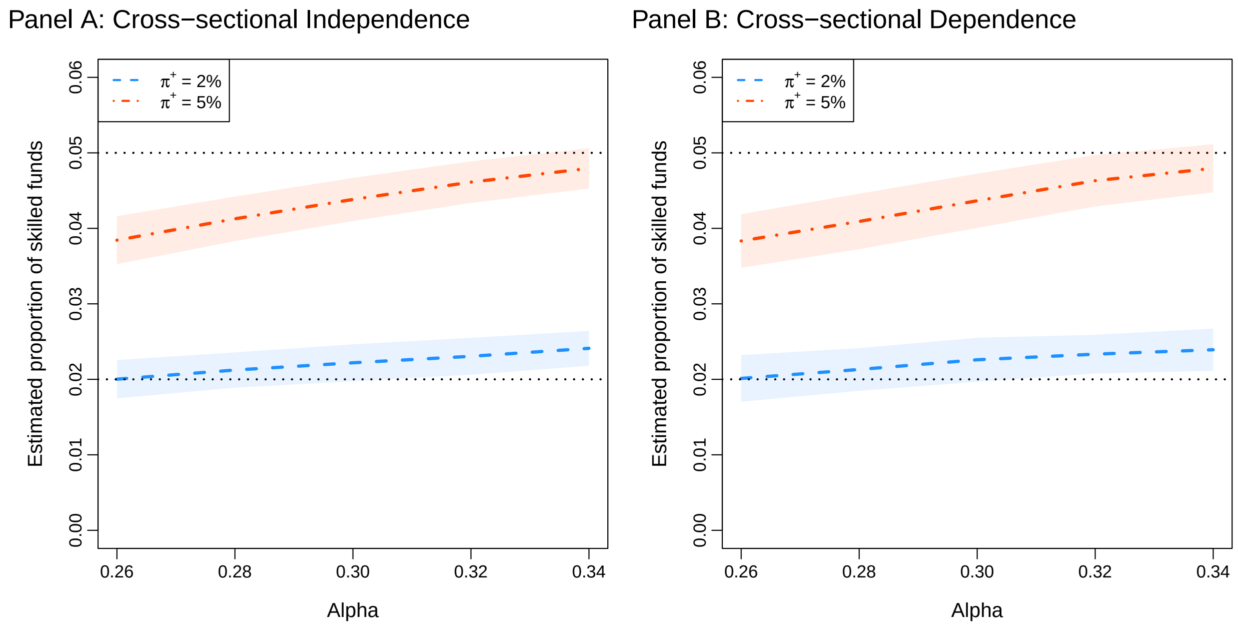
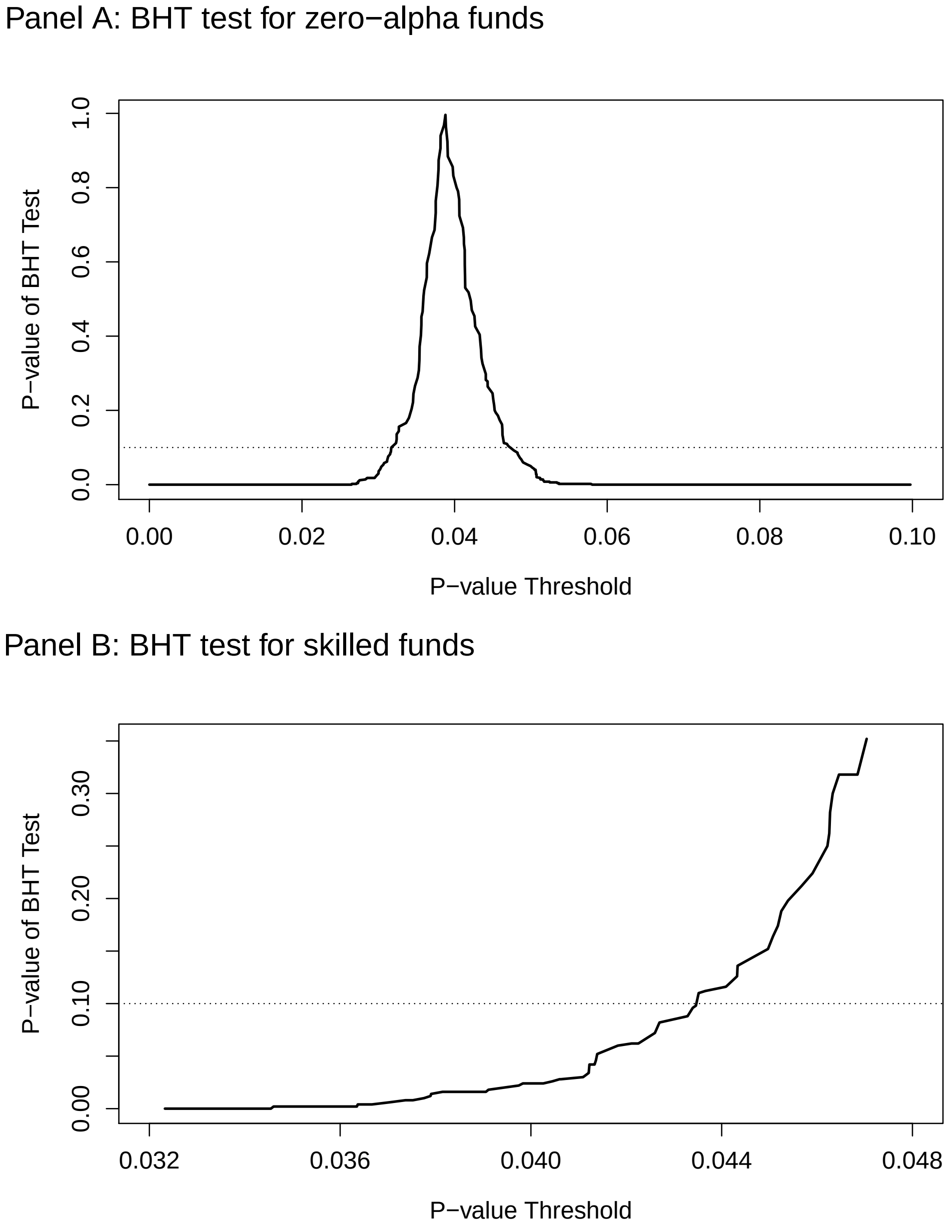


Figure 4.A.1. Empirical accuracy of the sequential testing procedure for selecting skilled funds.

In this figure, we plot the average estimated proportions of skilled funds for a range of alphas. In each simulated unbalanced panel of mutual funds ($N = 2650$ and T_i is the same as in the data), a small proportion ($\pi^+ = 2\%$ or 5%) of funds have an alpha from $\{0.26, 0.28, 0.30, 0.32, 0.34\}$, 20% of funds have an alpha of -0.30, and all other funds are zero-alpha. The dashed line is the average estimated proportion of skilled funds when $\pi^+ = 2\%$, and the dash-dotted line is the average estimated proportion when $\pi^+ = 5\%$. The shaded area surrounding each line is the average proportion plus/minus its standard deviation. The simulated fund returns are cross-sectionally independent for Panel A, and cross-sectionally dependent for Panel B ($\zeta_i = 0.1$ for all funds).



Panel A: BHT test for zero−alpha funds

Panel B: BHT test for skilled funds

Figure 4.A.2. Bootstrap Hotelling's T-squared test for fund selection – $\widetilde{HT}$ test.

In this figure, we plot the p-values for the bootstrap Hotelling's T-squared (BHT) test for a range of p-value thresholds in the sequential fund selection procedure. The BHT test using $\widetilde{HT}$ accounts for possible serial correlation by modeling fund residuals as AR processes and estimating the AR processes. Panel A shows the BHT test for a zero-alpha fund set, and Panel B shows the test for confirming a skilled fund set relative to the zero-alpha fund set. The data sample is monthly returns from January 1980 to December 2018 for all U.S. actively-managed equity funds with at least 60 return observations.

Table 4.A.1. Empirical size of bootstrap tests under normal fund residuals

The table presents the empirical size of the bootstrap tests in Kosowski et al. (2006) (KTWW) and Fama and French (2010) (FF) at the 10% significance level when fund residuals are normal. In Panels A and B, the simulated mutual fund data are balanced panels with the number of funds $N = 500$, $N = 2650$ and the number of time series observations $T = 60,\ 200,\ 468$; in Panel C, the simulated mutual fund data are an unbalanced panel with the number of funds $N = 2650$ and the number of time series observations for each fund matched to real data. Residuals of each fund are drawn independently from $\mathcal{N}(0, 1.613^2)$.

p	0.60	0.70	0.80	0.85	0.90	0.95	0.97	0.99
Panel A: Balanced Panel, $N = 500$								
$T = 60$								
KTWW	0.100	0.100	0.095	0.099	0.107	0.119	0.107	0.102
FF	0.061	0.028	0.007	0.003	0.001	0.005	0.007	0.020
$T = 200$								
KTWW	0.095	0.092	0.091	0.099	0.077	0.081	0.099	0.099
FF	0.087	0.061	0.050	0.040	0.029	0.033	0.043	0.066
$T = 468$								
KTWW	0.098	0.107	0.104	0.105	0.115	0.105	0.104	0.086
FF	0.094	0.094	0.070	0.074	0.076	0.081	0.079	0.072
Panel B: Balanced Panel, $N = 2650$								
$T = 60$								
KTWW	0.108	0.096	0.104	0.098	0.082	0.085	0.076	0.058
FF	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$T = 200$								
KTWW	0.101	0.091	0.111	0.110	0.115	0.101	0.095	0.091
FF	0.075	0.019	0.004	0.001	0.001	0.000	0.001	0.010
$T = 468$								
KTWW	0.098	0.104	0.099	0.099	0.101	0.098	0.099	0.116
FF	0.086	0.054	0.025	0.025	0.016	0.014	0.019	0.030
Panel C: Unbalanced Panel, $N = 2650$								
KTWW	0.085	0.110	0.109	0.111	0.116	0.108	0.119	0.130
FF	0.056	0.020	0.013	0.011	0.006	0.004	0.005	0.009

Table 4.A.2. Empirical size of bootstrap tests under moderately skewed fund residuals

The table presents the empirical size of the bootstrap tests in Kosowski et al. (2006) (KTWW) and Fama and French (2010) (FF) at the 10% significance level when fund residuals are moderately skewed. In Panels A and B, the simulated mutual fund data are balanced panels with the number of funds $N = 500$, $N = 2650$ and the number of time series observations $T = 60,\ 200,\ 468$; in Panel C, the simulated mutual fund data are an unbalanced panel with the number of funds $N = 2650$ and the number of time series observations for each fund matched to real data. Residuals of each fund are drawn independently from a standardized $\log \mathcal{N}(0, 0.010)$ distribution with a mean of 0 and a standard deviation of 1.613, so that the residual skewness of each fund is 0.302.

p	0.60	0.70	0.80	0.85	0.90	0.95	0.97	0.99
Panel A: Balanced Panel, $N = 500$								
$T = 60$								
KTWW	0.090	0.100	0.100	0.106	0.100	0.090	0.084	0.094
FF	0.057	0.030	0.007	0.003	0.002	0.001	0.006	0.009
$T = 200$								
KTWW	0.102	0.099	0.112	0.109	0.098	0.093	0.106	0.100
FF	0.085	0.073	0.049	0.043	0.039	0.045	0.052	0.061
$T = 468$								
KTWW	0.090	0.105	0.110	0.126	0.104	0.105	0.104	0.094
FF	0.081	0.094	0.085	0.100	0.071	0.060	0.071	0.077
Panel B: Balanced Panel, $N = 2650$								
$T = 60$								
KTWW	0.092	0.087	0.092	0.086	0.064	0.063	0.067	0.039
FF	0.011	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$T = 200$								
KTWW	0.100	0.092	0.090	0.095	0.110	0.089	0.091	0.088
FF	0.068	0.024	0.003	0.000	0.000	0.000	0.000	0.011
$T = 468$								
KTWW	0.094	0.102	0.097	0.109	0.095	0.085	0.096	0.111
FF	0.084	0.055	0.030	0.021	0.019	0.015	0.018	0.032
Panel C: Unbalanced Panel, $N = 2650$								
KTWW	0.090	0.080	0.085	0.090	0.099	0.095	0.102	0.093
FF	0.052	0.021	0.007	0.004	0.003	0.002	0.002	0.005

Table 4.A.3. Empirical size of bootstrap tests under heavily skewed fund residuals

The table presents the empirical size of the bootstrap tests in Kosowski et al. (2006) (KTWW) and Fama and French (2010) (FF) at the 10% significance level when fund residuals are heavily skewed. In Panels A and B, the simulated mutual fund data are balanced panels with the number of funds $N = 500$, $N = 2650$ and the number of time series observations $T = 60,\ 200,\ 468$; in Panel C, the simulated mutual fund data are an unbalanced panel with the number of funds $N = 2650$ and the number of time series observations for each fund matched to real data. Residuals of each fund are drawn independently from a standardized $\log\mathcal{N}(0, 0.250)$ distribution with a mean of 0 and a standard deviation of 1.613, so that the residual skewness of each fund is 1.750.

p	0.60	0.70	0.80	0.85	0.90	0.95	0.97	0.99
Panel A: Balanced Panel, $N = 500$								
$T = 60$								
KTWW	0.071	0.065	0.065	0.052	0.056	0.035	0.036	0.044
FF	0.048	0.023	0.005	0.000	0.002	0.000	0.002	0.002
$T = 200$								
KTWW	0.097	0.109	0.086	0.070	0.092	0.067	0.071	0.061
FF	0.080	0.083	0.039	0.023	0.035	0.018	0.017	0.029
$T = 468$								
KTWW	0.097	0.082	0.088	0.095	0.100	0.094	0.094	0.112
FF	0.091	0.076	0.073	0.070	0.062	0.062	0.065	0.086
Panel B: Balanced Panel, $N = 2650$								
$T = 60$								
KTWW	0.044	0.044	0.024	0.024	0.011	0.006	0.007	0.003
FF	0.017	0.001	0.000	0.000	0.000	0.000	0.000	0.000
$T = 200$								
KTWW	0.068	0.073	0.077	0.074	0.071	0.057	0.047	0.063
FF	0.055	0.024	0.001	0.001	0.001	0.000	0.000	0.003
$T = 468$								
KTWW	0.089	0.085	0.087	0.091	0.085	0.079	0.096	0.085
FF	0.078	0.050	0.026	0.020	0.011	0.009	0.011	0.024
Panel C: Unbalanced Panel, $N = 2650$								
KTWW	0.079	0.071	0.075	0.068	0.067	0.061	0.056	0.062
FF	0.061	0.022	0.006	0.008	0.003	0.000	0.000	0.000

Table 4.A.4. Empirical power of existing bootstrap tests

The table presents the empirical power of the bootstrap tests in Kosowski et al. (2006) (KTWW) and Fama and French (2010) (FF) at the 10% significance level. π^+ and π^- denote the proportions of skilled and unskilled funds, respectively. α^+ and α^- denote the alphas of skilled and unskilled funds, respectively. The simulated mutual fund data are an unbalanced panel with the number of funds $N = 2650$ and the number of time series observations for each fund matched to real data. Residuals of each fund are drawn independently from a normal distribution with a mean of 0 and a standard deviation of 1.613.

p	0.60	0.70	0.80	0.85	0.90	0.95	0.97	0.99
Panel A: $\pi^+ = 1\%, \alpha^+ = 0.35$								
$\pi^- = 0$								
KTWW	0.250	0.302	0.383	0.453	0.533	0.729	0.844	0.986
FF	0.182	0.112	0.077	0.059	0.058	0.136	0.293	0.861
$\pi^- = 20\%, \alpha^- = -0.30$								
KTWW	0.000	0.000	0.000	0.000	0.000	0.054	0.290	0.918
FF	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.665
Panel B: $\pi^+ = 2\%, \alpha^+ = 0.30$								
$\pi^- = 0$								
KTWW	0.480	0.554	0.691	0.806	0.902	0.983	0.998	1.000
FF	0.387	0.323	0.280	0.283	0.375	0.672	0.892	0.997
$\pi^- = 20\%, \alpha^- = -0.30$								
KTWW	0.000	0.000	0.000	0.002	0.026	0.579	0.934	1.000
FF	0.000	0.000	0.000	0.000	0.000	0.072	0.462	0.994
Panel C: $\pi^+ = 20\%, \alpha^+ = 0.05$								
$\pi^- = 0$								
KTWW	0.944	0.980	0.997	1.000	1.000	1.000	1.000	1.000
FF	0.914	0.916	0.934	0.941	0.966	0.996	0.999	1.000
$\pi^- = 20\%, \alpha^- = -0.10$								
KTWW	0.010	0.062	0.216	0.337	0.433	0.528	0.529	0.456
FF	0.005	0.013	0.025	0.030	0.038	0.057	0.059	0.097

Table 4.A.5. Empirical size of existing bootstrap tests under cross-sectional dependence

The table presents the empirical size of the bootstrap tests in Kosowski et al. (2006) (KTWW) and Fama and French (2010) (FF) at the 10% significance level in two different cases of cross-sectional dependence in fund residuals. Fund residuals are simulated using $\varepsilon_{i,t} = \zeta_i U_t + e_{i,t}$. In Panel A, all funds are cross-sectionally weakly dependent by setting $\zeta_i = 0.1$ for $i = 1, \ldots, N$. In Panel B, a small fraction of funds are cross-sectionally relatively strongly dependent by setting $\zeta_i = 0.4$ for 5% of funds and $\zeta_i = 0$ for the rest. U_t is drawn from $\mathcal{N}(0, 1)$ and $e_{i,t}$ is drawn independently from a zero-mean normal distribution. The standard deviation of $\varepsilon_{i,t}$ is fixed at 1.613 by adjusting the standard deviation of $e_{i,t}$. The simulated mutual fund data are an unbalanced panel with the number of funds $N = 2650$ and the sample size for each fund matched to real data.

p	0.60	0.70	0.80	0.85	0.90	0.95	0.97	0.99
Panel A: $\zeta_i = 0.1$ for All Funds								
KTWW	0.279	0.284	0.263	0.234	0.212	0.218	0.185	0.128
FF	0.098	0.078	0.061	0.046	0.035	0.025	0.023	0.024
Panel B: $\zeta_i = 0.4$ for 5% of Funds								
KTWW	0.118	0.113	0.112	0.118	0.118	0.104	0.105	0.092
FF	0.079	0.041	0.018	0.011	0.008	0.007	0.003	0.020

Table 4.A.6. Empirical size and power of bootstrap Hotelling's T-squared test

The table presents the empirical size and power of the bootstrap Hotelling's T-squared test. Fund residuals are simulated from $\varepsilon_{i,t} = \phi_{i,1}\varepsilon_{i,t-1} + \eta_{i,t}$, $\eta_{i,t} = \zeta_i U_t + e_{i,t}$. For HT test, $\phi_{i,1} = 0$ so that funds residuals are not serially correlated. For $\widetilde{HT}$ test, $\phi_{i,1} = 0.1$ so that funds residuals follow an AR(1) process and are serially correlated. $CSD1$, $CSD2$ and $CSD3$ refer to $\zeta_i = 0$ for all funds, $\zeta_i = 0.1$ for all funds and $\zeta_i = 0.4$ for 5% of funds, respectively. U_t is drawn from $\mathcal{N}(0,1)$. $e_{i,t}$ is drawn independently from a zero-mean normal distribution (*Normal*), standardized $\log\mathcal{N}(0, 0.010)$ (*SLN1*), or standardized $\log\mathcal{N}(0, 0.250)$ (*SLN2*). For Panel A, all funds have an alpha of zero. For Panel B and Panel C, π^+ of funds have an alpha α^+ and all other funds are zero-alpha, where π^+ and α^+ stand for the proportion and alpha of skilled funds, respectively. $\varepsilon_{i,t}$ has a standard deviation of 1.613 for all funds. The simulated mutual fund data are an unbalanced panel with the number of funds $N = 2650$ and the sample size for each fund matched to real data.

$e_{i,t} \sim$	HT Test			$\widetilde{HT}$ Test		
	CSD1	*CSD2*	*CSD3*	*CSD1*	*CSD2*	*CSD3*
			Panel A: Empirical Size			
Normal	0.088	0.102	0.083	0.103	0.106	0.109
SLN1	0.102	0.088	0.112	0.096	0.115	0.095
SLN2	0.098	0.117	0.105	0.100	0.118	0.102
			Panel B: Empirical Power, $\pi^+ = 1\%, \alpha^+ = 0.35$			
Normal	0.923	0.922	0.930	0.860	0.865	0.850
SLN1	0.897	0.900	0.945	0.880	0.850	0.880
SLN2	0.870	0.884	0.898	0.850	0.850	0.880
			Panel C: Empirical Power, $\pi^+ = 2\%, \alpha^+ = 0.30$			
Normal	1.000	0.989	0.995	0.995	1.000	1.000
SLN1	0.997	0.983	1.000	0.990	1.000	0.990
SLN2	0.983	0.993	0.990	0.985	0.995	0.986

Table 4.A.7. Stock holding characteristics

The table details the definitions of acronyms for stock characteristics, which are from CRSP, Compustat, and I/B/E/S based on Green et al. (2017). The stock characteristics are used to calculate the holding value-weighted average of stock characteristics for each fund each month based on shares of mutual fund portfolio holdings from the Thomson Reuters Mutual Fund Holdings database.

absacc/acc:	accrual ratio and its absolute value.
aeavol:	abnormal 3-day trading volume around earnings announcement day.
age:	firm age based on the coverage of Compustat.
agr:	total asset growth.
baspread:	monthly average of relative bid-ask spread.
beta:	market Beta based on weekly return of 3 years.
bm/bm_ia:	book to market ratio/industry adjusted book to market ratio.
cash:	cash to assets.
cashdebt:	cash to debt.
cashpr:	cash productivity.
cfp/cfp_ia:	cash to market capitalization / industry adjusted cash to market capitalization.
chatoia:	change in sales to asset (industry adjusted).
chcsho:	change in shares outstanding.
chempia:	change in employee number (industry adjusted).
chfeps:	changes in earnings per share forecast.
chinv:	change in inventory scaled by total asset.
chmom:	change in 6-month cumulative return.
chpmia:	change in profit margin (industry adjusted).
chtx:	change in tax.
cinvest:	investment.
convind:	dummy for convertible bond.
currat:	current ratio.
depr:	depreciation to property, plant, and equipment ratio.
disp:	standard deviation of earnings per share forecast.
divi/divo:	dummy for dividend initiation and dummy for dividend omission.
dy:	dividend to market capitalization.
ear*100:	3-day return around announcement day.
egr:	book value of equity growth.
ep:	earning to market capitalization.
fgr5yr:	forecasted growth in 5-year earnings per share.
gma:	profitability.
grcapx:	capital expenditure growth.
grltnoa:	long-term net operating assets growth.
herf:	sales concentration of the industry where the firms are.
hire:	employee growth.
idiovol:	idiosyncratic standard deviation based on weekly return of 3 years.
ill*1000000:	Amihud ratio.
indmom:	12-month cumulative return of the industry where the firms are.
invest:	invest to asset.
ipo:	dummy for new equity issue.

Table 4.A.7 (cont'd): Stock holding characteristics

mve/ mve_ia/1000:	log of market capitalization and market capitalization (industry adjusted).
nanalyst:	number of analysts.
nincr:	number of consecutive quarters of earnings increases over the same quarter of last year.
operprof:	operating profitability.
orgcap:	organizational capital.
pchcapx ia:	change in capital expenditures in percentage (industry adjusted).
pchcurrat:	change in current ratio (in percentage).
pchdepr:	change in depreciation ratio (in percentage).
pchgm_pchsale:	difference of change in gross margin and change in sales (in percentage).
pchsale_pchinvt:	difference of change in sales and change in inventory (in percentage).
pchsale_pchrect:	difference of change in sales and change in receivables (in percentage).
pchsale_pchxsga:	difference of change in sales (in percentage) and change in Selling, General and Administrative Expenses (in percentage).
pchsaleinv:	change in sales to inventory in percentage.
pctacc:	Percentage accruals.
pricedelay:	price delay.
ps:	financial health score.
rd:	dummy for more than 5% increase in R&D expense to total asset ratio.
rd_mve/rd_sale:	R&D expense to market capitalization ratio and R&D expense to sales ratio.
realestate:	buildings and capitalized leases to property, plant, and equipment ratio.
retvol:	standard deviation based on daily return of the month.
roaq:	return on assets.
roavol:	standard deviation of return on assets.
roeq:	return on equity.
roic:	return on invested capital.
rsup:	unexpected sales to market capitalization ratio.
salecash/saleinv/salerec:	sale to cash ratio, sale to inventory ratio, and sale to account receivable ratio.
secured:	total liability to secured debt ratio.
securedind:	dummy for secured debt obligations.
sfe:	earning forecast to price per share.
sgr:	sales growth.
sin:	dummy for smoke, tobacco, beer, alcohol, or gaming industry.
sp:	sales to market capitalization.
std_dolvol:	standard deviation of daily dollar trading volume of the month.
std_turn:	standard deviation of daily turnover of the month.
stdcf:	standard deviation of cash flows to sales.
sue*100:	standardized unexpected earnings.
tang:	debt capacity to firm tangibility ratio.
tb:	tax income to book income ratio.
turn:	turnover ratio.
zerotrade:	number of zero trading days of the month.

Table 4.A.8. Mutual fund characteristics based on $\widetilde{HT}$ test

From January 1980 to December 2018, for all U.S. actively-managed equity funds with at least 60 valid observations, we compute alphas using the four-factor model and generate mutual fund portfolios with significantly positive alphas, zero alphas, and negative alphas, respectively, based on the $\widetilde{HT}$ test statistic. The AR orders for fund residuals are automatically selected with a maximum order of 5 using *auto.arima* from the R package *forecast*. We report the time-series averages of the monthly cross-sectional means in each portfolio and the difference in means between the two extreme portfolios. We compute t-statistics of the differences with Newey and West (1987) correction for time-series correlation with 6 lags. The variables include fund level stock holding characteristics (using the same variable abbreviations as in table one of Green et al. (2017)), fund characteristics, and fund performance/active management measures. For ease of reading, *ear* and *sue* are scaled by 100, Amihud ratio by 1000000, and *mve_ia* by 1/1000. We take log for the total net asset ($ million), for the age of the fund's oldest share class (in years), and for the family total net asset ($ million). Annual turnover and expense ratio (both in percentage point) are the value weighted averages across all fund share classes. Fund flow (%) is the average monthly net growth in fund assets beyond reinvested dividends and portfolio returns. Return gap is in percentage point. Active weight is scaled by 2. The hypothetical excess returns are in percentage. The variables are defined in Section 4.C.1 of this Appendix. Statistical significance of 1, 5, and 10 percent are indicated by ***, **, and *, respectively.

Variable	Neg. Alpha	Zero Alpha	Pos. Alpha	Pos.−Neg.	t-stat
(Number of funds)	(286)	(2,236)	(28)		
Holding characteristics					
absacc	0.0657	0.0693	0.0758	0.010***	(12.43)
acc	-0.0218	-0.0213	-0.0214	0.000	(0.31)
aeavol	0.6264	0.6793	0.7706	0.144***	(14.05)
age	17.1648	16.0537	14.6465	-2.518***	(-19.22)
agr	0.1755	0.1949	0.2346	0.059***	(10.83)
baspread	0.0264	0.0277	0.0294	0.003***	(8.71)
beta	1.0174	1.0660	1.1480	0.131***	(7.50)
bm	0.4966	0.4825	0.3797	-0.117***	(-9.05)
bm_ia	31.7992	32.4697	45.6806	13.881**	(2.05)
cash	0.1263	0.1349	0.1792	0.053***	(16.55)
cashdebt	0.2547	0.2661	0.3263	0.072***	(10.43)
cashpr	5.7195	6.7016	14.4123	8.693***	(10.35)
cfp	0.0757	0.0731	0.0620	-0.014***	(-4.43)
cfp_ia	15.2713	15.7682	20.7236	5.452**	(2.16)
chatoia	-0.0076	-0.0083	-0.0130	-0.005***	(-3.03)
chcsho	0.1641	0.1686	0.2025	0.038***	(5.40)
chempia	-0.1078	-0.1060	-0.0953	0.013*	(1.84)
chfeps	0.0208	0.0201	0.0192	-0.002	(-0.68)
chinv	0.0108	0.0126	0.0155	0.005***	(6.03)
chmom	-0.0103	-0.0050	0.0020	0.012**	(2.19)
chpmia	0.2384	0.2389	0.2182	-0.020	(-0.12)
chtx	0.0019	0.0022	0.0031	0.001***	(5.74)
cinvest	-0.0020	-0.0027	-0.0077	-0.006**	(-2.47)
convind	0.1768	0.1817	0.1866	0.010*	(1.67)
currat	2.4120	2.5441	2.7767	0.365***	(15.68)
depr	0.2105	0.2210	0.2489	0.038***	(14.56)
disp	0.0631	0.0735	0.0760	0.013***	(7.06)
divi	0.0159	0.0194	0.0254	0.010***	(8.15)
divo	0.0111	0.0146	0.0182	0.007***	(8.96)

Table 4.A.8 (cont'd): Mutual fund characteristics based on $\widetilde{HT}$ test

Variable (Number of funds)	Neg. Alpha (286)	Zero Alpha (2,236)	Pos. Alpha (28)	Pos.−Neg.	t-stat
Holding characteristics					
dy	0.0230	0.0197	0.0115	-0.012***	(-10.93)
ear	0.7286	0.7954	0.9144	0.186***	(3.59)
egr	0.1849	0.2051	0.2473	0.062***	(9.08)
ep	0.0555	0.0509	0.0415	-0.014***	(-12.28)
fgr5yr	14.2452	15.1710	16.9270	2.682***	(12.53)
gma	0.3876	0.4114	0.5051	0.117***	(13.30)
grcapx	0.5942	0.6828	0.8112	0.217***	(11.58)
grltnoa	0.0973	0.1031	0.1126	0.015***	(11.43)
herf	0.0763	0.0781	0.0779	0.002	(1.28)
hire	0.0869	0.1038	0.1372	0.050***	(11.81)
idiovol	0.0402	0.0424	0.0459	0.006***	(12.31)
ill	0.0263	0.0436	0.0495	0.023***	(3.89)
indmom	0.1644	0.1638	0.1671	0.003	(0.52)
invest	0.0798	0.0875	0.0997	0.020***	(7.21)
ipo	0.0174	0.0244	0.0315	0.014***	(8.06)
lev	1.9954	1.7218	0.8607	-1.135***	(-17.19)
mom12m	0.2380	0.2556	0.2847	0.047***	(4.34)
mom1m	0.0174	0.0196	0.0237	0.006***	(8.10)
mom36m	0.5447	0.5736	0.6464	0.102***	(5.28)
ms	4.7557	4.7282	5.1420	0.386***	(15.91)
mve	15.4597	15.1361	14.9922	-0.467***	(-20.94)
mve_ia	13.3481	10.2732	9.5073	-3.841***	(-9.75)
nanalyst	18.8560	16.9063	17.3839	-1.472***	(-10.06)
nincr	1.2012	1.2380	1.3619	0.161***	(5.06)
operprof	0.9004	0.9185	1.0049	0.104***	(7.20)
orgcap	0.0081	0.0084	0.0095	0.001***	(9.54)
pchcapx_ia	7.6784	7.1983	9.7775	2.099*	(1.78)
pchcurrat	0.0334	0.0367	0.0354	0.002	(0.63)
pchdepr	0.0415	0.0446	0.0534	0.012***	(4.78)
pchgm_pchsale	0.0084	0.0025	-0.0094	-0.018***	(-4.78)
pchsale_pchinvt	-0.0320	-0.0318	-0.0130	0.019***	(2.78)
pchsale_pchrect	-0.0301	-0.0334	-0.0316	-0.001	(-0.47)
pchsale_pchxsga	0.0054	0.0086	0.0084	0.003	(1.23)
pchsaleinv	0.0930	0.0917	0.0842	-0.009	(-1.14)
pctacc	-0.8380	-0.9171	-0.9065	-0.068	(-1.63)

Table 4.A.8 (cont'd): Mutual fund characteristics based on $\widetilde{HT}$ test

Variable	Neg. Alpha	Zero Alpha	Pos. Alpha	Pos.−Neg.	*t*-stat
(Number of funds)	(286)	(2,236)	(28)		
Holding characteristics					
pricedelay	0.1057	0.0953	0.0900	-0.016***	(-4.38)
ps	4.8338	4.8157	4.8217	-0.012	(-0.67)
rd	0.0963	0.0952	0.0997	0.003	(1.09)
rd_mve	0.0339	0.0344	0.0393	0.005***	(5.90)
rd_sale	0.1117	0.1196	0.2236	0.112***	(6.18)
realestate	0.3041	0.3011	0.3052	0.001	(0.50)
retvol	0.0200	0.0210	0.0226	0.003***	(9.69)
roaq	0.0172	0.0173	0.0205	0.003***	(5.94)
roavol	0.0127	0.0141	0.0166	0.004***	(16.27)
roeq	0.0420	0.0403	0.0411	-0.001	(-0.84)
roic	0.1153	0.1172	0.1259	0.011**	(2.24)
rsup	0.0225	0.0249	0.0243	0.002	(1.18)
salecash	42.4397	43.7860	35.3757	-7.064***	(-6.05)
saleinv	28.1824	28.2670	28.1341	-0.048	(-0.08)
salerec	10.6053	11.5634	11.0288	0.424**	(2.38)
secured	0.2069	0.2537	0.3006	0.095***	(10.87)
securedind	0.3765	0.3995	0.4203	0.044***	(6.85)
sfe	0.0543	0.0498	0.0390	-0.015***	(-10.01)
sgr	0.1568	0.1751	0.2167	0.060***	(10.80)
sin	0.0142	0.0146	0.0083	-0.006***	(-9.72)
sp	1.2097	1.2279	0.8894	-0.320***	(-7.31)
std_dolvol	0.4904	0.5240	0.5496	0.059***	(13.21)
std_turn	3.4288	3.8379	4.3220	0.893***	(10.06)
stdcf	1.5950	2.0164	3.6246	2.030***	(4.60)
sue	0.0275	0.0315	0.0340	0.007	(0.45)
tang	0.4880	0.4929	0.5047	0.017***	(7.43)
tb	0.1224	0.1262	0.1794	0.057***	(3.15)
turn	1.3150	1.4054	1.5239	0.209***	(6.92)
zerotrade	0.0174	0.0401	0.0398	0.022***	(5.13)
Fund characteristics					
logtna	4.9235	5.4971	6.8447	1.921***	(16.00)
logage	2.4279	2.4486	2.3842	-0.044	(-0.58)
logtna_family	7.5525	8.5142	10.7384	3.186***	(22.05)
turn_ratio	78.3772	79.6128	70.6058	-7.771**	(-2.09)
flow_pct	0.4486	0.8393	1.5823	1.134***	(5.60)
exp_ratio	1.2079	1.1296	0.9998	-0.208***	(-15.33)

Table 4.A.8 (cont'd): Mutual fund characteristics based on $\widetilde{HT}$ test

Variable	Neg. Alpha	Zero Alpha	Pos. Alpha	Pos.−Neg.	*t*-stat
(Number of funds)	(286)	(2,236)	(28)		
Performance/Active management measures					
rsq	0.9168	0.8764	0.8709	-0.046***	(-8.78)
idiovolm	0.0130	0.0167	0.0182	0.005***	(13.38)
retgap	-0.1066	-0.0323	-0.0060	0.101***	(4.51)
active_share	0.7785	0.8465	0.8923	0.114***	(31.02)
aw	0.8336	0.8467	0.9132	0.080***	(10.32)
hrex	0.6274	0.7791	1.1348	0.477***	(6.60)

Bibliography

Ali, A., L. Hwang, and M. A. Trombley. 2003. Arbitrage risk and the book-to-market anomaly. *Journal of Financial Economics* 69:355–373.

Amihud, Y. 2002. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets* 5:31–56.

Amihud, Y., and R. Goyenko. 2013. Mutual fund's R^2 as predictor of performance. *Review of Financial Studies* 26:667–694.

Amihud, Y., and H. Mendelson. 1989. The effects of beta, bid-ask spread, residual risk, and size on stock returns. *Journal of Finance* 44:479–486.

Andrikogiannopoulou, A., and F. Papakonstantinou. 2019. Reassessing false discoveries in mutual fund performance: skill, luck, or lack of power? *Journal of Finance* 74:2667–2688.

Ang, A., and G. Bekaert. 2007. Stock return predictability: Is it there? *Review of Financial Studies* 20:651–707.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang. 2006. The cross-section of volatility and expected returns. *Journal of Finance* 61:259–299.

Bajgrowicz, P., and O. Scaillet. 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics* 106:473–491.

Barillas, F., and J. Shanken. 2018. Comparing asset pricing models. *Journal of Finance* 73:715–754.

Barras, L., O. Scaillet, and R. Wermers. 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance* 65:179–216.

Barras, L., O. Scaillet, and R. Wermers. 2020. Reassessing false discoveries in mutual fund performance: skill, luck, or lack of power? A reply. *Journal of Finance, Forthcoming.*

Basrak, B., R. A. Davis, and T. Mikosch. 2002. Regular variation of GARCH processes. *Stochastic Processes and Their Applications* 99:95–115.

Berk, J., and R. Green. 2004. Mutual Fund Flows and Performance in Rational Markets. *Journal of Political Economy* 112:1269–1295.

Berk, J. B., and J. H. van Binsbergen. 2015. Measuring skill in the mutual fund industry. *Journal of Financial Economics* 118:1–20.

Billingsley, P. 1999. *Convergence of probability measures.* John Wiley & Sons.

Blake, D., T. Caulfield, C. Ioannidis, and I. Tonks. 2014. Improved inference in the evaluation of mutual fund performance using panel bootstrap methods. *Journal of Econometrics* 183:202–210.

Blake, D., T. Caulfield, C. Ioannidis, and I. Tonks. 2017. New evidence on mutual fund performance: a comparison of alternative bootstrap methods. *Journal of Financial and Quantitative Analysis* 52:1279–1299.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 31:307–327.

Busse, J. A., and Q. Tong. 2012. Mutual fund industry selection and persistence. *Review of Asset Pricing Studies* 2:245–274.

Campbell, J. Y., and R. J. Shiller. 1991. Yield spreads and interest rate movements: A bird's eye view. *Review of Economic Studies* 58:495–514.

Campbell, J. Y., and M. Yogo. 2006. Efficient tests of stock return predictability. *Journal of Financial Economics* 81:27–60.

Carhart, M. M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52:57–82.

Cavaliere, G., I. Georgiev, and A. R. Taylor. 2016. Sieve-based inference for infinite-variance linear processes. *Annals of Statistics* 44:1467–1494.

Cavaliere, G., I. Georgiev, and A. R. Taylor. 2018. Unit root inference for non-stationary linear processes driven by infinite variance innovations. *Econometric Theory* 34:302–348.

Cavaliere, G., and A. R. Taylor. 2007. Testing for unit roots in time series models with non-stationary volatility. *Journal of Econometrics* 140:919–947.

Cavaliere, G., and A. R. Taylor. 2009. Heteroskedastic time series with a unit root. *Econometric Theory* pp. 1228–1276.

Chan, N. H. 1990. Inference for near-integrated time series with infinite variance. *Journal of the American Statistical Association* 85:1069–1074.

Chan, N. H., D. Li, and L. Peng. 2012. Toward a unified interval estimation of autoregressions. *Econometric Theory* pp. 705–717.

Chan, N. H., and L. T. Tran. 1989. On the first-order autoregressive process with infinite variance. *Econometric Theory* pp. 354–362.

Chan, N. H., and R.-M. Zhang. 2010. Inference for unit-root models with infinite variance GARCH errors. *Statistica Sinica* pp. 1363–1393.

Chordia, T., A. Goyal, and A. Saretto. 2017. p-hacking: Evidence from two million trading strategies. Working Paper.

Chordia, T., A. Goyal, and A. Saretto. 2020. Anomalies and false rejections. *Review of Financial Studies* 33:2134–2179.

Chordia, T., A. Subrahmanyam, and V. R. Anshuman. 2001. Trading activity and expected stock returns. *Journal of Financial Economics* 59:3–32.

Cochrane, J. H. 2008. The dog that did not bark: A defense of return predictability. *Review of Financial Studies* 21:1533–1575.

Cohen, J., C. C. Coughlin, and D. Soques. 2021. House Price Growth Interdependencies and Comovement. Federal Reserve Bank of St. Louis Working Paper 2019-028.

Cotter, J., S. Gabriel, and R. Roll. 2011. Integration and contagion in US housing markets. *Working Paper* .

Coval, J., J. Jurek, and E. Stafford. 2009. The economics of structured finance. *Journal of Economic Perspectives* 23:3–25.

Crane, A. D., and K. Crotty. 2018. Passive versus active fund performance: Do index funds have skill? *Journal of Financial and Quantitative Analysis* 53:33–64.

Cremers, K. M., J. A. Fulkerson, and T. B. Riley. 2019. Challenging the conventional wisdom on active management: A review of the past 20 years of academic literature on actively managed mutual funds. *Financial Analysts Journal* 75:8–35.

Cremers, K. M., and A. Petajisto. 2009. How active is your fund manager? A new measure that predicts performance. *Review of Financial Studies* 22:3329–3365.

Croux, C., M. Forni, and L. Reichlin. 2001. A measure of comovement for economic variables: Theory and empirics. *Review of Economics and Statistics* 83:232–241.

Cujean, J. 2019. Idea Sharing and the Performance of Mutual Funds. *Journal of Financial Economics, Forthcoming.*

Davis, M. A., and J. Heathcote. 2005. Housing and the business cycle. *International Economic Review* 46:751–784.

Davis, R., and S. Resnick. 1985. More limit theory for the sample correlation function of moving averages. *Stochastic Processes and Their Applications* 20:257–279.

Davis, R., and S. Resnick. 1986. Limit theory for the sample covariance and correlation functions of moving averages. *Annals of Statistics* pp. 533–558.

Del Negro, M., and C. Otrok. 2007. 99 Luftballons: Monetary policy and the house price boom across US states. *Journal of Monetary Economics* 54:1962–1985.

Diether, K. B., C. J. Malloy, and A. Scherbina. 2002. Differences of opinion and the cross section of stock returns. *Journal of Finance* 57:2113–2141.

Doshi, H., R. Elkamhi, and M. Simutin. 2015. Managerial activeness and mutual fund performance. *Review of Asset Pricing Studies* 5:156–184.

Drees, H. 2003. Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli* 9:617–657.

Elton, E. J., M. J. Gruber, and C. R. Blake. 2001. A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases. *Journal of Finance* 56:2415–2430.

Engle, R. F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* pp. 987–1007.

Evans, R. B. 2010. Mutual fund incubation. *Journal of Finance* 65:1581–1611.

Fama, E. F. 1965. The behavior of stock-market prices. *Journal of Business* 38:34–105.

Fama, E. F. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance* 25:383–417.

Fama, E. F., and K. R. French. 1988. Permanent and temporary components of stock prices. *Journal of Political Economy* 96:246–273.

Fama, E. F., and K. R. French. 1996. Multifactor explanations of asset pricing anomalies. *Journal of Finance* 51:55–84.

Fama, E. F., and K. R. French. 2010. Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance* 65:1915–1947.

Fan, J., Y. Liao, and J. Yao. 2015. Power enhancement in high-dimensional cross-sectional tests. *Econometrica* 83:1497–1541.

Fan, J., L. Qi, and D. Xiu. 2014. Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *Journal of Business & Economic Statistics* 32:178–191.

Feng, G., S. Giglio, and D. Xiu. 2020. Taming the factor zoo: A test of new factors. *Journal of Finance* 75:1327–1370.

Ferretti, N., and J. Romo. 1996. Unit root bootstrap tests for AR (1) models. *Biometrika* 83:849–860.

Ferson, W., and Y. Chen. 2015. How many good and bad fund managers are there, really? Working Paper.

Ferson, W., and J. Lin. 2014. Alpha and performance measurement: The effects of investor disagreement and heterogeneity. *Journal of Finance* 69:1565–1596.

Francq, C., and J.-M. Zakoian. 2004. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* 10:605–637.

Fuller, W. A. 1996. *Introduction to statistical time series*, vol. 428. John Wiley & Sons.

Gagliardini, P., E. Ossola, and O. Scaillet. 2016. Time-varying risk premium in large cross-sectional equity data sets. *Econometrica* 84:985–1046.

Giglio, S., Y. Liao, and D. Xiu. 2020. Thousands of alpha tests. *Review of Financial Studies, Forthcoming* .

Glaeser, E. L., and J. Gyourko. 2006. Housing dynamics. Tech. rep., National Bureau of Economic Research.

Goldstein, I., W. Jiang, and G. A. Karolyi. 2019. To FinTech and beyond. *Review of Financial Studies* 32:1647–1661.

Goldstein, I., C. S. Spatt, and M. Ye. 2021. Big Data in Finance. Tech. rep., National Bureau of Economic Research.

Gonçalves, S., and B. Perron. 2014. Bootstrapping factor-augmented regression models. *Journal of Econometrics* 182:156–173.

Gonçalves, S., and B. Perron. 2020. Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics* 218:476–495.

Green, J., J. R. Hand, and X. F. Zhang. 2017. The characteristics that provide independent information about average US monthly stock returns. *Review of Financial Studies* 30:4389–4436.

Grønborg, N. S., A. Lunde, A. Timmermann, and R. Wermers. 2021. Picking funds with confidence. *Journal of Financial Economics* 139:1–28.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies* 33:2223–2273.

Guo, R., B. Lev, and C. Shi. 2006. Explaining the Short-and Long-Term IPO Anomalies in the US by R&D. *Journal of Business Finance & Accounting* 33:550–579.

Hall, P. 1992. *The bootstrap and Edgeworth expansion*. Springer-Verlag, New York.

Hall, P., and C. C. Heyde. 1980. *Martingale limit theory and its application*. Academic press.

Hamilton, J. D., and M. T. Owyang. 2012. The propagation of regional recessions. *Review of Economics and Statistics* 94:935–947.

Hansen, P. R. 2005. A test for superior predictive ability. *Journal of Business & Economic Statistics* 23:365–380.

Hansen, P. R., A. Lunde, and J. M. Nason. 2011. The model confidence set. *Econometrica* 79:453–497.

Harvey, C. R., and Y. Liu. 2019. Lucky factors. Working Paper.

Harvey, C. R., and Y. Liu. 2020a. False (and missed) discoveries in financial economics. *Journal of Finance, Forthcoming* .

Harvey, C. R., and Y. Liu. 2020b. Luck versus skill in the cross-section of mutual fund returns: reexamining the rvidence. *Working Paper* .

Harvey, C. R., Y. Liu, and A. Saretto. 2020. An evaluation of alternative multiple testing methods for finance applications. *Review of Asset Pricing Studies* 10:199–248.

Hill, B. M. 1975. A simple general approach to inference about the tail of a distribution. *Annals of Statistics* pp. 1163–1174.

Hill, J., D. Li, and L. Peng. 2016. Uniform interval estimation for an AR (1) process with AR errors. *Statistica Sinica* pp. 119–136.

Hill, J. B. 2015. Robust generalized empirical likelihood for heavy tailed autoregressions with conditionally heteroscedastic errors. *Journal of Multivariate Analysis* 135:131–152.

Hill, J. B., and A. Prokhorov. 2016. GEL estimation for heavy-tailed GARCH models with robust empirical likelihood inference. *Journal of Econometrics* 190:18–45.

Ho, A. T., K. P. Huynh, and D. T. Jacho-Chávez. 2016. Flexible estimation of copulas: An application to the US housing crisis. *Journal of Applied Econometrics* 31:603–610.

Ho, A. T., K. P. Huynh, and D. T. Jacho-Chávez. 2019. Using nonparametric copulas to measure crude oil price co-movements. *Energy Economics* 82:211–223.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28:650–705.

Jach, A., and P. Kokoszka. 2004. Subsampling unit root tests for heavy-tailed observations. *Methodology and Computing in Applied Probability* 6:73–97.

Jensen, M. C. 1968. The performance of mutual funds in the period 1945–1964. *Journal of Finance* 23:389–416.

Jones, C. S., and J. Shanken. 2005. Mutual fund performance with learning across funds. *Journal of Financial Economics* 78:507–552.

Jordan, B. D., and T. B. Riley. 2015. Volatility and mutual fund manager skill. *Journal of Financial Economics* 118:289–298.

Kacperczyk, M., C. Sialm, and L. Zheng. 2008. Unobserved actions of mutual funds. *Review of Financial Studies* 21:2379–2416.

Kallberg, J. G., C. H. Liu, and P. Pasquariello. 2014. On the price comovement of US residential real estate markets. *Real Estate Economics* 42:71–108.

Kole, E., K. Koedijk, and M. Verbeek. 2007. Selecting copulas for risk management. *Journal of Banking & Finance* 31:2405–2423.

Kosowski, R., N. Y. Naik, and M. Teo. 2007. Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. *Journal of Financial Economics* 84:229–264.

Kosowski, R., A. Timmermann, R. Wermers, and H. White. 2006. Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *Journal of Finance* 61:2551–2595.

Kostakis, A., T. Magdalinos, and M. P. Stamatogiannis. 2015. Robust econometric inference for stock return predictability. *Review of Financial Studies* 28:1506–1553.

Kourogenis, N., and N. Pittis. 2008. Testing for a unit root under errors with just barely infinite variance. *Journal of Time Series Analysis* 29:1066–1087.

Kuethe, T. H., and V. O. Pede. 2011. Regional housing price cycles: a spatio-temporal analysis using US state-level data. *Regional studies* 45:563–574.

Landier, A., D. Sraer, and D. Thesmar. 2017. Banking integration and house price co-movement. *Journal of Financial Economics* 125:1–25.

Lange, T. 2011. Tail behavior and OLS estimation in AR-GARCH models. *Statistica Sinica* pp. 1191–1200.

Lange, T., A. Rahbek, and S. T. Jensen. 2011. Estimation and asymptotic inference in the AR-ARCH model. *Econometric Reviews* 30:129–153.

Leamer, E. E. 2007. Housing is the business cycle. Tech. rep., National Bureau of Economic Research.

Leamer, E. E. 2015. Housing really is the business cycle: what survives the lessons of 2008–09? *Journal of Money, Credit and Banking* 47:43–50.

Lettau, M., and S. Van Nieuwerburgh. 2008. Reconciling the return predictability evidence. *Review of Financial Studies* 21:1607–1652.

Li, B., and A. G. Rossi. 2021. Selecting Mutual Funds from the Stocks They Hold: A Machine Learning Approach. *WorkingPaper* .

Li, D. 2011. Financial constraints, R&D investment, and stock returns. *Review of Financial Studies* 24:2974–3007.

Li, D., N. Chan, and L. Peng. 2014. Empirical likelihood test for causality of bivariate AR (1) processes. *Econometric Theory* pp. 357–371.

Li, D., X. Zhang, K. Zhu, and S. Ling. 2018. The ZD-GARCH model: A new way to study heteroscedasticity. *Journal of Econometrics* 202:1–17.

Li, K., F. Mai, R. Shen, and X. Yan. 2021. Measuring corporate culture using machine learning. *Review of Financial Studies* 34:3265–3315.

Ling, S. 2007. Self-weighted and local quasi-maximum likelihood estimators for ARMA-GARCH/IGARCH models. *Journal of Econometrics* 140:849–873.

Liu, W. 2006. A liquidity-augmented capital asset pricing model. *Journal of Financial Economics* 82:631–671.

Liu, W., and Q.-M. Shao. 2014. Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics* 42:2003–2025.

Liu, X., B. Yang, Z. Cai, and L. Peng. 2019. A unified test for predictability of asset returns regardless of properties of predicting variables. *Journal of Econometrics* 208:141–159.

Loughran, T., and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66:35–65.

McElroy, T., and A. Jach. 2019. Subsampling Inference for the Autocorrelations of GARCH Processes. *Journal of Financial Econometrics* 17:495–515.

McLean, R. D., and J. Pontiff. 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71:5–32.

Miao, H., S. Ramchander, and M. W. Simpson. 2011. Return and volatility transmission in US housing markets. *Real Estate Economics* 39:701–741.

Miles, W. 2011. Long-range dependence in US home price volatility. *Journal of Real Estate Finance and Economics* 42:329–347.

Newey, W. K., and K. D. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55:703–708.

Owen, A. B. 1990. Empirical likelihood ratio confidence regions. *The Annals of Statistics* 18:90–120.

Owen, A. B. 2001. *Empirical likelihood*. CRC press.

Pesaran, M. H., and T. Yamagata. 2017. Testing for alpha in linear factor pricing models with a large number of securities. CESifo Working Paper Series No. 6432.

Phillips, P. C. 1987. Towards a unified asymptotic theory for autoregression. *Biometrika* 74:535–547.

Phillips, P. C. 1990. Time series regression with a unit root and infinite-variance errors. *Econometric Theory* pp. 44–62.

Phillips, P. C., and J. H. Lee. 2013. Predictive regression under various degrees of persistence and robust long-horizon regression. *Journal of Econometrics* 177:250–264.

Phillips, P. C., and P. Perron. 1988. Testing for a unit root in time series regression. *Biometrika* 75:335–346.

Qin, J., and J. Lawless. 1994. Empirical likelihood and general estimating equations. *Annals of Statistics* pp. 300–325.

Rapach, D. E., M. C. Ringgenberg, and G. Zhou. 2016. Short interest and aggregate stock returns. *Journal of Financial Economics* 121:46–65.

Samarakoon, D. M., and K. Knight. 2009. A note on unit root tests with infinite variance noise. *Econometric Reviews* 28:314–334.

Shiller, R. J. 2007. Understanding recent trends in house prices and home ownership. Tech. rep., National Bureau of Economic Research.

Siburg, K. F., P. Stoimenov, and G. N. Weiß. 2015. Forecasting portfolio-Value-at-Risk with nonparametric lower tail dependence estimates. *Journal of Banking & Finance* 54:129–140.

Welch, I., and A. Goyal. 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21:1455–1508.

Wermers, R. 1999. Mutual fund herding and the impact on stock prices. *Journal of Finance* 54:581–622.

White, H. 2000. A reality check for data snooping. *Econometrica* 68:1097–1126.

White, H., T.-H. Kim, and S. Manganelli. 2015. VAR for VaR: Measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics* 187:169–188.

Xiao, Z. 2014. Unit roots: A selective review of the contributions of Peter CB Phillips. *Econometric Theory* pp. 775–814.

Yan, X., and L. Zheng. 2017. Fundamental analysis and the cross-section of stock returns: A data-mining approach. *Review of Financial Studies* 30:1382–1423.

Zhang, R., and S. Ling. 2015. Asymptotic inference for AR models with heavy-tailed G-GARCH noises. *Econometric Theory* pp. 880–890.

Zhu, B., R. Füss, and N. B. Rottke. 2013. Spatial linkages in returns and volatilities among US regional housing markets. *Real Estate Economics* 41:29–64.

Zhu, K., and S. Ling. 2011. Global self-weighted and local quasi-maximum exponential likelihood estimators for ARMA–GARCH/IGARCH models. *Annals of Statistics* 39:2131–2163.

Zhu, K., and S. Ling. 2015. LADE-based inference for ARMA models with unspecified and heavy-tailed heteroscedastic noises. *Journal of the American Statistical Association* 110:784–794.

Zimmer, D. M. 2012. The role of copulas in the housing crisis. *Review of Economics and Statistics* 94:607–620.

Zimmer, D. M. 2015. Asymmetric dependence in house prices: evidence from USA and international data. *Empirical Economics* 49:161–183.