Essays on Model Selection



Yingyu Guo

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Acknowledgements

I would like to thank the many people without who I could not have completed this journey.

First and foremost, I wish to express my gratitude to my supervisors, Professor Bent Nielsen and Professor Vanessa Berenguer-Rico, for their endless support, patient discussion and ongoing guidance over my MPhil and DPhil years. This dissertation would not have been possible without them — they have always given me the inspiration and encouragement to continue. I feel fortunate to have them as my supervisors, and that everything I learnt from them will stay with me for the rest of my life.

I am grateful to Professor James Duffy, Professor Steve Bond, Professor Sophocles Mavroeidis, and Dr. Jurgen Doornik for their insightful suggestion and helpful comments. I would also like to thank staffs in the Department of Economics, Wolfson College and St Anne's College for their help and guidance.

I gratefully thank all my fellow student colleagues and friends, in particular Dr. Xiyu Jiao, Le Wu, Jonas Kurle and Angie Gao, for sharing their invaluable experience and continuous solid support.

I am deeply indebted to my parents and relatives for their unconditional love and support. Any kid would consider themselves fortunate to have been grown up as part of this family, and I am obliged that I was the lucky one to enjoy such a privilege. To my father: Thank you for so much understanding and patience during the hard times, and so much love every minute.

Abstract

This dissertation has two goals. Firstly, we consider the problem of model selection for nonlinear time series models and study an auxiliary-regression-based test for serial correlation. Under the assumption of martingale difference errors, rejection of the test would imply model misspecification in the adopted functional form or insufficient dynamic specification. The proposed test allows for the presence of stationary and/or non-stationary variables, and the test statistic follows a standard chi-square distribution asymptotically. Monte Carlo simulations show that the proposed test has good finite sample properties.

Secondly, we investigate how to choose the correct model in least trimmed squares (LTS) regressions when the underlying data is contaminated with outliers. This amounts to the problem of choosing the number of outliers to be trimmed. We propose conditions under which information criteria based and cumulant based methods yield consistent estimation of the true model under a location-scale framework. We also extend the classical Jarque-Bera test to test for the normality of non-outlying error terms. Numerical studies are conducted to compare the performance of these methods.

TABLE OF CONTENTS

General Introduction **1**

I Testing for Residual Autocorrelation in General Non-linear Time Series Models **4**

1 Introduction for Part I **5**

1.1 Misspecification and Serial Correlation 5

1.2 Auxiliary Regression 7

1.3 Overview of Some Key Results 10

1.4 Literature Review 12

1.4.1 Linear Stationary Models 12

1.4.2 Non-stationary and/or Non-linear Models 16

1.5 Correlation Functions 17

1.6 Contribution of Part I 19

1.7 Organisation of Part I 21

2 Asymptotic Equivalence to a Partial TR^2 **22**

2.1 Introduction 22

2.2 The Model and Test Statistic 24

2.3 Assumptions and Theorems 26

2.4 Examples 33

2.4.1 A Simple Linear Random Walk Model 34

2.4.2 A Stationary AR(1) Model 38

2.4.3 An AR(2) Model with One Unit Root 40
2.4.4 A Power-Curve Model 48
2.4.5 A Semilogarithmic Growth Model 50
2.5 Conclusion 52

Appendices **54**

Appendix 2.A Notations and Preliminaries 54
Appendix 2.B Norms 56
Appendix 2.C Intermediate Results 57
Appendix 2.D Proof of Theorems 75

3 Asymptotic Irrelevance of Non-stationary Components **83**

3.1 Introduction 83
3.2 Assumptions and Theorems 86
3.3 Example — The AR(2) Model with One Unit Root 94
3.4 Conclusion 97

Appendices **101**

Appendix 3.A Notations and Preliminaries 101
Appendix 3.B Key Intermediate Result—Decomposing Q 105
Appendix 3.C Analysis of $D_{\widehat{u},\widehat{v}}$ and $D_{\widehat{\mathbf{H}}_u,\widehat{v}}$ 108
Appendix 3.D Decomposing Stationary and Random Walk Components . . 115
Appendix 3.E Proof of Theorem 3.1 116

4 Asymptotic Distribution of the Test Statistic and Simulation Studies **120**

4.1 Introduction 120
4.2 Assumptions and Theorems 123
4.3 Examples 129
4.3.1 The AR(2) Model with One Unit Root 129
4.3.2 The Power-Curve Model 133

4.4 Finite Sample Performance of TR^2 136
4.5 Conclusion 147

Appendices **150**

Appendix 4.A Notations and Preliminaries 150
Appendix 4.B Intermediate Results 151
Appendix 4.C Asymptotics of $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$ 167
Appendix 4.D Proof of Theorem 4.1 176
Appendix 4.E Equivalence of the Technical Assumptions — Assumption 2.5 and Assumption 3.3(a) 177
Appendix 4.F Proof of Theorem 4.2 186

II Least Trimmed Squares **187**

5 Introduction for Part II **188**

5.1 The Sensitivity of OLS to Outliers 188
5.2 Robust Regressions 190
5.3 The Error Term Distribution and the Normality Test 192
5.4 The Least Trimmed Squares Method and Model Selection 193

6 Normality Test for LTS Regressions **196**

6.1 Introduction 196
6.2 The LTS Location-Scale Model 197
6.3 The Normality Test 200
6.4 Simulation — The Normality Test in the LTS Location-Scale Model . . 202
6.5 Conclusion 208

Appendices **209**

Appendix 6.A Proof of Theorem 6.1 209
6.A.1 Limiting Distribution Conditional on δ 209

6.A.2 Vanishing Estimation Error 217
6.A.3 Limiting Distribution Under Estimated δ 222
Appendix 6.B Monte Carlo Standard Errors for Simulations 222

7 Model Selection in LTS Regressions 226
7.1 Introduction 226
7.2 The General LTS Model 227
7.3 The Location-Scale Model and the Model Selection Problem 228
7.3.1 The Information Criteria Method 231
7.3.2 The Cumulant Based Normality Test Statistic Method 234
7.4 Simulation — Consistent Model Selection 236
7.4.1 A Location-Scale Model 236
7.4.2 A Linear Regression Model 241
7.5 Conclusion 243

Appendices 245
Appendix 7.A Proof of Theorem 7.1 245
Appendix 7.B Proof of Theorem 7.2 247
Appendix 7.C Proof of Theorem 7.3 249

Bibliography 250

LIST OF FIGURES

5.1 Star Data with OLS Fit 189
5.2 Star Data with OLS and LTS Fits 194

LIST OF TABLES

4.1 Data Generating Processes 138

4.2 Empirical Size of the Residual Autocorrelation Test $(T-p)R^2$ . . . 141

4.3 Size-corrected Power of the Residual Autocorrelation Test 141

4.4 Empirical Power of the Test Under Misspecification 142

4.5 Performance of the Test With Different Number of Lags 144

4.6 Size and Power under Violation of Assumption 2.5 146

6.1 DGP for ε — Normality Test 202

6.2 Exact Critical Values in Location Scale Models 204

6.3 Exact Critical Values for LS.2 with Varying Distances 206

6.4 Exact Critical Values for LS.5 with Varying Distances 207

6.5 Monte Carlo Standard Errors for Table 6.2 223

6.6 Monte Carlo Standard Errors for Table 6.3 224

6.7 Monte Carlo Standard Errors for Table 6.4 225

7.1 Comparison of Different $f(n)$ Penalties 233

7.2 DGP for ε — Model Selection 236

7.3 Simulated Bias and Standard Deviation for $\widehat{h}$ — Comparison 1 . . . 239

7.4 Logarithm Values 239

7.5 Simulated Bias and Standard Deviation for $\widehat{\gamma}$ — Comparison 1 . . . 239

7.6 Simulated Bias and Standard Deviation for $\widehat{h}$ — Comparison 2 . . . 240

7.7 Simulated Bias and Standard Deviation for $\widehat{\gamma}$ — Comparison 2 . . . 240

7.8 Simulated Bias and Standard Deviation for $\widehat{h}$ — Regression 242

7.9 Simulated Bias and Standard Deviation for $\widehat{\gamma}$ — Regression 242

General Introduction

The selection of an appropriate model among suitable candidates is crucial to applied statistical studies. One needs to determine the optimal mathematical structure to describe a system, which involves choosing the right functional form of the model and identifying the correct number of parameters to be estimated. The body of this dissertation consists of two parts, which explore different model selection issues under two contexts. The first part contributes to the econometrics literature on misspecification tests for non-linear time series models, and the second is an excursion into the field of robust estimation.

Part I comprises a collection of essays completed over the first three years of my DPhil study, which extends the results in my MPhil thesis. Together, these essays represent the results of my research on an auxiliary-regression-based test for serial correlation in non-linear time series models, which is motivated by Breusch (1978) and Godfrey (1978a,b). A key feature that characterises time series models is the serial dependence or persistence in the underlying variables. Under the assumption that the error terms are martingale differences, rejecting the null hypothesis of no residual autocorrelation would reflect the serial dependence properties of the underlying data generating process and indicate that the specified model is incompatible with data. One then faces the problem of either including more lag terms in the model or choosing a different functional form. The main goal of this part is to construct a simple and unified procedure for testing serial correlation that can exploit the possibly non-linear structure of time series data. The testing method studied is shown to be consistent

whether or not the model includes non-stationary variables.

There is a natural progression in the chapters in Part I. In Chapter 1, we discuss the problem of serial correlation and review recent literature on this topic. In Chapter 2, we propose the auxiliary-regression-based test that incorporates the non-linearity of the model and derive several asymptotically equivalent forms of the test statistic. Building on results contained in Chapter 2, Chapter 3 further simplifies analyses of the asymptotic behaviour of the test statistic, namely that the effects of non-stationary components on the test statistic die out asymptotically. It turns out that only the stationary components of the underlying data generating process affect the asymptotic properties of the test statistic. Chapter 4 derives the limiting distribution of the test statistic being a standard chi-square distribution. Numerical studies demonstrate the effectiveness of the proposed method.

Part II shifts away from serial correlation, and instead examines some research questions related to robust estimation. The standard ordinary least squares approach may be prone to bias and inconsistent estimates in the case of using data that contains significant outliers. This has motivated the extensive literature on various methods that are 'robust', or more precisely, insensitive to the presence of outliers. In this part, we focus on a robust estimation method called the least trimmed squares (LTS), which was originally introduced by Rousseeuw (1984). In this method, the estimates of regression parameters are obtained by fitting sub-samples, which requires pre-setting the number of outliers to be trimmed.

The purpose of Part II is twofold. The first purpose is to study the omnibus test for normality in the LTS framework, which is based on skewness and kurtosis. Assuming that the number of outliers is known, Chapter 6 extends the well-known Jarque and Bera (1980, 1987) test to a location-scale LTS model where the non-outlying error terms are assumed to be normally distributed. The second purpose is to derive some limit properties for consistent estimation of the true proportion of non-outlying observations in the data, which is usually assumed known in the literature. This is a model

selection problem specific to the LTS context. In Chapter 7, we consider two different approaches — one is information criteria based and the other is testing based — to evaluate the performance of different candidate models and choose the optimal one. Theoretical properties of the proposed methods are investigated through Monte Carlo simulations.

There are two main distinctions between the analyses in Part I and Part II, which lie in the number of candidate models under comparison and the relationship between them. In Part I, we consider model selection between two nested models, whereby the model under the null hypothesis is the restricted model and the model under the alternative hypothesis is the unrestricted model. On the other hand, in Part II, we compare a sequence of models with different numbers of outliers. In general, these models are non-nested but maybe overlapping.

Part I

Testing for Residual Autocorrelation in General Non-linear Time Series Models

CHAPTER 1

Introduction for Part I

In time series analysis, autocorrelated disturbances of the estimated model are often suspected, and will cause misleading inferences. The objective of Part I is to propose a test procedure for testing the null hypothesis that the error terms are serially uncorrelated. In particular, we construct an auxiliary-regression-based test that is applicable to a general class of time series models which may be non-linear in parameters. We shall assume that the error term follows a martingale difference sequence, i.e., a stochastic process whose conditional expectation given the past information is zero.

The plan of this chapter is as follows. In Section 1.1, we briefly discuss the problem of serial correlation. Section 1.2 presents the idea of the auxiliary-regression-based test. Section 1.4 reviews developments on tests for serial correlation. Section 1.5 compares two types of correlation functions that are used to construct different serial correlation test statistics. Section 1.6 briefly discusses the main contributions of this part. The last section provides a brief overview of the remaining chapters in Part I.

1.1 Misspecification and Serial Correlation

Part I is concerned with whether certain assumptions of the error terms have been satisfied. Specification tests play a prominent role in econometric modelling, as all

estimation methods are based on a set of assumptions underlying the model of interest. In applications, these assumptions have to be tested, as any violation would bias estimates of standard errors and invalidate inferences, which would result in faulty conclusions. For instance, when making statistical inferences that are based on maximum likelihood estimates or ordinary least squares estimates of the parameters, we implicitly assume that we have appropriately specified an adequate and reliable parametric model. The properties of these estimators and subsequent inferential procedures can be very dependent upon the validity of these assumptions. In the presence of un-modelled autocorrelation, the ordinary least squares estimation produces biased estimates of the standard errors, and in such case, one option is to re-specify the model. It is therefore of paramount importance to check whether all the necessary assumptions are fulfilled before performing inferences.

We draw our attention to specific diagnostic checking for the presence of serial correlation, which occurs when the disturbances associated with past observations carry over into the future. Serially correlated residuals contain extra information which should itself be modelled. In standard regression analysis, it is often assumed that we have independently and identically distributed (i.i.d.) disturbances with mean zero and constant variance. This implies that after identification and estimation of the parameters, a well-fitted model should generate asymptotically uncorrelated residuals. Thus, it is a standard practice to check whether the residuals display serially correlated behaviour after running a regression. Residual autocorrelation can arise because of the dependence structure of disturbances or the omission of relevant variables. Although the former does not necessarily lead to bias or inconsistent estimates of parameters, it does inflate the estimated standard errors, which results in unreliable hypothesis testing for restrictions on the parameters. The latter is the most common cause of residual serial correlation and may render parameter estimates bias and inconsistent. It is often the case where one or more lagged regressors and/or other important variables have been incorrectly omitted from the model equation. Alternatively, it could also be the

case where included variables appear in an incorrect functional form.

1.2 Auxiliary Regression

In this section, we briefly explain the intuition and motivation behind our auxiliary-regression-based test.

Consider the following non-linear regression model,

$$y_t = g(x_t; \boldsymbol{\theta_0}) + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{1.1}$$

where $\boldsymbol{\theta_0}$ is an unknown parameter vector in $\mathbb{R}^q$ and $g(x; \boldsymbol{\theta}) : \mathbb{R}^k \times \mathbb{R}^q \mapsto \mathbb{R}$ is a pre-specified non-linear function. The k-dimensional vector of regressors, x_t, may consist of stationary and/or non-stationary variables. The error term $\{\varepsilon_t\}_{t=1}^T$ is assumed to be a martingale difference sequence with respect to the filtration $\mathscr{F}_t$ in that $\mathbb{E}(\varepsilon_t | \mathscr{F}_{t-1}) = 0$. A consequence of the martingale difference assumption is that the errors are serially uncorrelated.

Suppose we are interested in testing the null hypothesis that the errors in the model equation are serially uncorrelated against the alternative that they follow a first-order autoregressive process. That is, we test the hypothesis $\mathsf{H}_0 : \rho = 0$ against the alternative $\mathsf{H}_1 : \rho \neq 0$ in the first-order autoregressive (AR) disturbance process

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \qquad t = 2, \ldots, T, \tag{1.2}$$

where $\{u_t\}$ is a martingale difference sequence with respect to $\mathscr{F}_t$.

Positing that we have obtained least squares estimates, $\widehat{\boldsymbol{\theta}}$, for the parameters and the corresponding residuals,

$$\widehat{\varepsilon}_t = y_t - g(x_t; \widehat{\boldsymbol{\theta}}), \qquad t = 1, \ldots, T, \tag{1.3}$$

our test relies on the construction of a linear auxiliary regression,

$$\widehat{\varepsilon}_t = \alpha\widehat{\varepsilon}_{t-1} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = 2,\ldots,T. \tag{1.4}$$

The regressors $\dot{g}_{\boldsymbol{\theta}}(\cdot;\widehat{\boldsymbol{\theta}})$ are the first-order partial derivatives of the function $g(\cdot;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at $\widehat{\boldsymbol{\theta}}$. Based on the sample multiple correlation, R^2, between the contemporaneous residual $\widehat{\varepsilon}_t$ and all the regressors included in the auxiliary regression, we can test the null hypothesis of serially uncorrelated errors using the test statistic TR^2. We show here that the test statistic has a χ^2 distribution with one degree of freedom. This auxiliary regression is for instance implemented in PcGive (Doornik and Hendry, 2013, Section 18.5.10).

It is worth pointing out that the auxiliary-regression test statistic, TR^2, implicitly assumes the orthogonality of the contemporaneous residual, $\widehat{\varepsilon}_t$, and the first-order partial derivatives, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. This follows immediately from the first-order conditions of the least squares estimation. However, this orthogonality holds only in an asymptotic sense, because one less observation is used in the auxiliary regression compared to the model equation. When the sample size is relatively small, say $T = 25$, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ and $\widehat{\varepsilon}_t$ may not necessarily be orthogonal to each other. Therefore, it is necessary to include $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ in the auxiliary regression so as to control for the size and power of the proposed test statistic.

The intention to perform a hypothesis test for disturbance autocorrelation based on auxiliary regressions is inspired by Breusch (1978) and Godfrey (1978a,b). Under the standard linear regression framework with stationary regressors, they discuss auxiliary-regression test statistics, which, in the one-lag case, coincide with our proposed statistic. They show that the auxiliary-regression test statistic can be computed directly from a two step procedure. The first step is to estimate the linear model

$$y_t = \boldsymbol{\theta}'x_t + \varepsilon_t, \qquad t = 1,\ldots,T, \tag{1.5}$$

where x_t is a k-dimensional vector, by ordinary least squares regression and obtain the corresponding residuals, $\widehat{\varepsilon}_t$. Next, regress the residual, $\widehat{\varepsilon}_t$ on its first lag, $\widehat{\varepsilon}_{t-1}$ as well as all the explanatory variables x_t,

$$\widehat{\varepsilon}_t = \alpha\widehat{\varepsilon}_{t-1} + \boldsymbol{\beta}' x_t + e_t, \qquad t = 2, \ldots, T.$$

The test statistic is then T times the coefficient of determination, R^2, from this auxiliary regression.

The differences between Breusch and Godfrey's auxiliary regression and our auxiliary regression come from the functional form of g and the types of regressors considered. In their analysis,

$$g(x_t; \boldsymbol{\theta}) = \boldsymbol{\theta}' x_t,$$

which is linear in parameters. So their auxiliary regression does not depend on the estimators $\widehat{\boldsymbol{\theta}}$, since

$$\dot{g}_{\boldsymbol{\theta}} = x_t.$$

And the regressors under consideration are strictly stationary.

On the contrary, our auxiliary regression does depend on the estimators $\widehat{\boldsymbol{\theta}}$, through the first-order partial derivatives of the non-linear function, $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$. The idea to include the $\dot{g}_{\boldsymbol{\theta}}$ functions is inspired by the software package PcGive [Doornik and Hendry (2013)] for OxMetrics. This runs the auxiliary regression of the Breusch-Godfrey test in non-linear models with $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$. Here, we derive the theory for it. Besides, we allow both stationary and non-stationary explanatory variables in our non-linear model.

Breusch and Godfrey referred to their test statistic as the Lagrange multiplier (LM) type test statistic, since the standard likelihood-based LM test statistic, and its equivalence, the score test statistic, can be characterised as TR^2 in linear models [see Breusch (1978)]. Note that the standard likelihood-based LM test statistic requires differentiation of the likelihood function under the alternative hypothesis. In the non-linear

framework, it is, however, left unclear what the model, and hence the likelihood function and the score are under the alternative H_1. This ambiguity with respect to the statistic vocabulary is enhanced because the number of observations differs in equation (1.2) and (1.3). Therefore, although this type of test is named the LM test in the literature, we shall instead refer to our test as an auxiliary-regression-based test.

Another motivation for our auxiliary-regression-based test is based on an artificial regression called Gauss-Newton regression (GNR); see MacKinnon (1992), and Davidson and MacKinnon (1993) for a full treatment of the GNR and artificial regression approaches for computing misspecification test statistics. Davidson and MacKinnon (1990) addressed that we can test any type of equality restrictions on the parameters of a non-linear model by evaluating the GNR at any root-T consistent estimates, i.e., estimators that are consistent and converge at a rate of $T^{1/2}$. It is shown that the test statistic is simply T times the R^2 from the GNR. They proved that such test statistics calculated based on the GNR are asymptotically equivalent to variants of the likelihood-based LM test statistic. They suggested constructing an auxiliary regression of the type in equation (1.4) to analyse if there are any relevant explanatory variables for the estimated residuals from the model equation under the null hypothesis, while explicitly taking the form of this model equation into account.

1.3 Overview of Some Key Results

In the remainder of this thesis, we shall consider a more general alternative hypothesis compared to the one discussed in the previous section. Taking the null hypothesis to be that of no serial correlation throughout, the proposed test has power against the alternative of serial correlation up to the p-th order.

To perform the test, we construct a more general auxiliary regression than that in equation (1.4)

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}' \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T, \tag{1.6}$$

where $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ is a p-dimensional vector of lag residuals, so $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} = (\widehat{\varepsilon}_{t-1}, \widehat{\varepsilon}_{t-2}, \ldots, \widehat{\varepsilon}_{t-p})'$. We test the null hypothesis of no serial correlation by examining whether the vector $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is zero. This is done by evaluating whether the test statistic TR^2 is sufficiently close to zero, where here the R^2 is the sample multiple correlation between $\widehat{\varepsilon}_t$ and the regressors $(\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}, \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})')$.

In Chapter 2, we show that this R^2 is asymptotically equivalent to a partial R^2 between $\widehat{\varepsilon}_t$ and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ after controlling for $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$. This means that testing the joint significance of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in equation (1.6) is asymptotically equivalent to testing the significance of the former alone. This gives rise to an (asymptotically) equivalent form of the auxiliary regression,

$$\widehat{\varepsilon}_t = \tilde{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})] + \tilde{e}_t, \qquad t = p+1, \ldots, T, \tag{1.7}$$

where $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]$ is the vector of residuals from regressing $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ on $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$. In other words, the R^2 between $\widehat{\varepsilon}_t$ and $(\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}, \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})')$ is asymptotically the same as the R^2 between $\widehat{\varepsilon}_t$ and $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]$.

We also show that replacing the contemporaneous residual on the left hand side (LHS) of equation (1.7) with ε_t does not affect the null asymptotic distribution. This implies the following asymptotically equivalent form of the auxiliary regression

$$\varepsilon_t = \dot{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})] + \dot{e}_t, \qquad t = p+1, \ldots, T. \tag{1.8}$$

Hence, the R^2 between ε_t and $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]$ is asymptotically equivalent to that between $\widehat{\varepsilon}_t$ and $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]$, which in turn is asymptotically equivalent to that between $\widehat{\varepsilon}_t$ and $(\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}, \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})')$.

In Chapter 3, we further exploit the ability to separate the stationary and non-stationary components of the model through a normalisation matrix. In particular, we show that we can decompose the q-dimensional vector of first-order partial derivatives, $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$, into a d_u-dimensional vector involving all the stationary components

$\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ and a d_v-dimensional vector consists of non-stationary components $\dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, where $d_u + d_v = q$. So, equation (1.8) can be rewritten as

$$\varepsilon_t = \boldsymbol{\alpha}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}),\dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \grave{e}_t, \qquad t = p+1,\ldots,T. \tag{1.9}$$

We demonstrate that the non-stationary components do not contribute to the limiting distribution of the test statistic, which results in another asymptotically equivalent form of the auxiliary regression

$$\varepsilon_t = \breve{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \breve{e}_t, \qquad t = p+1,\ldots,T, \tag{1.10}$$

where $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ are the residuals from regressing $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ on $\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ only and the non-stationary components $\dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ have been dropped. Therefore, the R^2 between ε_t and $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ is asymptotically equivalent to that between ε_t and $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$.

Chapter 4 is concerned with the limiting distribution of the test statistic. The results in Chapters 2 and 3 indicate that the TR^2 from the auxiliary regression (1.6) is asymptotically equivalent to that computed using equation (1.10). Then, using equation (1.10), we show that the proposed test statistic follows a χ^2 distribution with p degrees of freedom asymptotically.

1.4 Literature Review

This section is split into two subsections to review developments in the literature on residual autocorrelation tests for linear stationary models and non-linear or non-stationary models separately.

1.4.1 Linear Stationary Models

The rich literature concerned with testing for serial correlation in linear regressions started from the 1940s, when the important work of Cochrane and Orcutt (1949) brought the serial correlation problem in linear models to the attention of econome-

tricians. Then followed the influential work of Durbin and Watson (1950, 1951), in which a test for first-order autoregression in univariate linear models is constructed. Recall the linear model established in equation (1.5). The Durbin-Watson (DW) test statistic for the residuals $\widehat{\varepsilon}_t$ is calculated using

$$DW = \frac{\sum_{t=2}^{T}(\widehat{\varepsilon}_t - \widehat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T}\widehat{\varepsilon}_t^2},$$

which takes the value between 0 and 4. Although this classical test statistic is perhaps the most widely used procedure in econometric textbooks [see Hayashi (2000), Dougherty (2011), Greene (2012), and Wooldridge (2013) for details], there are some caveats about its applicability. First, the DW statistic cannot be applied in models without an intercept, or models with lagged dependent variables as it requires explanatory variables to be strongly exogenous. Second, the test might be inconclusive if the value of the test statistic falls in between the lower and upper critical values. Lastly, the DW statistic is only applicable to test for first-order autocorrelation.

There are various extensions and refinements to the original DW statistic. Durbin (1970) proposed a modification to the original DW statistic, which is known as Durbin's h-test, to take account of the fact that in the presence of lagged dependent variables, the DW test statistic tends to bias towards the null hypothesis [Nerlove and Wallis (1966)]. Later, King (1985) and Inder (1986, 1990) suggested more powerful modifications to the DW h-test. For models fitted through the origin, Farebrother (1980) provided tables of bounds that can be used for the DW statistic. Concerning the inconclusive range of the DW statistic, various papers proposed approximations to the critical values; see, for instance, Durbin and Watson (1971), Harrison (1972), Ali (1984), Jeong (1985), and Theil and Shonkwiler (1986) for details. While the original DW statistic was formulated to test against the specific alternative hypothesis of first-order autocorrelation, Wallis (1972) modified the procedure to test for fourth-order autocorrelation. Vinod (1973) and Ali (1987) further generalised the DW statistic to test for

autocorrelation of any order. Silvapulle and King (1993), and more recently, Srianan-thakumar and King (2006) extended the procedure to test autoregressive disturbances against moving-average disturbances.

Alternative to the DW test, there are three types of tests for serial correlation favoured in the literature: the portmanteau-type tests based on Box and Pierce (1970) (hereafter BP) lack-of-fit test, the so-called LM tests addressed in Breusch (1978) and Godfrey (1978a,b) that are based on auxiliary regressions as in equation (1.4), and the LM tests that are associated with restricted maximum likelihood (i.e. likelihood equations with Lagrangian multipliers) estimation explored by Aitchison and Silvey (1958, 1960) and Silvey (1959). A brief review of some important contributions to these types of tests is provided below.

Box and Pierce (1970) addressed test procedures for linear autoregressive regressions without exogenous regressors. Under the assumption that the error terms are independently normally distributed with mean zero and constant variance, they introduced the portmanteau-type test statistic

$$Q_p = T \sum_{h=1}^{p} \widehat{r}_h^2,$$

where

$$\widehat{r}_h = \frac{\sum_{t=h+1}^{T} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-h}}{\sum_{t=1}^{T} \widehat{\varepsilon}_t^2}.$$

Here $\widehat{r}_h$ is the autocorrelation coefficient between $\widehat{\varepsilon}_t$ and $\widehat{\varepsilon}_{t-h}$ for $h = 1, \ldots, p$, and Q_p is asymptotically chi-square distributed with p degrees of freedom under the null hypothesis. This test statistic allows checking for higher order autocorrelations in stationary and linear autoregressive moving average (ARMA) models with Gaussian errors. However, results of Davies et al. (1977) suggested that the finite sample exact distribution differs from that implied by the asymptotic theory and a more accurate scaling is needed. Considering linear models, Ljung and Box (1978) later improved the BP test statistic to be better approximated by a chi-squared distribution for smaller

sample sizes. Their test statistic takes the form

$$\tilde{Q}_m = T(T+2)\sum_{h=1}^{p}\frac{\widehat{r}_h^2}{T-h}, \tag{1.11}$$

so that the squared autocorrelation at lag h is weighted by $\frac{T+2}{T-h}$. Monti (1994) proposed a test statistic which replaces the residual autocorrelation in the Ljung-Box statistic by the standardised partial autocorrelation.

As mentioned above, Breusch (1978) and Godfrey (1978a,b) developed an auxiliary-regression-based LM test statistic using the residuals from equation (1.5), assuming stationary processes and normality of disturbances. They considered the auxiliary regression

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}' x_t + \sum_{h=1}^{p}\beta_h\widehat{\varepsilon}_{t-h} + v_t, \qquad t = p+1,\ldots,T.$$

Under the null hypothesis of no serial correlation up to order p, the test statistic TR^2 is asymptotically chi-square distributed with p degrees of freedom. This statistic allows us to test against both autoregressive and moving-average types alternatives, and works under the presence of lagged dependent variables. Concerning the finite sample distribution of the test statistic, Belsley (1997, 2000) provided a small-sample correction and a transformed statistic to deal with small sample distortions. Kyriazidou (1998) extended the test to multivariate linear stationary models with conditional heteroskedasticity. Edgerton and Shukur (1999) generalised the auxiliary-regression-based test to cover systems of equations. Godfrey (2007) further extended the results to cope with models with non-normal disturbances.

There is a growing body of literature favouring the efficient scores procedure. Poskitt and Tremayne (1980, 1981) and Pötscher (1985) developed diagnostic tests against ARMA disturbances using the likelihood ratio principle for linear stationary models.

1.4.2 Non-stationary and/or Non-linear Models

Compared to the extensive literature on linear stationary models, there are not many studies concerning non-stationary and/or non-linear models.

Regarding the DW statistic, Dufour and King (1991) refined the test procedure so that it can be applied to both stationary and non-stationary linear models. White (1992) then approximated the DW statistic to cope with first-order autoregressive errors in non-linear models.

Extending analyses of the portmanteau-type test statistic to non-stationary autoregressive models, Shin and Lee (1996) considered a portmanteau test under the assumption of i.i.d. errors for linear models. A more recent paper by Ling et al. (2013) studied a portmanteau test for model adequacy under the context of non-stationary ARMA models by making the same i.i.d. assumption on the disturbances. Fitting the portmanteau test statistic into non-linear dynamic regressions, Li (1992) derived the asymptotic distribution of the residual auto-covariances in univariate stationary models. Extending the work of Li (1992), Chabot-Hallé and Duchesne (2008) obtained a modified portmanteau statistic to account for martingale difference errors. In the same year, Chen (2008) proposed a portmanteau test for generalised autoregressive conditional heteroskedasticity (GARCH) type models with i.i.d. standardised innovations, which can be applied to both linear and non-linear regression functions. A later study by Fisher and Gallagher (2012) introduced weighted versions of the Ljung-Box and Monti statistics in the context of non-linear stationary ARMA models whose error terms are i.i.d. with mean zero and variance one. Their test statistics attach more weight to lower order lags and put less weight on higher order lags. A more recent paper by Wang and Zhu (2018) extended the portmanteau test to non-linear models with an endogenous random walk regressor, which implies that their functional form cannot be adjusted to include stationary components.

Considering the auxiliary-regression-based LM tests, Nielsen (2006b) proved that the test statistic is asymptotically chi-squared distributed regardless of the location of

characteristic roots in linear vector-autoregressive models with martingale difference innovations. Brüggemann et al. (2006) explored the test in co-integrated vector error correction models with i.i.d. errors.

Within the scope of likelihood-based LM tests, Andrews and Ploberger (1996) (hereafter AP) introduced a sup LM test and an exponential average LM test which can be applied to both linear and non-linear stationary models in the presence of i.i.d. errors, with the exception of deterministically trending non-linear regression functions. Based on the score test statistic, the AP tests consider the null hypothesis of serially uncorrelated errors against the ARMA$(1,1)$ alternative. However, a strong exogeneity assumption is imposed on the regressors in the sense that, conditional on lagged innovations as well as past, current and future values of the regressors, the innovations have mean zero. Hence, the test is not suitable for dynamic models where lagged dependent variables are included as regressors. Nankervis and Savin (2010) later generalised the AP tests to cope with models with uncorrelated but statistically dependent innovations, maintaining the same strong exogeneity requirement on the explanatory variables. Recently, Cao et al. (2014) considered a likelihood-based LM test with the score statistic for detecting disturbance autocorrelation in non-linear models on the basis of i.i.d. errors with scale mixtures of skew-normal distributions.

1.5 Correlation Functions

In this section, we discuss the difference between the sample correlation function on which our test statistic is based and the sample autocorrelation function that is used in the portmanteau test statistic.

The residual autocorrelation in the portmanteau test statistic can be computed using a Yule-Walker type approach. Consider the following mean zero AR(p) model for the error terms:

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_p \varepsilon_{t-p} + w_t. \tag{1.12}$$

In order to derive the autocorrelation function, we first multiply both sides of the above autoregressive equation by ε_{t-h} ($h \leq p$) to obtain

$$\varepsilon_{t-h}\varepsilon_t = \phi_1\varepsilon_{t-h}\varepsilon_{t-1} + \phi_2\varepsilon_{t-h}\varepsilon_{t-2} + \cdots + \phi_p\varepsilon_{t-h}\varepsilon_{t-p} + \varepsilon_{t-h}w_t.$$

Taking expectations gives

$$\gamma(h) = \phi_1\gamma(h-1) + \phi_2\gamma(h-2) + \cdots + \phi_p\gamma(h-p),$$

where $\gamma(h) = \mathbb{E}(\varepsilon_{t-h}\varepsilon_t) = Cov(\varepsilon_{t-h}, \varepsilon_t)$ and Cov denotes the covariance. Note that $\mathbb{E}(\varepsilon_{t-h}w_t) = 0$ because the random perturbation w_t is assumed to be unrelated to previous values of the process. When h is zero, we have

$$\gamma(0) = Var(\varepsilon_t),$$

where Var denotes the variance. The function

$$r_h = \frac{\gamma(h)}{\gamma(0)} = \frac{Cov(\varepsilon_{t-h}, \varepsilon_t)}{Var(\varepsilon_t)}$$

is sometimes called the "autocorrelation function", although it is a scaled "auto-covariance function", since it implicitly assumes $Var(\varepsilon_t) = Var(\varepsilon_{t-h})$. Its sample analogue,

$$\widehat{r}_h = \frac{\sum_{t=h+1}^{T} \widehat{\varepsilon}_t\widehat{\varepsilon}_{t-h}}{\sum_{t=1}^{T} \widehat{\varepsilon}_t^2}, \tag{1.13}$$

is used to construct the portmanteau test statistic, see equation (1.11).

Notice that this autocorrelation function is different from the usual definition of the correlation between ε_t and its h-th lag ε_{t-h}. The usual definition of correlation coefficient requires us to standardise the covariance between two variables by each of

their standard deviations,

$$\rho_h = \frac{Cov(\varepsilon_{t-h}, \varepsilon_t)}{\sqrt{Var(\varepsilon_t)Var(\varepsilon_{t-h})}},$$

whose sample analogue is

$$\widehat{\rho}_h = \frac{\sum_{t=h+1}^{T} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-h}}{\sqrt{\sum_{t=h+1}^{T} \widehat{\varepsilon}_t^2 \sum_{t=h+1}^{T} \widehat{\varepsilon}_{t-h}^2}}. \tag{1.14}$$

Although the difference between ρ_h and r_h does not matter when we have stationary time series, as $Var(\varepsilon_t) = Var(\varepsilon_{t-h})$, Nielsen (2006a) showed that the difference can be large for non-stationary time series. He illustrated that the usual correlation, ρ_h, indicates the persistence in the data clearer than the autocorrelation function, r_h, when non-stationarity presents. As a result, the portmanteau test, which makes use of the sample autocorrelation function, $\widehat{r}_h$, could be problematic to the extent that the power of the test would bias towards 0 as the magnitude of the slope coefficients in equation (1.12) grows to infinity. In contrast, our test statistic is constructed using the usual sample correlation function in equation (1.14) (after partialling out the effect of $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$). Therefore, one advantage of our proposed test is that while having the same asymptotic behaviour as the portmanteau test statistic in the stationary world, the power of the test increases to 1 as the magnitude of ϕ in equation (1.12) increases in the non-stationary case.

1.6 Contribution of Part I

Testing for serial correlation in non-linear models has piqued our interest since models that are non-linear in parameters have emerged as an indispensable area of econometrics with its intensive use in economic and financial applications. However, there is little work being done in the literature that tests for serial correlation in non-linear and non-stationary models using an auxiliary regression. Nielsen (2006b) showed that the

test statistic TR^2 has the standard chi-squared distribution in linear VAR models that contain both stationary and non-stationary variables. To the best of our knowledge, there is no existing literature specifically taking both nonlinearity and non-stationarity into account in deriving the theory for the auxiliary regression approach. This gap is being filled in this thesis. We generalise the results in Nielsen (2006b), and show that even with nonlinear regression models, one could still conduct the standard chi-squared inferences in the presence of non-stationary regressors.

The auxiliary-regression approach is favoured in the current study because the test statistic TR^2 is relatively easy to calculate, as we simply compute the R^2 as a measure of fit in a linear auxiliary regression. Thus, we can avoid the heavy computations required to obtain explicit numerical solutions to the maximum likelihood equations in likelihood-based LM tests. Another advantage of the proposed test is that it allows for testing first and higher order serial correlation. Furthermore, our test can be applied to non-linear dynamic models of a general form (such as ARMA and ECM models, power function models, exponential models, etc.) and any explanatory variables of interest — stationary, non-stationary or even co-integrated variables — can be included in the regression. Simulation results suggest that the finite sample distribution of our test statistic can be well approximated by the asymptotic chi-square distribution regardless of the location of the unit root. In addition, as discussed in the previous section, the R^2 in our test statistic uses correlation functions of the type in equation (1.14) rather than equation (1.13). Thus, the size of our test statistic is approximately correct in a wider class of models than the portmanteau-type test statistic.

A further contribution of the proposed test is that the assumption of independent Gaussian errors is relaxed, and the constructed test statistic can be applied to non-linear models with a broader class of disturbance structures. Uncorrelated but dependent time series hold wide appeal in empirical economic and financial work because of the unpredictable nature of certain financial data after demeaning, for example stock prices, which is often referred to as the martingale difference sequence hypothesis.

Based on the martingale difference assumption, our test relies on the fact that, if the model is correctly specified, ε_t must be serially uncorrelated with $\varepsilon_{t-1}, \ldots, \varepsilon_1$, since past information does not have any impact on current prediction. The martingale difference assumption is less stringent than the classical i.i.d. assumption on the error terms, because a martingale difference sequence can be serially dependent although uncorrelated.

1.7 Organisation of Part I

The rest of Part I is structured as follows. In Chapter 2, we set up the model formally and construct the auxiliary-regression-based test for serial correlation. The test statistic is simplified and we show that the F-test for overall significance in the auxiliary regression is asymptotically equivalent to a partial F-test on the joint significance of a subset of slope parameters. We then dig deeper into the first-order partial derivative functions $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$ in Chapter 3, and show that they can be decomposed into stationary and non-stationary components. We further simplify the test statistic and prove that the explosive component has no impact on the asymptotic distribution of the test statistic. In Chapter 4, we derive the limiting distribution of the test statistic and present some preliminary Monte Carlo simulation results to characterise its finite sample performance.

CHAPTER 2

Asymptotic Equivalence to a Partial TR^2

2.1 Introduction

In this chapter, we prepare the ground for the analysis of an auxiliary-regression-based diagnostic test for serial correlation. On the basis of Breusch (1978) and Godfrey (1978a,b), we extend the auxiliary-regression-based test to incorporate non-linear regression models, while maintaining the linearity property (in parameters) of the auxiliary regression. As mentioned in Chapter 1, this type of misspecification test statistic is also favoured in Davidson and MacKinnon (1990, 1993) and MacKinnon (1992).

Take the general univariate non-linear regression model we proposed in the previous chapter,

$$y_t = g(x_t; \boldsymbol{\theta_0}) + \varepsilon_t, \qquad t = 1, \ldots, T,$$

where the innovation term is assumed to be a martingale difference sequence with respect to the filtration $\mathscr{F}_t$. The null hypothesis to be tested is that there is no autocorrelation in the error term, while the alternative hypothesis is that the error terms are serially correlated up to the p-th order.

Suppose we have run the non-linear least squares estimation and obtained the estimates of the parameters $\boldsymbol{\theta_0}$, denoted as $\widehat{\boldsymbol{\theta}}$, and the corresponding residuals $\widehat{\varepsilon}_t$. As given in equation (1.6), the linear auxiliary regression is constructed by regressing the contemporaneous residual $\widehat{\varepsilon}_t$ on its own lags up to the p-th order and the $\dot{g}_{\boldsymbol{\theta}}(\cdot; \widehat{\boldsymbol{\theta}})$

functions,

$$\widehat{\varepsilon}_t = \alpha_1\widehat{\varepsilon}_{t-1} + \alpha_2\widehat{\varepsilon}_{t-2} + \cdots + \alpha_p\widehat{\varepsilon}_{t-p} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T. \quad (2.1)$$

Again, $\dot{g}_{\boldsymbol{\theta}}(\cdot;\widehat{\boldsymbol{\theta}})$ are the first-order partial derivatives of the non-linear function $g(\cdot;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at the NLLS estimators $\widehat{\boldsymbol{\theta}}$. Testing the null hypothesis is then equivalent to testing whether all the elements of the vector $(\alpha_1,\alpha_2,\cdots,\alpha_p,\boldsymbol{\beta})'$ are zero.

Two asymptotically equivalent forms of the auxiliary regression (2.1) are derived in this chapter, under general assumptions on the data generating processes and the non-linear functional form. The key output is that testing for the overall significance in the auxiliary regression is asymptotically equivalent to testing the joint significance of a subset of slope parameters on lag residuals. That is testing whether the vector $(\alpha_1,\alpha_2,\cdots,\alpha_p,\boldsymbol{\beta})'$ is zero is the same as testing whether the sub-vector $(\alpha_1,\alpha_2,\cdots,\alpha_p)'$ is zero asymptotically. Stacking the p lag residuals into the p-dimensional vector $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, we derive an asymptotically equivalent form of the auxiliary regression which is given by

$$\widehat{\varepsilon}_t = \tilde{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \tilde{e}_t, \qquad t = p+1,\ldots,T, \quad (2.2)$$

where $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ are the residuals from the regression of $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ on $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. So testing the significance of $(\alpha_1,\alpha_2,\cdots,\alpha_p,\boldsymbol{\beta})'$ in equation (2.1) is asymptotically equivalent to testing the significance of $\tilde{\boldsymbol{\alpha}}$ in equation (2.2). We also show that replacing the residual $\widehat{\varepsilon}_t$ on the left hand side (LHS) of the above equation by the error term ε_t has no asymptotic influence on the proposed test.

This chapter proceeds as follows. Section 2 contains a discussion of the model and constructs the auxiliary-regression-based test. Section 3 describes the model assumptions and presents the main results of this chapter. Several examples linking the general assumptions to specific time series models are presented in Section 4. Some

directions of future works are given in the last section. All proofs are provided in the appendix.

2.2 The Model and Test Statistic

We observe the data $\{(y_t, x_t)\}_{t=1}^{T}$, where y_t is a random variable and $x_t \in \mathbb{R}^k$ is a random vector. We are interested in a non-linear regression model that could be represented as follows:

$$y_t = g(x_t; \boldsymbol{\theta_0}) + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{2.3}$$

where $g(\cdot)$ is some known real-valued function, $\boldsymbol{\theta_0}$ corresponds to a q-dimensional vector of unknown parameters that belongs to some parameter space $\Theta \subset \mathbb{R}^q$, and ε_t represents the disturbance. The random sequence $\{\varepsilon_t\}$ adapted to the filtration $\mathscr{F}_t$ is a martingale difference sequence with mean zero and variance σ_ε^2. The sequence of explanatory variables $\{x_t\}$ is $\mathscr{F}_{t-1}$-adapted, and the process x_t may consist of stationary and/or non-stationary components.

We are concerned with checking the assumptions that validate the model of interest, following a certain estimating procedure, and we are not interested in hypothesis testing on the estimates of $\boldsymbol{\theta_0}$. Therefore it is assumed that the parameters of model (2.3) has been estimated elsewhere by the non-linear least squares (NLLS) procedure or some other asymptotically equivalent estimation method. Specifically, we presume that we have obtained the $\widehat{\boldsymbol{\theta}}$ by minimising the sum of squared residuals with respect to $\boldsymbol{\theta}$ over the parameter space Θ,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, S_T(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min} \sum_{t=1}^{T} \Big\{ y_t - g(x_t; \boldsymbol{\theta}) \Big\}^2, \tag{2.4}$$

and the corresponding residuals are

$$\widehat{\varepsilon}_t = y_t - g(x_t; \widehat{\boldsymbol{\theta}}), \qquad t = 1, \ldots, T.$$

In what follows, we consider testing the null hypothesis that the the error term ε_t is not serially correlated against the alternative that the error term follows a pth-order autoregressive (AR) disturbance process (assuming p is fixed),

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \cdots + \rho_p\varepsilon_{t-p} + u_t = \boldsymbol{\rho}'\boldsymbol{\varepsilon_{t-1}} + u_t, \qquad t = p+1,\ldots,T, \quad (2.5)$$

where $\boldsymbol{\varepsilon_{t-1}} = (\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots, \varepsilon_{t-p})'$ is a p-dimensional vector consisting of lag error terms up to the p-th order, $\boldsymbol{\rho} = (\rho_1, \rho_2, \cdots, \rho_p)' \in \mathbb{R}^p$, and $\{u_t\}$ is a martingale difference sequence with respect to $\mathscr{F}_t$. In other words, we are interested in testing $\mathsf{H}_0 : \boldsymbol{\rho} = \mathbf{0}$, where $\mathbf{0}$ is a p-dimensional column vector with all elements equal to zero, against the alternative $\mathsf{H}_1 : \boldsymbol{\rho} \neq \mathbf{0}$.

Observe that, despite having T observations of the residuals, only $T - p$ equations are available for testing autocorrelation in the disturbances. Drawing inspiration from Breusch (1978) and Godfrey (1978a,b) as well as PcGive [Doornik and Hendry (2013)], we construct an auxiliary regression of the form

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\boldsymbol{\widehat{\varepsilon}_{t-1}} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T, \quad (2.6)$$

to compute the test statistic, where e_t is the error term, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are p-and q-dimensional vectors of unknown parameters, respectively. $\boldsymbol{\widehat{\varepsilon}_{t-1}}$ is a p-dimensional vector of lag residuals up to the p-th order. $\dot{g}_{\boldsymbol{\theta}}(\cdot;\widehat{\boldsymbol{\theta}})$ is a q-dimensional vector with the i-th element $\dot{g}_{\boldsymbol{\theta}}(\cdot;\widehat{\theta}_i) = [\partial g(x_t;\boldsymbol{\theta}) \setminus \partial\theta_i]_{\theta_i=\widehat{\theta}_i}$. Analogously, the $q \times q$ symmetric matrix $\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta})$ denotes the second-order partial derivatives of $g(\cdot;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ twice. We test the hypothesis that $\boldsymbol{\rho} = \mathbf{0}$ by testing whether the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are jointly zero.

In a non-linear world, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ is included in the auxiliary regression (2.6) to take into account the possibility that x_t has some explanatory power for $\boldsymbol{\widehat{\varepsilon}_{t-1}}$. Unless $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ and the lag residuals $\boldsymbol{\widehat{\varepsilon}_{t-1}}$ are uncorrelated, we will not reject the null hypothesis of no serial correlation often enough if the former is excluded from the auxiliary regression. When $g(x_t;\boldsymbol{\theta})$ is a linear function, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ is simply the regressors, x_t,

themselves. Consequently, equation (2.6) will have exactly the same regressors as the usual auxiliary regression for linear models.

Stack the regressors in the auxiliary regression (2.6) into a $(p+q)$-dimensional column vector, which, throughout this study, is denoted as $\widehat{\mathbf{H}}_t$. Therefore,

$$\widehat{\mathbf{H}}_t := \begin{bmatrix} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \end{bmatrix}. \tag{2.7}$$

Let $\widehat{var}(\widehat{\varepsilon}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2$ and $\widehat{var}(\widehat{\mathbf{H}}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_t\widehat{\mathbf{H}}_t'$ denote the sample variances of the residual $\widehat{\varepsilon}_t$ and $\widehat{\mathbf{H}}_t$, respectively. Let $\widehat{cov}(\widehat{\varepsilon}_t,\widehat{\mathbf{H}}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t\widehat{\mathbf{H}}_t$ denote the sample multiple covariance between $\widehat{\varepsilon}_t$ and the vector $\widehat{\mathbf{H}}_t$. The test statistic for the null hypothesis H_0 is

$$TR^2 = T\frac{\widehat{cov}(\widehat{\varepsilon}_t,\widehat{\mathbf{H}}_t)'[\widehat{var}(\widehat{\mathbf{H}}_t)]^{-1}\widehat{cov}(\widehat{\mathbf{H}}_t,\widehat{\varepsilon}_t)}{\widehat{var}(\widehat{\varepsilon}_t)}, \tag{2.8}$$

which is based on the sample of size $T-p$, and the squared sample multiple correlation R^2 of $\widehat{\varepsilon}_t$ and the vector $\widehat{\mathbf{H}}_t$.

2.3 Assumptions and Theorems

Two theorems regarding the test statistic are established in this section. The analysis involves introducing various assumptions on ε_t, the NLLS estimates $\widehat{\boldsymbol{\theta}}$, and the function $g(x_t;\boldsymbol{\theta})$.

The first assumption concerns the non-predictability of the disturbance term. This assumption suggests that if the null hypothesis were rejected, there must have been some un-modelled dependence structure of the time series considered in the regression model, since a martingale difference sequence must be serially uncorrelated. As we are interested in the asymptotic results, it is unnecessary to impose the normality assumption on the disturbance, and we only need to assume the existence of finite higher (conditional) moments. The martingale difference assumption implies that the

error term forms an uncorrelated sequence of random variables with mean zero.

Assumption 2.1. *Assume that* $\{\varepsilon_t\}$ *adapted to the filtration* $\mathscr{F}_t$ *is a martingale difference sequence such that*

(a) $\mathbb{E}(\varepsilon_t|\mathscr{F}_{t-1}) = 0$ *a.s.,*

(b) $\mathbb{E}(\varepsilon_t^2|\mathscr{F}_{t-1}) = \sigma_\varepsilon^2 > 0$ *a.s.,*

(c) $sup_t\mathbb{E}(|\varepsilon_t|^\varphi|\mathscr{F}_{t-1}) < \infty$ *a.s. for some* $\varphi > 2$.

In order to discuss the asymptotics, we need a normalisation matrix, denoted as $N_{T,\boldsymbol{\theta_0}}$, that is compatible with both stationary and non-stationary explanatory variables. For this chapter, it suffices to know that this normalisation matrix can be chosen so that $N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$ is stochastically bounded. For example, $N_{T,\boldsymbol{\theta_0}}^{-1}$ can be $T^{\frac{1}{2}}$ if the regressors are stationary, or T if the regressors are random walks/near unit roots. Besides, this normalisation matrix does not have to be deterministic. Suppose we have a multiple linear model such that $g(x_t;\boldsymbol{\theta}) = x_t'\boldsymbol{\theta}$. We can make $N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$ self-normalised by setting $N_{T,\boldsymbol{\theta_0}}^{-1}$ equal to $(\sum_{t=1}^T x_t x_t')^{\frac{1}{2}}$. In general, we make the following assumption.

Assumption 2.2. *Suppose* $N_{T,\boldsymbol{\theta_0}}$ *is a non-singular normalisation matrix which depends on both the number of observations* T *and the true parameter* $\boldsymbol{\theta_0}$. *Then*

(a) $\inf\{T : N_{T,\boldsymbol{\theta_0}}$ *is non-singular*$\} < \infty$ *a.s.,*

(b) $N_{T,\boldsymbol{\theta_0}}^{-1} = \mathbf{O}(T^\iota)$ *a.s. for some* $\iota > 0$.

The next assumption relates to the asymptotic behaviour of $\widehat{\boldsymbol{\theta}}$. We assume that $\widehat{\boldsymbol{\theta}}$ is a consistent estimator for the true parameter vector, $\boldsymbol{\theta_0}$. However, instead of assuming that the normalised vector, $N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$, is $\mathbf{O_p}(1)$ so that it converges in distribution, we here give it some flexibility to diverge at a slight rate.

Assumption 2.3. $N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = \mathbf{o_p}(T^\tau)$, *where* $0 < \tau < \frac{1}{4}$ *and* $\tau < \iota$.

This assumption allows for both stationary and non-stationary components and allows the normalised sequence to diverge slightly so as to incorporate slowly varying functions [see Phillips (2007) for details]. The idea is that, if we knew the exact rate of convergence, we could have assumed that directly. Here, we allow the normalised rate to be slightly loose, so we do not need to pin down the exact rate. This is because the slowly varying components are dominated by the power function.

To derive asymptotic results of the test statistic, we need to ensure that the first and second-order terms in a Taylor expansion of $g(\cdot;\boldsymbol{\theta})$ around $\boldsymbol{\theta_0}$ are under control. Due to the extensive use of Taylor's Theorem, we next impose smoothness requirements on the $g(\cdot;\boldsymbol{\theta})$ function such that it possesses first and second-order partial derivatives with respect to $\boldsymbol{\theta}$. In addition, these partial derivatives need to satisfy appropriate boundedness conditions.

Recall that the first-order partial derivative function $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta})$ is vector-valued and the second-order partial derivative function $\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta})$ is matrix-valued. In order to measure the size of vector and matrix spaces, we need to introduce a norm. Vector norms and matrix norms are the yardsticks with which we measure the convergence of the test statistic. In the appendix, we define the spectral norm of matrices and the Euclidean norm of vectors (see Appendix 2.B for details), and use them to prove the main result. However, it is worth noting that any pair of compatible matrix and vector norm can be used.

Assumption 2.4. *Suppose the convex parameter space is $\Theta \subset \mathbb{R}^q$, $\{x_t\}$ is $\mathscr{F}_{t-1}$-adapted, the true q-dimensional parameter vector is $\boldsymbol{\theta_0}$, and the function $g(x_t;\boldsymbol{\theta})$ is thrice continuously differentiable with respect to $\boldsymbol{\theta}$ on Θ. In addition, the functions $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ and $\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta_0})$ are continuous in x_t. Then*

(a) $\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\right\|^2 = O_p(1)$,

(b) $\inf\{T : \sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})' \text{ is non-singular}\} < \infty$ *a.s.*,

(c) For every $K > 0$, we have

$$\sum_{t=p+1}^{T} \sup_{\tilde{\boldsymbol{\theta}}:\left\|N_{T,\boldsymbol{\theta_0}}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta_0})\right\|\leq K} \left\|N_{T,\boldsymbol{\theta_0}}' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\tilde{\boldsymbol{\theta}}) N_{T,\boldsymbol{\theta_0}}\right\|^2 = o_p(T^{-4\tau}).$$

Assuming that Θ is a convex subset of $\mathbb{R}^q$ and that $g : \mathbb{R}^k \times \Theta \to \mathbb{R}$ is thrice continuously differentiable on Θ validate the Mean Value Theorem [see Flett (1958) for details], and therefore allow us to take Taylor expansion up to the second-order. *Assumption 2.4 (b)* ensures that the matrix $\sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})'$ is finitely invertible. *Assumption 2.4 (c)* guarantees that higher order terms of the $\ddot{g}$ function are so small that they are insignificant to our analysis. In fact, *Assumption 2.4* serves as an identification condition for the parameter vector and implies that estimates of $\boldsymbol{\theta_0}$ are well defined when the sample size is large enough.

The next assumption is a technical assumption that we have to maintain for now to facilitate the proof of the second theorem in this chapter.

Assumption 2.5. *Let*

$$[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} - \sum_{s=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{s-1}}\dot{g}_{\boldsymbol{\theta}}(x_s;\widehat{\boldsymbol{\theta}})'[\sum_{s=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_s;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_s;\widehat{\boldsymbol{\theta}})']^{-1}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}).$$

Then

$$\{\sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\varepsilon_t = \mathbf{O_p}(1).$$

Before diving into the technical details of *Assumption 2.5*, notice that this assumption is made to preclude certain cases from analysis. For example, consider the following AR(1) model

$$y_t = \theta y_{t-1} + \varepsilon_t,$$

and suppose we are interested in testing against the first-order alternative. In this example, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ and $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ become y_{t-1} and $[\widehat{\varepsilon}_{t-1}|y_{t-1}]$, respectively. When

$\theta = 0$ and provided that Assumption 2.3 is satisfied so $\widehat{\theta}$ is a consistent estimator for θ, $[\widehat{\varepsilon}_{t-1}|y_{t-1}]$ becomes $[\varepsilon_{t-1}|\varepsilon_{t-1}]$, which are the residuals from regressing ε_{t-1} on itself, and are therefore zeros. Consequently,

$$\sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]' = \sum_{t=p+1}^{T} [\varepsilon_{t-1}|\varepsilon_{t-1}]^2 = 0.$$

So the inverse of $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'$ does not exist in this case and *Assumption 2.5* is violated. This assumption thus excludes the scenario where $\theta = 0$ from our analysis. *Assumption 2.5* has to be maintained for now to unify the arguments in the proof, but it is possible to relax it, as supported by the simulation results reported at the end of Section 4.4 in Chapter 4.

Now let's take a closer look at *Assumption 2.5*. First, observe that the left-hand-side (LHS) consists of two parts — a vector component, $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\varepsilon_t$, and a matrix component, $\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}$. Consider the vector term, which involves the contemporaneous error term, ε_t. Under *Assumption 2.1* and *2.4*, both $\boldsymbol{\varepsilon_{t-1}}$ and x_t are adapted to the filtration $\mathscr{F}_{t-1}$. Therefore, given the martingale difference assumption, the vector $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\varepsilon_t$ is approximately a martingale, which is likely to be stochastically bounded and thus, not restricted by *Assumption 2.5*. Instead, it is the matrix term, $\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}$, that is bound by this assumption. As we shall see below, *Assumption 2.5* is the most difficult assumption to deal with. It imposes some restrictions on the data generating process underlying the model, so that the relationship between the first-order partial derivatives, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, and the lagged residuals, $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, cannot be of certain form.

It might be easier to understand this assumption in the simple case where we test against the alternative that the error term follows an $AR(1)$ process. Notice that the term $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ is essentially a vector of least squares residuals from regressing the vector, $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, on $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. Suppose for now that $p = 1$. Thus, the p-dimensional vector, $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, which consists of lag residuals up to the p-th order, reduces to the scalar

$\widehat{\varepsilon}_{t-1}$. In this case, the $p \times p$ covariance matrix, $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'$, becomes a scalar, $\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]^2$, which is essentially the sum of squared residuals from the regression of $\widehat{\varepsilon}_{t-1}$ on $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. *Assumption 2.5* in this case places the restriction that $\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]^2$ is different from zero, which in turn prohibits the residuals from the regression of $\widehat{\varepsilon}_{t-1}$ on $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ being zero exactly. Consequently, this implies that $\widehat{\varepsilon}_{t-1}$ and $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ cannot be perfectly linearly correlated. *Assumption 2.5* is a restrictive assumption, and we will work on to relax it in the future. (See also the next section where we discuss this technical assumption in more detail using specific model examples.)

Having established assumptions that govern the asymptotic behaviour of the denominator and the numerator of the test statistic TR^2, we are now ready to formulate the main theorems. The mathematical details of the proofs are relegated to the appendix.

Theorem 2.1. *Consider the auxiliary regression*

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T.$$

If Assumptions 2.1 to 2.4 are satisfied, then

$$TR^2 = T\frac{\widehat{cov}[\widehat{\varepsilon}_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\{\widehat{var}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\}^{-1}\widehat{cov}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \widehat{\varepsilon}_t|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{\widehat{var}(\widehat{\varepsilon}_t)} + o_p(1).$$

Recall the test statistic

$$\begin{aligned} TR^2 &= T\frac{\widehat{cov}(\widehat{\varepsilon}_t, \widehat{\mathbf{H}}_t)'[\widehat{var}(\widehat{\mathbf{H}}_t)]^{-1}\widehat{cov}(\widehat{\mathbf{H}}_t, \widehat{\varepsilon}_t)}{\widehat{var}(\widehat{\varepsilon}_t)} \\ &= T\frac{\widehat{cov}(\widehat{\varepsilon}_t, \begin{bmatrix} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \end{bmatrix})'\{\widehat{var}(\begin{bmatrix} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \end{bmatrix})\}^{-1}\widehat{cov}(\begin{bmatrix} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \end{bmatrix}, \widehat{\varepsilon}_t)}{\widehat{var}(\widehat{\varepsilon}_t)}. \end{aligned}$$

Recall that the null hypothesis of no serial correlation can be tested by evaluat-

ing the joint significance of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the auxiliary regression. Therefore, the test statistic TR^2 is equivalent to an F-statistic on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. *Theorem 2.1* implies that an F-test on the overall significance of the slope parameters in the auxiliary regression is asymptotically equivalent to a partial F-test on the coefficients of the lag residuals alone. That is, the F-statistic on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ has the same limiting distribution as the partial F-statistic on $\boldsymbol{\alpha}$ alone. Under *Theorem 2.1*, testing for serial correlation with the auxiliary regression of the form given in equation (2.6),

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T,$$

is asymptotically equivalent to

$$\widehat{\varepsilon}_t = \tilde{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \tilde{e}_t, \qquad t = p+1,\ldots,T, \tag{2.9}$$

where $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ are the residuals from the regression of $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ on $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$.

This asymptotic equivalence comes from the asymptotic orthogonality of the nonlinear least squares residuals to the regressors in the model equation (2.3). Therefore, although the first-order partial derivative functions, $\dot{g}_{\boldsymbol{\theta}}(\cdot;\widehat{\boldsymbol{\theta}})$, are needed in the auxiliary regression, they do not count toward the degrees of freedom of the serial correlation test, because they were entered unrestrictedly. So the squared sample partial correlation between $\widehat{\varepsilon}_t$ and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, after controlling for $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, is asymptotically equivalent to the squared sample multiple correlation between $\widehat{\varepsilon}_t$ and $(\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}, \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})')$.

Consider the interesting case where we conduct an auxiliary regression for testing first-order serial correlation ($p = 1$),

$$\widehat{\varepsilon}_t = \alpha\widehat{\varepsilon}_{t-1} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = 2,\ldots,T.$$

In general inferential analysis, the F-test and the t-test on testing a single coefficient yield equivalent results. So we should deduce the same conclusion whether we use

an F-test or a t-test to test the significance of α in the above equation. The result in *Theorem 2.1* further suggests that the F-statistic for testing the joint significance of α and $\boldsymbol{\beta}$ is asymptotically equal to the square of the t-statistic on α.

Building on *Theorem 2.1*, the following theorem relates the test statistic to the true error term.

Theorem 2.2. *If Assumptions 2.1 to 2.5 are satisfied, then*

$$TR^2 = T\frac{\widehat{cov}[\varepsilon_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]' \{\widehat{var}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]\}^{-1} \widehat{cov}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \varepsilon_t | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]}{\widehat{var}(\varepsilon_t)} + o_p(1).$$

Comparing to *Theorem 2.1*, the difference is that we have now replaced the contemporaneous residual, $\widehat{\varepsilon}_t$, by the true error term, ε_t. *Theorem 2.2* suggests that the test statistic constructed from the auxiliary regression in equation (2.6) is asymptotically equivalent to a test statistic constructed based on the following equation

$$\varepsilon_t = \boldsymbol{\alpha}' [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})] + \grave{e}_t, \qquad t = p+1, \ldots, T, \tag{2.10}$$

where the contemporaneous error term replaces the contemporaneous residual in equation (2.9). This implies that conditioning on $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$, $\widehat{\varepsilon}_t$ contains the same information as ε_t asymptotically. In other words, the squared sample partial correlation between ε_t and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ is asymptotically the same as that between $\widehat{\varepsilon}_t$ and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, after controlling for $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$.

2.4 Examples

In this section, we adopt several examples to illustrate the general applicability of the results in this chapter. We start with two simple linear models — one is a random walk model and the other is a stationary AR(1) model. We then mix the two models together and consider an AR(2) process with one unit root. The fourth example to be analysed is a non-linear model studied in Wu (1981) — a power-curve model. And

finally, we consider the semilogarithmic growth model examined in Phillips (2007) to further explore *Assumption 2.3*. For simplicity, we take $p = 1$ here. While keeping in mind that some of the conditions stated in the assumptions only hold almost surely, we shall suppress the letters a.s. throughout this dissertation to simplify the notation.

2.4.1 A Simple Linear Random Walk Model

For a linear, purely non-stationary model, assumptions are relatively easy to check. In particular, it is relatively easy to verify *Assumption 2.5*.

Consider the following linear AR(1) model

$$y_t = \theta y_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{2.11}$$

where the particular data generating process under analysis is $\theta = 1$. The error term is a martingale difference sequence as specified in *Assumption 2.1*, and is independent of the initial value y_0, which is assumed to be zero here. Since this is a linear model, we can estimate θ using ordinary least squares (OLS), which gives the estimator $\widehat{\theta}$.

The first and second-order partial derivatives of the function $g(x_t; \theta) = \theta y_{t-1}$ with respect to θ are

$$\dot{g}_\theta(x_t; \theta) = y_{t-1} \qquad \text{and} \qquad \ddot{g}_{\theta\theta}(x_t; \theta) = 0,$$

respectively. It is straightforward that both functions are continuous in x and neither of them depends on θ.

The normalisation $N_{T,\boldsymbol{\theta_0}}$ is a scalar in this case. As there is a unit root, we choose $N^{-1}_{T,\boldsymbol{\theta_0}} = T$ so that the conditions in *Assumption 2.2* are satisfied. We now proceed to verify *Assumptions 2.3* to *2.5*.

Checking Assumption 2.3: We need to show that $T(\widehat{\theta} - 1)$ is no larger than $o_p(T^\tau)$,

where $0 < \tau < \frac{1}{4}$. Following from the OLS estimation and the model equation, we have

$$T(\widehat{\theta} - 1) = \frac{T^{-1}\sum_{t=1}^{T} y_{t-1}\varepsilon_t}{T^{-2}\sum_{t=1}^{T} y_{t-1}^2}.$$

From Chan and Wei (1988) we have: the numerator and denominator converge in distribution to $\frac{1}{2}\sigma_\varepsilon^2[W(1)^2 - 1]$ and $\sigma_\varepsilon^2 \int_0^1 W^2(r)dr$, respectively, by the functional central limit theorem. Here $W(r)$ denotes the Wiener process, with argument $r \in [0,1]$. Since $\mathbb{P}\{\int_0^1 W^2(r)dr \neq 0\} = 1$, applying the Continuous Mapping Theorem gives us the Dickey-Fuller distribution [see Dickey and Fuller (1979)]. Thus we have $T(\widehat{\theta} - 1) = O_p(1)$, which satisfies the assumption. For more details on the functional central limit theorem, see Phillips (1987).

Checking Assumption 2.4: The linear g function is thrice differentiable with respect to θ on its entire domain Θ. It remains to show that both the first and second-order partial derivatives are asymptotically bounded.

We can check *Assumption 2.4(a)* and *Assumption 2.4(b)* jointly by evaluating

$$\sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0}) \right\|^2 = T^{-2} \sum_{t=2}^{T} y_{t-1}^2 = T^{-2} \sum_{t=1}^{T} y_{t-1}^2 - T^{-2} y_0^2.$$

From above, the first term on the right-hand side (RHS), $T^{-2}\sum_{t=1}^{t} y_{t-1}^2$, is $O_p(1)$, as it converges in distribution to $\sigma_\varepsilon^2 \int_0^1 W^2(r)dr$. Given the zero-value assumption on the initial observation, y_0, it immediately follows that $T^{-2}\sum_{t=2}^{T} y_{t-1}^2$ is bounded in probability and invertible.

Assumption 2.4(c) involves taking the supremum over some $\tilde{\theta}$ satisfying the condition that $(\tilde{\theta} - \theta_0) \leq T^{-1}K$, for every $K > 0$. However, in our case, the second-order partial derivative of the linear function g does not depend on θ and is equal to the constant 0. This implies that, whatever value θ takes, the sum of the supremum of $[\ddot{g}_{\theta\theta}(x_t; \tilde{\theta})]^2$ is always bounded in probability, and *Assumption 2.4(c)* is satisfied.

Checking Assumption 2.5: In this example, we have $\dot{g}_\theta(x_t; \widehat{\theta}) = y_{t-1}$, and therefore

the leading term to be analysed is

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_\theta(x_t;\widehat{\theta})] = [\widehat{\varepsilon}_{t-1}|y_{t-1}] = [y_{t-1} - \widehat{\theta}y_{t-2}|y_{t-1}], \tag{2.12}$$

where the second equality follows from the model equation. Noting that the OLS residuals from regressing y_{t-1} on itself are just zero, we have

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_\theta(x_t;\widehat{\theta})] = -\widehat{\theta}[y_{t-2}|y_{t-1}]. \tag{2.13}$$

Recall that the underlying data generating process is a random walk, so we have $y_{t-2} = y_{t-1} - \varepsilon_{t-1}$. Substituting this into the above expression gives

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_\theta(x_t;\widehat{\theta})] = -\widehat{\theta}[y_{t-1} - \varepsilon_{t-1}|y_{t-1}] = \widehat{\theta}[\varepsilon_{t-1}|y_{t-1}], \tag{2.14}$$

where the second equality follows again from $(y_{t-1}|y_{t-1}) = 0$. *Assumption 2.5* imposes the restriction that

$$\frac{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_\theta(x_t;\widehat{\theta})]\varepsilon_t}{\sqrt{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_\theta(x_t;\widehat{\theta})]^2}} = O_p(1),$$

which is equivalent to

$$\frac{T^{-\frac{1}{2}}\sum_{t=2}^{T}[\varepsilon_{t-1}|y_{t-1}]\varepsilon_t}{\sqrt{T^{-1}\sum_{t=2}^{T}[\varepsilon_{t-1}|y_{t-1}]^2}} = O_p(1), \tag{2.15}$$

by plugging in equation (2.14) to the LHS. Note that we can cancel $\widehat{\theta}$ out in the above equation because it is the common factor in the numerator and the denominator.
To verify that equation (2.15) holds, we first consider the denominator. Expand the term to get

$$T^{-1}\sum_{t=2}^{T}[\varepsilon_{t-1}|y_{t-1}]^2 = T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2 - T^{-1}\frac{(T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}y_{t-1})^2}{T^{-2}\sum_{t=2}^{T}y_{t-1}^2}. \tag{2.16}$$

By *Assumption 2.1* and the martingale Laws of Large Numbers [see (Davidson, 1994, page 315)], the first term on the RHS, $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2$, converges in probability to a positive constant, σ_ε^2. Using the same Dickey-Fuller type arguments as in the discussion of *Assumption 2.3*, we can show that the second term is $O_p(T^{-1})$. To see this, notice that the numerator of the second term, $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}y_{t-1}$, is the same as $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2+T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}y_{t-2}$, which converges in distribution to $\sigma_\varepsilon^2+\frac{1}{2}\sigma_\varepsilon^2[W(1)^2-1]$, under the functional central limit theorem, martingale Laws of Large Numbers and the Continuous Mapping Theorem, while the denominator $T^{-2}\sum_{t=2}^{T}y_{t-1}^2$ converges in distribution to $\sigma_\varepsilon^2\int_0^1 W^2(r)dr$ by the functional central limit theorem. Since T^{-1} converges deterministically to zero, noting that $\mathbb{P}\{\int_0^1 W^2(r)dr\neq 0\}=1$, the Continuous Mapping Theorem implies that the second term on the RHS of equation (2.16) is small. Combining above results yields

$$T^{-1}\sum_{t=2}^{T}[\varepsilon_{t-1}|y_{t-1}]^2\xrightarrow{D}\sigma_\varepsilon^2. \tag{2.17}$$

Now take the numerator on the LHS of equation (2.15) and expand the term to get

$$T^{-\frac{1}{2}}\sum_{t=2}^{T}[\varepsilon_{t-1}|y_{t-1}]\varepsilon_t=T^{-\frac{1}{2}}\sum_{t=2}^{T}\varepsilon_{t-1}\varepsilon_t-T^{-\frac{1}{2}}\frac{T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}y_{t-1}T^{-1}\sum_{t=2}^{T}\varepsilon_t y_{t-1}}{T^{-2}\sum_{t=2}^{T}y_{t-1}^2}. \tag{2.18}$$

Assumption 2.1 together with the martingale central limit theorem [see Brown (1971)] imply that the first term on the RHS, $T^{-\frac{1}{2}}\sum_{t=2}^{T}\varepsilon_{t-1}\varepsilon_t$, converges in distribution to $N(0,\sigma_\varepsilon^4)$. We can use similar arguments as above to show that the second term is $O_p(T^{-\frac{1}{2}})$. Recall we have shown that $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}y_{t-1}\xrightarrow{D}\sigma_\varepsilon^2+\frac{1}{2}\sigma_\varepsilon^2[W(1)^2-1]$ and $T^{-1}\sum_{t=2}^{T}\varepsilon_t y_{t-1}\xrightarrow{D}\frac{1}{2}\sigma_\varepsilon^2[W(1)^2-1]$ in the numerator, while $T^{-2}\sum_{t=2}^{T}y_{t-1}^2\xrightarrow{D}\sigma_\varepsilon^2\int_0^1 W^2(r)dr$ in the denominator. Again, so long as $\mathbb{P}\{\int_0^1 W^2(r)dr\neq 0\}=1$, applying the Continuous Mapping Theorem gives the desired result that the second term on

the RHS of equation (2.18) is of order $T^{-\frac{1}{2}}$, and we have

$$T^{-\frac{1}{2}} \sum_{t=2}^{T} [\varepsilon_{t-1}|y_{t-1}]\varepsilon_t \xrightarrow{D} N(0, \sigma_\varepsilon^4). \tag{2.19}$$

The proof of *Assumption 2.5* is completed by substituting equations (2.17) and (2.19) back into equation (2.15) and again applying the Continuous Mapping Theorem so that

$$\frac{T^{-\frac{1}{2}} \sum_{t=2}^{T} [\varepsilon_{t-1}|y_{t-1}]\varepsilon_t}{\sqrt{T^{-1} \sum_{t=2}^{T} [\varepsilon_{t-1}|y_{t-1}]^2}} \xrightarrow{D} \frac{N(0, \sigma_\varepsilon^4)}{\sqrt{\sigma_\varepsilon^2}} = N(0, \sigma_\varepsilon^2).$$

2.4.2 A Stationary AR(1) Model

In the previous example, we saw that for a pure random walk process, the verification of *Assumption 2.5* is relatively simple. However, when this AR(1) model is a stationary one, the situation is slightly complicated, as we shall see below.

Recall the simple univariate AR(1) model in equation (2.11)

$$y_t = \theta y_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, T,$$

where $\{\varepsilon_t\}$ is a martingale difference sequence and the regressor $x_t = y_{t-1}$. The true θ considered in this example now has absolute value strictly less than one. The normalisation $N_{T,\boldsymbol{\theta_0}}$ is $T^{\frac{1}{2}}$ and *Assumption 2.2* is naturally satisfied. Let θ_0 denote the true population parameter. *Assumption 2.3* could be checked by using the results in Phillips and Solo (1992) so that $T^{\frac{1}{2}}(\widehat{\theta} - \theta_0)$ is stochastically bounded by invoking the martingale Law of Large Numbers and Central Limit Theorem. The verification of *Assumption 2.4* is somewhat similar to the analysis used in the random walk example above and we shall omit it here. We now proceed to check *Assumption 2.5*.

Checking Assumption 2.5: Recall from equations (2.12) and (2.13) that the leading

term to be analysed is

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_{\theta}(x_t;\widehat{\theta})] = (\widehat{\varepsilon}_{t-1}|y_{t-1}) = (y_{t-1} - \widehat{\theta}y_{t-2}|y_{t-1}) = -\widehat{\theta}(y_{t-2}|y_{t-1}). \tag{2.20}$$

Assumption 2.5 assumes that

$$\frac{\sum_{t=2}^{T}(\widehat{\varepsilon}_{t-1}|y_{t-1})\varepsilon_t}{\sqrt{\sum_{t=2}^{T}(\widehat{\varepsilon}_{t-1}|y_{t-1})^2}} = \frac{T^{-\frac{1}{2}}\sum_{t=2}^{T}(\widehat{\varepsilon}_{t-1}|y_{t-1})\varepsilon_t}{\sqrt{T^{-1}\sum_{t=2}^{T}(\widehat{\varepsilon}_{t-1}|y_{t-1})^2}} = O_p(1), \tag{2.21}$$

which requires the inverse of the denominator to exist and be bounded in probability. Because of equation (2.20), the denominator here is $\sqrt{\widehat{\theta}^2 T^{-1}\sum_{t=2}^{T}(y_{t-2}|y_{t-1})^2}$. While $T^{-1}\sum_{t=2}^{T}(y_{t-2}|y_{t-1})^2$ has a positive limit for all $|\theta_0| < 1$, the consistent estimator $\widehat{\theta}^2$ may converge in probability to zero when $\theta_0 = 0$. Thus the denominator $\sqrt{\widehat{\theta}^2 T^{-1}\sum_{t=2}^{T}(y_{t-2}|y_{t-1})^2}$ has a positive limit when $0 < |\theta_0| < 1$ but vanishes when θ_0 is zero. This seems to imply that *Assumption 2.5* is violated when $\theta_0 = 0$ in this example.

Yet, given that the model is linear in parameters, we have a solution. Note that when $\theta_0 = 0$, we have $y_t = \varepsilon_t$. Further, following from equation (2.20), we can rewrite *Assumption 2.5* in equation (2.21) as

$$-\frac{\sum_{t=2}^{T}(y_{t-2}|y_{t-1})\varepsilon_t}{\sqrt{\sum_{t=2}^{T}(y_{t-2}|y_{t-1})^2}} = O_p(1), \tag{2.22}$$

where the common factor $\widehat{\theta}$ is cancelled out.

When the underlying data generating process is $y_t = \theta_0 y_{t-1} + \varepsilon_t$, equation (2.22) follows from two different arguments when $\theta_0 = 0$ and when $\theta_0 \neq 0$.

When $\theta_0 = 0$, this implies that y_t follows a martingale difference structure itself, and we have

$$-\frac{\sum_{t=2}^{T}(y_{t-2}|y_{t-1})\varepsilon_t}{\sqrt{\sum_{t=2}^{T}(y_{t-2}|y_{t-1})^2}} = -\frac{\sum_{t=2}^{T}(\varepsilon_{t-2}|\varepsilon_{t-1})\varepsilon_t}{\sqrt{\sum_{t=2}^{T}(\varepsilon_{t-2}|\varepsilon_{t-1})^2}}. \tag{2.23}$$

So the key term to be analysed becomes $(\varepsilon_{t-2}|\varepsilon_{t-1})$.

On the other hand, when $\theta_0 \neq 0$, we can write y_{t-2} as $\frac{1}{\theta_0}y_{t-1} - \frac{1}{\theta_0}\varepsilon_{t-1}$. Noting that

$$(y_{t-2}|y_{t-1}) = (\frac{1}{\theta_0}y_{t-1} - \frac{1}{\theta_0}\varepsilon_{t-1}|y_{t-1}) = -\frac{1}{\theta_0}(\varepsilon_{t-1}|y_{t-1}),$$

we have

$$-\frac{\sum_{t=2}^{T}(y_{t-2}|y_{t-1})\varepsilon_t}{\sqrt{\sum_{t=2}^{T}(y_{t-2}|y_{t-1})^2}} = \frac{\sum_{t=2}^{T}(\varepsilon_{t-1}|y_{t-1})\varepsilon_t}{\sqrt{\sum_{t=2}^{T}(\varepsilon_{t-1}|y_{t-1})^2}}. \tag{2.24}$$

So the key term to be discussed in this situation is $(\varepsilon_{t-1}|y_{t-1})$.

In both cases, we can show that *Assumption 2.5* holds, using the martingale Law of Large Numbers and Central Limit Theorem. Therefore, provided that $|\theta_0| < 1$, whether θ_0 is zero or not does not affect the validity of *Assumption 2.5*, although this does affect the key terms to be analysed.

2.4.3 An AR(2) Model with One Unit Root

We next combine the two examples discussed above and study a linear model with both stationary and non-stationary components. In particular, we show that it is the stationary component that complicates the analysis of *Assumption 2.5*.

Consider the following AR(2) process

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{2.25}$$

where $\gamma_1 + \gamma_2 = 1$ and $|\gamma_2| < 1$, so that y_t is integrated of order one. Here, μ is a finite constant whose true value is zero, and ε_t is again a martingale difference sequence satisfying *Assumption 2.1*. We also have the error term being independent to the initial observations, y_{-1} and y_0, both of which are assumed to be zero.

Notice that the model equation is the same as

$$y_t = \mu + (\gamma_1 + \gamma_2)y_{t-1} - \gamma_2(y_{t-1} - y_{t-2}) + \varepsilon_t.$$

Let $\theta_1 = -\gamma_2$, $\theta_2 = \mu$, and $\theta_3 = \gamma_1 + \gamma_2$, so we have $|\theta_1| < 1$, θ_2 is a finite constant, and $\theta_3 = 1$. We can rewrite the above equation as

$$y_t = \theta_1 \triangle y_{t-1} + \theta_2 + \theta_3 y_{t-1} + \varepsilon_t \equiv x_t' \boldsymbol{\theta_0} + \varepsilon_t, \tag{2.26}$$

where $\triangle y_{t-1} = y_{t-1} - y_{t-2}$, $x_t = (\triangle y_{t-1}, 1, y_{t-1})'$ and $\boldsymbol{\theta_0} = (\theta_1, \theta_2, \theta_3)'$. Again, as this model is linear in parameters, we can obtain the usual OLS estimates, $\widehat{\boldsymbol{\theta}}$.

The first and second-order partial derivatives of the function $g(x_t; \boldsymbol{\theta}) = x_t' \boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}$ are

$$\dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta}) = x_t = \begin{bmatrix} \triangle y_{t-1} \\ 1 \\ y_{t-1} \end{bmatrix} \qquad \text{and} \qquad \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t; \boldsymbol{\theta}) = \mathbf{0},$$

respectively. Here $\mathbf{0}$ is a 3×3 square matrix with all elements being 0. None of these partial derivative functions depends on θ, and they are all continuous in x. In the analysis of $\dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta})$, we consider the vector Euclidean norm, which can be obtained by summing the square of each element and then taking the square root.

Notice that $\triangle y_{t-1}$ is stationary, so its coefficient is $T^{\frac{1}{2}}$ convergent, whereas y_{t-1} is integrated of order one and therefore its coefficient should be normalised by T. The normalisation matrix $N_{T,\boldsymbol{\theta_0}}$ is of dimension 3×3 in this example and

$$N_{T,\boldsymbol{\theta_0}}^{-1} = \begin{bmatrix} T^{\frac{1}{2}} & 0 & 0 \\ 0 & T^{\frac{1}{2}} & 0 \\ 0 & 0 & T \end{bmatrix}.$$

Because $\mathrm{N}_{T,\boldsymbol{\theta_0}}$ is a diagonal matrix with nonzero values on the diagonal, it must be invertible and the conditions in *Assumption 2.2* are fulfilled. We now turn to check *Assumptions 2.3* to *2.5*.

Checking Assumption 2.3: This assumption requires we have $N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) =$

$\mathbf{o_p}(T^\tau)$, where $0 < \tau < \frac{1}{4}$. Consider the deviation of $\widehat{\boldsymbol{\theta}}$ from $\boldsymbol{\theta_0}$

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0} = (\sum_{t=1}^{T} x_t x_t')^{-1} \sum_{t=1}^{T} x_t \varepsilon_t,$$

where we have

$$\sum_{t=1}^{T} x_t x_t' = \begin{bmatrix} \sum_{t=1}^{T} \triangle^2 y_{t-1} & \sum_{t=1}^{T} \triangle y_{t-1} & \sum_{t=1}^{T} \triangle y_{t-1} \cdot y_{t-1} \\ \sum_{t=1}^{T} \triangle y_{t-1} & T & \sum_{t=1}^{T} y_{t-1} \\ \sum_{t=1}^{T} y_{t-1} \cdot \triangle y_{t-1} & \sum_{t=1}^{T} y_{t-1} & \sum_{t=1}^{T} y_{t-1}^2 \end{bmatrix},$$

and

$$\sum_{t=1}^{T} x_t \varepsilon_t = \begin{bmatrix} \sum_{t=1}^{T} \triangle y_{t-1} \cdot \varepsilon_t \\ \sum_{t=1}^{T} \varepsilon_t \\ \sum_{t=1}^{T} y_{t-1} \varepsilon_t \end{bmatrix}.$$

Premultiplying $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}$ with the inverse of the normalisation matrix gives

$$N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = [N_{T,\boldsymbol{\theta_0}}'(\sum_{t=1}^{T} x_t x_t') N_{T,\boldsymbol{\theta_0}}]^{-1} N_{T,\boldsymbol{\theta_0}}' \sum_{t=1}^{T} x_t \varepsilon_t. \tag{2.27}$$

First consider the inverse matrix in the above equation. Recall that $\mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2$ by *Assumption 2.1*. Let $\kappa = \frac{\sigma_\varepsilon}{1-\theta_1}$. Applying the functional central limit theorem and the martingale Laws of Large Numbers together with the Continuous Mapping Theorem gives,

$$N_{T,\boldsymbol{\theta_0}}'(\sum_{t=1}^{T} x_t x_t') N_{T,\boldsymbol{\theta_0}} \xrightarrow{D} \begin{bmatrix} \mathbb{E}[(\triangle y_t)^2] & 0 & 0 \\ 0 & 1 & \kappa \int_0^1 W(r)dr \\ 0 & \kappa \int_0^1 W(r)dr & \kappa^2 \int_0^1 W(r)^2 dr \end{bmatrix} \equiv \begin{bmatrix} V & \mathbf{0} \\ \mathbf{0} & Q \end{bmatrix}, \tag{2.28}$$

where $V = \mathbb{E}[(\triangle y_t)^2]$ and

$$Q = \begin{bmatrix} 1 & \kappa \int_0^1 W(r)dr \\ \kappa \int_0^1 W(r)dr & \kappa^2 \int_0^1 W(r)^2 dr \end{bmatrix}.$$

Therefore both V and Q are stochastically bounded.

Next, take the second term in equation (2.27). By the martingale central limit theorem as well as the functional central limit theorem,

$$N'_{T,\boldsymbol{\theta_0}} \sum_{t=1}^{T} x_t \varepsilon_t \xrightarrow{D} \begin{bmatrix} N(0, \sigma_\varepsilon^2 V) \\ \sigma_\varepsilon W(1) \\ \frac{1}{2}\sigma_\varepsilon \kappa [W(1)^2 - 1] \end{bmatrix} \equiv \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}, \tag{2.29}$$

where $H_1 = N(0, \sigma_\varepsilon^2 V)$ and $H_2 = (\sigma_\varepsilon W(1), \frac{1}{2}\sigma_\varepsilon \kappa [W(1)^2 - 1])'$.

Substituting equations (2.28) and (2.29) back into equation (2.27) and applying the Continuous Mapping Theorem yields

$$N^{-1}_{T,\boldsymbol{\theta_0}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \xrightarrow{D} \begin{bmatrix} V & \mathbf{0} \\ \mathbf{0} & Q \end{bmatrix}^{-1} \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} = \begin{bmatrix} V^{-1} H_1 \\ Q^{-1} H_2 \end{bmatrix}.$$

Therefore, the normalised vector $N^{-1}_{T,\boldsymbol{\theta_0}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$ is $\mathbf{O_p}(1)$, and *Assumption 2.3* is satisfied.

Checking Assumption 2.4: As the function $g(x_t; \boldsymbol{\theta})$ is linear in parameters, it is thrice differentiable with respect to $\boldsymbol{\theta}$ on its entire domain Θ. *Assumption 2.4(c)* is naturally satisfied, because the second-order partial derivatives are all exactly zero. We now check that the first-order partial derivatives, x_t, are asymptotically bounded. For *Assumption 2.4(a)*, it is sufficient that the summation term $\sum_{t=2}^{T} ||N'_{T,\boldsymbol{\theta_0}} x_t||^2$ converges in distribution. First, notice that analysing $\sum_{t=2}^{T} ||N'_{T,\boldsymbol{\theta_0}} x_t||^2$ is the same as analysing $\sum_{t=1}^{T} ||N'_{T,\boldsymbol{\theta_0}} x_t||^2$, since the initial observations, y_{-1} and y_0, are zero by assumption. Next, note that the summand $N'_{T,\boldsymbol{\theta_0}} x_t$ is a vector. For any vector $\mathbf{z} \in \mathbb{R}^n$, we

have the following equality

$$\|\mathbf{z}\|^2 = \mathbf{z}'\mathbf{z} = tr(\mathbf{z}\mathbf{z}'),$$

where $tr(\cdot)$ denotes the trace of the matrix. Therefore, the square of the Euclidean norm of the vector $N'_{T,\boldsymbol{\theta_0}}x_t$ equals the trace of its own outer product, and we have

$$\sum_{t=1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}x_t\right\|^2 = tr(\sum_{t=1}^{T}N'_{T,\boldsymbol{\theta_0}}x_t x'_t N_{T,\boldsymbol{\theta_0}}).$$

Recall it is shown in equation (2.28) that $\sum_{t=1}^{T}N'_{T,\boldsymbol{\theta_0}}x_t x'_t N_{T,\boldsymbol{\theta_0}}$ converges in distribution. It must follow that $tr(\sum_{t=1}^{T}N'_{T,\boldsymbol{\theta_0}}x_t x'_t N_{T,\boldsymbol{\theta_0}})$ is $O_p(1)$, and thus *Assumption 2.4(a)* holds. It can also be shown that the matrix $\sum_{t=2}^{T}x_t x'_t$ has a non-zero determinant, and therefore is finitely invertible.

Checking Assumption 2.5: In this example, we have $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = x_t$. Thus the leading term to be analysed is

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = [\widehat{\varepsilon}_{t-1}|x_t] = [y_{t-1} - \widehat{\mu} - \widehat{\gamma}_1 y_{t-2} - \widehat{\gamma}_2 y_{t-3}|x_t], \tag{2.30}$$

where the second equality follows from equation (2.25). Notice that partialling out the vector $x_t = (\triangle y_{t-1}, 1, y_{t-1})'$ is the same as partialling out the vector $(y_{t-1}, 1, y_{t-2})'$, since

$$x_t = \begin{bmatrix} \triangle y_{t-1} \\ 1 \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ 1 \\ y_{t-2} \end{bmatrix}. \tag{2.31}$$

Therefore we can rewrite equation (2.30) as

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = [y_{t-1} - \widehat{\mu} - \widehat{\gamma}_1 y_{t-2} - \widehat{\gamma}_2 y_{t-3}|y_{t-1}, 1, y_{t-2}].$$

Since the residuals from regressing y_{t-1}, y_{t-2} and 1 on themselves are just zeros, we

can further simplify the above equation to get

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = -\widehat{\gamma}_2[y_{t-3}|y_{t-1},1,y_{t-2}]. \tag{2.32}$$

From the data generating process, we have

$$y_{t-1} = \mu + \gamma_1 y_{t-2} + \gamma_2 y_{t-3} + \varepsilon_{t-1},$$

where $\mu = 0$, $\gamma_1 + \gamma_2 = 1$ and $|\gamma_2| < 1$. The unit root is not a particular problem here, but the stationary root is. As in the stationary AR(1) case above, we need to distinguish between $\gamma_2 = 0$ and $\gamma_2 \neq 0$. If $\gamma_2 = 0$, this AR(2) model reduces to a random walk with a drift and the analysis of *Assumption 2.5* is similar to that of the pure random walk case above in Section 2.4.1. Here we consider the situation where $\gamma_2 \neq 0$.
Provided that $\gamma_2 \neq 0$, rearranging gives

$$y_{t-3} = \frac{1}{\gamma_2}y_{t-1} - \frac{\mu}{\gamma_2} - \frac{\gamma_1}{\gamma_2}y_{t-2} - \frac{1}{\gamma_2}\varepsilon_{t-1}.$$

Substitute the above equation back into equation (2.32) and further simplify to get

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \frac{\widehat{\gamma}_2}{\gamma_2}[\varepsilon_{t-1}|y_{t-1},1,y_{t-2}].$$

Recall from equation (2.31) that conditioning on the vector $(y_{t-1},1,y_{t-2})'$ is same as conditioning on the vector $x_t = (\triangle y_{t-1},1,y_{t-1})'$. We can therefore rewrite the above equation as

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \frac{\widehat{\gamma}_2}{\gamma_2}[\varepsilon_{t-1}|x_t]. \tag{2.33}$$

Assumption 2.5 requires that $\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]^2\}^{-1/2}\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\varepsilon_t\}$ is $O_p(1)$. This is the same as requiring

$$\frac{T^{-\frac{1}{2}}\sum_{t=2}^{T}[\varepsilon_{t-1}|x_t]\varepsilon_t}{\sqrt{T^{-1}\sum_{t=2}^{T}[\varepsilon_{t-1}|x_t]^2}} = O_p(1) \tag{2.34}$$

by substituting in equation (2.33). Note that the common factor $\frac{\widehat{\gamma}_2}{\gamma_2}$ is cancelled out.

To prove equation (2.34), we first take the denominator and expand it to get

$$T^{-1}\sum_{t=2}^{T}[\varepsilon_{t-1}|x_t]^2 = T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2 - T^{-1}\{\sum_{t=2}^{T}\varepsilon_{t-1}x_t'N_{T,\boldsymbol{\theta_0}}(\sum_{t=2}^{T}N_{T,\boldsymbol{\theta_0}}'x_tx_t'N_{T,\boldsymbol{\theta_0}})^{-1}\sum_{t=2}^{T}N_{T,\boldsymbol{\theta_0}}'x_t\varepsilon_{t-1}\}. \quad (2.35)$$

Following from *Assumption 2.1* and the martingale Laws of Large Numbers, the first term on the RHS, $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2$, converges in probability to σ_ε^2.

Using the same type of arguments as in the discussion of *Assumption 2.3*, we can show that the second term is $O_p(T^{-1})$. To see this, note that following from equation (2.28) and the assumption that the initial observations, y_{-1} and y_0, are zero, we have for the middle matrix

$$\sum_{t=2}^{T}N_{T,\boldsymbol{\theta_0}}'x_tx_t'N_{T,\boldsymbol{\theta_0}} \xrightarrow{D} \begin{bmatrix} V & \mathbf{0} \\ \mathbf{0} & Q \end{bmatrix},$$

where V and Q are both invertible and stochastically bounded.

Next consider the vector $\sum_{t=2}^{T}N_{T,\boldsymbol{\theta_0}}'x_t\varepsilon_{t-1}$. From the data generating process, we have

$$x_t = \begin{bmatrix} \triangle y_{t-1} \\ 1 \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3-1 \\ 0 & 1 & 0 \\ \theta_1 & \theta_2 & \theta_3 \end{bmatrix}\begin{bmatrix} \triangle y_{t-2} \\ 1 \\ y_{t-2} \end{bmatrix} + \varepsilon_{t-1}\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} =: \Pi x_{t-1} + \varepsilon_{t-1}\mathbf{l},$$

where $|\theta_1| < 1$, θ_2 is a finite constant, and $\theta_3 = 1$. This implies that the parameter matrix Π is $\mathbf{O}(1)$. Substituting the above equation for x_t in the vector $\sum_{t=2}^{T}N_{T,\boldsymbol{\theta_0}}'x_t\varepsilon_{t-1}$ gives

$$\sum_{t=2}^{T}N_{T,\boldsymbol{\theta_0}}'x_t\varepsilon_{t-1} = \Pi\sum_{t=2}^{T}N_{T,\boldsymbol{\theta_0}}'x_{t-1}\varepsilon_{t-1} + \sum_{t=2}^{T}N_{T,\boldsymbol{\theta_0}}'\varepsilon_{t-1}^2\mathbf{l}. \quad (2.36)$$

Following from equation (2.29) and the Continuous Mapping Theorem, we have for

the first term on the RHS that

$$\Pi \sum_{t=2}^{T} N'_{T,\boldsymbol{\theta_0}} x_{t-1} \varepsilon_{t-1} \xrightarrow{D} \Pi \begin{bmatrix} H_1 \\ H_2 \end{bmatrix},$$

where both H_1 and H_2 are stochastically bounded. Therefore $\Pi \sum_{t=2}^{T} N'_{T,\boldsymbol{\theta_0}} x_{t-1} \varepsilon_{t-1}$ is $\mathbf{O_p}(1)$. According to *Assumption 2.1* and the martingale central limit theorem, the second term on the RHS of equation (2.36), $\sum_{t=2}^{T} N'_{T,\boldsymbol{\theta_0}} \varepsilon_{t-1}^2 \mathbf{l}$, is also $\mathbf{O_p}(1)$.

Given that T^{-1} vanishes deterministically as T goes to infinity, combining the above results together with the Continuous Mapping Theorem implies that the second term on the RHS of equation (2.35) is small, and therefore

$$T^{-1} \sum_{t=2}^{T} [\varepsilon_{t-1} | x_t]^2 \xrightarrow{D} \sigma_\varepsilon^2. \tag{2.37}$$

Next consider the numerator, $T^{-\frac{1}{2}} \sum_{t=2}^{T} [\varepsilon_{t-1} | x_t] \varepsilon_t$, in equation (2.34). Expanding the term gives

$$\begin{aligned} T^{-\frac{1}{2}} \sum_{t=2}^{T} [\varepsilon_{t-1} | x_t] \varepsilon_t = T^{-\frac{1}{2}} \sum_{t=2}^{T} \varepsilon_{t-1} \varepsilon_t \\ - T^{-\frac{1}{2}} \{ \sum_{t=2}^{T} \varepsilon_{t-1} x_t' N_{T,\boldsymbol{\theta_0}} (\sum_{t=2}^{T} N'_{T,\boldsymbol{\theta_0}} x_t x_t' N_{T,\boldsymbol{\theta_0}})^{-1} \sum_{t=2}^{T} N'_{T,\boldsymbol{\theta_0}} x_t \varepsilon_t \}. \end{aligned} \tag{2.38}$$

By *Assumption 2.1* and the martingale central limit theorem, the first term converges in distribution to $N(0, \sigma_\varepsilon^4)$. We can use similar arguments as above to show that the second term is small. Recall we have shown that both the row vector, $\sum_{t=2}^{T} \varepsilon_{t-1} x_t' N_{T,\boldsymbol{\theta_0}}$, and the middle matrix, $(\sum_{t=2}^{T} N'_{T,\boldsymbol{\theta_0}} x_t x_t' N_{T,\boldsymbol{\theta_0}})^{-1}$, are stochastically bounded. Consider the column vector, $\sum_{t=2}^{T} N'_{T,\boldsymbol{\theta_0}} x_t \varepsilon_t$, in the second term, and from equation (2.29), it converges in distribution. Therefore, applying the Continuous Mapping Theorem gives the desired result that the second term on the RHS of equation (2.38) is $O_p(T^{-\frac{1}{2}})$, and

we have

$$T^{-\frac{1}{2}} \sum_{t=2}^{T} [\varepsilon_{t-1}|x_t]\varepsilon_t \xrightarrow{D} N(0, \sigma_\varepsilon^4). \tag{2.39}$$

Assumption 2.5 is asserted by combining equations (2.37) and (2.39) together with the Continuous Mapping Theorem such that

$$\frac{T^{-\frac{1}{2}} \sum_{t=2}^{T} [\varepsilon_{t-1}|x_t]\varepsilon_t}{\sqrt{T^{-1} \sum_{t=2}^{T} [\varepsilon_{t-1}|x_t]^2}} \xrightarrow{D} \frac{N(0, \sigma_\varepsilon^4)}{\sqrt{\sigma_\varepsilon^2}} = N(0, \sigma_\varepsilon^2).$$

2.4.4 A Power-Curve Model

Having discussed two linear examples, we now turn our attention to a non-linear model. Consider the following non-linear model, which is provided in Example 3 of Wu (1981),

$$y_t = (\theta + x_t)^2 + \varepsilon_t, \qquad t = 1, \ldots, T,$$

where θ belongs to the convex parameter space $\Theta \subset \mathbb{R}$ and the true parameter is θ_0. For simplicity, we shall assume that $\{x_t\}$ is an independently and identically distributed (i.i.d.) sequence with mean 0 and variance $\sigma_X^2 < \infty$, and is adapted to the filtration $\mathscr{F}_{t-1}$. We also impose the restriction that the x_t has finite fourth moment. The error term, ε_t, is a martingale difference sequence satisfying the conditions specified in *Assumption 2.1*.

The first and second-order partial derivatives of the non-linear function $g(x_t; \theta) = (\theta + x_t)^2$ with respect to θ are

$$\dot{g}_\theta(x_t; \theta) = 2(\theta + x_t) \qquad \text{and} \qquad \ddot{g}_{\theta\theta}(x_t; \theta) = 2,$$

respectively, and they are both continuous in x. The norm notation is not needed here, since both partial derivatives are scalar-valued.

Given that there is only one parameter in the model and x_t is i.i.d., the normalisation matrix $N_{T,\boldsymbol{\theta_0}}$ reduces to a scalar and we set $N_{T,\boldsymbol{\theta_0}}^{-1} = T^{\frac{1}{2}}$. *Assumption 2.2* is

naturally satisfied, since $N_{T,\boldsymbol{\theta_0}}$ is deterministic. We now proceed to verify *Assumptions 2.3* to *2.5*.

Checking Assumption 2.3: Suppose we have estimated the model by non-linear least squares and obtained the non-linear least squares estimator, $\widehat{\theta}$. To verify *Assumption 2.3*, we need to show that $T^{\frac{1}{2}}(\widehat{\theta}-\theta_0) = o_p(T^{\tau})$, where $0<\tau<\frac{1}{4}$.
Provided that *Assumption 2.1* is satisfied and that the $\mathscr{F}_{t-1}$-adapted x_t is i.i.d. with mean 0, finite variance σ_X^2 and finite fourth moment, applying *Theorem 4.16* of Berenguer-Rico and Nielsen (2020) gives

$$\{\sum_{t=1}^{T}(x_t+\theta_0)^2\}^{\frac{1}{2}}(\widehat{\theta}-\theta_0) = o_p\{(\log T)^2\}.$$

Notice that $\log T$ is dominated by T^{η}, for any $\eta>0$, since by L'Hôpital's rule,

$$\lim_{T\to\infty}\frac{\log T}{T^{\eta}} = \lim_{T\to\infty}\frac{\frac{1}{T}}{\eta T^{\eta-1}} = \lim_{T\to\infty}\frac{1}{\eta T^{\eta}} = 0.$$

It follows that $\{\sum_{t=1}^{T}(x_t+\theta_0)^2\}^{\frac{1}{2}}(\widehat{\theta}-\theta_0)$ is $o_p(T^{\tau})$, where $0<\tau<\frac{1}{4}$. In addition, *Lemma A.18* of Berenguer-Rico and Nielsen (2020) implies that $\{T^{-1}\sum_{t=1}^{T}(x_t+\theta_0)^2\}^{-\frac{1}{2}}$ is $O_p(1)$. Combining these results yields

$$\{T^{-1}\sum_{t=1}^{T}(x_t+\theta_0)^2\}^{-\frac{1}{2}}\{\sum_{t=1}^{T}(x_t+\theta_0)^2\}^{\frac{1}{2}}(\widehat{\theta}-\theta_0) = T^{\frac{1}{2}}(\widehat{\theta}-\theta_0) = o_p(T^{\tau}).$$

Checking Assumption 2.4: The differentiability of the g function is straightforward. Given that the function is quadratic, it is thrice differentiable with respect to θ on its entire domain Θ. The boundedness and invertibility requirements on the first-order partial derivative have been checked by Berenguer-Rico and Nielsen (2020) in their proof of *Theorem 4.17*. Again, because the second-order partial derivative, $\ddot{g}_{\theta\theta}$, is exactly 2 regardless of the value of θ, *Assumption 2.4(c)* always holds.

Checking Assumption 2.5: We will see that this assumption implicitly assumes that there is no perfect linear correlation, positive or negative, between the first-order

partial derivative, $\dot{g}_\theta(x_t;\widehat{\theta})$, and the lagged residual, $\widehat{\varepsilon}_{t-1}$. Recall that $\dot{g}_\theta(x_t;\widehat{\theta}) = 2(\widehat{\theta}+x_t)$ and $\widehat{\varepsilon}_{t-1} = y_{t-1} - (x_{t-1}+\widehat{\theta})^2$. The leading term to be analysed in this example is

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = [y_{t-1} - (x_{t-1}+\widehat{\theta})^2|2(\widehat{\theta}+x_t)].$$

Assumption 2.5 requires this term, which is the residual from regressing $y_{t-1} - (x_{t-1} + \widehat{\theta})^2$ on $2(\widehat{\theta}+x_t)$, to be nonzero. The validity of *Assumption 2.5* will depend on the assumptions to the regressor x_t and its relation to the error process. We have seen that for linear stationary auto-regressions, *Assumption 2.5* is satisfied but the proof is not straight forward. For nonlinear models, this has to be checked with care. Here, we will leave *Assumption 2.5* as an assumption to be checked in particular examples.

2.4.5 A Semilogarithmic Growth Model

We next study a particular regression model with non-linear trends that is discussed in Phillips (2007) to further explore the slightly loose convergence rate assumed in *Assumption 2.3*. Consider the following model

$$y_t = \theta_1 + \theta_2 \log t + \varepsilon_t, \qquad t = 1,\ldots,T,$$

where the error term ε_t satisfies *Assumption 2.1*. Let $x_t = (1, \log t)'$ and $\boldsymbol{\theta_0} = (\theta_1, \theta_2)'$. This model involves a deterministic non-linear trend function, $\log t$, but is linear in parameter. We can therefore obtain the usual OLS estimates, $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2)'$.

The first and second-order partial derivatives of the linear function $g(x_t;\boldsymbol{\theta}) = x_t'\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}$ are

$$\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}) = x_t = \begin{bmatrix} 1 \\ \log t \end{bmatrix} \qquad \text{and} \qquad \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}) = \mathbf{0},$$

respectively. Here $\mathbf{0}$ is a 2×2 square matrix with all elements being 0. These partial derivative functions are continuous in x and independent of $\boldsymbol{\theta}$. To analyse these

functions, we again consider the Euclidean norm.

The normalisation matrix $N_{T,\boldsymbol{\theta_0}}$ is a 2×2 square matrix in this example and we choose

$$N_{T,\boldsymbol{\theta_0}}^{-1} = \begin{bmatrix} T^{\frac{1}{2}} & 0 \\ 0 & T^{\frac{1}{2}} \log T \end{bmatrix}.$$

so that the conditions in *Assumption 2.2* are met. We now discuss *Assumptions 2.3* to *2.5* one after another.

Checking Assumption 2.3: We need to verify that $N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = \mathbf{o_p}(T^{\tau})$, where $0 < \tau < \frac{1}{4}$. From *Theorem 3.1* of Phillips (2007), we have

$$N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = \begin{bmatrix} T^{\frac{1}{2}}(\widehat{\theta}_1 - \theta_1) \\ T^{\frac{1}{2}} \log T(\widehat{\theta}_2 - \theta_2) \end{bmatrix} = \begin{bmatrix} O(\log T) \\ O(\log T) \end{bmatrix}. \tag{2.40}$$

Note that $O(\log T)$ is dominated by $o(T^{\tau})$ for any $\tau > 0$ by L'Hôpital's rule. It follows that $N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$ is $\mathbf{o}(T^{\tau})$.

Checking Assumption 2.4: The differentiability of $g(x_t;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and the boundedness requirement on the second-order partial derivatives are naturally satisfied, since we have a linear model. We proceed to check that the first-order partial derivatives are asymptotically bounded. Consider

$$\Sigma_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \right\|^2 = \Sigma_{t=2}^{T} \left\| \begin{bmatrix} T^{-\frac{1}{2}} & 0 \\ 0 & T^{-\frac{1}{2}}(\log T)^{-1} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \log t \end{bmatrix} \right\|^2 = \Sigma_{t=2}^{T} \left\| \begin{bmatrix} T^{-\frac{1}{2}} \\ T^{-\frac{1}{2}}(\log T)^{-1} \log t \end{bmatrix} \right\|^2.$$

Following from the definition of the Euclidean norm,

$$\left\| N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \right\|^2 = T^{-1} + T^{-1}(\log T)^{-2}(\log t)^2.$$

Therefore, we have

$$\sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \right\|^2 = \sum_{t=2}^{T} [T^{-1} + T^{-1}(\log T)^{-2}(\log t)^2].$$

Rearranging the above equation, and noting that $\log 1$ is just zero, gives

$$\sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \right\|^2 = 1 - \frac{1}{T} + T^{-1}(\log T)^{-2} \sum_{t=1}^{T} (\log t)^2.$$

Notice that $\sum_{t=1}^{T}(\log t)^2 = O(T(\log T)^2)$. It then follows that the last term on the RHS is $O(1)$, and so is $\sum_{t=p+1}^{T} ||N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})||^2$.

Checking Assumption 2.5: In this example, the first-order partial derivatives are $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = x_t = (1,\log t)'$. Thus the leading term to be analysed is

$$[\widehat{\varepsilon}_{t-1}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = [\widehat{\varepsilon}_{t-1}|x_t] = [y_{t-1} - \widehat{\theta}_1 - \widehat{\theta}_2 \log(t-1)|(1,\log t)'].$$

where the second equality comes from $\widehat{\varepsilon}_{t-1} = y_{t-1} - x'_{t-1}\widehat{\boldsymbol{\theta}}$. *Assumption 2.5* requires no perfect linear correlation between $\widehat{\varepsilon}_{t-1}$ and $x_t = (1,\log t)'$, so that the residuals from regressing $\widehat{\varepsilon}_{t-1}$ on the vector $(1,\log t)'$ are different from zero. This needs to be proved in future studies.

2.5 Conclusion

In this chapter, an auxiliary-regression-based test is proposed for detecting innovation autocorrelation in non-linear regression models with martingale difference error terms. Consider the auxiliary regression

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T.$$

Two main theorems are established in this chapter based on assumptions regarding the error terms, the estimators $\widehat{\boldsymbol{\theta}}$, and the g functions. We first show that the test statistic TR^2 has the same limiting distribution as T times the partial R^2 for testing $\boldsymbol{\alpha} = \mathbf{0}$ in the above auxiliary regression. This partial R^2 is the squared sample partial correlation between the contemporaneous residual $\widehat{\varepsilon}_t$ and the lag residuals $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, conditioning on the first-order partial derivatives, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$.

We then proceed to show that the partial R^2 for testing $\boldsymbol{\alpha} = \mathbf{0}$ in the above equation above is asymptotically the same as the partial R^2 for testing $\grave{\boldsymbol{\alpha}} = \mathbf{0}$ in the following regression

$$\varepsilon_t = \grave{\boldsymbol{\alpha}}' \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \grave{\boldsymbol{\beta}}' \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) + \grave{e}_t, \qquad t = p+1, \ldots, T.$$

Thus, conditioning on the first order partial derivatives evaluated at the NLLS estimates $\widehat{\boldsymbol{\theta}}$, $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$, the partial correlation between $\widehat{\varepsilon}_t$ and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ is asymptotically equivalent to that between the true error term, ε_t, and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$.

Both *Theorems 2.1* and *2.2* hold regardless of the regressors x_t being stationary or not, with proofs provided in the appendix. However, there is one restriction being imposed on the data generating process underlying the model by *Assumption 2.5*. This assumption is rather involved and it would be desirable to have a better understanding.

As illustrated in the examples above, *Assumption 2.5* is not binding on non-stationary components. However, when the underlying process involves stationary components, different key terms need to be analysed depending on the values of the stationary roots. Therefore, it is a bit difficult to unify the asymptotic analysis of our test statistic under *Assumption 2.5*. Nonetheless, for linear (vector) auto-regressions, Nielsen (2006b) has shown that the value of stationary and unit roots does not affect the asymptotic behaviour of the test statistic TR^2.

Appendix

This appendix gives the derivation of *Theorem 2.1* and *Theorem 2.2*, with some useful results broken down into several lemmas. We begin by introducing some notation that is used throughout the appendix to simplify the equations. In Appendix 2.B, the definition of spectral norm is provided. Key intermediate results are established and proved in Appendix 2.C, which prepare the ground for proofs of the main theorems in Appendix 2.D.

2.A Notations and Preliminaries

Notation 2.1. *Let* $\nabla_t = g(x_t;\widehat{\boldsymbol{\theta}}) - g(x_t;\boldsymbol{\theta_0})$.

Notation 2.2. *Let* $\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}} = N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) - N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$.

Notation 2.3. *Let* $\mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}} = \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})' N_{T,\boldsymbol{\theta_0}}$.

Notation 2.4. *Let* $\widehat{\boldsymbol{\phi}}_T = N^{-1}_{T,\boldsymbol{\theta_0}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$.

Following from *Assumption 2.3*, we have $\widehat{\boldsymbol{\phi}}_T = \mathbf{o_p}(T^{\tau})$.

Notation 2.5. *We shall follow the notation of Nielsen (2006b) that, for any time series vector* $\mathbf{Z_t}$,

$$Q(\mathbf{Z_t}) = \sum_{t=p+1}^{T} \varepsilon_t \mathbf{Z_t}' \Big(\sum_{t=p+1}^{T} \mathbf{Z_t}\mathbf{Z_t}' \Big)^{-1} \sum_{t=p+1}^{T} \mathbf{Z_t}\varepsilon_t,$$

and

$$\widehat{Q}(\mathbf{Z_t}) = \sum_{t=p+1}^{T} \widehat{\varepsilon}_t \mathbf{Z_t}' \Big(\sum_{t=p+1}^{T} \mathbf{Z_t}\mathbf{Z_t}' \Big)^{-1} \sum_{t=p+1}^{T} \mathbf{Z_t}\widehat{\varepsilon}_t.$$

Recall the test statistic in equation (2.8)

$$TR^2 = T\frac{\widehat{cov}(\widehat{\varepsilon}_t, \widehat{\mathbf{H}}_t)'[\widehat{var}(\widehat{\mathbf{H}}_t)]^{-1}\widehat{cov}(\widehat{\mathbf{H}}_t, \widehat{\varepsilon}_t)}{\widehat{var}(\widehat{\varepsilon}_t)},$$

where $\widehat{var}(\widehat{\varepsilon}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2$, $\widehat{var}(\widehat{\mathbf{H}}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_t\widehat{\mathbf{H}}_t'$, and $\widehat{cov}(\widehat{\varepsilon}_t, \widehat{\mathbf{H}}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t\widehat{\mathbf{H}}_t$. Substituting these expressions for each of the term in TR^2 gives

$$TR^2 = T\frac{\sum_{t=p+1}^{T}\widehat{\varepsilon}_t\widehat{\mathbf{H}}_t'\Big(\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_t\widehat{\mathbf{H}}_t'\Big)^{-1}\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_t\widehat{\varepsilon}_t}{\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}.$$

Noting that the numerator can be written as $\widehat{Q}(\widehat{\mathbf{H}}_t)$ as defined in Notation 2.5,

$$\widehat{Q}(\widehat{\mathbf{H}}_t) = \sum_{t=p+1}^{T}\widehat{\varepsilon}_t\widehat{\mathbf{H}}_t'\Big(\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_t\widehat{\mathbf{H}}_t'\Big)^{-1}\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_t\widehat{\varepsilon}_t, \tag{2.41}$$

we can further simply the test statistic as

$$TR^2 = T\frac{\widehat{Q}(\widehat{\mathbf{H}}_t)}{\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}. \tag{2.42}$$

Notation 2.6. *Suppose there are two sequences of random vectors* $\{\mathbf{g_t}\}_{t=p+1}^{T}$ *and* $\{\mathbf{h}_t\}_{t=p+1}^{T}$. *Let*

$$[\mathbf{g_t}|\mathbf{h_t}] = \mathbf{g_t} - \sum_{s=p+1}^{T}\mathbf{g_s}\mathbf{h_s}'(\sum_{s=p+1}^{T}\mathbf{h_s}\mathbf{h_s}')^{-1}\mathbf{h}_t.$$

In fact, the notation $[\mathbf{g_t}|\mathbf{h_t}]$ denotes the ordinary least squares residuals from the regression of $\mathbf{g_t}$ on $\mathbf{h}_t$. This notation represents the extra information that $\mathbf{g_t}$ possesses after partialling out the effect of $\mathbf{h}_t$ on it. It therefore contains the variation of $\mathbf{g_t}$ that is orthogonal to $\mathbf{h}_t$.

In order to prove the main results of this chapter, we need to invoke Laws of Large Numbers for sums involving martingale difference sequences. The following martingale Laws of Large Numbers is a slightly modified version of Theorem 5 of Chow (1965).

Lemma 2.1 (Martingale Laws of Large Numbers)**.** *Suppose* $\{\zeta_t\}_{t=1}^{\infty}$ *is a martingale difference sequence for* $\{\mathscr{F}_t\}_{t=0}^{\infty}$ *that satisfies*

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}(|\zeta_t|^2|\mathscr{F}_{t-1})}{t^2} < \infty \quad a.s., \tag{2.43}$$

then as $T \to \infty$

$$T^{-1}\sum_{t=1}^{T} \zeta_t \xrightarrow{p} 0.$$

2.B Norms

Hereafter, we shall adopt the spectral norm for matrices. For a matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$, denote its spectral norm by

$$\|\mathbf{S}\| := \sqrt{\lambda_{\max}(\mathbf{S}'\mathbf{S})},$$

where λ denotes the eigenvalue. According to *Theorem 1.6* of Schwarz and Waldvogel (1989), the spectral norm is the smallest matrix norm that is compatible with the Euclidean norm of a vector in that, for any vector $\mathbf{z} \in \mathbb{R}^n$, the Euclidean norm of $\mathbf{Sz}$ is equal to the product of the spectral norm of the matrix $\mathbf{S}$ and the Euclidean norm of the vector $\mathbf{z}$. In other words, the compatibility between the spectral norm and the Euclidean norm ensures that

$$\|\mathbf{Sz}\| = \|\mathbf{S}\| \cdot \|\mathbf{z}\|.$$

Furthermore, the spectral norm satisfies the triangle inequality and the norm inequality, such that, for any two matrices of the same size $\mathbf{L}, \mathbf{S} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{L}+\mathbf{S}\| \leq \|\mathbf{L}\| + \|\mathbf{S}\|$ and $\|\mathbf{LS}'\| \leq \|\mathbf{L}\| \cdot \|\mathbf{S}\|$.

2.C Intermediate Results

From equation (2.42), we have

$$TR^2 = T\frac{\widehat{Q}(\widehat{\mathbf{H}}_t)}{\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}.$$

In order to prove the *Theorems*, we need to analyse both $\widehat{Q}(\widehat{\mathbf{H}}_t)$ and $\widehat{var}(\widehat{\varepsilon}_t)$. Some useful intermediate results are provided in this section, and being proved with applications of assumptions stated in Section 2.3. Let us first sketch out the lemmas for tackling the *Theorems*.

To start with, we decompose the numerator of the test statistic, $\widehat{Q}(\widehat{\mathbf{H}}_t)$, into two terms, a conditional term, $\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$, and a marginal term $\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$, in *Lemma 2.2*. So that the test statistic TR^2 in equation (2.42) can be written as

$$TR^2 = T\frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}.$$

This decomposition plays a central role in our analysis, because later on we show that the marginal term has no influence on the test statistic asymptotically — only the conditional term matters.

Basing upon *Assumption 2.4 (c)*, we next establish an important result on the second-order partial derivative functions (see *Lemma 2.3*). We then proceed to formulate some high level results for ∇_t and $\dot{\mathbf{\nabla}}_t$ (defined in Notation 1 and 2) that are frequently used in subsequent proofs (see *Lemmas 2.4* and *2.5*). In *Lemma 2.6*, we show that the denominator of the test statistic, $\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2$, is asymptotically equivalent to $\sum_{t=p+1}^{T}\varepsilon_t^2$. The last two lemmas, *Lemma 2.8* and *2.9*, concern the key matrix and vector components of the marginal term in the numerator of the test statistic, $\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$, respectively. These results are later used in the proof of *Theorem 2.1* to show that $\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ is small.

Recall from equation (2.41) that

$$\widehat{Q}(\widehat{\mathbf{H}}_t) = \sum_{t=p+1}^{T} \widehat{\varepsilon}_t \widehat{\mathbf{H}}_t' \Big(\sum_{t=p+1}^{T} \widehat{\mathbf{H}}_t \widehat{\mathbf{H}}_t' \Big)^{-1} \sum_{t=p+1}^{T} \widehat{\mathbf{H}}_t \widehat{\varepsilon}_t,$$

and that $\widehat{\mathbf{H}}_t = [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}', \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})']'$. A well known result of partitioned inversion regarding $\widehat{Q}(\widehat{\mathbf{H}}_t)$ [see Nielsen (2006b) for details] is stated in the following lemma. This lemma is extremely useful in the analysis of the test statistic as it helps to separate the numerator into a conditional term and a marginal term. We adapt this decomposition for our test statistic here with a formal proof.

Lemma 2.2. $\widehat{Q}(\widehat{\mathbf{H}}_t) = \widehat{Q}\Big(\begin{bmatrix} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \end{bmatrix} \Big) = \widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})].$

Proof of Lemma 2.2.

It is sufficient to prove that for any vector $\mathbf{Z_t} = [\mathbf{z_{1t}}', \mathbf{z_{2t}}']'$, it holds that

$$Q(\mathbf{Z_t}) = Q\Big(\begin{bmatrix} \mathbf{z_{1t}} \\ \mathbf{z_{2t}} \end{bmatrix} \Big) = Q[\mathbf{z_{1t}} | \mathbf{z_{2t}}] + Q(\mathbf{z_{2t}}),$$

where $(\mathbf{z_{1t}}|\mathbf{z_{2t}})$ denotes the residuals from the regression of $\mathbf{z_{1t}}$ on $\mathbf{z_{2t}}$. Recall from Notation 2.5 the definition of $Q(\cdot)$:

$$Q(\mathbf{Z_t}) = \sum_{t=p+1}^{T} \varepsilon_t \mathbf{Z_t}' \Big(\sum_{t=p+1}^{T} \mathbf{Z_t} \mathbf{Z_t}' \Big)^{-1} \sum_{t=p+1}^{T} \mathbf{Z_t} \varepsilon_t.$$

Partitioning $\mathbf{Z_t}$ into $\mathbf{z_{1t}}$ and $\mathbf{z_{2t}}$ and rearranging yields

$$Q(\mathbf{Z_t}) = \sum_{t=p+1}^{T} \varepsilon_t \begin{bmatrix} \mathbf{z_{1t}} \\ \mathbf{z_{2t}} \end{bmatrix}' \begin{bmatrix} \sum_{t=p+1}^{T} \mathbf{z_{1t}} \mathbf{z_{1t}}' & \sum_{t=p+1}^{T} \mathbf{z_{1t}} \mathbf{z_{2t}}' \\ \sum_{t=p+1}^{T} \mathbf{z_{2t}} \mathbf{z_{1t}}' & \sum_{t=p+1}^{T} \mathbf{z_{2t}} \mathbf{z_{2t}}' \end{bmatrix}^{-1} \sum_{t=p+1}^{T} \begin{bmatrix} \mathbf{z_{1t}} \\ \mathbf{z_{2t}} \end{bmatrix} \varepsilon_t.$$

Let

$$\mathbf{L} = \begin{bmatrix} \mathbf{I} & -\sum_{s=p+1}^{T} \mathbf{z}_{1s} \mathbf{z}_{2s}' (\sum_{s=p+1}^{T} \mathbf{z}_{2s} \mathbf{z}_{2s}')^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

where the upper off-diagonal element, $\sum_{s=p+1}^{T} \mathbf{z}_{1s}\mathbf{z}_{2s}'(\sum_{s=p+1}^{T} \mathbf{z}_{2s}\mathbf{z}_{2s}')^{-1}$, is the estimate of the slope parameter from a simple regression of $\mathbf{z_{1t}}$ on $\mathbf{z_{2t}}$. Recognise that

$$\mathbf{L}\begin{bmatrix}\mathbf{z_{1t}}\\ \mathbf{z_{2t}}\end{bmatrix} = \begin{bmatrix}\mathbf{z_{1t}}|\mathbf{z_{2t}}\\ \mathbf{z_{2t}}\end{bmatrix}. \tag{2.44}$$

Pre- and post-multiplying the middle matrix in $Q(\mathbf{Z_t})$ with $\mathbf{L}'[\mathbf{L}']^{-1}$ and its transpose gives

$$Q(\mathbf{Z_t}) = \sum_{t=p+1}^{T} \varepsilon_t \mathbf{Z_t}'\mathbf{L}'\Big(\sum_{t=p+1}^{T} \mathbf{L}\mathbf{Z_t}\mathbf{Z_t}'\mathbf{L}'\Big)^{-1} \sum_{t=p+1}^{T} \mathbf{L}\mathbf{Z_t}\varepsilon_t. \tag{2.45}$$

Using equation (2.44), we can rewrite the middle matrix, $\sum_{t=p+1}^{T} \mathbf{L}\mathbf{Z_t}\mathbf{Z_t}'\mathbf{L}'$, as

$$\begin{aligned}\sum_{t=p+1}^{T} \mathbf{L}\mathbf{Z_t}\mathbf{Z_t}'\mathbf{L}' &= \sum_{t=p+1}^{T} \begin{bmatrix}\mathbf{z_{1t}}|\mathbf{z_{2t}}\\ \mathbf{z_{2t}}\end{bmatrix}\begin{bmatrix}\mathbf{z_{1t}}|\mathbf{z_{2t}}\\ \mathbf{z_{2t}}\end{bmatrix}'\\ &= \begin{bmatrix}\sum_{t=p+1}^{T}[\mathbf{z_{1t}}|\mathbf{z_{2t}}][\mathbf{z_{1t}}|\mathbf{z_{2t}}]' & \sum_{t=p+1}^{T}[\mathbf{z_{1t}}|\mathbf{z_{2t}}]\mathbf{z_{2t}}'\\ \sum_{t=p+1}^{T}\mathbf{z_{2t}}[\mathbf{z_{1t}}|\mathbf{z_{2t}}]' & \sum_{t=p+1}^{T}\mathbf{z_{2t}}\mathbf{z_{2t}}'\end{bmatrix}.\end{aligned}$$

Observe that the off-diagonal elements, $\sum_{t=p+1}^{T}[\mathbf{z_{1t}}|\mathbf{z_{2t}}]\mathbf{z_{2t}}'$, are $\mathbf{0}$, as the residuals from regressing $\mathbf{z_{1t}}$ on $\mathbf{z_{2t}}$ are orthogonal to $\mathbf{z_{2t}}$. Hence, the above equation becomes

$$\sum_{t=p+1}^{T} \mathbf{L}\mathbf{Z_t}\mathbf{Z_t}'\mathbf{L}' = \begin{bmatrix}\sum_{t=p+1}^{T}[\mathbf{z_{1t}}|\mathbf{z_{2t}}][\mathbf{z_{1t}}|\mathbf{z_{2t}}]' & \mathbf{0}\\ \mathbf{0} & \sum_{t=p+1}^{T}\mathbf{z_{2t}}\mathbf{z_{2t}}'\end{bmatrix}. \tag{2.46}$$

Plugging equations (2.44) and (2.46) into equation (2.45) and rearranging yields

$$\begin{aligned}Q(\mathbf{Z_t}) = \sum_{t=p+1}^{T} \varepsilon_t[\mathbf{z_{1t}}|\mathbf{z_{2t}}]'\Big[\sum_{t=p+1}^{T}[\mathbf{z_{1t}}|\mathbf{z_{2t}}][\mathbf{z_{1t}}|\mathbf{z_{2t}}]'\Big]^{-1} \sum_{t=p+1}^{T}[\mathbf{z_{1t}}|\mathbf{z_{2t}}]\varepsilon_t\\ + \sum_{t=p+1}^{T} \varepsilon_t\mathbf{z_{2t}}'\Big(\sum_{t=p+1}^{T}\mathbf{z_{2t}}\mathbf{z_{2t}}'\Big)^{-1} \sum_{t=p+1}^{T}\mathbf{z_{2t}}\varepsilon_t,\end{aligned}$$

where the RHS is precisely the sum of $Q[\mathbf{z_{1t}}|\mathbf{z_{2t}}]$ and $Q(\mathbf{z_{2t}})$. Therefore,

$$Q(\mathbf{Z_t}) = Q[\mathbf{z_{1t}}|\mathbf{z_{2t}}] + Q(\mathbf{z_{2t}}).$$

The proof is completed by replacing ε_t and the vector $\mathbf{Z_t} = [\mathbf{z_{1t}}', \mathbf{z_{2t}}']'$ with $\widehat{\varepsilon}_t$ and $\widehat{\mathbf{H}}_t = [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}', \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})']'$, respectively.

□

Following from *Lemma 2.2*, we can now rewrite equation (2.42) as

$$TR^2 = T\frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}. \tag{2.47}$$

We next show that higher order partial derivatives of the $g(\cdot)$ functions (with respect to $\boldsymbol{\theta}$) are so small that they are negligible for the asymptotic results.

Recall from *Assumption 2.4 (c)* that the sum of the suprema of $||N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta_0}}||^2$ is bounded in probability for any intermediate point, $\boldsymbol{\theta}_t^*$, between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta_0}$. The subscript t implies that this intermediate point is dependent on x_t. Furthermore, by *Assumption 2.3*, the NLLS estimator $\widehat{\boldsymbol{\theta}}$ is consistent. Therefore $\boldsymbol{\theta}_t^*$ must lie within the neighbourhood of the true parameter vector $\boldsymbol{\theta_0}$. We now proceed to show that the random series $\sum_{t=p+1}^{T}||N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta_0}}||^2$ is itself bounded by combining the aforementioned assumptions.

Lemma 2.3. *Let*

$$R(x_t;\widehat{\boldsymbol{\theta}}) = g(x_t;\widehat{\boldsymbol{\theta}}) - g(x_t;\boldsymbol{\theta_0}) - (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})'\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}).$$

If Assumption 2.3 and 2.4 are satisfied, then

$$\sum_{t=p+1}^{T}[R(x_t;\widehat{\boldsymbol{\theta}})]^2 = o_p(1).$$

Proof of Lemma 2.3.

On Taylor expanding $g(x_t;\widehat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$ around $\boldsymbol{\theta_0}$, we have

$$g(x_t;\widehat{\boldsymbol{\theta}}) = g(x_t;\boldsymbol{\theta_0}) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}),$$

where $\boldsymbol{\theta}_t^*$ is an intermediate point between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta_0}$, and is dependent on x_t. Recall that

$$R(x_t;\widehat{\boldsymbol{\theta}}) = g(x_t;\widehat{\boldsymbol{\theta}}) - g(x_t;\boldsymbol{\theta_0}) - (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}).$$

Combining above equations gives

$$R(x_t;\widehat{\boldsymbol{\theta}}) = \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}),$$

So $R(x_t;\widehat{\boldsymbol{\theta}})$ denotes the higher order expansion terms of $g(x_t;\widehat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$. Take the squares and then sum both sides of the above equation from $t = p+1$ to T to get

$$\sum_{t=p+1}^{T} [R(x_t;\widehat{\boldsymbol{\theta}})]^2 = \frac{1}{4} \sum_{t=p+1}^{T} \{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})\}^2.$$

Pre-and post-multiplying the $\ddot{g}_{\boldsymbol{\theta\theta}}$ matrix with $N_{T,\boldsymbol{\theta_0}}'^{-1} N_{T,\boldsymbol{\theta_0}}'$ and its transpose gives

$$\sum_{t=p+1}^{T} [R(x_t;\widehat{\boldsymbol{\theta}})]^2 = \frac{1}{4} \sum_{t=p+1}^{T} \{\widehat{\boldsymbol{\phi}}_T' N_{T,\boldsymbol{\theta_0}}' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*) N_{T,\boldsymbol{\theta_0}} \widehat{\boldsymbol{\phi}}_T\}^2,$$

where $\widehat{\boldsymbol{\phi}}_T = N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$ following from Notation 2.4.

Apply the norm compatibility to get

$$\left| \sum_{t=p+1}^{T} [R(x_t;\widehat{\boldsymbol{\theta}})]^2 \right| \leq \frac{1}{4} \left\| \widehat{\boldsymbol{\phi}}_T \right\|^4 \sum_{t=p+1}^{T} \left\| N_{T,\boldsymbol{\theta_0}}' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*) N_{T,\boldsymbol{\theta_0}} \right\|^2.$$

Recognise that $\widehat{\boldsymbol{\phi}}_T = \mathbf{o_p}(T^{\tau})$ by *Assumption 2.3*, so $||\widehat{\boldsymbol{\phi}}_T||^4$ is $o_p(T^{4\tau})$. It remains to show that the sum $\sum_{t=p+1}^{T} ||N_{T,\boldsymbol{\theta_0}}' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*) N_{T,\boldsymbol{\theta_0}}||^2$ is of order $T^{-4\tau}$.

Let

$$G(\boldsymbol{\theta}^*) = T^{4\tau} \sum_{t=p+1}^{T} \left\| N_{T,\boldsymbol{\theta_0}}' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*) N_{T,\boldsymbol{\theta_0}} \right\|^2.$$

To prove the lemma, it is equivalent to show that, for every $\varepsilon > 0$, there exists a T_0 such that $\forall T > T_0$,

$$\mathbb{P}\{G(\boldsymbol{\theta}^*) > \varepsilon\} \to 0.$$

Suppose *Assumption 2.4* is satisfied, so we can apply *Lemma A.9* of Berenguer-Rico and Nielsen (2020) on $G(\boldsymbol{\theta}^*)$ to get

$$\mathbb{P}\{G(\boldsymbol{\theta}^*) > \varepsilon\} \leq \mathbb{P}\{T^{4\tau} \sup_{\tilde{\boldsymbol{\theta}}:||N_{T,\boldsymbol{\theta}_0}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)||\leq K} \sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta}_0} \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}}) N_{T,\boldsymbol{\theta}_0} \right\|^2 > \varepsilon\}. \tag{2.48}$$

Observe that the supremum of a sum is bounded by the sum of suprema so that

$$\sup_{\tilde{\boldsymbol{\theta}}:||N_{T,\boldsymbol{\theta}_0}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)||\leq K} \sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta}_0} \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}}) N_{T,\boldsymbol{\theta}_0} \right\|^2 \leq \sum_{t=p+1}^{T} \sup_{\tilde{\boldsymbol{\theta}}:||N_{T,\boldsymbol{\theta}_0}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)||\leq K} \left\| N'_{T,\boldsymbol{\theta}_0} \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}}) N_{T,\boldsymbol{\theta}_0} \right\|^2.$$

It then follows

$$\mathbb{P}\{T^{4\tau} \sup_{\tilde{\boldsymbol{\theta}}:||N_{T,\boldsymbol{\theta}_0}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)||\leq K} \sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta}_0} \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}}) N_{T,\boldsymbol{\theta}_0} \right\|^2 > \varepsilon\} \leq \mathbb{P}\{T^{4\tau} \sum_{t=p+1}^{T} \sup_{\tilde{\boldsymbol{\theta}}:||N_{T,\boldsymbol{\theta}_0}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)||\leq K} \left\| N'_{T,\boldsymbol{\theta}_0} \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}}) N_{T,\boldsymbol{\theta}_0} \right\|^2 > \varepsilon\}. \tag{2.49}$$

From *Assumption 2.4 (c)*, for every $K > 0$,

$$\sum_{t=p+1}^{T} \sup_{\tilde{\boldsymbol{\theta}}:||N_{T,\boldsymbol{\theta}_0}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)||\leq K} \left\| N'_{T,\boldsymbol{\theta}_0} \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}}) N_{T,\boldsymbol{\theta}_0} \right\|^2 = o_p(T^{-4\tau}),$$

and thus

$$T^{4\tau} \sum_{t=p+1}^{T} \sup_{\tilde{\boldsymbol{\theta}}:||N_{T,\boldsymbol{\theta}_0}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)||\leq K} \left\| N'_{T,\boldsymbol{\theta}_0} \ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}}) N_{T,\boldsymbol{\theta}_0} \right\|^2 = o_p(1).$$

This indicates that

$$\mathbb{P}\{T^{4\tau}\sum_{t=p+1}^{T}\sup_{\tilde{\boldsymbol{\theta}}:||N_{T,\boldsymbol{\theta_0}}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta_0})||\leq K}\left\|N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\tilde{\boldsymbol{\theta}})N_{T,\boldsymbol{\theta_0}}\right\|^2>\varepsilon\}\to 0. \tag{2.50}$$

The proof is completed by combining equations (2.48) to (2.50).

□

From the proof above, we can form the following sub-lemma regarding the second-order partial derivative functions, $\ddot{g}_{\boldsymbol{\theta\theta}}$.

Lemma 2.3a. *If Assumption 2.4 is satisfied, then*

$$\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta_0}}\right\|^2=\mathbf{o_p}(T^{-4\tau}).$$

The next lemma concerns $\nabla_t = g(x_t;\widehat{\boldsymbol{\theta}})-g(x_t;\boldsymbol{\theta_0})$ defined in *Notation 2.1*.

Lemma 2.4. *If Assumptions 2.3 and 2.4 are satisfied, then* $\sum_{t=p+1}^{T}\nabla_t^2=o_p(T^{2\tau})$.

Proof of Lemma 2.4.

On expanding the $g(\cdot;\widehat{\boldsymbol{\theta}})$ in ∇_t with respect to $\boldsymbol{\theta}$ around $\boldsymbol{\theta_0}$ by Taylor's Theorem, we have

$$g(x_t;\widehat{\boldsymbol{\theta}})=g(x_t;\boldsymbol{\theta_0})+(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})'\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})+R(x_t;\widehat{\boldsymbol{\theta}}).$$

Substitute this equation back into ∇_t and rearrange to get

$$\sum_{t=p+1}^{T}\nabla_t^2=\sum_{t=p+1}^{T}[g(x_t;\widehat{\boldsymbol{\theta}})-g(x_t;\boldsymbol{\theta_0})]^2=\sum_{t=p+1}^{T}\{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})'\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})+R(x_t;\widehat{\boldsymbol{\theta}})\}^2.$$

For any two variables a and b, it holds that $(a+b)^2\leq 2(a^2+b^2)$. Applying this inequality to the above equation and rearranging gives

$$\sum_{t=p+1}^{T}\nabla_t^2\leq 2\{\sum_{t=p+1}^{T}[(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})'\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})]^2+\sum_{t=p+1}^{T}[R(x_t;\widehat{\boldsymbol{\theta}})]^2\}. \tag{2.51}$$

1. Take the first term on the RHS. Pre-multiply $\dot{g}_{\boldsymbol{\theta}}(\cdot)$ by $N'^{-1}_{T,\boldsymbol{\theta_0}} N'_{T,\boldsymbol{\theta_0}}$ and rearrange to get

$$\sum_{t=p+1}^{T} \left| (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0}) \right|^2 = \sum_{t=p+1}^{T} \left| \widehat{\boldsymbol{\phi}}'_T N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0}) \right|^2,$$

where $\widehat{\boldsymbol{\phi}}_T = N^{-1}_{T,\boldsymbol{\theta_0}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$ by Notation 2.4. By the norm inequality we have

$$\sum_{t=p+1}^{T} \left| \widehat{\boldsymbol{\phi}}'_T N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0}) \right|^2 \leq \left\| \widehat{\boldsymbol{\phi}}_T \right\|^2 \sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0}) \right\|^2.$$

According to *Assumption 2.3*, $||\widehat{\boldsymbol{\phi}}_T||$ is $o_p(T^{\tau})$. By *Assumption 2.4 (a)*, the bound on the squared spectral norm $||N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0})||^2$ applies, so that $\sum_{t=p+1}^{T} ||N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0})||^2$ is $O_p(1)$. Hence,

$$\sum_{t=p+1}^{T} \left| (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0}) \right|^2 = \sum_{t=p+1}^{T} \left| \widehat{\boldsymbol{\phi}}'_T N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0}) \right|^2 \leq o_p(T^{2\tau}) \cdot O_p(1) = o_p(T^{2\tau}). \tag{2.52}$$

2. Consider the second term on the RHS of equation (2.51). The application of *Lemma 2.3*, which holds under *Assumptions 2.3* and *2.4* , yields

$$\sum_{t=p+1}^{T} [R(x_t; \widehat{\boldsymbol{\theta}})]^2 = o_p(1). \tag{2.53}$$

Substituting results in equations (2.52) and (2.53) back into equation (2.51) completes the proof of *Lemma 2.4*.

□

This following lemma deals with $\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}} = N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) - N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0})$ defined in *Notation 2.2*.

Lemma 2.5. *Under Assumptions 2.3 and 2.4 ,* $\sum_{t=p+1}^{T} ||\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}}||^2 = o_p(T^{-2\tau})$.

Proof of Lemma 2.5.

By definition,

$$\sum_{t=p+1}^{T}\left\|\dot{\nabla}_t\right\|^2=\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})-N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\right\|^2.$$

Taylor expand the $\dot{g}_{\boldsymbol{\theta}}(\cdot;\widehat{\boldsymbol{\theta}})$ term with respect to $\widehat{\boldsymbol{\theta}}$ around $\boldsymbol{\theta_0}$ and rearrange to get

$$\sum_{t=p+1}^{T}\left\|\dot{\nabla}_t\right\|^2=\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}^*_t)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})\right\|^2,$$

where $\boldsymbol{\theta}^*_t$ is an intermediate point between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta_0}$, and is dependent on x_t. Post-multiplying the $\ddot{g}_{\boldsymbol{\theta\theta}}$ matrix by $N_{T,\boldsymbol{\theta_0}}N^{-1}_{T,\boldsymbol{\theta_0}}$ leaves the equation unchanged and we have

$$\sum_{t=p+1}^{T}\left\|\dot{\nabla}_t\right\|^2=\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}^*_t)N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T\right\|^2,$$

where, again, $\widehat{\boldsymbol{\phi}}_T=N^{-1}_{T,\boldsymbol{\theta_0}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})$ by Notation 2.4. By the compatibility property of norms, we have

$$\sum_{t=p+1}^{T}\left\|\dot{\nabla}_t\right\|^2=\left\|\widehat{\boldsymbol{\phi}}_T\right\|^2\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}^*_t)N_{T,\boldsymbol{\theta_0}}\right\|^2.$$

Lemma 2.3a holds following from *Assumption 2.4* , which gives

$$\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}^*_t)N_{T,\boldsymbol{\theta_0}}\right\|^2=o_p(T^{-4\tau}).$$

Furthermore, *Assumption 2.3* shows that $||\widehat{\boldsymbol{\phi}}_T||$ is $o_p(T^{\tau})$. So

$$\sum_{t=p+1}^{T}\left\|\dot{\nabla}_t\right\|^2=o_p(T^{2\tau})\cdot o_p(T^{-4\tau})=o_p(T^{-2\tau}).$$

□

Having established some important preliminary lemmas, we are now ready to analyse the key terms in our test statistic.

Recall from equation (2.47) that

$$TR^2 = \frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}.$$

We begin by showing that the denominator, the sample variance of the residuals, is asymptotically equivalent to the sample variance of the true error terms, which, by the martingale Laws of Large Numbers, converges in probability to its population counterparts.

Lemma 2.6. *If Assumptions 2.1, 2.3 and 2.4 are satisfied, then*

$$T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2 = T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2 + o_p(T^{-\frac{1}{2}}).$$

Proof of Lemma 2.6.

Consider the difference between $\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2$ and $\sum_{t=p+1}^{T}\varepsilon_t^2$. Provided that *Assumptions 2.1*, *2.3* and *2.4* are satisfied and *Lemma 2.4* holds, we can apply *Assumption A.2* of Berenguer-Rico and Nielsen (2020), which is proved to be true in their proof of *Theorem 3.6*, to get

$$\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2 - \sum_{t=p+1}^{T}\varepsilon_t^2 = o_p(T^{\frac{1}{2}}).$$

Rearranging the above equation gives

$$T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2 = T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2 + o_p(T^{-\frac{1}{2}}).$$

□

Recall the martingale Laws of Large Numbers stated in *Lemma 2.1*. We now show that the sample variance of the true error terms converges to the population variance using this asymptotic result.

Lemma 2.7. *If Assumption 2.1 is satisfied, then*

$$T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2 \xrightarrow{p} \sigma_\varepsilon^2.$$

Proof of Lemma 2.7.

Notice that

$$T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2 = T^{-1}\sum_{t=1}^{T}\varepsilon_t^2 - T^{-1}\sum_{t=1}^{p}\varepsilon_t^2. \tag{2.54}$$

First, consider the second term on the RHS, $T^{-1}\sum_{t=1}^{p}\varepsilon_t^2$. Note that for each t, the random variable ε_t is $O_p(1)$. Thus, by Slutsky's theorem, ε_t^2 must also be stochastically bounded. Given that p is fixed and finite, $\sum_{t=1}^{p}\varepsilon_t^2$ is $O_p(1)$, and we obtain

$$T^{-1}\sum_{t=1}^{p}\varepsilon_t^2 = T^{-1}\cdot O_p(1) = O_p(T^{-1}).$$

It remains to show that the first term on the RHS of equation (2.54), $T^{-1}\sum_{t=1}^{T}\varepsilon_t^2$, has probability limit σ_ε^2. Observe that $\{\varepsilon_t^2 - \sigma_\varepsilon^2\}$ is a martingale difference sequence, since by *Assumption 2.1* $\mathbb{E}(\varepsilon_t^2 - \sigma_\varepsilon^2|\mathscr{F}_{t-1}) = 0$. So we could apply the Laws of Large Numbers for martingale differences in *Lemma 2.1* to $\varepsilon_t^2 - \sigma_\varepsilon^2$ after verifying the condition in equation (2.43).

Using the inequality that $|a-b|^2 \leq a^2 - b^2$ we get

$$\sum_{t=1}^{\infty}\frac{\mathbb{E}(|\varepsilon_t^2 - \sigma_\varepsilon^2|^2|\mathscr{F}_{t-1})}{t^2} \leq \sum_{t=1}^{\infty}\frac{\mathbb{E}(\varepsilon_t^4 + \sigma_\varepsilon^4|\mathscr{F}_{t-1})}{t^2}.$$

Given that σ_ε is a constant and note that $\sum_{t=1}^{\infty}\frac{1}{t^2} = \frac{\pi^2}{6}$, we can simplify the RHS of the above equation such that

$$\sum_{t=1}^{\infty}\frac{\mathbb{E}(|\varepsilon_t^2 - \sigma_\varepsilon^2|^2|\mathscr{F}_{t-1})}{t^2} \leq \sum_{t=1}^{\infty}\frac{\mathbb{E}(\varepsilon_t^4|\mathscr{F}_{t-1})}{t^2} + \sigma_\varepsilon^4\frac{\pi^2}{6}.$$

Using the condition in *Assumption 2.1(c)* that $\sup_t \mathbb{E}[|\varepsilon_t|^\varphi|\mathscr{F}_{t-1}] < \infty$ for some

$\varphi > 2$ and letting $\varphi = 4$, we obtain

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}(|\varepsilon_t^2 - \sigma_\varepsilon^2|^2 | \mathscr{F}_{t-1})}{t^2} \leq \{\sup_t \mathbb{E}[\varepsilon_t^4 | \mathscr{F}_{t-1}]\} \frac{\pi^2}{6} + \sigma_\varepsilon^4 \frac{\pi^2}{6} < \infty.$$

Therefore the condition specified in equation (2.43) is satisfied, and we can safely apply *Lemma 2.1* to $\varepsilon_t^2 - \sigma_\varepsilon^2$ such that

$$T^{-1} \sum_{t=1}^{T} \varepsilon_t^2 - \sigma_\varepsilon^2 \xrightarrow{p} 0. \tag{2.55}$$

□

Let's turn our attention now to the marginal $\widehat{Q}$ term in the numerator of the test statistic, $\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]$. By Notation 2.5,

$$\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})] = \sum_{t=p+1}^{T} \widehat{\varepsilon}_t \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})' [\sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})']^{-1} \sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) \widehat{\varepsilon}_t.$$

The next lemma concerns the middle matrix of $\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]$, which is the normalised second moment of $\dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})$. Recall from *Notation 2.3* that

$$\mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}} = \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta}_0) \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta}_0)' N_{T,\boldsymbol{\theta}_0}.$$

We next show that $\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0} \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})' N_{T,\boldsymbol{\theta}_0}$ is asymptotically equivalent to $\mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}}$.

Lemma 2.8. *Under Assumptions 2.3 and 2.4 ,*

$$\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0} \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})' N_{T,\boldsymbol{\theta}_0} = \mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}} \left\{ \mathbf{I_q} + \mathbf{o_p}(T^{-\tau}) \right\},$$

where $\mathbf{I_q}$ *is an identity matrix of size q.*

Proof of Lemma 2.8.

Recall that $\mathbf{S}_{\dot{g}\dot{g}} = \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})' N_{T,\boldsymbol{\theta_0}}$. The first step is to show that the difference between $\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})' N_{T,\boldsymbol{\theta_0}}$ and $\mathbf{S}_{\dot{g}\dot{g}}$ is small. Add and subtract $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ from the term $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ and rearrange to get

$$\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})' N_{T,\boldsymbol{\theta_0}} - \mathbf{S}_{\dot{g}\dot{g}}$$
$$= \sum_{t=p+1}^{T} \{N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) + \dot{\mathbf{V}}_{\boldsymbol{t}}\} \{\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})' N_{T,\boldsymbol{\theta_0}} + \dot{\mathbf{V}}_{\boldsymbol{t}}'\} - \mathbf{S}_{\dot{g}\dot{g}},$$

where $\dot{\mathbf{V}}_{\boldsymbol{t}} = N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) - N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ following from Notation 2.2. Rearranging the above equation gives

$$\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})' N_{T,\boldsymbol{\theta_0}} - \mathbf{S}_{\dot{g}\dot{g}} = \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \dot{\mathbf{V}}_{\boldsymbol{t}}'$$
$$+ \sum_{t=p+1}^{T} \dot{\mathbf{V}}_{\boldsymbol{t}} \dot{\mathbf{V}}_{\boldsymbol{t}}' + \sum_{t=p+1}^{T} \dot{\mathbf{V}}_{\boldsymbol{t}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})' N_{T,\boldsymbol{\theta_0}}.$$

To show that the difference of the two matrices is small, it suffices to show that the norm of this difference is small. Notice that the spectral norm of a matrix is equal to that of its transpose. Therefore, taking the norm of the difference and applying the triangle inequality gives

$$\left\| \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})' N_{T,\boldsymbol{\theta_0}} - \mathbf{S}_{\dot{g}\dot{g}} \right\| \leq 2 \sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \right\| \left\| \dot{\mathbf{V}}_{\boldsymbol{t}} \right\|$$
$$+ \sum_{t=p+1}^{T} \left\| \dot{\mathbf{V}}_{\boldsymbol{t}} \right\|^2. \quad (2.56)$$

We only need to show that the RHS is small.

1. Following from *Assumptions 2.3* and *2.4* , *Lemma 2.5* holds, and we have

$$\sum_{t=p+1}^{T} \left\| \dot{\mathbf{V}}_{\boldsymbol{t}} \right\|^2 = o_p(T^{-2\tau}).$$

2. Consider the cross-product term between the norms of $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ and $\dot{\mathbf{V}}_{\boldsymbol{t}}$. Apply

the Cauchy-Schwarz inequality to get

$$\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\right\|\left\|\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}}\right\| \leq [\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\right\|^2]^{\frac{1}{2}}\cdot[\sum_{t=p+1}^{T}\left\|\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}}\right\|^2]^{\frac{1}{2}}.$$

We have $\sum_{t=p+1}^{T}||N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})||^2 = O_p(1)$ under *Assumption 2.4 (a)*. Combining *Assumption 2.4 (a)* and *Lemma 2.5* gives

$$\sum_{t=p+1}^{T}\left\|N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\right\|\left\|\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}}\right\| \leq \left[O_p(1)\right]^{\frac{1}{2}}\cdot\left[o_p(T^{-2\tau})\right]^{\frac{1}{2}} = o_p(T^{-\tau}).$$

Substituting above results back into equation (2.56) gives

$$\left\|\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'N_{T,\boldsymbol{\theta_0}} - \mathbf{S_{\dot{g}\dot{g}}}\right\| = o_p(T^{-\tau}),$$

and thus,

$$\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'N_{T,\boldsymbol{\theta_0}} - \mathbf{S_{\dot{g}\dot{g}}} = \mathbf{o_p}(T^{-\tau}).$$

Moreover, under *Assumption 2.4 (b)*, the inverse of $\mathbf{S_{\dot{g}\dot{g}}}$, $\mathbf{S^{-1}_{\dot{g}\dot{g}}}$, exists and is stochastically bounded. For this reason, the above equation can be rewritten as

$$\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'N_{T,\boldsymbol{\theta_0}} = \mathbf{S_{\dot{g}\dot{g}}}\{\mathbf{I_q} + \mathbf{o_p}(T^{-\tau})\}.$$

□

We now move on to analysing the vector component of $\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$, which is $\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t$.

Lemma 2.9. *Under Assumptions 2.2 to 2.4 ,*

$$\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t = \mathbf{O_p}(1).$$

Proof of Lemma 2.9.

Note that

$$\sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t = \sum_{t=1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t - \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t. \tag{2.57}$$

We shall show that the asymptotic behaviour of $\sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t$ depends solely on the second term of the RHS, because the first term is exactly zero by the first-order conditions (FOCs) of the non-linear least squares estimation.

1. Take the first term on the RHS of equation (2.57). From equation (2.4), the FOCs with respect to $\boldsymbol{\theta}$ are

$$\dot{S}_T(\boldsymbol{\theta}) = \frac{\partial S_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\sum_{t=1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta})\Big[y_t - g(x_t;\boldsymbol{\theta})\Big].$$

The NLLS estimators are derived by setting the FOCs to zero. Hence,

$$\sum_{t=1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\Big[y_t - g(x_t;\widehat{\boldsymbol{\theta}})\Big] = \mathbf{0}.$$

Notice that the term between the square brackets is exactly the NLLS residuals, $\widehat{\varepsilon}_t$. It is immediate that

$$\sum_{t=1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t = \mathbf{0}. \tag{2.58}$$

2. Plugging equation (2.58) back into equation (2.57) yields

$$\sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t = -\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t.$$

To complete the proof, it suffices to show that $\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t$ is stochastically bounded. By the definition of $\widehat{\varepsilon}_t$ and the model equation, we have

$$\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t = \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\{\varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\widehat{\boldsymbol{\theta}})\}. \tag{2.59}$$

Recall from Notation 2.4 that

$$\widehat{\boldsymbol{\phi}}_T = N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}).$$

Rearranging gives

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T,$$

and we can rewrite equation (2.59) as

$$\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t = \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\{\varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\}.$$

Next subtract $\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\varepsilon_t$ from $\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t$. Following from the above equation, we have

$$\begin{aligned}
&\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t - \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\varepsilon_t \\
&= \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\{\varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\} - \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\varepsilon_t.
\end{aligned}$$

Rearranging the above equation gives

$$\begin{aligned}
&\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t - \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\varepsilon_t \\
&= \sum_{t=1}^{p} [\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})]\{\varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\} \\
&\quad + \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\{g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\}. \qquad (2.60)
\end{aligned}$$

We next consider analysing the differences $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ and $g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)$ in the above equation, and show that they are small. Under *Assumptions 2.2* and *2.3*, we have $N_{T,\boldsymbol{\theta_0}} = \mathbf{O}(T^{-\iota})$ and $\widehat{\boldsymbol{\phi}}_T = \mathbf{o_p}(T^{\tau})$ $(0 < \tau < \iota)$, respectively. Thus, for all $\xi > 0$, there exits T_0 such that, for every $T > T_0$, it holds

that

$$\mathbb{P}(\left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\| > \xi) < \xi.$$

Moreover, *Assumption 2.4* guarantees that both the functions g and $\dot{g}_{\boldsymbol{\theta}}$ are continuous in $\boldsymbol{\theta}$. Then, given t, by the Mean Value Theorem, we have for $g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)$ and $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)$,

$$\left|g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\right| \leq \left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\| \cdot \|\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}^*)\|, \tag{2.61}$$

and

$$\left\|\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\right\| \leq \left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\| \cdot \|\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}^{**})\|, \tag{2.62}$$

where $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{**}$ are intermediate points between $\boldsymbol{\theta_0}$ and $\widehat{\boldsymbol{\theta}}$, and dependent on the functions g and $\dot{g}_{\boldsymbol{\theta}}$, respectively.

As both $||\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}^*)||$ and $||\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}^{**})||$ are bounded by their suprema, we can further write equations (2.61) and (2.62) as

$$\left|g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\right| \leq \left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\| \cdot \sup_{\tilde{\boldsymbol{\theta}}:\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\|\leq\left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\|} \left\|\dot{g}_{\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}})\right\|, \tag{2.63}$$

and

$$\left\|\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\right\| \leq \left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\| \cdot \sup_{\tilde{\boldsymbol{\theta}}:\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\|\leq\left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\|} \left\|\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\tilde{\boldsymbol{\theta}})\right\|. \tag{2.64}$$

Notice that $N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}} = \mathbf{o_p}(T^{\tau-\iota}) = o_p(1)$, as τ is smaller than ι by *Assumption 2.3*. In addition, *Assumption 2.4* ensures that both the functions $\dot{g}_{\boldsymbol{\theta}}$ and $\ddot{g}_{\boldsymbol{\theta\theta}}$ are continuous in x_t and $\boldsymbol{\theta}$. By the boundedness theorem, continuous functions are bounded on closed bounded intervals, and their suprema with respect to the intervals must be finite. It

then follows that, for the two suprema in equations (2.63) and (2.64),

$$\sup_{\tilde{\boldsymbol{\theta}}:\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\|\leq\left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\|}\left\|\dot{g}_{\boldsymbol{\theta}}(x_t;\tilde{\boldsymbol{\theta}})\right\| = O_p(1),$$

and

$$\sup_{\tilde{\boldsymbol{\theta}}:\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\|\leq\left\|N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}\right\|}\left\|\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\tilde{\boldsymbol{\theta}})\right\| = O_p(1).$$

Combining the results above and substituting back into equations (2.61) and (2.62) gives

$$\left|g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\right| = o_p(1), \tag{2.65}$$

and

$$\left\|\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\right\| = o_p(1). \tag{2.66}$$

We are now ready to analyse the terms in equation (2.60),

$$\begin{aligned}
&\sum_{t=1}^{p}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t - \sum_{t=1}^{p}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\varepsilon_t \\
=&\sum_{t=1}^{p}[\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})]\{\varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\} \\
&+\sum_{t=1}^{p}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\{g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\}.
\end{aligned}$$

Consider the second summation term on the RHS. By the norm inequality we have

$$\begin{aligned}
&\left\|\sum_{t=1}^{p}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\{g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\}\right\| \\
&\qquad\qquad\leq\sum_{t=1}^{p}\|\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\|\left|g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\right|.
\end{aligned}$$

Notice that for each t, $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ is $\mathbf{O_p}(1)$, and following from equation (2.65), $|g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)|$ is $o_p(1)$. Thus

$$\|\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\|\left|g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0}+N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\right| = o_p(1),\ \forall t.$$

Provided that there are finitely many summands, that is p is finite, we have

$$\left\| \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\{g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\} \right\| = o_p(1). \tag{2.67}$$

Now take the first term on the RHS of equation (2.60). Apply the norm inequality to get

$$\begin{aligned} &\left\| \sum_{t=1}^{p} [\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})]\{\varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\} \right\| \\ &\quad \leq \sum_{t=1}^{p} \left\| \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \right\| \left| \varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) \right|. \end{aligned}$$

Note that for each t, the random variable ε_t is $O_p(1)$. Equation (2.65) together with equation (2.66) ensure that the two difference terms, $||\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})||$ and $g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)$, are both $o_p(1)$. Since p is given, for finitely many summands, it holds that

$$\left\| \sum_{t=1}^{p} [\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T) - \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})]\{\varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\boldsymbol{\theta_0} + N_{T,\boldsymbol{\theta_0}}\widehat{\boldsymbol{\phi}}_T)\} \right\| = o_p(1). \tag{2.68}$$

Substituting equations (2.67) and (2.68) back into equation (2.60) and rearranging yields

$$\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t = \sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\varepsilon_t + o_p(1).$$

For each t, both $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ and ε_t are stochastically bounded. Given fixed p, $\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\varepsilon_t$ is $\mathbf{O_p}(1)$. As a result, $\sum_{t=1}^{p} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t$ is also $\mathbf{O_p}(1)$.

□

2.D Proof of Theorems

In this section, we restate the *Theorems* established in Section 2.3 in an essentially equivalent form based on Notation 2.5 and Lemma 2.2, and give a direct proof based on these restatements.

Theorem 2.1. *If Assumptions 2.1 to 2.4 are satisfied, then*

$$TR^2 = \frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2} + o_p(1).$$

Proof of Theorem 2.1.

Recall from equation (2.47) that

$$TR^2 = \frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}. \tag{2.69}$$

Take the $\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ term in the numerator and by Notation 2.5,

$$\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \sum_{t=p+1}^{T}\widehat{\varepsilon}_t\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'[\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})']^{-1}\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t.$$

Pre- and post-multiplying the middle matrix with $N_{T,\boldsymbol{\theta_0}}N_{T,\boldsymbol{\theta_0}}^{-1}$ and its transpose gives

$$\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \sum_{t=p+1}^{T}\widehat{\varepsilon}_t\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'N_{T,\boldsymbol{\theta_0}}[\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'N_{T,\boldsymbol{\theta_0}}]^{-1}$$
$$\cdot N'_{T,\boldsymbol{\theta_0}}\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t.$$

Apply the norm compatibility to the above equation to obtain

$$\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] \leq \left\|N_{T,\boldsymbol{\theta_0}}\right\|^2\left\|\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t\right\|^2\cdot\left\|[\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'N_{T,\boldsymbol{\theta_0}}]^{-1}\right\|. \tag{2.70}$$

Under *Assumptions 2.3* and *2.4* , *Lemma 2.8* is satisfied and we have

$$[\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'N_{T,\boldsymbol{\theta_0}}]^{-1} = \{\mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}}\Big\{\mathbf{I_q} + \mathbf{o_p}(T^{-\tau})\Big\}\}^{-1}.$$

According to *Assumptions 2.2* to *2.4* , *Lemma 2.9* holds, and therefore

$$\sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\widehat{\varepsilon}_t = \mathbf{O_p}(1).$$

From *Assumption 2.2*, the normalisation matrix $N_{T,\boldsymbol{\theta_0}}$ is $\mathbf{O}(T^{-\iota})$. Combining above results and substituting back into equation (2.70) gives

$$\widehat{Q}\left[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] \leq O(T^{-2\iota})\cdot \left\|\mathbf{O_p}(1)\right\|^2 \cdot \left\|\{\mathbf{S_{\dot{g}\dot{g}}}\left\{\mathbf{I_q}+\mathbf{o_p}(T^{-\tau})\right\}\}^{-1}\right\|.$$

Rearrange the above equation and note that, by *Assumption 2.4* , $\mathbf{S_{\dot{g}\dot{g}}^{-1}}$ exists and is $\mathbf{O_p}(1)$ to get

$$\widehat{Q}\left[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] \leq O(T^{-2\iota})O_p(1)\cdot\{1+o_p(T^{-\tau})\}\cdot O_p(1) = o_p(1). \tag{2.71}$$

Plugging equation (2.71) back into the test statistic in equation (2.69) gives

$$TR^2 = \frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]+\widehat{Q}[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2} = \frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]+o_p(1)}{T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}. \tag{2.72}$$

Recall from *Lemma 2.6*, which holds under *Assumptions 2.1*, *2.3* and *2.4* , that

$$T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2 = T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2 + o_p(T^{-\frac{1}{2}}).$$

Since the first term on the RHS is $O_p(1)$ under *Lemma 2.7* if *Assumption 2.1* is satisfied, $T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2$ must also be stochastically bounded. So we can further rewrite equation (2.72) as

$$TR^2 = \frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2} + o_p(1),$$

which is precisely the assertion of *Theroem 2.1*.

□

We now focus on the leading term that remains in the test statistic, $\frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2}$.

Notice that by Notation 2.5, we have for the numerator

$$\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \sum_{t=p+1}^{T} \widehat{\varepsilon}_t [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]' \{ \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]' \}^{-1} \cdot \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t.$$

We will show that the current residuals in the leading term, $\widehat{\varepsilon}_t$, can be replaced by the true error terms, ε_t, such that $\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ and $\sum_{t=p+1}^{T} \widehat{\varepsilon}_t^2$ are asymptotically equivalent to $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$ and $\sum_{t=p+1}^{T} \varepsilon_t^2$, respectively.

Theorem 2.2. *If Assumptions 2.1 to 2.5 are satisfied, then*

$$TR^2 = \frac{Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T} \varepsilon_t^2} + o_p(1).$$

Proof of Theorem 2.2.

From *Theorem 2.1*, which is satisfied under *Assumptions 2.1* to *2.4* , we have,

$$TR^2 = \frac{\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T} \widehat{\varepsilon}_t^2} + o_p(1). \tag{2.73}$$

Take the $\widehat{Q}$ term and recall that

$$\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \sum_{t=p+1}^{T} \widehat{\varepsilon}_t [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]' \{ \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]' \}^{-1} \cdot \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t.$$

Noting that the middle matrix can be written as $\{\sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}$ multiplying by itself, we can rear-

range the equation to get

$$\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \sum_{t=p+1}^{T} \widehat{\varepsilon}_t[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}$$
$$\cdot\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t.$$
$$=\widehat{\mathbf{R}}'_{\widehat{\varepsilon}_{\mathbf{t-1}}\dot{\mathbf{g}}}\cdot\widehat{\mathbf{R}}_{\widehat{\varepsilon}_{\mathbf{t-1}}\dot{\mathbf{g}}}, \tag{2.74}$$

where we introduce the notation

$$\widehat{\mathbf{R}}_{\widehat{\varepsilon}_{\mathbf{t-1}}\dot{\mathbf{g}}} = \{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t. \tag{2.75}$$

Notice that the 'hat' of $\widehat{R}$ emphasises that it is a function of the current residuals, $\widehat{\varepsilon}_t$.

To start with, consider the column vector $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t$ in $\widehat{\mathbf{R}}_{\widehat{\varepsilon}_{\mathbf{t-1}}\dot{\mathbf{g}}}$. Notice that it is invariant whether we condition $\widehat{\varepsilon}_t$ or $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ on $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. Therefore,

$$\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t = \sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}[\widehat{\varepsilon}_t|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$$
$$= \sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}[\varepsilon_t + g(x_t;\boldsymbol{\theta_0}) - g(x_t;\widehat{\boldsymbol{\theta}})|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})], \tag{2.76}$$

where the second equality follows from the definition of $\widehat{\varepsilon}_t$ and the model equation. Taylor expanding $g(\cdot;\boldsymbol{\theta_0})$ with respect to $\boldsymbol{\theta}$ around $\widehat{\boldsymbol{\theta}}$ yields

$$g(x_t;\boldsymbol{\theta_0}) = g(x_t;\widehat{\boldsymbol{\theta}}) + (\boldsymbol{\theta_0}-\widehat{\boldsymbol{\theta}})'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + R(x_t;\widehat{\boldsymbol{\theta}}).$$

Substitute this equation back into equation (2.76) and rearrange to obtain

$$\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t = \sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}[\varepsilon_t + (\boldsymbol{\theta_0}-\widehat{\boldsymbol{\theta}})'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + R(x_t;\widehat{\boldsymbol{\theta}})|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})].$$

Since the residual of regressing $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ on itself is just zero, we can safely drop

the middle term between the square brackets in the above equation, and hence

$$\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t = \sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}[\varepsilon_t + R(x_t;\widehat{\boldsymbol{\theta}})|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$$
$$= \sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}[\varepsilon_t|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}[R(x_t;\widehat{\boldsymbol{\theta}})|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})],$$

where the second equality follows from the additive separability.

Note that we can rewrite the above equation as

$$\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\widehat{\varepsilon}_t = \sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\varepsilon_t + \sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]R(x_t;\widehat{\boldsymbol{\theta}}). \tag{2.77}$$

Plugging equation (2.77) back into $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}}$ defined in equation (2.75) and rearranging yields

$$\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}} = \{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\varepsilon_t$$
$$+\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]R(x_t;\widehat{\boldsymbol{\theta}}).$$

Let

$$\mathbf{R}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}} = \{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\varepsilon_t,$$

where we drop the 'hat' from $\widehat{R}$ here to emphasise that R is a function of the current error terms, ε_t.

Then we can rewrite the equation for $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}}$ as

$$\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}} = \mathbf{R}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}} + \{\textstyle\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\textstyle\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]R(x_t;\widehat{\boldsymbol{\theta}}). \tag{2.78}$$

Consider the second term on the RHS of the above equation and we need to show that it is small. By the norm compatibility together with the triangle inequality, we

have for the second term,

$$\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]R(x_t;\widehat{\boldsymbol{\theta}})$$
$$\leq\{\sum_{t=p+1}^{T}\left\|[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\right\|^2\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}\left\|[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\right\|\left|R(x_t;\widehat{\boldsymbol{\theta}})\right|.$$

Apply the Cauchy-Schwarz Inequality to the second summation term to get

$$\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]R(x_t;\widehat{\boldsymbol{\theta}})$$
$$\leq\{\sum_{t=p+1}^{T}\left\|[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\right\|^2\}^{-\frac{1}{2}}\{\sum_{t=p+1}^{T}\left\|[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\right\|^2\}^{\frac{1}{2}}\{\sum_{t=p+1}^{T}[R(x_t;\widehat{\boldsymbol{\theta}})]^2\}^{\frac{1}{2}}.$$

Noting that the product of the first two terms on the RHS is just 1, this gives

$$\{\textstyle\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]R(x_t;\widehat{\boldsymbol{\theta}})\leq\sqrt{\textstyle\sum_{t=p+1}^{T}[R(x_t;\widehat{\boldsymbol{\theta}})]^2}.$$

Under *Assumptions 2.3* and *2.4* , *Lemma 2.3* is satisfied, and hence

$$\sqrt{\sum_{t=p+1}^{T}[R(x_t;\widehat{\boldsymbol{\theta}})]^2}=o_p(1).$$

Combining above results and substituting back into equation (2.78) gives

$$\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}}=\mathbf{R}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}}+\mathbf{o_p}(1).$$

Plugging this back into equation (2.74) yields

$$\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]=[\mathbf{R}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}}+\mathbf{o_p}(1)]'[\mathbf{R}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}}+\mathbf{o_p}(1)].$$

Under *Assumption 2.5*, $\mathbf{R}_{\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{\mathbf{g}}}$ is $\mathbf{O_p}(1)$, and we can rearrange the above equation to

obtain

$$\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = \mathbf{R}'_{\widehat{\varepsilon}_{t-1}\dot{\mathbf{g}}}\mathbf{R}_{\widehat{\varepsilon}_{t-1}\dot{\mathbf{g}}} + o_p(1).$$

Notice that the first term is just $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]$. Thus we have

$$\widehat{Q}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] = Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + o_p(1). \tag{2.79}$$

Plugging equation (2.79) back into equation (2.73) yields

$$TR^2 = \frac{Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + o_p(1)}{T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2} + o_p(1). \tag{2.80}$$

Consider the denominator of the first term. According to *Lemma 2.6*, which holds under *Assumptions 2.1*, *2.3* and *2.4* ,

$$T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2 = T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2 + o_p(T^{-\frac{1}{2}}).$$

By *Assumption 2.1*, $T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2$ converges in probability to σ_ε^2 under *Lemma 2.7*. Thus we can further rewrite the above equation as

$$T^{-1}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2 = T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2[1 + o_p(T^{-\frac{1}{2}})].$$

Substituting these results back into equation (2.80) and rearranging gives

$$TR^2 = \frac{Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2} + o_p(1).$$

□

CHAPTER 3

Asymptotic Irrelevance of Non-stationary Components

3.1 Introduction

Valid inferences on estimated models heavily rely on the assumption that there is no serial correlation in the error terms. This assumption may be violated typically in time series analysis since past disturbances may carry information over time into the future. This thesis addresses the problem of testing the presence of serial correlation in dynamic econometric models. In Chapter 2, we proposed an extension of the auxiliary-regression-based test, which was originally developed by Breusch (1978) and Godfrey (1978a,b) in linear models, to nonlinear time series models, and derive some asymptotic equivalent forms of the test statistic. One of the most important features underlying our model is that both stationary and highly persistent variables are allowed to be included as components, which caters for a variety of applications. The main goal of this chapter is to show that, despite the appearance of both stationary and integrated regressors, only the former has an effect on the asymptotic distribution of the proposed test statistic.

Nielsen (2006b) showed that the test statistic TR^2 has the standard χ^2-distribution regardless of the location of characteristic roots in linear VAR models. In his proof, he first decomposed the test statistic into various components with different values of characteristic roots. He then illustrated that those terms with characteristic roots less than or equal to one vanish asymptotically. In other words, terms involving non-stationary variables die out asymptotically. In this chapter, we generalise his results to

nonlinear regression models and show that an asymptotically equivalent form of the test statistic contains only stationary variables.

Recall the nonlinear regression model in equation (2.3) discussed in Chapter 2:

$$y_t = g(x_t;\boldsymbol{\theta_0}) + \varepsilon_t, \qquad t = 1,\ldots,T,$$

where $g(\cdot)$ is some scalar-valued function and $\boldsymbol{\theta_0} \in \Theta \subset \mathbb{R}^q$. The error term ε_t is assumed to follow a martingale difference structure and is adapted to the filtration $\mathscr{F}_t$. The regressors $x_t \in \mathbb{R}^k$ are allowed to have deterministic time trends, stationary and/or integrated variables as components.

We are interested in testing the null hypothesis of no serial correlation against the alternative that the error term follows a pth-order autoregressive (AR) process. Positing that we have obtained the NLLS estimates (or other asymptotically equivalent estimators), $\widehat{\boldsymbol{\theta}}$, for the parameters $\boldsymbol{\theta_0}$, and the corresponding residuals, $\widehat{\varepsilon}_t$, our test statistic relies on the construction of the following linear regression

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha'}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T, \tag{3.1}$$

where $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ is the p-dimensional vector of lag residuals $\widehat{\varepsilon}_{t-1},\ldots,\widehat{\varepsilon}_{t-p}$, and $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, as defined in the previous chapters, are the first-order partial derivatives of the nonlinear function $g(x_t;\boldsymbol{\theta})$ with respect to the q-dimensional vector $\boldsymbol{\theta}$ evaluated at $\widehat{\boldsymbol{\theta}}$. We test the null hypothesis of no autocorrelation by testing the hypothesis that the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are jointly zero. Our test statistic then depends on the contemporaneous residual $\widehat{\varepsilon}_t$, the $p \times 1$ vector of lag residuals $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and the $q \times 1$ vector of first-order partial derivatives $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. We demonstrated in Chapter 2 that this auxiliary regression is asymptotically equivalent to

$$\varepsilon_t = \grave{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \grave{e}_t, \qquad t = p+1,\ldots,T. \tag{3.2}$$

So testing the joint significance of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in equation (3.1) is asymptotically the same as testing the significance of $\grave{\boldsymbol{\alpha}}$ in equation (3.2).

In this chapter, we propose that the first-order partial derivative functions, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, can be decomposed into stationary and non-stationary components, upon employment of a normalisation matrix. In particular, the normalisation matrix under consideration takes account of the different types of stationarity/non-stationarity possessed by each component. Assuming that the stationary and integrated components are asymptotically uncorrelated, this decomposition allows us to split the test statistic into two parts — one involving only stationary first-order partial derivatives and the other involving only non-stationary ones — and derive the asymptotic irrelevance of the non-stationary components.

Let

$$\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) =: \begin{bmatrix} \dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \\ \dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \end{bmatrix},$$

where $\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ and $\dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ denote the stationary and non-stationary components of the first-order partial derivatives $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, respectively. [In the case of co-integration, it could be the rotated $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ that can be decomposed into $\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ and $\dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$; see the next section for details.] We can then rewrite the auxiliary regression in equation (3.2) as

$$\varepsilon_t = \grave{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}),\dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \grave{e}_t, \qquad t = p+1,\ldots,T. \tag{3.3}$$

To test the null hypothesis that there is no autocorrelation in the error term, we need to check whether the parameters in the vector $\grave{\boldsymbol{\alpha}}$ are all equal to zero. The main result of this chapter indicates that testing the overall significance in the above equation is asymptotically equivalent to testing the overall significance in the following regression,

$$\varepsilon_t = \breve{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \breve{e}_t, \qquad t = p+1,\ldots,T. \tag{3.4}$$

Notice that in equation (3.4), we only regress the contemporaneous error, ε_t, on the lag residuals, $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, after controlling for the stationary components, $\dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. The non-stationary components $\dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ are omitted from this auxiliary regression. We show that the test statistic computed using equation (3.4) is asymptotically the same as the one computed from equation (3.3). This implies that the non-stationary part of the underlying model does not matter in the limit, so testing $\dot{\boldsymbol{\alpha}}$ being zero is asymptotically equivalent to testing the insignificance of $\breve{\boldsymbol{\alpha}}$.

The remainder of this chapter is as follows. In Section 3.2, we introduce some new assumptions needed to establish the results in this chapter and present the main theorem. The empirical application to an AR(2) model with one unit root is given in Section 3.3. Some concluding comments are in Section 3.4. All proofs are relegated to the Appendix.

3.2 Assumptions and Theorems

In this section, we show that the asymptotic properties of our test statistic do not depend on the non-stationary components, if any, of the first-order partial derivative functions, $\dot{g}_{\boldsymbol{\theta}}$.

Recall from equation (2.7) that

$$\widehat{\mathbf{H}}_t := \begin{bmatrix} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \end{bmatrix},$$

which is a $(p+q)$-dimensional column vector of all the regressors included in the auxiliary regression in equation (3.1). We can test the null hypothesis of no residual autocorrelation using the test statistic TR^2, where the R^2 is the squared sample multiple correlation between the contemporaneous residual, $\widehat{\varepsilon}_t$ and $\widehat{\mathbf{H}}_t$. From equation (2.8),

TR^2 can be computed using

$$TR^2 = T\frac{\widehat{cov}(\widehat{\varepsilon}_t, \widehat{\mathbf{H}}_t)' \left[\widehat{var}(\widehat{\mathbf{H}}_t)\right]^{-1} \widehat{cov}(\widehat{\mathbf{H}}_t, \widehat{\varepsilon}_t)}{\widehat{var}(\widehat{\varepsilon}_t)}. \tag{3.5}$$

where $\widehat{cov}(\widehat{\varepsilon}_t, \widehat{\mathbf{H}}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T} \widehat{\varepsilon}_t \widehat{\mathbf{H}}_t$ is the squared sample multiple covariance between $\widehat{\varepsilon}_t$ and $\widehat{\mathbf{H}}_t$, $\widehat{var}(\widehat{\varepsilon}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T} \widehat{\varepsilon}_t^2$ is the sample variance of the contemporaneous residual, and $\widehat{var}(\widehat{\mathbf{H}}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T} \widehat{\mathbf{H}}_t \widehat{\mathbf{H}}_t'$ is the sample variance-covariance matrix of $\widehat{\mathbf{H}}_t$.

Having the expression for the test statistic in hand, we proceed to make a few additional assumptions that the theorem in this chapter rests on. [The main results also rely on some of the assumptions that have previously been used in Chapter 2, specifically, *Assumptions 2.1*, *2.3*, *2.4* and *2.5*, and we would not restate them here.]

In Chapter 2, we emphasised the necessity of a nonsingular normalisation matrix, $N_{T,\boldsymbol{\theta_0}}$, to analyse the asymptotic properties of our test statistic. Specifically, we require this normalisation matrix to be compatible with both stationary and integrated regressors. We now extend the *Assumption 2.2* in Chapter 2 such that we further require $N_{T,\boldsymbol{\theta_0}}$ to have different normalisation blocks on the diagonal. In particular, $N_{T,\boldsymbol{\theta_0}}$ block-diagonalises the $q \times q$ matrix $\sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})'$ so that the normalised matrix, $\sum_{t=p+1}^{T} N_{T,\boldsymbol{\theta_0}}' \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})' N_{T,\boldsymbol{\theta_0}}$, is a block diagonal matrix. See the next section for an AR(2) example.

One important application of this extension is that it enables us to distinguish between the stationary and non-stationary components of $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, which will converge to their population counterpart evaluated at the true $\boldsymbol{\theta}$. To see this, consider Taylor expanding $N_{T,\boldsymbol{\theta_0}}' \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$ around the true parameter $\boldsymbol{\theta_0}$:

$$N_{T,\boldsymbol{\theta_0}}' \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = N_{T,\boldsymbol{\theta_0}}' \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) + N_{T,\boldsymbol{\theta_0}}' \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}^*) N_{T,\boldsymbol{\theta_0}} N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}),$$

where $\boldsymbol{\theta}^*$ is an intermediate point between $\boldsymbol{\theta_0}$ and $\widehat{\boldsymbol{\theta}}$, and is dependent on x_t. The

second term on the right-hand side (RHS) can be shown to be small. By *Assumption 2.3*, $N_{T,\boldsymbol{\theta_0}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$ is $\mathbf{o_p}(T^{\tau})$. In addition, *Assumption 2.4(c)* ensures that the norm of $N'_{T,\boldsymbol{\theta_0}}\ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}^*)N_{T,\boldsymbol{\theta_0}}$ must be no greater than $o_p(T^{-4\tau})$. Combining these assumptions implies that the second term on the RHS is infinitesimal and therefore negligible in our analysis, and we have

$$N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) + \mathbf{o_p}(1).$$

Now, let

$$\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) =: \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix},$$

where $\dot{g}_{\widehat{u},t} = \dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \in \mathbb{R}^{d_u}$ and $\dot{g}_{\widehat{v},t} = \dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \in \mathbb{R}^{d_v}$ $(d_u + d_v = q)$ denote the stationary and non-stationary components of the $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ functions, respectively.

The notation here could be extended to cover co-integrated variables. If x_t has two components with a common stochastic trend, a linear combination is stationary. Suppose there is an invertible non-stochastic rotation matrix $M \in \mathbb{R}^q$. It suffices that the rotated vector, $M\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, could be decomposed into $\dot{g}_{\widehat{v},t}$, which is the common stochastic trend, and $\dot{g}_{\widehat{u},t}$, the stationary component. Since $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ is a regressor in the auxiliary regression (3.1) and is partialled out in the analysis of the asymptotic theory for the test statistic, it is equivalent to regress on $M\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. Given that the rotation matrix M does not affect $\boldsymbol{\alpha}$ asymptotically in equation (3.1), we leave it out for notational simplicity.

Thus, we have

$$N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} N_{T,u} & 0 \\ 0 & N_{T,v} \end{bmatrix}' \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix} = \begin{bmatrix} N'_{T,u}\dot{g}_{\widehat{u},t} \\ N'_{T,v}\dot{g}_{\widehat{v},t} \end{bmatrix}.$$

Instead of discussing the stationary and non-stationary parts of the $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ functions jointly, this normalisation matrix allows us to separate them and discuss their individ-

ual asymptotic convergence. The assumption regarding the normalisation matrix $N_{T,\boldsymbol{\theta_0}}$ is stated formally below. Note that *Assumption 2.2* is implied by this assumption.

Assumption 3.1. *Suppose $N_{T,\boldsymbol{\theta_0}} \in \mathbb{R}^{q\times q}$ is a nonsingular normalisation matrix which depends on both the number of observations T and the true parameter $\boldsymbol{\theta_0}$. Then*

(a) $\inf\{T : N_{T,\boldsymbol{\theta_0}}$ *is nonsingular*$\} < \infty$ *a.s.,*

(b) $N_{T,\boldsymbol{\theta_0}}^{-1} = \begin{bmatrix} N_{T,u} & 0 \\ 0 & N_{T,v} \end{bmatrix}^{-1} = \mathbf{O}(T^{\iota})$ *a.s. for some* $\iota > 0$*, where* $N_{T,u} \in \mathbb{R}^{d_u\times d_u}$ *and* $N_{T,v} \in \mathbb{R}^{d_v\times d_v}$ *are normalisation blocks associated with stationary and non-stationary components, respectively.*

Notice that *Assumption 3.1(a)* requires the normalisation matrix to be finitely invertible. Here, *Assumption 3.1(b)* restricts the normalisation matrix to be block diagonal. The upper diagonal block is associated with stationary components, and in general we have $N_{T,u} = \mathbf{O}(T^{-\frac{1}{2}})$. The lower diagonal block is associated with non-stationary components, and we may have, for example, $N_{T,v} = \mathbf{O}(T^{-1})$ if the underlying model equation is linear and we have near unit root or random walk regressors.

The next assumption is vital to the ability to split the test statistic into two parts — one involving the stationary components of the first-order partial derivatives $\dot{g}_{\boldsymbol{\theta}}$ and the other involving their non-stationary components — when the g function includes more than one regressors.

Assumption 3.2. *Let*

$$\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) =: \begin{bmatrix} \dot{g}_{u,t} \\ \dot{g}_{v,t} \end{bmatrix},$$

where $\dot{g}_{u,t} = \dot{g}_{u,\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \in \mathbb{R}^{d_u}$ and $\dot{g}_{v,t} = \dot{g}_{v,\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \in \mathbb{R}^{d_v}$ ($d_u + d_v = q$) denote the stationary and non-stationary components of $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$, respectively.

Suppose $\dot{g}_{u,t}$ and $\dot{g}_{v,t}$ are uncorrelated such that

$$\{\sum_{t=p+1}^{T} \dot{g}_{u,t}\dot{g}_{u,t}'\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{u,t}\dot{g}_{v,t}' \{\sum_{t=p+1}^{T} \dot{g}_{v,t}\dot{g}_{v,t}'\}^{-\frac{1}{2}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}}),$$

for some $0 \leq \psi < \frac{1}{2}$.

If the first-order partial derivative functions, $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$, consist of both stationary and integrated components, *Assumption 3.2* ensures that the correlation between these two components, when evaluated at the true parameters $\boldsymbol{\theta_0}$, converges in probability to zero. In other words, the stationary and non-stationary parts of $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ are asymptotically uncorrelated.

The motivation underlying the asymptotic uncorrelation assumption between the stationary and non-stationary components comes from the Brownian motion theory [see Phillips (1987), Chan and Wei (1988) for details]. The Brownian motion theory suggests that properties of stationary and non-stationary components are fundamentally different. Given that the two components are asymptotically unrelated, distinguishing between them allows us to discuss their asymptotic behaviour separately.

The following assumption is a modified version of *Assumption 2.5*.

Assumption 3.3. *Let* $\dot{g}_{\widehat{u},t} = \dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ *and* $\dot{g}_{\widehat{v},t} = \dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ *denote the stationary and non-stationary components of the first order partial derivative functions evaluated at the NLLS estimators* $\widehat{\boldsymbol{\theta}}$*, respectively.*

Let

$$[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] = \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} - \sum_{s=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{s-1}} \dot{g}'_{\widehat{u},s} [\sum_{s=p+1}^{T} \dot{g}_{\widehat{u},s} \dot{g}'_{\widehat{u},s}]^{-1} \dot{g}_{\widehat{u},t}.$$

Then

(a)

$$\left\{ \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\varepsilon_t = \mathbf{O_p}(1).$$

(b)

$$\left\{ \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}}).$$

Different from *Assumption 2.5*, where we consider the least squares residuals from

regressing $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ on $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, *Assumption 3.3* focuses on the residuals from regressing $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ on the stationary components of $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, $\dot{g}_{\widehat{u},t}$, which form the vector $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$. The relation between *Assumptions 2.5* and *3.3* will be discussed in the conclusion. In particular, we argue that *Assumptions 2.5* is implied by *Assumption 3.3* together with an extra condition.

There is a similar consideration in *Assumption 3.3(a)* as that in *Assumptions 2.5*, with $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ in the latter being replaced by its stationary component in the former. Here we require the asymptotic invertibility of the sample covariance matrix $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'$.

Assumption 3.3(b) imposes the restriction that the sample partial correlation coefficients of the lag residuals, $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, and the non-stationary components, $\dot{g}_{\widehat{v},t}$, are asymptotically zero, after partialling out the effect of the stationary components $\dot{g}_{\widehat{u},t}$. To see this, notice that the middle matrix, $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t}$, is just the sample partial covariances between $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and $\dot{g}_{\widehat{v},t}$, partialling out the effect of $\dot{g}_{\widehat{u},t}$. The matrices inside the curly brackets, $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'$ and $\sum_{t=p+1}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}$ are the sample variances of $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$ and $\dot{g}_{\widehat{v},t}$, respectively. Provided that these two variance-covariance matrices are of full rank and their inverses are bounded, *Assumption 3.3(b)* ensures that residuals from regressing $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ on the stationary components $\dot{g}_{\widehat{u},t}$, $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$, are asymptotically uncorrelated with the non-stationary components $\dot{g}_{\widehat{v},t}$.

One can also treat *Assumption 3.3(b)* as an analogous assumption to *Assumption 3.2*, where we assume the asymptotic uncorrelatedness between the stationary $\dot{g}_{\widehat{u},t}$ and the non-stationary $\dot{g}_{\widehat{v},t}$. To see this, recall from *Assumption 2.1* that the error terms follow a martingale difference structure with identical mean and finite variance. Therefore, the error terms are covariance stationary. *Assumption 3.3(b)* then assumes that $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$ is asymptotically uncorrelated with non-stationary components, $\dot{g}_{\widehat{v},t}$.

Assumption 3.3 is technically very demanding, and we leave its simplification for future research.

Based on the results in Chapter 2 and a moderately expanded set of assumptions,

we are ready to state the main theorem of this chapter.

Theorem 3.1. *Consider the auxiliary regression*

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}' \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T,$$

where $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ can be broken down into two components — one involving only stationary processes, $\dot{g}_{\widehat{u},t}$, and the other involving only integrated processes, $\dot{g}_{\widehat{v},t}$.

If Assumptions 2.1, 2.3 to 2.5, 3.1 to 3.3 are satisfied, then

$$TR^2 = T\frac{\widehat{cov}\left[\varepsilon_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}\right]' \left\{\widehat{var}\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}\right]\right\}^{-1} \widehat{cov}\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \varepsilon_t | \dot{g}_{\widehat{u},t}\right]}{\widehat{var}(\varepsilon_t)} + o_p(1).$$

This theorem relies heavily on the assumptions that $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ can be decomposed into stationary and non-stationary components upon utilising the normalisation matrix and that these two components are asymptotically uncorrelated with each other.

Given that we can split $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ into two parts, $\dot{g}_{\widehat{u},t}$ and $\dot{g}_{\widehat{v},t}$, the auxiliary regression can be rewritten as

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta_1}' \dot{g}_{\widehat{u},t} + \boldsymbol{\beta_2}' \dot{g}_{\widehat{v},t} + e_t, \qquad t = p+1,\ldots,T. \tag{3.6}$$

Under *Assumptions 2.1, 2.3 to 2.5*, and *3.1*, the main theorems of Chapter 2 give the following asymptotic equivalent form of our test statistic

$$TR^2 = T\frac{\widehat{cov}\left[\varepsilon_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t}\right]' \left\{\widehat{var}\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t}\right]\right\}^{-1} \widehat{cov}\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \varepsilon_t | \dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t}\right]}{\widehat{var}(\varepsilon_t)} + o_p(1).$$

This implies that, the test statistic TR^2 constructed from the auxiliary regression in equation (3.6) is asymptotically equivalent to that calculated from the following regression

$$\varepsilon_t = \grave{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{u,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}), \dot{g}_{v,\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})] + \grave{e}_t, \qquad t = p+1,\ldots,T. \tag{3.7}$$

So the squared sample multiple correlation between $\widehat{\varepsilon}_t$ and $(\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}, \dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})$ is asymptotically the same as the squared sample partial correlation between ε_t and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ after controlling for $\dot{g}_{\widehat{u},t}$ and $\dot{g}_{\widehat{v},t}$.

Theorem 3.1 then further shows that the squared sample partial correlation between ε_t and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ after controlling for $\dot{g}_{\widehat{u},t}$ and $\dot{g}_{\widehat{v},t}$ is asymptotically equivalent to that controlling for $\dot{g}_{\widehat{u},t}$ only. So the test statistic calculated using equation (3.7) is asymptotically the same as that constructed from the following regression

$$\varepsilon_t = \breve{\boldsymbol{\alpha}}'[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{u,\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})] + \breve{e}_t, \qquad t = p+1, \ldots, T, \tag{3.8}$$

where we do not partial out the effect of the non-stationary components $\dot{g}_{\widehat{v},t}$ on $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$. *Theorem 3.1* implies that the asymptotic behaviour of the test statistic does not depend on the random walk components, and only the stationary components contribute to the final distribution. As shown in the Appendix, the non-stationary components get cancelled out, so that they do not affect our test statistic asymptotically.

Consider the interesting case where we have a nonlinear model with integrated regressors only. So $\boldsymbol{\beta_1}$ in equation (3.6) is zero. Then the auxiliary regression takes the form

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta_2}'\dot{g}_{v,\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1, \ldots, T.$$

Theorem 3.1 indicates that the test statistic constructed from the above regression is asymptotically equivalent to that computed from the regression of the contemporaneous error term ε_t on the lag residuals $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ only. In other words, the limiting distribution of the test statistic only depends on the squared sample correlation between $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and ε_t. In this case, the expression for the test statistic can be further simplified to

$$TR^2 = T\frac{\widehat{cov}(\varepsilon_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}})' \left\{\widehat{var}(\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}})\right\}^{-1} \widehat{cov}(\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \varepsilon_t)}{\widehat{var}(\varepsilon_t)} + o_p(1),$$

such that it does not depend on any first-order partial derivative functions (directly).

3.3 Example — The AR(2) Model with One Unit Root

In this section, we use an example to interpret and assess the applicability of the assumptions introduced in this chapter. In particular, we consider the AR(2) model with one unit root that was studied in Section 2.4.3 of Chapter 2. To simplify the expressions, we will suppress the drift term here. We focus on testing against the alternative hypothesis that the error terms follow an AR(1) process, so we take $p = 1$.

Consider the following AR(2) process

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t, \qquad t = 1, \ldots, T,$$

where $\gamma_1 + \gamma_2 = 1$, $|\gamma_2| < 1$ and ε_t satisfies the martingale difference assumption in *Assumption 2.1*. Both the initial observations, y_{-1} and y_0, are assumed to be zero and independent of the error term. We here give an illustration of *Assumptions 3.1* and *3.2*. One could check *Assumption 3.3* in a manner similar to how we check *Assumption 3.2*. But this requires a longer calculation which is omitted for now. In addition, given that *Assumption 2.5* has been checked in Chapter 2, *Assumption 3.3* should also hold, since these two assumptions are almost equivalent to each other (see the discussion in the next section for details).

Checking Assumption 3.1: First, add and subtract $\gamma_2 y_{t-1}$ to the RHS of the model equation to get

$$y_t = (\gamma_1 + \gamma_2) y_{t-1} - \gamma_2 (y_{t-1} - y_{t-2}) + \varepsilon_t.$$

Let $\theta_1 = -\gamma_2$ and $\theta_2 = \gamma_1 + \gamma_2$. We can rewrite the above equation as

$$y_t = \theta_1 \triangle y_{t-1} + \theta_2 y_{t-1} + \varepsilon_t \equiv x_t' \boldsymbol{\theta_0} + \varepsilon_t, \tag{3.9}$$

where $x_t = (\triangle y_{t-1}, y_{t-1})'$ and $\boldsymbol{\theta_0} = (\theta_1, \theta_2)'$. Since this model is linear in parameters, we can obtain the usual ordinary least squares (OLS) estimates, $\widehat{\boldsymbol{\theta}}$.

The first-order partial derivatives of the function $g(x_t;\boldsymbol{\theta}) = x_t'\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}$ are

$$\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}) = x_t = \begin{bmatrix} \triangle y_{t-1} \\ y_{t-1} \end{bmatrix}.$$

Note that the first element, $\triangle y_{t-1}$, is stationary, while the last element, y_{t-1}, is integrated of order 1. Thus we can split the first-order partial derivatives, $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta})$, into $\dot{g}_{\widehat{u},t}$ and $\dot{g}_{\widehat{v},t}$, where

$$\dot{g}_{\widehat{u},t} = \triangle y_{t-1} \qquad \text{and} \qquad \dot{g}_{\widehat{v},t} = y_{t-1}.$$

Since $\dot{g}_{\widehat{u},t}$ corresponds to the stationary regressor, it must be $T^{\frac{1}{2}}$ convergent. As $\dot{g}_{\widehat{v},t}$ corresponds to the integrated variable and we have a linear model, it should be normalised by T. The block diagonal normalisation matrix $N_{T,\boldsymbol{\theta_0}}$ is then

$$N_{T,\boldsymbol{\theta_0}}^{-1} = \begin{bmatrix} T^{\frac{1}{2}} & 0 \\ 0 & T \end{bmatrix} = \begin{bmatrix} N_{T,u} & 0 \\ 0 & N_{T,v} \end{bmatrix}^{-1},$$

where,

$$N_{T,u} = T^{-\frac{1}{2}} \qquad \text{and} \qquad N_{T,v} = T^{-1}.$$

Since we have a diagonal normalisation matrix with nonzero values on the diagonal, it must be invertible and *Assumption 3.1* is satisfied.

Checking Assumption 3.2: This assumption requires the stationary and non-stationary components of $\dot{g}_{\boldsymbol{\theta}}$ to be asymptotically uncorrelated when evaluated at the true parameters $\boldsymbol{\theta_0}$. Notice that the elements of $\dot{g}_{\boldsymbol{\theta}}$ do not depend on $\boldsymbol{\theta}$, because the model is linear in parameters. Therefore, we have

$$\dot{g}_{u,t} = \triangle y_{t-1} \qquad \text{and} \qquad \dot{g}_{v,t} = y_{t-1}.$$

To verify *Assumption 3.2*, we need to ensure that

$$\{\sum_{t=2}^{T}\dot{g}_{u,t}\dot{g}'_{u,t}\}^{-\frac{1}{2}}\sum_{t=2}^{T}\dot{g}_{u,t}\dot{g}'_{v,t}\{\sum_{t=2}^{T}\dot{g}_{v,t}\dot{g}'_{v,t}\}^{-\frac{1}{2}}=\mathbf{O_p}(T^{\psi-\frac{1}{2}}),$$

for some $0\leq\psi<\frac{1}{2}$. Since the initial observations y_{-1} and y_0, are assumed to be zero, this is equivalent to proving

$$\{\sum_{t=1}^{T}\dot{g}_{u,t}\dot{g}'_{u,t}\}^{-\frac{1}{2}}\sum_{t=1}^{T}\dot{g}_{u,t}\dot{g}'_{v,t}\{\sum_{t=1}^{T}\dot{g}_{v,t}\dot{g}'_{v,t}\}^{-\frac{1}{2}}=\mathbf{O_p}(T^{\psi-\frac{1}{2}}),$$

where we have $\sum_{t=1}^{T}\dot{g}_{u,t}\dot{g}'_{u,t}=\sum_{t=1}^{T}(\triangle y_{t-1})^2$, $\sum_{t=1}^{T}\dot{g}_{u,t}\dot{g}'_{v,t}=\sum_{t=1}^{T}\triangle y_{t-1}\cdot y_{t-1}$ as well as $\sum_{t=1}^{T}\dot{g}_{v,t}\dot{g}'_{v,t}=\sum_{t=1}^{T}y_{t-1}^2$.

Let

$$\tilde{D}_{u,v}=\{\sum_{t=1}^{T}\dot{g}_{u,t}\dot{g}'_{u,t}\}^{-\frac{1}{2}}\sum_{t=1}^{T}\dot{g}_{u,t}\dot{g}'_{v,t}\{\sum_{t=1}^{T}\dot{g}_{v,t}\dot{g}'_{v,t}\}^{-\frac{1}{2}}.$$

As in the proof of *Lemma 3.2* in the Appendix, it can be shown that

$$\tilde{D}_{u,v}=\left\{\sum_{t=1}^{T}N'_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u}\right\}^{-\frac{1}{2}}\sum_{t=1}^{T}N'_{T,u}\dot{g}_{u,t}\dot{g}'_{v,t}N_{T,v}\left\{\sum_{t=1}^{T}N'_{T,v}\dot{g}_{v,t}\dot{g}'_{v,t}N_{T,v}\right\}^{-\frac{1}{2}}. \tag{3.10}$$

First consider the first term in the above equation. By the martingale Laws of Large Numbers [see (Davidson, 1994, page 315)],

$$\sum_{t=1}^{T}N'_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u}=T^{-1}\sum_{t=1}^{T}(\triangle y_{t-1})^2\xrightarrow{p}\mathbb{E}[(\triangle y_t)^2].$$

Provided that $\mathbb{E}[(\triangle y_t)^2]$ is a positive finite constant, the inverse of $\sum_{t=1}^{T}N'_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u}$ exists and is bounded in probability. Thus,

$$\{\sum_{t=1}^{T}N'_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u}\}^{-\frac{1}{2}}=O_p(1). \tag{3.11}$$

Next, take the middle term in equation (3.10). Under the functional central limit theorem [see Phillips (1987)], martingale Laws of Large Numbers and the Continuous

Mapping Theorem,

$$\sum_{t=1}^{T} N'_{T,u}\dot{g}_{u,t}\dot{g}'_{v,t}N_{T,v} = T^{-\frac{3}{2}}\sum_{t=1}^{T} \triangle y_{t-1}\cdot y_{t-1} \xrightarrow{p} 0.$$

Therefore, we have

$$\sum_{t=1}^{T} N'_{T,u}\dot{g}_{u,t}\dot{g}'_{v,t}N_{T,v} = o_p(1). \tag{3.12}$$

Now, consider the last term in equation (3.10). Recall from *Assumption 2.1* that $\mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2$. Let $\kappa = \frac{\sigma_\varepsilon}{1-\theta_1}$. By the functional central limit theorem,

$$\sum_{t=1}^{T} N'_{T,v}\dot{g}_{v,t}\dot{g}'_{v,t}N_{T,v} = T^{-2}\sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{D} \kappa^2 \int_0^1 W(r)^2 dr,$$

where $W(r)$ denotes the Wiener process, with argument $r \in [0,1]$. Since $\mathbb{P}\{\int_0^1 W^2(r)dr \neq 0\} = 1$, the inverse of $\sum_{t=1}^{T} N'_{T,v}\dot{g}_{v,t}\dot{g}'_{v,t}N_{T,v}$ exists and is stochastically bounded. This implies,

$$\left\{\sum_{t=1}^{T} N'_{T,v}\dot{g}_{v,t}\dot{g}'_{v,t}N_{T,v}\right\}^{-\frac{1}{2}} = O_p(1). \tag{3.13}$$

Substituting equations (3.11) to (3.13) back into equation (3.10) together with the Continuous Mapping Theorem verifies *Assumption 3.2*.

3.4 Conclusion

The main result of this chapter is that, even when integrated variables are included in the non-linear regression model, the asymptotic behaviour of the proposed test statistic is not affected by these non-stationary components. This observation suggests that, regardless of the location of the characteristic roots of the underlying dynamic model, the test statistic TR^2 follows a standard chi-square distribution. This limiting distribution of the test statistic will be proved in the next chapter.

Reconsider the auxiliary regression in equation (3.1):

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T.$$

Breaking the first-order partial derivatives, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, down into stationary and non-stationary components gives:

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta_1}'\dot{g}_{\widehat{u},t} + \boldsymbol{\beta_2}'\dot{g}_{\widehat{v},t} + e_t, \qquad t = p+1,\ldots,T.$$

Results from Chapter 2 imply that an F-test on the overall significance of $(\boldsymbol{\alpha},\boldsymbol{\beta_1},\boldsymbol{\beta_2})$ in the above equation is asymptotically equivalent to a partial F-test on $\grave{\boldsymbol{\alpha}}$ alone from the following regression:

$$\varepsilon_t = \grave{\boldsymbol{\alpha}}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \grave{\boldsymbol{\beta}}_{\mathbf{1}}'\dot{g}_{\widehat{u},t} + \grave{\boldsymbol{\beta}}_{\mathbf{2}}'\dot{g}_{\widehat{v},t} + \grave{e}_t, \qquad t = p+1,\ldots,T. \tag{3.14}$$

That is, the sample multiple correlation coefficient between the contemporaneous residual $\widehat{\varepsilon}_t$ and $(\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}, \dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})$ is asymptotically equivalent to the sample partial correlation coefficient between the contemporaneous error ε_t and the vector of lag residuals $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, after partialling out the effect of both the stationary and non-stationary components, $\dot{g}_{\widehat{u},t}$ and $\dot{g}_{\widehat{v},t}$.

The main result of this chapter further simplifies this sample partial correlation coefficient such that only the effect of the stationary components $\dot{g}_{\widehat{u},t}$ needs to be partialled out. To see this, notice that *Theorem 3.1* implies that the partial F-statistic on $\grave{\boldsymbol{\alpha}}$ in equation (3.14) is asymptotically the same as the partial F-statistic on $\breve{\boldsymbol{\alpha}}$ from the following regression

$$\varepsilon_t = \breve{\boldsymbol{\alpha}}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \breve{\boldsymbol{\beta}}_{\mathbf{1}}'\dot{g}_{\widehat{u},t} + \breve{e}_t, \qquad t = p+1,\ldots,T.$$

Observe that the non-stationary components, $\dot{g}_{\widehat{v},t}$, do not enter the above regression.

Therefore, whether we include $\dot{g}_{\widehat{v},t}$ in the auxiliary regression or not has no asymptotic impact on our test statistic.

Theorem 3.1 holds for a general set of assumptions on the error term, the NLLS estimators and the nonlinear g functions. However, there are two technical assumptions that we impose at this stage, *Assumptions 2.5* and *3.3*. These two assumptions place some stringent restrictions on the correlation between the lag residuals, $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, and the first-order partial derivative functions $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$.

In fact, there is some relationship between the two assumptions. To see this, recall we assume in *Assumption 2.5* that the scaled sample partial covariance between $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and ε_t is bounded in probability, after partialling out the effect of $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. That is, we require

$$\left\{\sum_{t=p+1}^{T}\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right]\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right]'\right\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right]\varepsilon_t=\mathbf{O_p}(1). \tag{3.15}$$

On the other hand, in *Assumption 3.3*, we assume that the scaled sample partial covariance between $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and ε_t is bounded in probability, after partialling out the effect of the stationary components of these first-order partial derivatives. That is, we impose the restriction

$$\left\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\right\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\varepsilon_t=\mathbf{O_p}(1). \tag{3.16}$$

One can see that equation (3.15) has the same structure as equation (3.16), except that we replace $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ in the former by $\dot{g}_{\widehat{u},t}$ in the latter.

Besides, it is also assumed in *Assumption 3.3* that the sample partial correlation between $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and the non-stationary components of the first-order partial derivatives is asymptotically zero, in the sense that

$$\left\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\right\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t}\left\{\sum_{t=p+1}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\right\}^{-\frac{1}{2}}=\mathbf{O_p}(T^{\psi-\frac{1}{2}}),$$

where $0 \leq \psi < \frac{1}{2}$.

It can, in fact, be shown that *Assumption 2.5* is always true if an additional assumption is being imposed together with *Assumptions 3.2* and *3.3*. Specifically, we need an additional assumption of the form

$$\left\{ \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}]' \right\}^{-\frac{1}{2}} = \mathbf{o_p} \left(\frac{T^{\frac{1}{2}-\psi}}{\log T} \right). \tag{3.17}$$

This assumption restricts how small the limit of the matrix $\sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}]'$ in equation (3.16) can get. A formal illustration is given in the next chapter, where *Assumption 2.5* becomes redundant and can be removed from our analyses.

We also wish to relax *Assumption 3.3*, since Nielsen (2006b) showed that, in linear vector auto-regression models, whether *Assumption 3.3* is satisfied or not does not affect the asymptotic properties of the test statistic TR^2. Future studies would relax this assumption in our nonlinear model set-up to make the test procedure more applicable if possible.

Appendix

In this appendix, we show how our test statistic can be decomposed into two parts — one involving only the stationary components and the other involving the non-stationary components of the first-order partial derivative functions, and then give the derivation of *Theorem 3.1* that the test statistic does not depend on the non-stationary parts.

We start with defining some notations and sketching out a very brief outline of the proof. Appendix 3.B consists of a general decomposition result (*Lemma 3.1*) that is critical to the proof of the main theorem. Appendix 3.C illustrates that the condition required in *Lemma 3.1* is satisfied in our model given the assumptions described in Section 3.2. Appendix 3.D applies *Lemma 3.1* to our key analysing terms and obtains the desired decompositions. Appendix 3.E presents a proof of *Theorem 3.1*.

3.A Notations and Preliminaries

Recall from Notation 2.3 that $\mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}} = \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})' N_{T,\boldsymbol{\theta_0}}$. We here extend this notation to cope with stationary and non-stationary components of the $\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})$ function.

Notation 3.1. *Let*

$$\mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}} = \begin{bmatrix} \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} & \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{v}} \\ \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{u}} & \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \dot{g}'_{u,t} N_{T,u} & \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \dot{g}'_{v,t} N_{T,v} \\ \sum_{t=p+1}^{T} N'_{T,v} \dot{g}_{v,t} \dot{g}'_{u,t} N_{T,u} & \sum_{t=p+1}^{T} N'_{T,v} \dot{g}_{v,t} \dot{g}'_{v,t} N_{T,v} \end{bmatrix}.$$

Notation 3.2. *Let*

$$D_{u,v} = \left\{ \sum_{t=p+1}^{T} \dot{g}_{u,t}\dot{g}'_{u,t} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{u,t}\dot{g}'_{v,t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{v,t}\dot{g}'_{v,t} \right\}^{-\frac{1}{2}},$$

where both $\dot{g}_{u,t}$ and $\dot{g}_{v,t}$ are evaluated at the true parameters, $\boldsymbol{\theta_0}$. Similarly, let

$$D_{\widehat{u},\widehat{v}} = \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}},$$

where both $\dot{g}_{\widehat{u},t}$ and $\dot{g}_{\widehat{v},t}$ are evaluated at the NLLS estimators, $\widehat{\boldsymbol{\theta}}$.

Following from *Assumption 3.2*, we have $D_{u,v} = \mathbf{O_p}(T^{\psi-\frac{1}{2}})$, for some $0 \leq \psi < \frac{1}{2}$.

Notation 3.3. *Let*

$$\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} = \begin{bmatrix} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\widehat{u},t} \end{bmatrix},$$

and

$$D_{\widehat{\mathbf{H}}_u,\widehat{v}} = \left\{ \sum_{t=p+1}^{T} \widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}},$$

where both $\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}$ and $\dot{g}_{\widehat{v},t}$ are evaluated at the NLLS estimators, $\widehat{\boldsymbol{\theta}}$.

According to *Assumption 3.3(b)*, $D_{\widehat{\mathbf{H}}_u,\widehat{v}}$ is also of order $\mathbf{O_p}(T^{\psi-\frac{1}{2}})$, for some $0 \leq \psi < \frac{1}{2}$.

Before proceeding to derive the main results of this chapter, we first briefly outline the proof.

Recall from equation (3.5) the expression for R^2

$$R^2 = \frac{\widehat{cov}(\widehat{\varepsilon}_t, \widehat{\mathbf{H}}_t)' \left[\widehat{var}(\widehat{\mathbf{H}}_t)\right]^{-1} \widehat{cov}(\widehat{\mathbf{H}}_t, \widehat{\varepsilon}_t)}{\widehat{var}(\widehat{\varepsilon}_t)},$$

where we have $\widehat{var}(\widehat{\varepsilon}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T} \widehat{\varepsilon}_t^2$, $\widehat{var}(\widehat{\mathbf{H}}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T} \widehat{\mathbf{H}}_t\widehat{\mathbf{H}}'_t$, and $\widehat{cov}(\widehat{\varepsilon}_t, \widehat{\mathbf{H}}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T} \widehat{\varepsilon}_t\widehat{\mathbf{H}}_t$.

We could use the notation $Q(\cdot)$ (Notation 2.5) introduced in Chapter 2 to simplify the expression of our test statistic. Recall that, for any time series vector $\mathbf{Z_t}$,

$$Q(\mathbf{Z_t}) = \sum_{t=p+1}^{T} \varepsilon_t \mathbf{Z_t}' \Big(\sum_{t=p+1}^{T} \mathbf{Z_t} \mathbf{Z_t}' \Big)^{-1} \sum_{t=p+1}^{T} \mathbf{Z_t} \varepsilon_t,$$

and

$$\widehat{Q}(\mathbf{Z_t}) = \sum_{t=p+1}^{T} \widehat{\varepsilon}_t \mathbf{Z_t}' \Big(\sum_{t=p+1}^{T} \mathbf{Z_t} \mathbf{Z_t}' \Big)^{-1} \sum_{t=p+1}^{T} \mathbf{Z_t} \widehat{\varepsilon}_t.$$

Therefore, the numerator of R^2 can be written as

$$\frac{1}{T-p} \widehat{Q}(\widehat{\mathbf{H}}_t) = \frac{1}{T-p} \sum_{t=p+1}^{T} \widehat{\varepsilon}_t \widehat{\mathbf{H}}_t' \Big(\sum_{t=p+1}^{T} \widehat{\mathbf{H}}_t \widehat{\mathbf{H}}_t' \Big)^{-1} \sum_{t=p+1}^{T} \widehat{\mathbf{H}}_t \widehat{\varepsilon}_t.$$

Since the denominator of R^2 is just $\frac{1}{T-p} \sum_{t=p+1}^{T} \widehat{\varepsilon}_t^2$, the expression of our test statistic can be further simplified to

$$TR^2 = \frac{\widehat{Q}(\widehat{\mathbf{H}}_t)}{T^{-1} \sum_{t=p+1}^{T} \widehat{\varepsilon}_t^2}. \tag{3.18}$$

Following from Theorem 2.2 in Chapter 2, which is satisfied under *Assumptions 2.1*, *2.3* to *2.5*, and *3.1*, we can rewrite the expression for TR^2 in equation (3.18) as

$$TR^2 = \frac{Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})\right]}{T^{-1} \sum_{t=p+1}^{T} \varepsilon_t^2} + o_p(1). \tag{3.19}$$

Note that both $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]$ and $\sum_{t=p+1}^{T} \varepsilon_t^2$ are now functions of the contemporaneous error term, ε_t, instead of the contemporaneous residuals $\widehat{\varepsilon}_t$.

The leading term of the test statistic to be analysed is $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}})]$. Since the first-order partial derivatives can be broken down into stationary and integrated

components such that $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = (\dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t})'$, we have

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = Q\left(\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}\right).$$

By the formula of partitioned inversion, we have

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = Q\begin{bmatrix}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}\\ \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\end{bmatrix} - Q\left[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right],$$

which is exactly

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = Q\begin{bmatrix}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}\\ \dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} - Q\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}.$$

Recall from Notation 3.3 that $\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} = (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \dot{g}_{\widehat{u},t})'$. Substituting into the above equation gives

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = Q\begin{bmatrix}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} - Q\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}. \tag{3.20}$$

The main idea of the proof is to show that the two Q functions on the RHS of the above equation can be further decomposed using the general decomposition result in *Lemma 3.1* such that

$$Q\begin{bmatrix}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} \approx Q\left(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\right) + Q\left(\dot{g}_{\widehat{v},t}\right). \tag{3.21}$$

and

$$Q\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} \approx Q\left(\dot{g}_{\widehat{u},t}\right) + Q\left(\dot{g}_{\widehat{v},t}\right). \tag{3.22}$$

As a result, both the Q functions can be split into two parts, one with the stationary components of $\dot{g}_{\boldsymbol{\theta}}$ only and one with its non-stationary components. Upon taking the

difference of $Q[(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}, \dot{g}_{\widehat{v},t})']$ and $Q[(\dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t})']$, the non-stationary parts cancel itself out and only the stationary components remain in the test statistic.

3.B Key Intermediate Result—Decomposing Q

In order to decompose the two Q functions on the RHS of equation (3.20) as in equations (3.21) and (3.22), we first derive the following general result that is applicable to decompose any Q functions whose arguments satisfy an asymptotic uncorrelatedness assumption as specified below.

Lemma 3.1. *Given any two processes m_t and n_t, we have*

$$Q\begin{bmatrix} m_t \\ n_t \end{bmatrix} = Q(m_t)\left\{1+O_p(T^{-\alpha})\right\} + Q(n_t)\left\{1+O_p(T^{-\alpha})\right\},$$

if

$$D_{m,n} = \frac{\sum_{t=p+1}^{T} m_t n_t}{\sqrt{\sum_{t=p+1}^{T} m_t^2 \sum_{t=p+1}^{T} n_t^2}} = O_p(T^{-\alpha}),$$

for some $\alpha > 0$.

Proof of Lemma 3.1.

From Notation 2.5, we have

$$Q\begin{bmatrix} m_t \\ n_t \end{bmatrix} = \begin{bmatrix} \sum_{t=p+1}^{T} \varepsilon_t m_t \\ \sum_{t=p+1}^{T} \varepsilon_t n_t \end{bmatrix}' \begin{bmatrix} \sum_{t=p+1}^{T} m_t^2 & \sum_{t=p+1}^{T} m_t n_t \\ \sum_{t=p+1}^{T} n_t m_t & \sum_{t=p+1}^{T} n_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=p+1}^{T} m_t \varepsilon_t \\ \sum_{t=p+1}^{T} n_t \varepsilon_t \end{bmatrix}. \tag{3.23}$$

First, take the middle matrix and consider the following decomposition

$$\begin{bmatrix} \sum_{t=p+1}^{T} m_t^2 & \sum_{t=p+1}^{T} m_t n_t \\ \sum_{t=p+1}^{T} n_t m_t & \sum_{t=p+1}^{T} n_t^2 \end{bmatrix} = \begin{bmatrix} \sum_{t=p+1}^{T} m_t^2 & 0 \\ 0 & \sum_{t=p+1}^{T} n_t^2 \end{bmatrix}^{\frac{1}{2}} \begin{bmatrix} 1 & D_{m,n} \\ D_{m,n} & 1 \end{bmatrix} \cdot \begin{bmatrix} \sum_{t=p+1}^{T} m_t^2 & 0 \\ 0 & \sum_{t=p+1}^{T} n_t^2 \end{bmatrix}^{\frac{1}{2}}, \quad (3.24)$$

where

$$D_{m,n} = \frac{\sum_{t=p+1}^{T} m_t n_t}{\sqrt{\sum_{t=p+1}^{T} m_t^2 \sum_{t=p+1}^{T} n_t^2}}.$$

Notice that multiplying the row vector on the RHS of equation (3.23), $(\sum_{t=p+1}^{T} \varepsilon_t m_t, \sum_{t=p+1}^{T} \varepsilon_t n_t)$, with the inverse of the first matrix on the RHS of equation (3.24) gives

$$\begin{bmatrix} \sum_{t=p+1}^{T} \varepsilon_t m_t \\ \sum_{t=p+1}^{T} \varepsilon_t n_t \end{bmatrix}' \begin{bmatrix} \sum_{t=p+1}^{T} m_t^2 & 0 \\ 0 & \sum_{t=p+1}^{T} n_t^2 \end{bmatrix}^{-\frac{1}{2}} = \begin{bmatrix} Q(m_t)^{\frac{1}{2}} & Q(n_t)^{\frac{1}{2}} \end{bmatrix},$$

where

$$Q(m_t) = \sum_{t=p+1}^{T} \varepsilon_t m_t \Big(\sum_{t=p+1}^{T} m_t^2 \Big)^{-1} \sum_{t=p+1}^{T} m_t \varepsilon_t,$$

and

$$Q(n_t) = \sum_{t=p+1}^{T} \varepsilon_t n_t \Big(\sum_{t=p+1}^{T} n_t^2 \Big)^{-1} \sum_{t=p+1}^{T} n_t \varepsilon_t,$$

by Notation 2.5. Plug equation (3.24) back into equation (3.23) to get

$$Q \begin{bmatrix} m_t \\ n_t \end{bmatrix} = \begin{bmatrix} Q(m_t)^{\frac{1}{2}} & Q(n_t)^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 1 & D_{m,n} \\ D_{m,n} & 1 \end{bmatrix}^{-1} \begin{bmatrix} Q(m_t)^{\frac{1}{2}} \\ Q(n_t)^{\frac{1}{2}} \end{bmatrix}.$$

Noting that $D_{m,n} = O_p(T^{-\alpha})$, the above equation can be rewritten as

$$Q\begin{bmatrix} m_t \\ n_t \end{bmatrix} = \begin{bmatrix} Q(m_t)^{\frac{1}{2}} & Q(n_t)^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 1 & O_p(T^{-\alpha}) \\ O_p(T^{-\alpha}) & 1 \end{bmatrix}^{-1} \begin{bmatrix} Q(m_t)^{\frac{1}{2}} \\ Q(n_t)^{\frac{1}{2}} \end{bmatrix}.$$

Further rearranging gives

$$Q\begin{bmatrix} m_t \\ n_t \end{bmatrix} = Q(m_t) + Q(n_t) + 2Q(m_t)^{\frac{1}{2}} Q(n_t)^{\frac{1}{2}} O_p(T^{-\alpha}). \tag{3.25}$$

Take the last term on the RHS. Use the inequality $[Q(m_t)^{\frac{1}{2}} - Q(n_t)^{\frac{1}{2}}]^2 \geq 0$ to get

$$2Q(m_t)^{\frac{1}{2}} Q(n_t)^{\frac{1}{2}} O_p(T^{-\alpha}) \leq [Q(m_t) + Q(n_t)] \cdot O_p(T^{-\alpha}).$$

Substituting this inequality back into equation (3.25) and rearranging gives

$$Q\begin{bmatrix} m_t \\ n_t \end{bmatrix} = Q(m_t)\left\{1 + O_p(T^{-\alpha})\right\} + Q(n_t)\left\{1 + O_p(T^{-\alpha})\right\}.$$

□

For any two processes m_t and n_t, *Lemma 3.1* gives the decomposition of $Q[(m_t, n_t)']$ provided that the two processes are asymptotically uncorrelated. This requirement is being imposed by the condition

$$D_{m,n} = \frac{\sum_{t=p+1}^{T} m_t n_t}{\sqrt{\sum_{t=p+1}^{T} m_t^2 \sum_{t=p+1}^{T} n_t^2}} = O_p(T^{-\alpha}),$$

for some $\alpha > 0$. This condition requires the correlation coefficient of m_t and n_t, $D_{m,n}$, to converge in probability to zero as T goes to infinity. This lemma can be generalised to vector processes. To see this, suppose we have two vector random processes $\mathbf{m_t} \in$

$\mathbb{R}^m$ and $\mathbf{n_t} \in \mathbb{R}^n$. So long as the condition

$$D_{\mathbf{m},\mathbf{n}} = \{ \sum_{t=p+1}^{T} \mathbf{m_t}\mathbf{m_t}' \}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \mathbf{m_t}\mathbf{n_t}' \{ \sum_{t=p+1}^{T} \mathbf{n_t}\mathbf{n_t}' \}^{-\frac{1}{2}} = \mathbf{O_p}(T^{-\alpha}),$$

for some $\alpha > 0$, is satisfied, we have

$$Q \begin{bmatrix} \mathbf{m_t} \\ \mathbf{n_t} \end{bmatrix} = Q(\mathbf{m_t}) \left\{ 1 + O_p(T^{-\alpha}) \right\} + Q(\mathbf{n_t}) \left\{ 1 + O_p(T^{-\alpha}) \right\}.$$

3.C Analysis of $D_{\widehat{u},\widehat{v}}$ and $D_{\widehat{\mathbf{H}}_u,\widehat{v}}$

Recall from equation (3.20) the numerator of our test statistic is

$$Q \left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) \right] = Q \begin{bmatrix} \widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} \\ \dot{g}_{\widehat{v},t} \end{bmatrix} - Q \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix}.$$

The key terms to be analysed are $Q[(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}, \dot{g}_{\widehat{v},t})']$ and $Q[(\dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t})']$. Before applying *Lemma 3.1* to these two Q functions, we need to check that their arguments, $(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}, \dot{g}_{\widehat{v},t})'$ and $(\dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t})'$, satisfy the asymptotic uncorrelatedness condition.

From Notation 3.2 and Notation 3.3, we have

$$D_{\widehat{u},\widehat{v}} = \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}},$$

and

$$D_{\widehat{\mathbf{H}}_u,\widehat{v}} = \left\{ \sum_{t=p+1}^{T} \widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} \widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} \dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}}.$$

We now verify that both $D_{\widehat{u},\widehat{v}}$ and $D_{\widehat{\mathbf{H}}_u,\widehat{v}}$ are small.

Lemma 3.2. *If Assumptions 2.3, 2.4, and 3.2 are satisfied, then* $D_{\widehat{u},\widehat{v}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}})$, *for some* $0 \leq \psi < \frac{1}{2}$.

Proof of Lemma 3.2.

To prove the lemma, we begin by showing that the following equality holds

$$D_{\widehat{u},\widehat{v}} = \left\{ \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} N_{T,v} \left\{ \sum_{t=p+1}^{T} N'_{T,v} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} N_{T,v} \right\}^{-\frac{1}{2}}.$$

Using this above expression for $D_{\widehat{u},\widehat{v}}$, we then prove that it is asymptotically equivalent to $D_{u,v}$, which is of order $T^{\psi-\frac{1}{2}}$ by *Assumption 3.2*.

1. From Notation 3.2,

$$D_{\widehat{u},\widehat{v}} = \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}}.$$

Premultiplying the middle matrix by $N'^{-1}_{T,u} N'_{T,u}$ leaves the above equation unchanged and we have

$$D_{\widehat{u},\widehat{v}} = \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \right\}^{-\frac{1}{2}} N'^{-1}_{T,u} N'_{T,u} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}}.$$

Rearranging gives

$$D_{\widehat{u},\widehat{v}} = \left\{ N'_{T,u} \left(\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \right)^{\frac{1}{2}} \right\}^{-1} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}}. \quad (3.26)$$

Take the first matrix on the RHS of the above equation, and notice that

$$\left\{ N'_{T,u} \left(\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \right)^{\frac{1}{2}} \right\}^{-1} = \left\{ N'_{T,u} \left(\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \right)^{\frac{1}{2}} \left(\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \right)^{\frac{1}{2}} N_{T,u} \right\}^{-\frac{1}{2}}.$$

Rearrange the RHS to get

$$\left\{ N'_{T,u} \left(\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \right)^{\frac{1}{2}} \right\}^{-1} = \left\{ N'_{T,u} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u} \right\}^{-\frac{1}{2}}. \quad (3.27)$$

Plugging this back into equation (3.26) gives

$$D_{\widehat{u},\widehat{v}} = \left\{ N'_{T,u} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}}.$$

Now post-multiplying the middle matrix of the above equation by $N_{T,v} N_{T,v}^{-1}$ and rearranging gives

$$D_{\widehat{u},\widehat{v}} = \left\{ N'_{T,u} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} N_{T,v} \left\{ \left(\sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right)^{\frac{1}{2}} N_{T,v} \right\}^{-1}. \tag{3.28}$$

Similar to equation (3.27), we can show that the last matrix on the RHS of the above equation $\{(\sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t})^{\frac{1}{2}} N_{T,v}\}^{-1}$ is equivalent to $\{N'_{T,v} \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} N_{T,v}\}^{-\frac{1}{2}}$. Thus, we can rewrite equation (3.28) as

$$D_{\widehat{u},\widehat{v}} = \left\{ \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} N_{T,v} \left\{ \sum_{t=p+1}^{T} N'_{T,v} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} N_{T,v} \right\}^{-\frac{1}{2}}, \tag{3.29}$$

which gives the desired result

Analogously, we have for $D_{u,v}$ that

$$\begin{aligned} D_{u,v} &= \left\{ \sum_{t=p+1}^{T} \dot{g}_{u,t} \dot{g}'_{u,t} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{u,t} \dot{g}'_{v,t} \left\{ \sum_{t=p+1}^{T} \dot{g}_{v,t} \dot{g}'_{v,t} \right\}^{-\frac{1}{2}} \\ &= \left\{ \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \dot{g}'_{u,t} N_{T,u} \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \dot{g}'_{v,t} N_{T,v} \left\{ \sum_{t=p+1}^{T} N'_{T,v} \dot{g}_{v,t} \dot{g}'_{v,t} N_{T,v} \right\}^{-\frac{1}{2}}. \end{aligned} \tag{3.30}$$

2. We now proceed to show that $D_{\widehat{u},\widehat{v}} \approx D_{u,v}$, where the latter is $\mathbf{O_p}(T^{\psi-\frac{1}{2}})$.

Recall from *Notation 3.1* that

$$\mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}} = \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0})' N_{T,\boldsymbol{\theta_0}},$$

and

$$\mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}} = \begin{bmatrix} \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} & \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{v}} \\ \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{u}} & \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u} & \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\dot{g}'_{v,t}N_{T,v} \\ \sum_{t=p+1}^{T} N'_{T,v}\dot{g}_{v,t}\dot{g}'_{u,t}N_{T,u} & \sum_{t=p+1}^{T} N'_{T,v}\dot{g}_{v,t}\dot{g}'_{v,t}N_{T,v} \end{bmatrix}.$$

Under *Assumptions 2.3* and *2.4*, *Lemma 2.8* in Chapter 2 is satisfied and we have

$$\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})'N_{T,\boldsymbol{\theta}_0} = \mathbf{S}_{\dot{\mathbf{g}}\dot{\mathbf{g}}}\Big\{\mathbf{I_q} + \mathbf{o_p}(T^{-\tau})\Big\},$$

which implies

$$\begin{bmatrix} \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}N_{T,u} & \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{v},t}N_{T,v} \\ \sum_{t=p+1}^{T} N'_{T,v}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{u},t}N_{T,u} & \sum_{t=p+1}^{T} N'_{T,v}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}N_{T,v} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} & \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{v}} \\ \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{u}} & \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{v}} \end{bmatrix} \cdot \Big\{\mathbf{I_q} + \mathbf{o_p}(T^{-\tau})\Big\}.$$

Recall from equation (3.29) that

$$D_{\widehat{u},\widehat{v}} = \left\{\sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}N_{T,u}\right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{v},t}N_{T,v} \left\{\sum_{t=p+1}^{T} N'_{T,v}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}N_{T,v}\right\}^{-\frac{1}{2}}.$$

Thus, we have, for each of the three matrices on the RHS,

$$\sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}N_{T,u} = \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}\Big\{\mathbf{I_{d_u}} + \mathbf{o_p}(T^{-\tau})\Big\},$$

and

$$\sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{v},t}N_{T,v} = \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{v}}\Big\{\mathbf{I_{d_v}} + \mathbf{o_p}(T^{-\tau})\Big\},$$

and

$$\sum_{t=p+1}^{T} N'_{T,v}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}N_{T,v} = \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{v}}\Big\{\mathbf{I_{d_v}} + \mathbf{o_p}(T^{-\tau})\Big\},$$

respectively.

Substituting above results back into equation (3.29) gives

$$D_{\widehat{u},\widehat{v}} = \left\{\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}\left[\mathbf{I}_{\mathbf{d_u}} + \mathbf{o_p}(T^{-\tau})\right]\right\}^{-\frac{1}{2}}\left\{\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{v}}\left[\mathbf{I}_{\mathbf{d_v}} + \mathbf{o_p}(T^{-\tau})\right]\right\}\left\{\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{v}}\left[\mathbf{I}_{\mathbf{d_v}} + \mathbf{o_p}(T^{-\tau})\right]\right\}^{-\frac{1}{2}}.$$

Rearrange to get

$$D_{\widehat{u},\widehat{v}} = D_{u,v}\left[\mathbf{I}_{\mathbf{d_v}} + \mathbf{o_p}(T^{-\tau})\right]. \tag{3.31}$$

where

$$D_{u,v} = \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}{}^{-\frac{1}{2}}\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{v}}\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{v}\dot{\mathbf{g}}_\mathbf{v}}{}^{-\frac{1}{2}}.$$

following from equation (3.30) and Notation 3.1.

Recall from *Assumption 3.2* that $D_{u,v}$ is $\mathbf{O_p}(T^{\psi-\frac{1}{2}})$. Substituting this back into equation (3.31) gives

$$D_{\widehat{u},\widehat{v}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}}).$$

□

Recall from Notation 3.3 that

$$D_{\widehat{\mathbf{H}}_u,\widehat{v}} = \left\{\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}}\right\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\dot{g}'_{\widehat{v},t}\left\{\sum_{t=p+1}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\right\}^{-\frac{1}{2}}.$$

We next show that $D_{\widehat{\mathbf{H}}_u,\widehat{v}}$ is also converging in probability to zero.

Lemma 3.3. *If Assumptions 2.3, 2.4, 3.2 and 3.3 are satisfied, then* $D_{\widehat{\mathbf{H}}_u,\widehat{v}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}})$, *for some* $0 \leq \psi < \frac{1}{2}$.

Proof of Lemma 3.3.

From Notation 3.3,

$$\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} = \begin{bmatrix}\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}} \\ \dot{g}_{\widehat{u},t}\end{bmatrix}.$$

Substitute this into $D_{\widehat{\mathbf{H}}_{u,\widehat{v}}}$ to get

$$D_{\widehat{\mathbf{H}}_{u,\widehat{v}}} = \begin{bmatrix} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} & \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} \\ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} & \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{v},t} \\ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \end{bmatrix} \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}}. \tag{3.32}$$

Take the first matrix on the RHS of the above equation and manipulate it to get

$$\begin{bmatrix} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} & \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} \\ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} & \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \end{bmatrix} = \begin{bmatrix} \mathbf{I_p} & \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} (\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t})^{-1} \\ 0 & \mathbf{I_{d_u}} \end{bmatrix}$$
$$\cdot \begin{bmatrix} \sum_{t=p+1}^{T} (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}) (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t})' & 0 \\ 0 & \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \end{bmatrix} \begin{bmatrix} \mathbf{I_p} & 0 \\ (\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t})^{-1} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}{}' & \mathbf{I_{d_u}} \end{bmatrix}.$$

Taking the inverse gives

$$\begin{bmatrix} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} & \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} \\ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} & \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I_p} & 0 \\ -(\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t})^{-1} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}{}' & \mathbf{I_{d_u}} \end{bmatrix}$$
$$\cdot \begin{bmatrix} \sum_{t=p+1}^{T} (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}) (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t})' & 0 \\ 0 & \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I_p} & -\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} (\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t})^{-1} \\ 0 & \mathbf{I_{d_u}} \end{bmatrix}.$$

Now take the square root to get

$$\begin{bmatrix} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} & \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} \\ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} & \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \end{bmatrix}^{-\frac{1}{2}} = \begin{bmatrix} \sum_{t=p+1}^{T} (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}) (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t})' & 0 \\ 0 & \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \end{bmatrix}^{-\frac{1}{2}}$$
$$\cdot \begin{bmatrix} \mathbf{I_p} & -\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} (\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t})^{-1} \\ 0 & \mathbf{I_{d_u}} \end{bmatrix}. \tag{3.33}$$

Before substituting the above equation back into equation (3.32), notice that if we multiply the second matrix on the RHS by the middle matrix in equation (3.32), we get

$$\begin{bmatrix} \mathbf{I_p} & -\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} (\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t})^{-1} \\ 0 & \mathbf{I_{d_u}} \end{bmatrix} \begin{bmatrix} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{v},t} \\ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \end{bmatrix} = \begin{bmatrix} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] \dot{g}'_{\widehat{v},t} \\ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \end{bmatrix}.$$

Thus, plug equation (3.33) back into equation (3.32) and rearrange to obtain

$$D_{\widehat{\mathbf{H}}_u,\widehat{v}} = \begin{bmatrix} \sum_{t=p+1}^{T} (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}) (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t})' & 0 \\ 0 & \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] \dot{g}'_{\widehat{v},t} \\ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \end{bmatrix} \cdot \left\{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right\}^{-\frac{1}{2}}.$$

Further rearranging the above equation yields

$$D_{\widehat{\mathbf{H}}_u,\widehat{v}} = \begin{bmatrix} [\sum_{t=p+1}^{T} (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}) (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t})']^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] \dot{g}'_{\widehat{v},t} (\sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t})^{-\frac{1}{2}} \\ D_{\widehat{u},\widehat{v}} \end{bmatrix}, \tag{3.34}$$

where

$$D_{\widehat{u},\widehat{v}} = (\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t})^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} \left(\sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} \right)^{-\frac{1}{2}}$$

from Notation 3.2.

Following from *Assumption 3.3(b)*, we have for the upper matrix components of $D_{\widehat{u},\widehat{v}}$,

$$[\sum_{t=p+1}^{T} (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}) (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t})']^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] \dot{g}'_{\widehat{v},t} (\sum_{t=p+1}^{T} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t})^{-\frac{1}{2}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}}).$$

According to *Lemma 3.2*, which is satisfied under *Assumptions 2.3*, *2.4* and *3.2*, we have for the lower matrix components $D_{\widehat{u},\widehat{v}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}})$.

Combining above results implies that $D_{\widehat{\mathbf{H}}_{u,\widehat{v}}}$ is $\mathbf{O_p}(T^{\psi-\frac{1}{2}})$.

□

3.D Decomposing Stationary and Random Walk Components

Recall again from equation (3.20)

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = Q\begin{bmatrix}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} - Q\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}. \tag{3.35}$$

Having verified that both the arguments of the two Q functions on the RHS, $Q[(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}, \dot{g}_{\widehat{v},t})']$ and $Q[(\dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t})']$, satisfy the condition required by *Lemma 3.1*, we can now apply this lemma to decompose these functions.

First, consider the second term on the RHS of equation (3.35), $Q[(\dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t})']$. We now show that it can be split into two parts— one involving only the stationary components, $\dot{g}_{\widehat{u},t}$, and the other involving only the non-stationary components, $\dot{g}_{\widehat{v},t}$.

Lemma 3.4. *If Assumptions 2.3, 2.4 and 3.2 are satisfied, then*

$$Q\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} = Q(\dot{g}_{\widehat{u},t})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\} + Q(\dot{g}_{\widehat{v},t})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\},$$

for some $0 \leq \psi < \frac{1}{2}$.

Proof of Lemma 3.4.

From *Lemma 3.2*, which holds under *Assumptions 2.3*, *2.4* and *3.2*, we have $D_{\widehat{u},\widehat{v}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}})$, where $0 \leq \psi < \frac{1}{2}$. Thus we can apply *Lemma 3.1* to get

$$Q\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} = Q(\dot{g}_{\widehat{u},t})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\} + Q(\dot{g}_{\widehat{v},t})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\}. \tag{3.36}$$

□

Next, take the first term on the RHS of equation (3.35), $Q[(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}, \dot{g}_{\widehat{v},t})']$, and again apply *Lemma 3.1*.

Lemma 3.5. *If Assumptions 2.3, 2.4, 3.2 and 3.3 are satisfied, then*

$$Q\begin{bmatrix}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} \\ \dot{g}_{\widehat{v},t}\end{bmatrix} = Q(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\} + Q(\dot{g}_{\widehat{v},t})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\},$$

for some $0 \leq \psi < \frac{1}{2}$.

Proof of Lemma 3.5.

Following from *Lemma 3.3*, which is satisfied under *Assumptions 2.3*, *2.4*, *3.2* and *3.3*, we have $D_{\widehat{\mathbf{H}}_{u,\widehat{v}}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}})$, where $0 \leq \psi < \frac{1}{2}$. Again applying *Lemma 3.1* completes the proof. □

Recall from Notation 3.3,

$$\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} = \begin{bmatrix}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\widehat{u},t}\end{bmatrix}.$$

So $\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}$ only contains the stationary part of the first-order partial derivative functions $\dot{g}_{\boldsymbol{\theta}}$. This indicates that we have effectively decomposed the stationary and integrated elements of $\dot{g}_{\boldsymbol{\theta}}$ in $Q[(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}, \dot{g}_{\widehat{v},t})']$ in *Lemma 3.5*.

3.E Proof of Theorem 3.1

In this section, we restate the main theorem established in Section 3.2 in a mathematically equivalent form based on Notation 2.5, and give a direct proof based on this restatement.

Theorem 3.1. *If Assumptions 2.1, 2.3 to 2.5, 3.1 to 3.3 are satisfied, then*

$$TR^2 = \frac{Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}\right]}{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2} + o_p(1).$$

Proof of Theorem 3.1.

According to Theorem 2.2 of Chapter 2, which holds under *Assumptions 2.1*, *2.3* to *2.5*, and *3.1*, we have

$$TR^2 = \frac{Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right]}{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2} + o_p(1). \tag{3.37}$$

Take the Q function in the numerator and by the formula of partitioned inversion,

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = Q\begin{bmatrix}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\end{bmatrix} - Q\left[\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right]. \tag{3.38}$$

Recall that we can separate out the stationary and explosive components in the $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ vector, such that

$$\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = \begin{bmatrix}\dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t}\end{bmatrix}.$$

Thus we can rewrite equation (3.38) as

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = Q\begin{bmatrix}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t}\end{bmatrix} - Q\begin{bmatrix}\dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t}\end{bmatrix} = Q\begin{bmatrix}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} \\ \dot{g}_{\widehat{v},t}\end{bmatrix} - Q\begin{bmatrix}\dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t}\end{bmatrix}, \tag{3.39}$$

where the second equality follows from Notation 3.3 that $\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} = (\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \dot{g}_{\widehat{u},t})'$.

Consider the first term on the RHS. Under *Assumptions 2.3*, *2.4*, *3.2* and *3.3*,

Lemma 3.5 implies that

$$Q\begin{bmatrix}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} = Q(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\} + Q(\dot{g}_{\widehat{v},t})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\}, \tag{3.40}$$

for some $0 \leq \psi < \frac{1}{2}$.

Now take the second term on the RHS of equation (3.39). Following from *Assumptions 2.3*, *2.4* and *3.2*, *Lemma 3.4* holds and we have

$$Q\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix} = Q(\dot{g}_{\widehat{u},t})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\} + Q(\dot{g}_{\widehat{v},t})\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\}. \tag{3.41}$$

Plugging equations (3.40) and (3.41) back in to equation (3.39), and rearranging yields

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = \left\{Q(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}) - Q(\dot{g}_{\widehat{u},t})\right\}\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\}. \tag{3.42}$$

Notice that

$$Q(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}) = Q\begin{bmatrix}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}\\ \dot{g}_{\widehat{u},t}\end{bmatrix}$$

from Notation 3.3. Thus the difference of $Q(\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}})$ and $Q(\dot{g}_{\widehat{u},t})$ is exactly $Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right]$ by the formula of partitioned inversion. And we can rewrite equation (3.42) as

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\right] = Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right]\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\}. \tag{3.43}$$

Substitute equation (3.43) back into equation (3.37) to get

$$TR^2 = \frac{Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right]\left\{1+O_p(T^{\psi-\frac{1}{2}})\right\}}{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2} + o_p(1). \tag{3.44}$$

Take the term $Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right]$ in the numerator. From Notation 2.5,

$$Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right] = \sum_{t=p+1}^{T} \varepsilon_t[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' \left\{ \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' \right\}^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\varepsilon_t.$$

According to *Assumption 3.3(a)*

$$\left\{Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right]\right\}^{\frac{1}{2}} = \left\{ \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' \right\}^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\varepsilon_t = \mathbf{O_p}(1).$$

It must follow that $Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right]$ is $O_p(1)$. Thus further rearranging equation (3.44) gives

$$TR^2 = \frac{Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right] + O_p(T^{\psi-\frac{1}{2}})}{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2} + o_p(1). \tag{3.45}$$

Now consider the denominator in the above equation. Given *Assumption 2.1*, *Lemma 2.7* of Chapter 2 shows that $T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2$ converges in probability to $\sigma_\varepsilon^2 > 0$, so that $\{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2\}^{-1}$ is $O_p(1)$. Substituting this result back into equation (3.45) and noting that $0 \leq \psi < \frac{1}{2}$ gives

$$TR^2 = \frac{Q\left[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}\right]}{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2} + o_p(1).$$

□

CHAPTER 4

Asymptotic Distribution of the Test Statistic and Simulation Studies

4.1 Introduction

In analysing time series models, the need to test for the presence of persistence in the error term routinely arises. If significant serial correlation is established in the data, one then needs to attempt to identify the type of correlation structure and to re-specify the time series model. This thesis focuses on an auxiliary-regression-based test for serial correlation introduced by Breusch (1978) and Godfrey (1978a,b). We extend the application of their test from linear regression models to nonlinear regression models. In Chapters 2 and 3, we illustrated several asymptotically equivalent forms of the test statistic. The key result is that, while the model allows for both stationary and highly persistent regressors, the latter does not affect the test statistic asymptotically and only the stationary components contribute to the limiting distribution of the test statistic. Building on these results, we show in this chapter that the test statistic has a standard chi-square distribution. The size and power properties of the test are evaluated in Monte-Carlo experiments.

Recall from equation (2.3) that the nonlinear dynamic model is given by

$$y_t = g(x_t; \boldsymbol{\theta_0}) + \varepsilon_t, \qquad t = 1, \ldots, T,$$

where x_t is a $k \times 1$ vector of variables that involve stationary and/or non-stationary

variables as components, $\boldsymbol{\theta_0}$ denotes the q-dimensional vector of unknown parameters that belongs to some parameter space $\Theta \subset \mathbb{R}^q$, $\{\varepsilon_t\}$ is a martingale difference error process and is adapted to the filtration $\mathscr{F}_t$. As in the previous chapter, $g(\cdot)$ is some know real-valued function mapping $\mathbb{R}^k \times \Theta$ into $\mathbb{R}$.

Suppose we have estimated $\boldsymbol{\theta_0}$ by the nonlinear least squares (NLLS) estimation. That is, we obtain their estimators, $\widehat{\boldsymbol{\theta}}$, by minimizing the residual sum of squares with respect to $\boldsymbol{\theta}$ over the parameter space Θ,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, S_T(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min} \sum_{t=1}^{T} \Big\{ y_t - g(x_t;\boldsymbol{\theta}) \Big\}^2, \tag{4.1}$$

and, accordingly, the NLLS residuals are

$$\widehat{\varepsilon}_t = y_t - g(x_t;\widehat{\boldsymbol{\theta}}), \qquad t = 1,\ldots,T.$$

Note that although we focus on the least squares estimator, other asymptotically equivalent estimates can be entertained as well.

In this thesis, we provide an auxiliary-regression based test statistic to test the null hypothesis of no serial correlation against the alternative that the error term follows an AR(p) process (assuming fixed p). To perform this hypothesis test, we stack the first p lags of NLLS residuals in a p-dimensional vector and denote it as $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$. That is, we have $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} = (\widehat{\varepsilon}_{t-1}, \widehat{\varepsilon}_{t-2}, \cdots, \widehat{\varepsilon}_{t-p})'$. Again, we facilitate the test by constructing a linear auxiliary regression of the form

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T, \tag{4.2}$$

where the regressors $\dot{g}_{\boldsymbol{\theta}}(\cdot;\widehat{\boldsymbol{\theta}})$ are the first-order partial derivatives of the nonlinear function $g(\cdot;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at the NLLS estimates $\widehat{\boldsymbol{\theta}}$. The unknown parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are then column vectors of dimension $p \times 1$ and $q \times 1$, respectively. We aim to test the null hypothesis of no autocorrelation by testing the vectors

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are jointly zero. Notice that only $T-p$ auxiliary regressions are available because we are testing against the alternative hypothesis that the disturbances are autocorrelated up to the p-th order. However, the loss of p observations does not affect the asymptotic behaviour of our test statistic when T is sufficiently large.

Recall from equation (2.7) that $\widehat{\mathbf{H}}_t = (\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}, \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})')'$, which is a $(p+q)$ dimensional vector whose elements are the explanatory variables included in the auxiliary regression (4.2). From equation (2.8), the test statistic TR^2 takes the form:

$$TR^2 = T\frac{\widehat{cov}(\widehat{\varepsilon}_t, \widehat{\mathbf{H}}_t)'[\widehat{var}(\widehat{\mathbf{H}}_t)]^{-1}\widehat{cov}(\widehat{\mathbf{H}}_t, \widehat{\varepsilon}_t)}{\widehat{var}(\widehat{\varepsilon}_t)}. \tag{4.3}$$

In Chapter 2, we derived an asymptotically equivalent form of the test statistic

$$TR^2 = T\frac{\widehat{cov}[\varepsilon_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\{\widehat{var}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\}^{-1}\widehat{cov}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \varepsilon_t|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]}{\widehat{var}(\varepsilon_t)} + o_p(1).$$

This suggests that the statistic TR^2 for testing the overall significance of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in equation (4.2) is asymptotically equivalent to the $TR^2_{partial}$ for testing $\grave{\boldsymbol{\alpha}} = \mathbf{0}$ in the following regression

$$\varepsilon_t = \grave{\boldsymbol{\alpha}}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \grave{\boldsymbol{\beta}}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + \grave{e}_t, \qquad t = p+1,\ldots,T. \tag{4.4}$$

In Chapter 3, we showed that the vector of first-order partial derivatives, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, can be broken down into two parts — one involving only stationary processes, $\dot{g}_{\widehat{u},t}$, and the other involving only integrated processes, $\dot{g}_{\widehat{v},t}$. According to Theorem 3.1, the latter does not contribute to the limiting distribution of the test statistic, so the TR^2 in equation (4.3) is then asymptotically the same as the $TR^2_{partial}$ for testing $\breve{\boldsymbol{\alpha}} = 0$ in the following regression:

$$\varepsilon_t = \breve{\boldsymbol{\alpha}}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \breve{\boldsymbol{\beta}}_{\mathbf{1}}'\dot{g}_{\widehat{u},t} + \breve{e}_t, \qquad t = p+1,\ldots,T, \tag{4.5}$$

where we have dropped the non-stationary components $\dot{g}_{\widehat{v},t}$. Hence, we can further write the test statistic TR^2 as

$$TR^2 = T\frac{\widehat{cov}[\varepsilon_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\{\widehat{var}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\}^{-1}\widehat{cov}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \varepsilon_t|\dot{g}_{\widehat{u},t}]}{\widehat{var}(\varepsilon_t)} + o_p(1). \tag{4.6}$$

Building on these results, we show in this chapter that the test statistic TR^2 follows a χ^2 distribution with p degrees of freedom asymptotically.

This chapter is structured as follows. In Section 4.2, we describe the extra assumptions that the results in this chapter rest on and present the main theorems. In Section 4.3, we discuss the assumptions using two specific example models. The properties of the Monte-Carlo studies on the size and empirical power of the test procedure are analysed in Section 4.4. Section 4.5 concludes.

4.2 Assumptions and Theorems

To lay out the asymptotic theory of the auxiliary-regression based test statistic, let us impose some additional moment conditions for the error processes, $\boldsymbol{\varepsilon_{t-1}}$, and the stationary components of the first-order partial derivative functions, $\dot{g}_{u,t}$. These assumptions are essential conditions that should be met to derive the limiting distribution of the test statistic.

We first make the assumption that there exists no perfect linear relationship between $\boldsymbol{\varepsilon_{t-1}}$ and $\dot{g}_{u,t}$.

Assumption 4.1. *Let*

$$\mathbf{H_{u,t}} = \begin{bmatrix} \boldsymbol{\varepsilon_{t-1}} \\ \dot{g}_{u,t} \end{bmatrix}.$$

Then the matrix $T^{-1}\sum_{t=p+1}^{T}\mathbf{H_{u,t}}\mathbf{H'_{u,t}}$ *is full rank w.p.1.*

Assumption 4.1 is the key assumption to establish the limiting distribution of the test statistic. The key feature of this assumption is the covariance matrix of $\mathbf{H_{u,t}}$

being positive definite. The implied data restrictions amount to a constraint on the relationship between the lag error terms, $\boldsymbol{\varepsilon_{t-1}}$, and the stationary part of the first-order partial derivatives, $\dot{g}_{u,t}$. In this model setup, we require that $\dot{g}_{u,t}$ cannot be an exact linear combination of $\boldsymbol{\varepsilon_{t-1}}$. This assumption also prevents $\dot{g}_{u,t}$ and $\boldsymbol{\varepsilon_{t-1}}$ from being linear functions of the same random vector. As a counter example, let $\mathbf{B_1} \in \mathbb{R}^{p\times m}$, $\mathbf{B_2} \in \mathbb{R}^{d_u\times m}$ and let $\mathbf{Z_{t-1}} \in \mathbb{R}^{m\times 1}$ i.i.d. with mean vector zero and variance matrix $\mathbf{I_m}$. Then, let $\boldsymbol{\varepsilon_{t-1}} = \mathbf{B_1 Z_{t-1}}$ and $\dot{g}_{u,t} = \mathbf{B_2 Z_{t-1}}$. In this setup, $\boldsymbol{\varepsilon_{t-1}}$ and $\dot{g}_{u,t}$ both depend linearly on the same random vector, and we find that

$$T^{-1}\sum_{t=p+1}^{T}\mathbf{H_{u,t}H_{u,t}}' = T^{-1}\sum_{t=p+1}^{T}\begin{bmatrix}\mathbf{B_1 Z_{t-1}Z_{t-1}'B_1'} & \mathbf{B_1 Z_{t-1}Z_{t-1}'B_2'}\\ \mathbf{B_2 Z_{t-1}Z_{t-1}'B_1'} & \mathbf{B_2 Z_{t-1}Z_{t-1}'B_2'}\end{bmatrix}.$$

Note that both $\mathbf{B_1}$ and $\mathbf{B_2}$ are deterministic and could be moved outside of the summation operator so that

$$T^{-1}\sum_{t=p+1}^{T}\mathbf{H_{u,t}H_{u,t}}' = \begin{bmatrix}\mathbf{B_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{B_2}\end{bmatrix} T^{-1}\sum_{t=p+1}^{T}\begin{bmatrix}\mathbf{Z_{t-1}Z_{t-1}'} & \mathbf{Z_{t-1}Z_{t-1}'}\\ \mathbf{Z_{t-1}Z_{t-1}'} & \mathbf{Z_{t-1}Z_{t-1}'}\end{bmatrix}\begin{bmatrix}\mathbf{B_1'} & \mathbf{0}\\ \mathbf{0} & \mathbf{B_2'}\end{bmatrix}.$$

Since the middle matrix converges in probability to a singular matrix,

$$T^{-1}\sum_{t=p+1}^{T}\begin{bmatrix}\mathbf{Z_{t-1}Z_{t-1}'} & \mathbf{Z_{t-1}Z_{t-1}'}\\ \mathbf{Z_{t-1}Z_{t-1}'} & \mathbf{Z_{t-1}Z_{t-1}'}\end{bmatrix} \xrightarrow{p} \begin{bmatrix}\mathbf{I_m} & \mathbf{I_m}\\ \mathbf{I_m} & \mathbf{I_m}\end{bmatrix},$$

it follows that the variance-covariance matrix $T^{-1}\sum_{t=p+1}^{T}\mathbf{H_{u,t}H_{u,t}}'$ has zero determinant and *Assumption 4.1* is violated.

In practice, we may face the singularity problem if the distribution of $\mathbf{H_{u,t}}$ is concentrated on a lower-dimensional subspace. In other words, if there are linear relationships between columns or rows of the covariance matrix, $T^{-1}\sum_{t=p+1}^{T}\mathbf{H_{u,t}H_{u,t}}'$, this results in zero determinant in its calculation. This also means that the inverse of the underlying matrix becomes intractable, which leads to a computational problem

in deriving the limiting distribution. To unravel the challenge of a singular covariance matrix, we therefore impose the full rank restriction in *Assumption 4.1*.

Note that *Assumption 4.1* rules out testing for AR errors in the AR(1) model:

$$y_t = \theta y_{t-1} + \varepsilon_t$$

when $\theta = 0$. This is the issue commented in Chapter 2. Recall that when the true θ is strictly less than one in absolute value, the key terms to be analysed are different depending on whether $\theta = 0$ or not, and there is no unified argument. Under *Assumption 4.1*, θ is restricted to be different from zero, and thus the case where $\theta = 0$ is not included in the analysis.

The next assumption is concerned with the probability limits of the variance of the stationary components, $\dot{g}_{u,t}$, as well as its covariance with the lagged error terms, $\boldsymbol{\varepsilon_{t-1}}$.

Assumption 4.2. *Suppose that*

(a) $\sum_{t=p+1}^{T} N_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u} \xrightarrow{p} \mathbf{M_u}$*, where* $\mathbf{M_u}$ *is a* $d_u \times d_u$ *positive definite, and possibly stochastic matrix.*

(b) $T^{-1}\sum_{t=p+1}^{T} \dot{g}_{u,t}\boldsymbol{\varepsilon'_{t-1}} \xrightarrow{p} \mathbf{A}$ *where* $\mathbf{A}$ *is a* $d_u \times p$*-dimensional matrix whose* $(i,j)^{th}$ *element* $\left|a_{ij}\right| < \infty$*, for* $1 \leq i \leq d_u$ *and* $1 \leq j \leq p$*.*

Note that $N_{T,u} = T^{-\frac{1}{2}}\mathbf{I_{d_u}}$ and we have

$$T^{-1}\sum_{t=p+1}^{T} \mathbf{H_{u,t}}\mathbf{H_{u,t}}' = \begin{bmatrix} T^{-1}\sum_{t=p+1}^{T} \boldsymbol{\varepsilon_{t-1}}\boldsymbol{\varepsilon'_{t-1}} & T^{-1}\sum_{t=p+1}^{T} \boldsymbol{\varepsilon_{t-1}}\dot{g}'_{u,t} \\ T^{-1}\sum_{t=p+1}^{T} \dot{g}_{u,t}\boldsymbol{\varepsilon'_{t-1}} & T^{-1}\sum_{t=p+1}^{T} \dot{g}_{u,t}\dot{g}'_{u,t} \end{bmatrix}.$$

Then by *Assumption 4.2*, the off-diagonal terms converge in probability to $\mathbf{A}$ and its transpose, and the second main diagonal element, $T^{-1}\sum_{t=p+1}^{T} \dot{g}_{u,t}\dot{g}'_{u,t}$ converges in probability to $\mathbf{M_u}$. In addition, using *Assumption 2.1*, one can show that the first diagonal term, $T^{-1}\sum_{t=p+1}^{T} \boldsymbol{\varepsilon_{t-1}}\boldsymbol{\varepsilon'_{t-1}}$, converges in probability to $\sigma_\varepsilon^2\mathbf{I_p}$. These assumptions

together implies that

$$T^{-1}\sum_{t=p+1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}' \xrightarrow{p} \begin{bmatrix} \sigma_\varepsilon^2\mathbf{I_p} & \mathbf{A}' \\ \mathbf{A} & \mathbf{M_u} \end{bmatrix}.$$

Since the inverse of $\mathbf{M_u}$ exists by *Assumption 4.2(a)*, we have

$$det(T^{-1}\sum_{t=p+1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}') = det(\sigma_\varepsilon^2\mathbf{I_p} - \mathbf{A}'\mathbf{M_u^{-1}}\mathbf{A})det(\mathbf{M_u}).$$

Under *Assumption 4.1*, the $(p+d_u)\times(p+d_u)$ square matrix, $T^{-1}\sum_{t=p+1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}'$ is invertible w.p.1. It follows immediately that $det(\sigma_\varepsilon^2\mathbf{I_p} - \mathbf{A}'\mathbf{M_u^{-1}}\mathbf{A}) \neq 0$ so the matrix $\sigma_\varepsilon^2\mathbf{I_p} - \mathbf{A}'\mathbf{M_u^{-1}}\mathbf{A}$ has full rank.

In order to apply the martingale Central Limit Theorem to our test statistic, our model should satisfy the conditional Lyapunov condition. To this end, we need to ensure that higher (conditional) order moments of $\dot{g}_{u,t}$ exist. Since under *Assumption 2.4*, $\dot{g}_{u,t}$ is adapted to the filtration $\mathscr{F}_{t-1}$, it suffices to consider unconditional moments.

Assumption 4.3. $\sup_t \mathbb{E}\|\dot{g}_{u,t}\|^{2+\delta} < \infty$ *a.s. for some* $\delta > 0$.

It has been identified that the non-stationary components of the model do not affect the limiting behaviour of the test statistic. In the following theorem, we show that the auxiliary-regression-based test statistic follows a standard χ^2 distribution.

Theorem 4.1. *If Assumptions 2.1, 2.3 to 2.5, 3.1 to 3.3, 4.1 to 4.3 are satisfied, then the residuals based test statistic* $TR^2 \xrightarrow{D} \chi_p^2$.

Theorem 4.1 indicates that standard inferences apply when using the auxiliary-regression-based method to test for autocorrelation in the error terms. The asymptotic properties of the test statistic TR^2 imply the robustness of the proposed test in the presence of non-stationarity. The essential advantage of using the auxiliary-regression-based test is that we do not need to distinguish between stationary and non-stationary

data to determine the limiting distribution of the test statistic under the null hypothesis. The information of non-stationarity is naturally contained in the first-order partial derivatives $\dot{g}_{v,t}$, whose effect on the test statistic dies away at the limits as we have illustrated in Chapter 3. We can make robust inferences by simply conducting this technique, given potentially non-stationary DGPs.

Using the results in Theorem 3.1, which are applicable under *Assumptions 2.1*, *2.3* to *2.5*, *3.1* to *3.3*, the test statistic TR^2 could be written as

$$TR^2 = T\frac{\widehat{cov}[\varepsilon_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\{\widehat{var}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\}^{-1}\widehat{cov}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \varepsilon_t|\dot{g}_{\widehat{u},t}]}{\widehat{var}(\varepsilon_t)} + o_p(1).$$

It has been shown in Appendix 2.C that the denominator, $\widehat{var}(\varepsilon_t)$, converges in probability to σ_ε^2 (see Lemma 2.7). The key component in the derivation of the chi-square distribution is the numerator of the test statistic. Specifically, we prove in the appendix of this chapter that $T^{\frac{1}{2}}\widehat{cov}[\varepsilon_t, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\{\widehat{var}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\}^{-\frac{1}{2}}$ is asymptotically normally distributed.

As discussed in Section 3.4, *Assumption 2.5* is redundant if *Assumptions 3.2* and *3.3* are satisfied together with an extra assumption on the variance-covariance matrix $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'$. We now present this assumption formally.

Assumption 4.4.

$$\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\}^{-\frac{1}{2}} = \mathbf{o_p}(\frac{T^{\frac{1}{2}-\psi}}{\log T}).$$

Recall that $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$ denotes the residuals from the regression of the lagged residuals, $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$, on the stationary components of the first-order partial derivatives, $\dot{g}_{\widehat{u},t}$. So $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'$ is the (un-normalised) sample variance of $[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$. Previously in *Assumption 3.3*, we impose the following condition

$$\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\varepsilon_t = \mathbf{O_p}(1),$$

which requires the matrix $\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'$ to be invertible. So *Assumption 3.3* precludes any exact linear relationship between $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and $\dot{g}_{\widehat{u},t}$. A natural question that arises here is the extent to which we allow $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and $\dot{g}_{\widehat{u},t}$ to be related so that we could still establish the asymptotic properties of the test statistic. This sufficient condition is given in *Assumption 4.4* .

One could try to understand these assumptions through the auxiliary regression. Recall from Section 4.1 that, asymptotically, our test statistic TR^2 constructed from the auxiliary regression in equation (4.2)

$$\widehat{\varepsilon}_t = \boldsymbol{\alpha}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T,$$

is equivalent the TR^2 computed using the regression in equation (4.5)

$$\varepsilon_t = \breve{\boldsymbol{\alpha}}'\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} + \breve{\boldsymbol{\beta}}_{\mathbf{1}}'\dot{g}_{\widehat{u},t} + \breve{e}_t, \qquad t = p+1,\ldots,T.$$

So testing the null hypothesis of no serial correlation is asymptotically equivalent to testing the significance of $\Upsilon = (\breve{\boldsymbol{\alpha}}', \breve{\boldsymbol{\beta}}_{\mathbf{1}}')'$. In standard regression analysis, to construct hypothesis testing on regression parameters, we need to first estimate the model equation. To pin down a unique solution for the estimate of Υ, we require the ability to calculate the inverse of $\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}}$, where $\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}} = (\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}, \dot{g}'_{\widehat{u},t})'$. This is essentially the restriction imposed by *Assumption 3.3*, which rules out the possibility that the matrix $\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}}$ is singular.

If the variables in $\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}$ are not perfectly linearly correlated, but there is near-linear dependence among them, $\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}}$ is close to singularity and thus an 'ill-conditioned" matrix, which causes problems in statistical analyses. Notice that for an invertible square matrix $\mathbf{A}$, if $\lambda \neq 0$ is an eigenvalue of $\mathbf{A}$, then λ^{-1} is an eigenvalue of $\mathbf{A}^{-1}$. If $\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}}$ is near-singular, one or more of its eigenvalues are close to zero, so some of the eigenvalues of $\{\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}}\}^{-1}$ may be very large such that the matrix is unbounded. Hence, *Assumption 4.4* imposes the condition beyond which

collinearity in $\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}$ begins to severely distort the auxiliary regression estimation and the distribution of the test statistic. One could treat *Assumption 4.4* as the threshold of correlation between $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and $\dot{g}_{\widehat{u},t}$ at which should remedies be considered.

In fact, *Assumptions 4.1* and *4.4* bear some relations to each other. In *Assumption 4.1*, we require that the inverse of $T^{-1}\sum_{t=p+1}^{T}\mathbf{H}_{\mathbf{u},\mathbf{t}}\mathbf{H}'_{\mathbf{u},\mathbf{t}}$ exists w.p.1, where $\mathbf{H}_{\mathbf{u},\mathbf{t}} = (\boldsymbol{\varepsilon}'_{\boldsymbol{t-1}}, \dot{g}'_{u,t})'$. In the appendix, we show that $\sum_{t=p+1}^{T}\widehat{\mathbf{H}}_{\mathbf{u},\mathbf{t}}\widehat{\mathbf{H}}'_{\mathbf{u},\mathbf{t}}$ is asymptotically equivalent to $\sum_{t=p+1}^{T}\mathbf{H}_{\mathbf{u},\mathbf{t}}\mathbf{H}'_{\mathbf{u},\mathbf{t}}$, so *Assumption 4.4* technically restricts how small the probability limit of $\sum_{t=p+1}^{T}\mathbf{H}_{\mathbf{u},\mathbf{t}}\mathbf{H}'_{\mathbf{u},\mathbf{t}}$ could get.

Since *Assumption 2.5* is implied by *Assumptions 3.2*, *3.3* and *4.4* jointly (see Appendix 4.E for a formal proof), this assumption is not needed to obtain the limiting distribution of the test statistic, which gives rise to the following theorem.

Theorem 4.2. *If Assumptions 2.1, 2.3, 2.4, 3.1 to 3.3, 4.1 to 4.4 are satisfied, then the residuals based test statistic* $TR^2 \xrightarrow{D} \chi^2_p$.

4.3 Examples

After having introduced some extra assumptions to our model, we take another look at some of the examples in Section 2.4 of Chapter 2. To illustrate, we consider the linear AR(2) model with one unit root and the nonlinear power-curve model studied in Wu (1981). Without loss of generality, we shall restrict $p = 1$ to simplify the notations.

4.3.1 The AR(2) Model with One Unit Root

A simple application is the linear AR(2) model with a single unit root. Reconsider Example 3 of Chapter 2 (which is also discussed in Chapter 3),

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t, \qquad t = 1, \ldots, T,$$

where $\gamma_1 + \gamma_2 = 1$ and $|\gamma_2| < 1$. For simplicity, we assume there is no drift so $\mu = 0$. The error terms satisfy *Assumption 2.1* such that $\{\varepsilon_t\}$ is an $\mathscr{F}_t$-adapted martingale dif-

ference sequence. We maintain the same assumptions that the initial observations, y_{-1} and y_0, are zero and independent of the error terms. We also assume that $\mathbb{E}|\triangle y_t|^{\varphi} < \infty$ for some $\varphi > 2$.

Recall from equation (3.9) that we can rewrite the above model equation (assuming $\mu = 0$) as

$$y_t = \theta_1 \triangle y_{t-1} + \theta_2 y_{t-1} + \varepsilon_t \equiv x_t' \boldsymbol{\theta_0} + \varepsilon_t, \tag{4.7}$$

where $x_t = (\triangle y_{t-1}, y_{t-1})'$, $\boldsymbol{\theta_0} = (\theta_1, \theta_2)'$, $\theta_1 = -\gamma_2$ and $\theta_2 = \gamma_1 + \gamma_2$. Given the linearity of the model, the estimators, $\widehat{\boldsymbol{\theta}}$, could be obtained by the usual OLS estimation. The first-order partial derivatives of the function $g(x_t; \boldsymbol{\theta}) = x_t' \boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}$ are

$$\dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta}) = x_t = \begin{bmatrix} \triangle y_{t-1} \\ \\ y_{t-1} \end{bmatrix}.$$

Note that $\triangle y_{t-1}$ has a stationary root whereas y_{t-1} contains a unit root. Thus, the stationary component of the first-order partial derivatives is $\dot{g}_{u,\boldsymbol{\theta}}(x_t; \boldsymbol{\theta}) = \triangle y_{t-1}$ and this does not depend on $\boldsymbol{\theta}$. So we have $\dot{g}_{\widehat{u},t} = \dot{g}_{u,t} = \triangle y_{t-1}$. The second-order partial derivatives of a model that is linear in parameters are zeros. In addition, for stationary terms, the normalisation is $N_{T,u} = T^{-\frac{1}{2}}$.

For verifying, we have to check *Assumptions 4.1* to *4.3*. We shall assume that *Assumption 4.4* holds. Recall that *Assumptions 3.2*, *3.3* and *4.4* jointly implies *Assumption 2.5*. Since we have verified *Assumptions 2.5*, *3.2* and *3.3* in previous chapters, *Assumption 4.4* should also hold. We begin by illustrating *Assumption 4.2*, and then go back to *Assumption 4.1*.

Checking Assumption 4.2: We first show that that the variance $T^{-1}\sum_{t=2}^{T} \dot{g}_{u,t}^2$ and the covariance $T^{-1}\sum_{t=2}^{T} \dot{g}_{u,t}\varepsilon_{t-1}$ are asymptotically bounded.

Recall from equation (2.28) that

$$T^{-1}\sum_{t=1}^{T} \dot{g}_{u,t}^2 = T^{-1}\sum_{t=1}^{T} (\triangle y_{t-1})^2 \xrightarrow{p} \mathbb{E}[(\triangle y_t)^2].$$

Since the initial observations, y_{-1} and y_0, are both assumed to be zero, we have $\triangle y_0 = 0$ and thus

$$T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}^2 = T^{-1}\sum_{t=1}^{T}\dot{g}_{u,t}^2 \xrightarrow{p} \mathbb{E}[(\triangle y_t)^2].$$

Given that $\{\triangle y_t\}$ is a stationary stochastic process, $\mathbb{E}[(\triangle y_t)^2] > 0$ is stochastically bounded and *Assumption 4.2(a)* is satisfied.

Next, consider the covariance $T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}\varepsilon_{t-1}$ which can be written as

$$T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}\varepsilon_{t-1} = T^{-1}\sum_{t=2}^{T}\triangle y_{t-1}\varepsilon_{t-1} = T^{-1}\sum_{t=2}^{T}(\theta_1\triangle y_{t-2}+\varepsilon_{t-1})\varepsilon_{t-1}$$

using equation (4.7) and noting that $\theta_2 = \gamma_1 + \gamma_2 = 1$. Rearranging the above equation gives

$$T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}\varepsilon_{t-1} = \theta_1 T^{-1}\sum_{t=2}^{T}\triangle y_{t-2}\varepsilon_{t-1} + T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2.$$

Take the first term on the RHS. Note that θ_1 is stochastically bounded and $\triangle y_{t-2}$ is stationary. By *Assumption 2.1*, $\{\triangle y_{t-2}\varepsilon_{t-1}\}$ is a martingale difference sequence. We could then invoke the martingale LLN to get $T^{-1}\sum_{t=2}^{T}\triangle y_{t-2}\varepsilon_{t-1} \xrightarrow{p} 0$ and thus the first term vanishes asymptotically.

Furthermore, by *Assumption 2.1*, the second term on the RHS of the above equation, $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2$, converges in probability to $\sigma_\varepsilon^2 > 0$ under the martingale LLN. Combining these results gives

$$T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}\varepsilon_{t-1} \xrightarrow{p} \sigma_\varepsilon^2$$

and *Assumption 4.2(b)* is satisfied.

Checking Assumption 4.1: In this model, we have

$$\mathbf{H_{u,t}} = \begin{bmatrix}\varepsilon_{t-1}\\ \dot{g}_{u,t}\end{bmatrix} = \begin{bmatrix}\varepsilon_{t-1}\\ \triangle y_{t-1}\end{bmatrix}.$$

With some calculations, we obtain

$$T^{-1}\sum_{t=2}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}' = \begin{bmatrix} T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2 & T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}\triangle y_{t-1} \\ T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}\triangle y_{t-1} & T^{-1}\sum_{t=2}^{T}(\triangle y_{t-1})^2 \end{bmatrix}.$$

From the discussion above, the probability limits of $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}\triangle y_{t-1}$ and $T^{-1}\sum_{t=2}^{T}(\triangle y_{t-1})^2$ are σ_ε^2 and $\mathbb{E}[(\triangle y_t)^2]$, respectively. By *Assumption 2.1*, we also have for the upper main-diagonal term, $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2$, converging in probability to σ_ε^2. Combining these results gives

$$T^{-1}\sum_{t=2}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}' \xrightarrow{p} \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 & \mathbb{E}[(\triangle y_t)^2] \end{bmatrix}.$$

Assumption 4.1 requires the probability limit of $T^{-1}\sum_{t=2}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}'$ has a non-zero determinant, so that it is positive definite. This in turn requires that $\sigma_\varepsilon^4 - \sigma_\varepsilon^2\mathbb{E}[(\triangle y_t)^2]$ is different from zero so that $\mathbb{E}[(\triangle y_t)^2] \neq \sigma_\varepsilon^2$. Note that $\mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2$. We see that *Assumption 4.1* seems to be equivalent to the requirement that $\triangle y_t \neq \varepsilon_t$. Below, we argue that this is equivalent to $\gamma_2 \neq 0$.

Recall the model equation

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t, \qquad t = 1, \ldots, T,$$

where $\gamma_1 + \gamma_2 = 1$ and $|\gamma_2| < 1$. As a result, for *Assumption 4.1* to hold, we would need to impose an additional constraint on the parameter space such that $\gamma_2 \neq 0$.

Note that *Assumption 4.1* breaks down when $\gamma_2 = 0$ and $T^{-1}\sum_{t=2}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}'$ becomes a singular matrix. In this case, the model equation becomes

$$y_t = y_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, T,$$

so we have a random walk process. This is the same model as that studied in Section

2.4.1 in Chapter 2. We could show that in this case, the test statistic still follows a χ^2-distribution with one degree of freedom asymptotically, although the key terms to be analysed are different from those when $\gamma_2 \neq 0$. Thus, whether γ_2 is zero or not does not affect the limiting distribution of TR^2, but the derivation.

Checking Assumption 4.3: This last assumption requires the suprema of higher order moments of $\dot{g}_{u,t}$ being finite. Recall that $\dot{g}_{u,t} = \triangle y_{t-1}$, so $\mathbb{E}|\dot{g}_{u,t}|^{2+\delta} = \mathbb{E}|\triangle y_{t-1}|^{2+\delta}$. Since it is assumed that $\mathbb{E}|\triangle y_t|^{\varphi} < \infty$ for some $\varphi > 2$, the supremum of $\mathbb{E}|\dot{g}_{u,t}|^{2+\delta}$ must exist for any δ greater than zero, so *Assumption 4.3* is confirmed.

4.3.2 The Power-Curve Model

Let us adopt the nonlinear model given in Example 3 of Wu (1981),

$$y_t = (\theta_0 + x_t)^2 + \varepsilon_t, \qquad t = 0, 1, \ldots, T.$$

where $\theta_0 \in \Theta \subset \mathbb{R}$, and Θ is assumed to be a convex space. The martingale difference errors, ε_t, satisfy *Assumption 2.1*. The random variable x_t is measurable with respect to $\mathscr{F}_{t-1}$. We also assume that x_t's are independent and identically distributed with mean zero, variance $\sigma_X^2 < \infty$ and finite higher moments such that $\mathbb{E}|x_t|^{\varphi} < \infty$ for some $\varphi > 2$. In addition, it is assumed that $[x_t, \varepsilon_{t-1}]'$ are jointly i.i.d. with $\mathbb{E}[x_t \varepsilon_{t-1}] = \frac{a}{2}$, where $|a| < \infty$.

Since an i.i.d. process is strictly stationary and the first two moments of x_t exist, x_t is both strictly stationary and covariance stationary. Then the first order partial derivative of the nonlinear function $g(x_t; \theta) = (\theta + x_t)^2$ with respect to θ is also stationary and we have with

$$\dot{g}_{u,t} = \dot{g}_{\theta}(x_t; \theta_0) = 2(\theta_0 + x_t).$$

The second-order partial derivative function is then $\ddot{g}_{\theta\theta}(x_t; \theta) = 2$. The normalisation matrix $N_{T,\boldsymbol{\theta_0}}$ in this example becomes a scalar and is the same as $N_{T,u}$, which is $T^{-\frac{1}{2}}$. Given that both partial derivatives are scalar-valued, the norm notation is unnecessary

for the discussion.

Our goal is to check *Assumptions 4.1* to *4.3*. Note that Assumption 4.4 restricts how small the variance of $[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]$ could get such that the inverse is not too big. Similar to the discussion of *Assumption 2.5* in Section 2.4.4 of Chapter 2, this assumption depends on the underlying structure of the data, which is very hard to check given the nonlinearity of the model and we shall assume that it holds. Again, we start with a discussion on *Assumption 4.2*, and then move back to *Assumption 4.1*.

Checking Assumption 4.2: We need to demonstrate that both the variance of $\dot{g}_{u,t}$ and the covariance between $\dot{g}_{u,t}$ and ε_{t-1} are bounded in probability.

First consider the variance of $\dot{g}_{u,t}$, and recall that $\dot{g}_{u,t} = \dot{g}_\theta(x_t;\theta_0) = 2(\theta_0 + x_t)$. So we have,

$$T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}^2 = 4T^{-1}\sum_{t=2}^{T}(x_t+\theta_0)^2 = 4\Big\{T^{-1}\sum_{t=2}^{T}x_t^2 + 2\theta_0 T^{-1}\sum_{t=1}^{T}x_t + \theta_0^2\Big\}.$$

Provided that $\mathbb{P}(x_t = -\theta_0) = 0$, $T^{-1}\sum_{t=1}^{T}\dot{g}_{u,t}^2$ is strictly positive. Since $\{x_t\}$ is an i.i.d. sequence with mean zero and finite variance, we could invoke the strong LLN to obtain $T^{-1}\sum_{t=1}^{T}x_t \xrightarrow{p} 0$ and $T^{-1}\sum_{t=1}^{T}x_t^2 \xrightarrow{p} \sigma_X^2$. The unknown true parameter θ_0 is a finite constant, and therefore it is $O_p(1)$. Using the Slutsky's theorem, the probability limit of $T^{-1}\sum_{t=1}^{T}\dot{g}_{u,t}^2$ is $4(\sigma_X^2+\theta_0^2)$ which is a positive finite constant.

Next, consider the covariance between $\dot{g}_{u,t}$ and ε_{t-1},

$$T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}\varepsilon_{t-1} = 2T^{-1}\sum_{t=2}^{T}x_t\varepsilon_{t-1} + 2\theta_0 T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}.$$

Since the error terms are assumed to be martingale differences, their sample mean, $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}$, converges in probability to 0 by the martingale LLN. Under the assumption that $[x_t,\varepsilon_{t-1}]'$ are jointly i.i.d. with $\mathbb{E}[x_t\varepsilon_{t-1}] = \frac{a}{2}$, where $|a|<\infty$, the first term on the RHS converges in probability to a according to the strong LLN. Therefore we have $T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}\varepsilon_{t-1} \xrightarrow{p} a$, and *Assumption 4.2* is satisfied.

Checking Assumption 4.1: In this example, we have

$$\mathbf{H_{u,t}} = \begin{bmatrix} \varepsilon_{t-1} \\ \dot{g}_{u,t} \end{bmatrix} = \begin{bmatrix} \varepsilon_{t-1} \\ 2(\theta_0 + x_t) \end{bmatrix}.$$

and

$$T^{-1}\sum_{t=2}^{T} \mathbf{H_{u,t}}\mathbf{H_{u,t}}' = \begin{bmatrix} T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2 & T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}\dot{g}_{u,t} \\ T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}\dot{g}_{u,t} & T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}^2 \end{bmatrix}.$$

From the discussion of *Assumption 4.2*, we have shown that $T^{-1}\sum_{t=1}^{T}\dot{g}_{u,t}^2 \xrightarrow{p} 4(\sigma_X^2 + \theta_0^2)$ and $T^{-1}\sum_{t=2}^{T}\dot{g}_{u,t}\varepsilon_{t-1} \xrightarrow{p} a$. In addition, under the martingale difference assumption, we also have that $T^{-1}\sum_{t=2}^{T}\varepsilon_{t-1}^2 \xrightarrow{p} \sigma_\varepsilon^2$. Combining these results gives

$$T^{-1}\sum_{t=2}^{T} \mathbf{H_{u,t}}\mathbf{H_{u,t}}' \xrightarrow{p} \begin{bmatrix} \sigma_\varepsilon^2 & a \\ a & 4(\sigma_X^2 + \theta_0^2) \end{bmatrix}.$$

Assumption 4.1 requires that the matrix $T^{-1}\sum_{t=2}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}'$ is invertible w.p.1, and therefore the probability limit needs to have a non-zero determinant. This is equivalent to the requirement that $a^2 - 4(\sigma_X^2 + \theta_0^2)\sigma_\varepsilon^2 \neq 0$. Therefore *Assumption 4.1* imposes the constraint that $\sigma_\varepsilon^2 \neq \frac{a^2}{4(\sigma_X^2+\theta_0^2)}$. Recall that the numerator a^2 is the squared covariance between $\dot{g}_{u,t}$ and ε_{t-1}, and that the denominator $4(\sigma_X^2 + \theta_0^2)$ is the variance of $\dot{g}_{u,t}$. *Assumption 4.1* implicitly restricts the squared correlation coefficient between $\dot{g}_{u,t}$ and ε_{t-1} cannot be exactly one. That is, these two terms cannot be an exact linear function of the other. Since this condition depends on the structure of the data, it is very difficult to verify given the nonlinear DGP and we shall assume that it holds.

Checking Assumption 4.3: This assumption requires the existence of suprema of higher order moments of $\dot{g}_{u,t}$. Recall that $\dot{g}_{u,t} = 2(x_t + \theta_0)$, and therefore we have $\mathbb{E}|\dot{g}_{u,t}|^{2+\delta} = \mathbb{E}|2(\theta_0 + x_t)|^{2+\delta}$. For some $0 < C < \infty$, the inequality $\mathbb{E}|2(\theta_0 + x_t)|^{2+\delta} \leq C(|\theta_0|^{2+\delta} + \mathbb{E}|x_t|^{2+\delta})$ always hold. Using this result, we have $\mathbb{E}|\dot{g}_{u,t}|^{2+\delta} < \infty$ for any $\delta > 0$, since θ_0 is a finite constant and $\mathbb{E}|x_t|^{\varphi} < \infty$ for some $\varphi > 2$ by assumption.

It is immediate that the supremum of $\mathbb{E}|\dot{g}_{u,t}|^{2+\delta}$ must exist for any $\delta > 0$, and hence *Assumption 4.3* is satisfied.

4.4 Finite Sample Performance of TR^2

In this section, we study the finite sample performance of the proposed test by assessing the size and power via some Monte Carlo simulations. Because the known asymptotic distribution of a test statistic may possibly provide an inaccurate approximation to its unknown finite sample distribution, it is of great importance to evaluate the effectiveness of the constructed test in finite samples. We hope that the test has an empirical size close to the nominal size, since decisions based on the test will be inappropriate when there is a substantial difference between the two sizes. In addition, the test should be powerful to reject the null hypothesis when the model under the null is misspecified.

We generate the dependent variable y_t with the following regressors and innovations:

$$\begin{aligned} x_t &= \gamma x_{t-1} + v_t, && t = 1,\ldots,T, \\ z_t &= \varphi z_{t-1} + u_t && t = 1,\ldots,T, \\ \varepsilon_t &= \rho \varepsilon_{t-1} + w_t. && t = 1,\ldots,T, \end{aligned}$$

where the error terms v_t, u_t and w_t are generated by a multivariate normal distribution:

$$(v_t, u_t, w_t)' \sim \mathsf{MN}(\mathbf{0}, \mathbf{I_3}).$$

The initial observations, x_0, z_0 and ε_0, are all set to zero.

Here, the processes generating x_t, z_t and ε_t are all AR(1). In particular, we set $\varphi = 0.6$ so that $\{z_t\}$ is a stationary process, whereas γ and ρ control for the degree of persistence in x_t and ε_t, respectively. Since the error terms u_t, v_t and w_t are mutually

independent and all initial values are set to be zero, it follows that x_t, z_t and ε_t are mutually independent.

The data generating processes (DGPs) for y_t are given in Table 4.1. For the sake of comparison, we generate y_t using the processes specified in column 2 of Table 4.1 but estimate the parameters using the regression models $g(x_t, z_t; \boldsymbol{\theta})$ in column 3.

There are four correctly specified models (DGP 1 - 4). In DGP 1, both the parameters and the explanatory variables enter the model linearly. In DGP 2, we consider a nonlinear functional form of the explanatory variables while maintaining the property of linearity in parameters. DGP 3 is based on the prototype of the power-curve model in Wu (1981) where he considered $x_t = t$. We use this example to evaluate the finite sample performance of the test statistic when the underlying model is nonlinear in parameters. We also study an example where there is no intercept (DGP 4) in the model equation to check whether the difference in the empirical sizes in the nonlinear model (DGP 3), in comparison to DGP 1 and 2 (as one shall see below in Table 4.2), is due to the nonlinearity in parameters or the omission of an explicit constant term. Note that DGP 4 is exactly DGP 2 if an intercept were included.

In addition, three misspecified models (DGP 5 - 7) are included in our analyses. In particular, we analyse misspecifications where an AR(1) regressor has been incorrectly omitted from $g(x_t, z_t; \boldsymbol{\theta})$ (DGP 5) and where the relevant regressor is included in the wrong functional form (DGP 6). In DGP 7, we consider a type of misspecification where the error term is integrated of order 2.

Except for DGP 3, where we estimate the model using the nonlinear least squares, the parameters in the linear models (DGP 1, 2, 4 - 7) are estimated by the standard OLS method.

We use these DGPs to test the null hypothesis of serial uncorrelation against the alternative that the error term is autocorrelated up to the sixth-order. To facilitate the

Table 4.1 Data Generating Processes

DGP	y_t	$g(x_t, z_t; \boldsymbol{\theta})$	ε_t
(1)	$1+0.5x_t+\varepsilon_t$	$\theta_0+\theta_1 x_t$	$\rho\varepsilon_{t-1}+w_t$
(2)	$1+0.5x_t^2+\varepsilon_t$	$\theta_0+\theta_1 x_t^2$	$\rho\varepsilon_{t-1}+w_t$
(3)	$(1+x_t)^2+\varepsilon_t$	$(\theta_0+x_t)^2$	$\rho\varepsilon_{t-1}+w_t$
(4)	$0.5x_t^2+\varepsilon_t$	$\theta_1 x_t^2$	$\rho\varepsilon_{t-1}+w_t$
(5)	$0.3z_t+0.5x_t^2+\varepsilon_t$	$\theta_2 z_t$	$\rho\varepsilon_{t-1}+w_t$
(6)	$0.5x_t^2+\varepsilon_t$	$\theta_1 x_t$	$\rho\varepsilon_{t-1}+w_t$
(7)	$1+0.5x_t^2+\varepsilon_t$	$\theta_0+\theta_1 x_t^2$	$2\varepsilon_{t-1}+\varepsilon_{t-2}+w_t$

test, we construct the following auxiliary regression

$$\widehat{\varepsilon}_t = \alpha_1\widehat{\varepsilon}_{t-1}+\alpha_2\widehat{\varepsilon}_{t-2}+\cdots+\alpha_6\widehat{\varepsilon}_{t-6}+\boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})+e_t, \qquad t=7,\ldots,T. \tag{4.8}$$

Testing the null hypothesis of no serial correlation is equivalent to testing the overall significance of the above auxiliary regression. Under $\mathbf{H}_0$, the auxiliary-regression-based test statistic TR^2 is asymptotically χ^2-distributed with six degrees of freedom. Since we focus on small samples in our Monte Carlo analyses, a small finite sample correction is applied to the test statistic, so we use $(T-6)R^2$ to study the size and power instead. This is because, although there are T observations for $\widehat{\varepsilon}_t$, we can only construct $T-6$ auxiliary regression equations. Note that while $(T-6)R^2$ and TR^2 are asymptotically equivalent, the differences between them do matter in finite samples and using $(T-6)R^2$ is more appropriate.

Using the data generating structures specified above, a number of simulations have been conducted by varying the processes generating x_t and ε_t to assess the effectiveness of the proposed test. Typically, we consider three alternative processes to generate the explanatory variable x_t: 1. $|\gamma|$ strictly less than one generates a stationary AR(1) process for x_t; 2. $|\gamma|$ close to one generates a near unit root process for x_t; 3. $|\gamma|=1$ generates a non-stationary process for x_t. The aim of our simulation experiments is to compare the exact distributions of the test statistic and its corresponding asymptotic $\chi^2_{(6)}$ distribution, and show that the effectiveness of the test is robust to any degree of

persistence in x_t.

The θ's are to be estimated in $g(x_t, z_t; \boldsymbol{\theta})$. We take $\gamma = 0.5$, 0.9, 1, and $\rho = 0$, 0.3, 0.5, 1. We study the significance level (size) of the test when $\rho = 0$, and investigate power properties of the test when ρ is different from zero. Sample sizes for the auxiliary regression are $T = 56$, 106, 506.

We examine the empirical rejection frequencies of the null hypothesis at the 5% nominal level. Let $\widehat{p}$ and m denote the empirical rejection frequency and the total number of replications, respectively. The p-value of rejecting the null hypothesis can be estimated using $\widehat{p} = m^{-1} \sum_{i=1}^{m} 1_{(T-6)R_i^2 > 12.592}$. The variance of $\widehat{p}$ is approximately $p(1-p)/m$ [see Chapter 18 of Hendry and Nielsen (2007) for details]. We choose the number of replicates to be $m = 100,000$ in order to minimise the estimation error. Therefore, with $p = 0.05$, the Monte Carlo standard error is approximately 0.0007.

The empirical sizes of the proposed test statistic are reported in Table 4.2. Overall, the proposed test is slightly undersized in finite samples. As the sample size increases, some improvements are observed. As expected, for both linear and nonlinear data generating processes, the estimated sizes under the null approach their nominal level, 0.05.

When the underlying data generating process is linear in parameters and an intercept term is included (DGP 1 and 2), the test statistic is computed from the centred R^2. Note that, our results suggest that the sizes of the test statistic computed from the nonlinear data generating process (DGP 3) are consistent with those computed using the uncentred R^2 in a linear model without a constant term (DGP 4). The empirical sizes based on two different groups of models, with an intercept (DGP 1 and 2) or without an intercept (DGP 3 and 4), deviate from the nominal size, 0.05, in different magnitude. But, for fixed T, the deviation within the same group is nearly uniform in γ across models.

Note that the R^2 considered in this thesis, which is the one defined in equation (2.8), is not the standard R^2 in the literature. Traditionally, the standard R^2, termed

the coefficient of determination, is used in an F-test of overall significance in which the restricted model is the intercept-only model. Contrast to the traditional case, the restricted model considered in this thesis, as opposed to the unrestricted model in equation (2.6), is that $\widehat{\varepsilon}_t = e_t$. If there were an intercept in equation (2.6), it would have been included in $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$. Thus, consistent with theory established in this thesis, under the restricted model where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are jointly zero, there should be no intercept in the auxiliary regression in our test. We shall refer to the R^2 derived with and without a constant in $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ here as the centred and uncentred R^2, respectively.

It is reasonable that the sizes based on the centred and uncentred R^2 differ because the denominator of the R^2 has different interpretations under the two situations.

When the original model contains an intercept (i.e. DGP 1 and 2), that is when there is a constant in $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$, the residuals $\widehat{\varepsilon}_t$ always have mean zero. On the contrary, when an intercept is excluded from the model equation (i.e. DGP 3 and 4), the residuals do not necessarily sum to zero unless the dependent variable has mean zero.

Recall from equation (2.8) that the denominator of the R^2 studied in this thesis is $\widehat{var}(\widehat{\varepsilon}_t) = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2$. For the centred R^2, the mean of $\widehat{\varepsilon}_t$, $\overline{\widehat{\varepsilon}}$, is zero, and thus $\frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2 = \frac{1}{T-p}\sum_{t=p+1}^{T}(\widehat{\varepsilon}_t - \overline{\widehat{\varepsilon}})^2$. In other words, the sample second moment is the same as the sample variance. However, for the uncentred R^2, $\overline{\widehat{\varepsilon}} \neq 0$ in general, and thus $\frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\varepsilon}_t^2 \neq \frac{1}{T-p}\sum_{t=p+1}^{T}(\widehat{\varepsilon}_t - \overline{\widehat{\varepsilon}})^2$. Only when $\widehat{\varepsilon}_t$ has (empirical) mean zero, that is, $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ contains a constant, will the uncentered R^2 coincide with the centred one.

Note that discrepancies of actual sizes and the nominal size of a test may possibly contaminate comparisons of the power of the test in different models, so we need to consider using size-corrected power to unify comparisons. Our results suggest a convenient way of adjusting the test so that its true size equals the nominal size and we can then estimate the power of this adjusted test. We only need to consider two approaches to adjust the size, one for models without an intercept, and one for those with an intercept.

Table 4.2 Empirical Size of the Residual Autocorrelation Test $(T-p)R^2$

	$\gamma=0.5$			$\gamma=0.9$			$\gamma=1$		
Model	$T=56$	$T=106$	$T=506$	$T=56$	$T=106$	$T=506$	$T=56$	$T=106$	$T=506$
$\rho=0$									
(1)	0.041	0.045	0.049	0.046	0.048	0.049	0.050	0.051	0.050
(2)	0.040	0.045	0.049	0.044	0.047	0.049	0.046	0.049	0.050
(3)	0.030	0.038	0.048	0.031	0.039	0.048	0.031	0.038	0.047
(4)	0.030	0.039	0.048	0.031	0.039	0.048	0.032	0.040	0.047

Table 4.3 Size-corrected Power of the Residual Autocorrelation Test

	$\gamma=0.5$			$\gamma=0.9$			$\gamma=1$		
Model	$T=56$	$T=106$	$T=506$	$T=56$	$T=106$	$T=506$	$T=56$	$T=106$	$T=506$
$\rho=0.3$									
(1)	0.213	0.515	1.000	0.192	0.493	1.000	0.191	0.489	1.000
(2)	0.223	0.527	1.000	0.198	0.499	1.000	0.193	0.490	1.000
(3)	0.251	0.567	1.000	0.249	0.557	1.000	0.248	0.550	1.000
(4)	0.264	0.568	1.000	0.252	0.554	1.000	0.245	0.548	1.000
$\rho=0.5$									
(1)	0.652	0.969	1.000	0.614	0.964	1.000	0.611	0.963	1.000
(2)	0.669	0.972	1.000	0.629	0.965	1.000	0.615	0.962	1.000
(3)	0.733	0.979	1.000	0.721	0.976	1.000	0.716	0.975	1.000
(4)	0.730	0.979	1.000	0.714	0.977	1.000	0.706	0.976	1.000
$\rho=1$									
(1)	1.000	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000
(2)	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000
(3)	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(4)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.3 reveals the size-adjusted power of the proposed test under correct specification of the functional form (DGP 1 to 4) but dependent error terms ($\rho \neq 0$). The size-adjusted power in this table are computed through two simulations. In the first simulation, the DGP satisfies the null hypothesis ($\rho = 0$) and the empirical critical values corresponding to the 5% size level are determined by varying the magnitude of γ and sample size n. In the second simulation, the DGP satisfies the alternative hypothesis ($\rho > 0$) and the rejection rates (the power) are computed using the empirical critical values obtained from the first simulation. Take DGP 1 as an illustration. We first derive 9 empirical critical values associated with the 5% size level for 9 different combinations of (γ, n)-values with $\rho = 0$. We then calculate the power by comparing the test statistic against the empirical critical values associated with different (γ, n)-combinations when ρ is different from zero.

From Table 4.3, we can see that the empirical power is close to 1 as the sample

size increases when compared to the empirical critical values. In addition, the larger the value of the coefficient ρ, the higher the power. This is reasonable since a higher ρ implies the martingale difference model specified under the null is far away from the true DGP. Furthermore, it appears that the power of the test is greater for models without a constant term (DGP 3 and 4) than for models with a constant term (DGP 1 and 2).

Table 4.4 Empirical Power of the Test Under Misspecification

	$\gamma = 0.5$			$\gamma = 0.9$			$\gamma = 1$		
Model	$T = 56$	$T = 106$	$T = 506$	$T = 56$	$T = 106$	$T = 506$	$T = 56$	$T = 106$	$T = 506$
$\rho = 0$									
(5)	0.294	0.709	1.000	0.958	1.000	1.000	0.995	1.000	1.000
(6)	0.260	0.659	1.000	0.757	0.983	1.000	0.933	0.999	1.000
$\rho = 0.3$									
(5)	0.478	0.886	1.000	0.976	1.000	1.000	0.997	1.000	1.000
(6)	0.439	0.863	1.000	0.841	0.996	1.000	0.958	1.000	1.000
$\rho = 0.5$									
(5)	0.676	0.971	1.000	0.985	1.000	1.000	0.998	1.000	1.000
(6)	0.655	0.966	1.000	0.902	0.999	1.000	0.975	1.000	1.000
$\rho = 1$									
(5)	0.998	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
(6)	0.998	1.000	1.000	0.997	1.000	1.000	0.999	1.000	1.000
$I(2)$									
(7)	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000

Table 4.4 presents the (size-uncorrected) power of the proposed test under functional form misspecification. Compared to Table 4.3, the power of the test under misspecification is higher than that under correct specification when serial correlation is present. As the explanatory variable x_t and the error term become more persistent, the empirical power of the test increases substantially and approaches one even when the sample size is rather small.

In addition, Table 4.4 shows that, when the regression model suffers from misspecification, the null hypothesis of no serial correlation is rejected even if the error terms are serially uncorrelated ($\rho = 0$). This implies that autocorrelated residuals may be a result of model misspecification rather than serial correlation in the error process. For example, when an AR(1) explanatory variable x_t has been incorrectly omitted from the regression (DGP 5) and $\rho = 0$, the autocorrelated residuals should be treated as

evidence of the autocorrelation structure of this omitted variable rather than autocorrelated errors. We can see that the power approaches one at a faster rate as the degree of persistence in the process generating the omitted x_t becomes greater. Thus, one should be aware that rejection of the null hypothesis could indicate an autocorrelated error process and/or model misspecification. Given that the error terms, ε_t, cannot be observed in practice, researchers need to be careful when interpreting the results of diagnostic tests based on regression residuals. One should not naively conclude the presence of serial correlation when the test statistic is greater than the critical value, unless they are confident that their model is correctly specified.

Overall, the finite sample distribution of the test statistic $(T-6)R^2$ is well approximated by a $\chi^2_{(6)}$ distribution. While there are some small size distortions in finite samples, the size of the test statistic eventually converges to 5%, not being affected by the degree of persistence in the regressors. The power of the test also approaches one as the model deviates from the one under the null hypothesis. These results suggest the effectiveness of our proposed test to correctly reject the null hypothesis for a given significance level.

In above studies, we consider testing against the alternative of autocorrelation up to the sixth-order to show that the auxiliary regression approach enjoys the flexibility to easily test for higher order serial correlation. In practice, a natural question that arises is what is the optimal number of lags (denoted as p) to be included in the auxiliary regression. When choosing the value of p, there is a trade-off between size distortions (which occur if p is too small) and a loss in power (which happens if p is too large). When too few lags are included in the auxiliary regression, one might have to contend with misspecification errors and this may bias the result of the test. On the other hand, when too many lags are included for a given sample size, the model soaks up too many degrees of freedom and the likelihood of multicollinearity also increases. The latter leads to inefficiency and the standard errors of the estimated coefficients tend to be inflated. Consequently, we may tend to commit a Type II error and incorrectly fail to

reject the null hypothesis when a lagged coefficient is in fact statistically significant.

We investigate this tradeoff using DGP 1 as an illustration. We compare the performance of the proposed test when different number of lag residuals are included in the auxiliary regression. Consider the following auxiliary regression

$$\widehat{\varepsilon}_t = \alpha_1\widehat{\varepsilon}_{t-1} + \alpha_2\widehat{\varepsilon}_{t-2} + \cdots + \alpha_p\widehat{\varepsilon}_{t-p} + \boldsymbol{\beta}'\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) + e_t, \qquad t = p+1,\ldots,T, \quad (4.9)$$

where $p = 1$, 2, 4, 6 or 12. Under the null hypothesis of no serial correlation, the test statistic $(T-p)R^2$ follows a χ^2-distribution with p degrees of freedom. The size and (size-uncorrected) power properties of the test are reported in Table 4.5.

Table 4.5 Performance of the Test With Different Number of Lags

	$\gamma = 0.5$			$\gamma = 0.9$			$\gamma = 1$		
Number of Lags	$T = 56$	$T = 106$	$T = 506$	$T = 56$	$T = 106$	$T = 506$	$T = 56$	$T = 106$	$T = 506$
$\rho = 0$									
1	0.053	0.050	0.050	0.055	0.051	0.050	0.056	0.051	0.050
2	0.053	0.051	0.051	0.057	0.053	0.051	0.059	0.054	0.051
4	0.048	0.048	0.050	0.053	0.051	0.050	0.056	0.052	0.051
6	0.041	0.045	0.049	0.046	0.048	0.049	0.050	0.051	0.050
12	0.017	0.032	0.046	0.020	0.035	0.047	0.022	0.038	0.048
$\rho = 0.3$									
1	0.485	0.818	1.000	0.459	0.804	1.000	0.457	0.802	1.000
2	0.377	0.725	1.000	0.358	0.710	1.000	0.359	0.709	1.000
4	0.257	0.595	1.000	0.248	0.581	1.000	0.255	0.584	1.000
6	0.185	0.497	1.000	0.180	0.486	1.000	0.187	0.490	1.000
12	0.056	0.290	0.999	0.059	0.284	0.999	0.065	0.293	0.999
$\rho = 0.5$									
1	0.916	0.998	1.000	0.902	0.997	1.000	0.902	0.997	1.000
2	0.855	0.994	1.000	0.838	0.993	1.000	0.838	0.994	1.000
4	0.739	0.983	1.000	0.719	0.981	1.000	0.723	0.981	1.000
6	0.621	0.966	1.000	0.603	0.962	1.000	0.612	0.963	1.000
12	0.266	0.870	1.000	0.262	0.863	1.000	0.275	0.866	1.000
$\rho = 1$									
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6	1.000	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000
12	0.995	1.000	1.000	0.990	1.000	1.000	0.987	1.000	1.000

From Table 4.5, we can see that there is some oversize for smaller values of p and undersize for larger values. For $p = 12$, the empirical size is significantly less than the nominal 5% when the sample size is relatively small, suggesting that comparing the test statistic to the nominal critical value may result in problematic conclusions in finite samples when too many lags are included. Furthermore, we can see that for a

given sample size, there is a negative relationship between the number of lag residuals included and the power of the test. The higher the value of p, the greater the loss in power. In particular, when $\rho = 0.3$ and $T = 56$, the test has very low power if 12 lag residuals are included in the auxiliary regression, and we barely reject the null hypothesis of no serial correlation. This indicate that when too many lags are included, there are fewer degrees of freedom left in the model, resulting in somewhat unreliable statistical inferences.

While one should bear in mind the tradeoff between inefficiency and misspecification when choosing the optimal lag length, one also needs to take the underlying property of the data into account when making this choice. For example, with annual data, it might suffice to include one or two lags, whereas with quarterly data, one could consider choosing up to 8 lags. For monthly data, one can consider including 6, 12 or 24 lags so long as there is sufficient number of observations. In addition, with the auxiliary regression approach, one does not need to restrict the analysis to successive lags. We could drop those intermediate lags that are expected to be insignificant and include only relevant lags in the test. For example, with quarterly data, one could include only the first four lags and the eighth lag in the auxiliary regression, dropping the fifth, sixth and seventh lags which are potentially irrelavant to the analysis.

We next study an example where the *Assumption 2.5* is violated. We evaluate the size and power of the proposed test under this violation, and compare the results with the portmanteau-type test statistic in Wang and Zhu (2018).

The particular data generating processes for y_t and x_t under consideration are

$$y_t = 0.5x_t + \varepsilon_t \qquad \text{and} \qquad x_t = \varepsilon_{t-1}, \qquad t = 1, \ldots T.$$

We study the size when ε_t is a white noise process and evaluate the power when ε_t follows a stationary moving average process of order one. We estimate the linear model equation $y_t = \theta_1 x_t + \varepsilon_t$ by ordinary least squares.

To simplify the analyses, we test here the null of no disturbance autocorrelation against the alternative that the error term follows an AR(1) process. Therefore, our proposed test statistic $(T-1)R^2$ is asymptotically χ^2-distributed with one degree of freedom under $\mathbf{H}_0$.

According to Wang and Zhu (2018), the portmanteau test statistic in this example is

$$\widehat{U}_T(1) = \frac{T(T+2)}{T-1} \frac{\sum_{t=2}^{T} \widehat{v}_t \widehat{v}_{t-1}}{\sum_{t=1}^{T} \widehat{v}_t^2},$$

where $\widehat{v}_t$ is the residual from the following regression

$$\widehat{\varepsilon}_t = \rho \widehat{\varepsilon}_{t-1} + v_t.$$

Under the null hypothesis, the portmanteau test statistic $\widehat{U}_T(1)$ is also asymptotically $\chi^2_{(1)}$-distributed.

Table 4.6 Size and Power under Violation of Assumption 2.5

	test statistic	$T=56$	$T=106$	$T=506$	$T=5006$
$\varepsilon_t \sim$ white noise					
Size	Auxilliary	0.046	0.049	0.050	0.050
	Portmanteau	0.000	0.000	0.000	0.000
$\varepsilon_t \sim \mathrm{MA}(1)$					
Power	Auxilliary	0.464	0.740	1.000	1.000
	Portmanteau	0.006	0.025	0.858	1.000
Size-corrected Power	Auxilliary	0.477	0.745	1.000	1.000
	Portmanteau	0.830	0.978	1.000	1.000

The simulation results are contained in Table 4.6. We can see that the size and power of our test statistic are not affected by the violation of *Assumption 2.5*. The empirical size of our proposed test is very close to 0.05, even when the sample size is small. The power of our test also approaches one as the sample size increases. On the contrary, the portmanteau-type test statistic in Wang and Zhu (2018) has very small sizes (0.000 means that the size is less than 0.001), which results in problematic empirical power, although they have better size-corrected power. The result that the

size is close to zero implies that the Monte Carlo uncertainty in the portmanteau-type test statistic is large.

The simulation results of this particular example suggest that *Assumption 2.5* is not binding, so it is possible to remove it from our analyses. Since this assumption is asymptotically equivalent to *Assumptions 3.3(a)* and *4.1*, the results here indicate that the latter two could also be dropped. We need to work on relaxing these assumptions in the next stage.

4.5 Conclusion

In Part I, we propose a linear-auxiliary-regression-based test for detecting innovation autocorrelation in nonlinear regression models with martingale difference errors. The limiting behaviour of this test statistic is shown to be robust to different degrees of persistence in the model.

In Chapter 2, we illustrate that the F-test for overall significance in the auxiliary regression is asymptotically equivalent to a partial F test on the slope parameters of the lag residuals, and simplify the test statistic. Chapter 3 shows that, by constructing a normalisation matrix, we can decompose the regressors $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ in the auxiliary regression into stationary and integrated components. Following from this decomposition and the assumption that the stationary and non-stationary components are asymptotically uncorrelated, we prove that the non-stationary terms do not affect the test statistic asymptotically. Therefore, we can safely get rid of the non-stationary components and further simplify the test statistic for asymptotic purposes. In this chapter, we showed that the proposed test statistic follows a standard χ^2 distribution asymptotically and investigated the finite sample properties through some Monte Carlo experiments. Simulation results suggest that, regardless of the underlying dynamic structure of explanatory variables — whether it is stationary or not — the test statistic has reasonable sizes and satisfactory powers, which confirms the usefulness of the proposed test in practice.

It is worth pointing out that rejecting the null hypothesis when the test statistic is large enough does not give us any further information about the real data generating process — whether it is in fact the one specified in the alternative hypothesis that the errors follow an AR(p) process or whether it is something else. One cannot preclude the possibilities that the disturbances may instead follow a moving average process or ARMA process when the error terms are indeed serially correlated. Furthermore, specification errors, such as the omission of relevant variables and misspecification of the model functional form, can also result in rejection of the null hypothesis. For example, if the true relationship between the dependent and explanatory variables is quadratic but we have adopted a regression model that is linear in variables, the residuals may be autocorrelated because of their dependency on the squared regressors.

There are two potential improvements to our proposed test that we will make in the next stage. First, simulation results in the previous section suggests that *Assumption 2.5* is not binding in our analyses. As discussed in Section 4.2, *Assumption 2.5* is asymptotically equivalent to *Assumption 3.3(a)* under *Assumptions 2.4*, *3.2*, *3.3(b)* and *4.4*. It is also shown in the appendix that *Assumption 3.3(a)* is asymptotically the same as *Assumption 4.1*. If *Assumption 2.5* is redundant, this implies that *Assumptions 3.3(a)* and *4.1* are not binding either and should be relaxed.

Second, recall that $\{x_t\}$ is assumed to be adapted to the filtration $\mathscr{F}_{t-1}$ by *Assumption 2.4*. In the next stage, we would like to relax this assumption such that $\{x_t\}$ is $\mathscr{F}_t$-adapted and endogeneity may present. The presence of endogeneity may affect the analysis of the limiting distribution of the test statistic. To see this, we need to consider separately the two scenarios where x_t contains only non-stationary regressors and where x_t contains stationary regressors.

As in Wang and Zhu (2018), when all the regressors are non-stationary, relaxing the $\mathscr{F}_{t-1}$-adapted assumption will not change the limiting distribution of the proposed test statistic even if we allow for contemporary correlation between x_t and the error term. This follows from the asymptotic irrelevance result of the non-stationary com-

ponents in Chapter 3 and that the correlation between the non-stationary x_t and the martingale difference ε_t is temporary.

However, the situation is more complicated when the regressors are stationary. If the $\mathscr{F}_t$-adapted stationary regressors are exogenous such that they are uncorrelated with ε_t, this will not change our current analysis of the limiting distribution, and we can still apply the martingale Laws of Large Numbers and martingale central limit theorem to analyse the test statistic. If, on the other hand, the $\mathscr{F}_t$-adapted stationary regressors are endogenous, the standard martingale results are not applicable. In this case, we would need to find different asymptotic theories that are appropriate to analyse the test statistic. We will try to work this out in the next stage.

Appendix

4.A Notations and Preliminaries

As before, we use the spectral norm for vectors and matrices. For a matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, its spectral norm is simply the square root of the largest eigenvalue of the matrix $\mathbf{S}'\mathbf{S}$, $\|\mathbf{S}\| = \sqrt{\lambda_{\max}(\mathbf{S}'\mathbf{S})}$, where λ denotes the eigenvalue.

Recall from Notation 2.1 that $\nabla_t = g(x_t; \widehat{\boldsymbol{\theta}}) - g(x_t; \boldsymbol{\theta_0})$. We will now further extend this notation to allow for lags of ∇_t.

Notation 4.1. *Let* $\nabla_{t-s} = g(x_{t-s}; \widehat{\boldsymbol{\theta}}) - g(x_{t-s}; \boldsymbol{\theta_0})$, *for* $s \in \{0, 1, \ldots, p\}$.

Stack the p lags of ∇_t *into a vector, and let* $\boldsymbol{\nabla_{t-1}}$ *denote the* $p \times 1$ *vector whose elements are* ∇_{t-s} *for* $s \in \{1, \ldots, p\}$

Recall from Notation 2.2 that $\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}} = N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \widehat{\boldsymbol{\theta}}) - N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t; \boldsymbol{\theta_0})$. The next notation extends the definition to both stationary and nonstationary components of $\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}}$.

Notation 4.2. *Let*

$$\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}} = \begin{bmatrix} N'_{T,u} \dot{g}_{\widehat{u},t} - N'_{T,u} \dot{g}_{u,t} \\ N'_{T,u} \dot{g}_{\widehat{v},t} - N'_{T,v} \dot{g}_{v,t} \end{bmatrix} = \begin{bmatrix} \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}} \\ \dot{\boldsymbol{\nabla}}_{\boldsymbol{v,t}} \end{bmatrix}.$$

Notation 4.3. *Let* $\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\boldsymbol{\varepsilon}} = \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \boldsymbol{\varepsilon}'_{\boldsymbol{t-1}}$.

Note that $N_{T,u} = T^{-\frac{1}{2}} \mathbf{I}_{\mathbf{d_u}}$, so $T^{-\frac{1}{2}} \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\boldsymbol{\varepsilon}} = T^{-1} \sum_{t=p+1}^{T} \dot{g}_{u,t} \boldsymbol{\varepsilon}'_{\boldsymbol{t-1}}$. By *Assumption 4.2*, $T^{-\frac{1}{2}} \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\boldsymbol{\varepsilon}}$ is $\mathbf{O_p}(1)$.

In order to prove the main results of this chapter, we need to formulate the Central Limit Theorem for sums involving martingale difference sequences. The following martingale Central Limit Theorem is a slightly simplified version of that given in

page 384 of Davidson (1994). The simplified version stated here suffices to prove the desired results. See *Chapter 3* of Hall and Heyde (1980) for more general martingale results.

Lemma 4.1 (Martingale Central Limit Theorem)**.** *Suppose* $\{\zeta_t\}_{t=1}^{\infty}$ *is a martingale difference sequence for* $\{\mathscr{F}_t\}_{t=0}^{\infty}$*. If the following conditions hold,*

1. $T^{-1}\sum_{t=1}^{T}\mathbb{E}[\zeta_t^2|\mathscr{F}_{t-1}] \xrightarrow{p} \sigma^2$*; and*
2. $T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}\mathbb{E}[|\zeta_t|^{2+\delta}|\mathscr{F}_{t-1}] \xrightarrow{p} 0$*, for some* $\delta > 0$*.*

Then

$$T^{-\frac{1}{2}}\sum_{t=1}^{T}\zeta_t \xrightarrow{D} N[0,\sigma^2].$$

4.B Intermediate Results

To work out the asymptotic distribution of the test statistic TR^2, start by recalling Theorem 3.1 in Chapter 3. Under *Assumptions 2.1*, *2.3* to *2.5*, *3.1* to *3.3*, Theorem 3.1 implies that the test statistic could be written as

$$TR^2 = \frac{Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]}{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2} + o_p(1),$$

where

$$Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] = \sum_{t=p+1}^{T}\varepsilon_t[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'\}^{-1}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]\varepsilon_t,$$

by Notation 2.5.

According to Lemma 2.7, the denominator $T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2$ converges in probability to σ_ε^2. So the key term to be analysed is the numerator, $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$. Before preceding to derive the asymptotics of $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$, we first establish some intermediate results on some of its key components.

The following lemma concerns the $p \times p$ matrix $\sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}'$.

Lemma 4.2. *If Assumptions 2.1, 2.3 and 2.4 are satisfied, then*

$$\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} = \sum_{t=p+1}^{T} \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} \boldsymbol{\varepsilon}'_{\boldsymbol{t-1}} + \mathbf{o_p}(T^{\frac{1}{2}+\tau}).$$

Proof of Lemma 4.2.

Recall that $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ is a vector of lag residuals up to the p-th order, whose elements are $\widehat{\varepsilon}_{t-s} = y_{t-s} - g(x_{t-s}; \widehat{\boldsymbol{\theta}})$ for $1 \leq s \leq p$. Therefore, the $p \times p$ matrix $\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}$ takes the form

$$\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} = \sum_{t=p+1}^{T} \begin{bmatrix} \widehat{\varepsilon}_{t-1}^2 & \widehat{\varepsilon}_{t-1}\widehat{\varepsilon}_{t-2} & \cdots & \widehat{\varepsilon}_{t-1}\widehat{\varepsilon}_{t-p} \\ \widehat{\varepsilon}_{t-2}\widehat{\varepsilon}_{t-1} & \widehat{\varepsilon}_{t-2}^2 & \cdots & \widehat{\varepsilon}_{t-2}\widehat{\varepsilon}_{t-p} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\varepsilon}_{t-p}\widehat{\varepsilon}_{t-1} & \widehat{\varepsilon}_{t-p}\widehat{\varepsilon}_{t-2} & \cdots & \widehat{\varepsilon}_{t-p}^2 \end{bmatrix}.$$

The main diagonal consists of the sum of squares of the lag residuals up to order p, $\sum_{t=p+1}^{T} \widehat{\varepsilon}_{t-s}^2$ for $s \in \{1, \ldots, p\}$, and all off-diagonal terms are the cross products between $\widehat{\varepsilon}_{t-s}$ and $\widehat{\varepsilon}_{t-j}$ for all $j, s \in \{1, \ldots, p\}$ and $j \neq s$.

Similarly, we have

$$\sum_{t=p+1}^{T} \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} \boldsymbol{\varepsilon}'_{\boldsymbol{t-1}} = \sum_{t=p+1}^{T} \begin{bmatrix} \varepsilon_{t-1}^2 & \varepsilon_{t-1}\varepsilon_{t-2} & \cdots & \varepsilon_{t-1}\varepsilon_{t-p} \\ \varepsilon_{t-2}\varepsilon_{t-1} & \varepsilon_{t-2}^2 & \cdots & \varepsilon_{t-2}\varepsilon_{t-p} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{t-p}\varepsilon_{t-1} & \varepsilon_{t-p}\varepsilon_{t-2} & \cdots & \varepsilon_{t-p}^2 \end{bmatrix}.$$

To show that the difference between the two matrices, $\sum_{t=p+1}^{T} \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} \boldsymbol{\varepsilon}'_{\boldsymbol{t-1}}$ and its sample analogue $\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}$, is of order $T^{\frac{1}{2}+\tau}$, it suffices to make element-wise comparison.

Since p is fixed and finite, we can apply *Lemma 2.6* in Chapter 2, which holds

under *Assumptions 2.1*, *2.3* and *2.4*, to the main diagonal terms so that

$$\sum_{t=p+1}^{T} \widehat{\varepsilon}_{t-s}^2 = \sum_{t=p+1}^{T} \varepsilon_{t-s}^2 + o_p(T^{\frac{1}{2}}) \quad \text{for } s = 1, \ldots, p. \tag{4.10}$$

Now consider the off-diagonal terms, $\sum_{t=p+1}^{T} \widehat{\varepsilon}_{t-s}\widehat{\varepsilon}_{t-j}$ $\forall j, s \in \{1, \ldots, p\}$ and $j \neq s$. Recall from Notation 4.1 that $\nabla_{t-s} = g(x_{t-s}; \widehat{\boldsymbol{\theta}}) - g(x_{t-s}; \boldsymbol{\theta_0})$. Using the model equation and Notation 4.1, we can write $\widehat{\varepsilon}_{t-s}$ as $\varepsilon_{t-s} - \nabla_{t-s}$, and thus

$$\begin{aligned}\sum_{t=p+1}^{T} \widehat{\varepsilon}_{t-s}\widehat{\varepsilon}_{t-j} &= \sum_{t=p+1}^{T} (\varepsilon_{t-s} - \nabla_{t-s})(\varepsilon_{t-j} - \nabla_{t-j}) \\ &= \sum_{t=p+1}^{T} \varepsilon_{t-s}\varepsilon_{t-j} - \sum_{t=p+1}^{T} \varepsilon_{t-s}\nabla_{t-j} - \sum_{t=p+1}^{T} \varepsilon_{t-j}\nabla_{t-s} + \sum_{t=p+1}^{T} \nabla_{t-j}\nabla_{t-s}.\end{aligned}$$

Rearranging gives

$$\sum_{t=p+1}^{T} \varepsilon_{t-s}\varepsilon_{t-j} - \sum_{t=p+1}^{T} \widehat{\varepsilon}_{t-s}\widehat{\varepsilon}_{t-j} = \sum_{t=p+1}^{T} \varepsilon_{t-s}\nabla_{t-j} + \sum_{t=p+1}^{T} \varepsilon_{t-j}\nabla_{t-s} - \sum_{t=p+1}^{T} \nabla_{t-j}\nabla_{t-s}.$$

Using the triangle inequality, we have

$$\begin{aligned}|\sum_{t=p+1}^{T} \varepsilon_{t-s}\varepsilon_{t-j} - \sum_{t=p+1}^{T} \widehat{\varepsilon}_{t-s}\widehat{\varepsilon}_{t-j}| \leq |\sum_{t=p+1}^{T} \varepsilon_{t-s}\nabla_{t-j}| + |\sum_{t=p+1}^{T} \varepsilon_{t-j}\nabla_{t-s}| \\ + |\sum_{t=p+1}^{T} \nabla_{t-j}\nabla_{t-s}|.\end{aligned} \tag{4.11}$$

Without loss of generality, we shall assume $j < s$.

First, consider the last term of the above equation. By the Cauchy-Schwarz Inequality we have

$$|\sum_{t=p+1}^{T} \nabla_{t-j}\nabla_{t-s}| \leq (\sum_{t=p+1}^{T} \nabla_{t-j}^2)^{\frac{1}{2}} (\sum_{t=p+1}^{T} \nabla_{t-s}^2)^{\frac{1}{2}} \tag{4.12}$$

Since $1 \leq j < s \leq p$, where p is finite and fixed, *Lemma 2.4* applies under *Assumptions 2.3* and *2.4* so that $\sum_{t=p+1}^{T} \nabla_{t-j}^2$ and $\sum_{t=p+1}^{T} \nabla_{t-s}^2$ are both of order $o_p(T^{2\tau})$, where

$0 < \tau < \frac{1}{4}$. Thus,

$$|\sum_{t=p+1}^{T} \nabla_{t-j}\nabla_{t-s}| \leq o_p(T^{2\tau}).$$

It remains to show that the first two terms on the RHS of the equation (4.11) is of order $T^{\frac{1}{2}+\tau}$. Take the first term on the RHS of equation (4.11), $|\sum_{t=p+1}^{T} \varepsilon_{t-s}\nabla_{t-j}|$. By the Cauchy-Schwarz Inequality,

$$|\sum_{t=p+1}^{T} \varepsilon_{t-s}\nabla_{t-j}| \leq (\sum_{t=p+1}^{T} \varepsilon_{t-s}^2)^{\frac{1}{2}}(\sum_{t=p+1}^{T} \nabla_{t-j}^2)^{\frac{1}{2}}.$$

Again, using *Lemma 2.4*, we have $\sum_{t=p+1}^{T} \nabla_{t-j}^2 = o_p(T^{2\tau})$. Following from *Lemma 2.7*, which holds under *Assumption 2.1*, $\sum_{t=p+1}^{T} \varepsilon_{t-s}^2$ is of order T. Hence,

$$|\sum_{t=p+1}^{T} \varepsilon_{t-s}\nabla_{t-j}| \leq O_p(T^{\frac{1}{2}}) \cdot o_p(T^{\tau}) = o_p(T^{\frac{1}{2}+\tau}).$$

Similarly, one can show that the second term in equation (4.11), $\sum_{t=p+1}^{T} \varepsilon_{t-j}\nabla_{t-s}$, is also of order $T^{\frac{1}{2}+\tau}$.

Substituting above results back into equation 4.11 gives

$$|\sum_{t=p+1}^{T} \varepsilon_{t-s}\varepsilon_{t-j} - \sum_{t=p+1}^{T} \widehat{\varepsilon}_{t-s}\widehat{\varepsilon}_{t-j}| = o_p(T^{\frac{1}{2}+\tau}). \tag{4.13}$$

The proof is completed by combining equations 4.10 and 4.13 □

The next lemma mirrors the result in *Lemma 2.7* and establishes the probability limit of the variance-covariance matrix of $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$.

Lemma 4.3. *If Assumptions 2.1 is satisfied, then*

$$T^{-1}\sum_{t=p+1}^{T} \boldsymbol{\varepsilon_{t-1}\varepsilon'_{t-1}} \xrightarrow{p} \sigma_{\varepsilon}^2 \mathbf{I_p}.$$

Proof of Lemma 4.3.

Observe that

$$T^{-1}\sum_{t=p+1}^{T}\boldsymbol{\varepsilon_{t-1}\varepsilon_{t-1}'}=T^{-1}\sum_{t=1}^{T}\boldsymbol{\varepsilon_{t-1}\varepsilon_{t-1}'}-T^{-1}\sum_{t=1}^{p}\boldsymbol{\varepsilon_{t-1}\varepsilon_{t-1}'} \tag{4.14}$$

First, take the second term on the RHS. Note that for each t, the random vector $\boldsymbol{\varepsilon_{t-1}}$ is $\mathbf{O_p}(1)$. Thus, by the Continuous Mapping Theorem, the matrix $\boldsymbol{\varepsilon_{t-1}\varepsilon_{t-1}'}$ must also be stochastically bounded. Provided that p is fixed and finite, $\sum_{t=1}^{p}\boldsymbol{\varepsilon_{t-1}\varepsilon_{t-1}'}$ is $\mathbf{O_p}(1)$, and we have $T^{-1}\sum_{t=1}^{p}\boldsymbol{\varepsilon_{t-1}\varepsilon_{t-1}'}=\mathbf{O_p}(T^{-1})$.

It remains to show that the first term on the RHS of equation (4.14) converges in probability to $\sigma_\varepsilon^2\mathbf{I_p}$. Recognise that

$$\sum_{t=1}^{T}\boldsymbol{\varepsilon_{t-1}\varepsilon_{t-1}'}=\sum_{t=1}^{T}\begin{bmatrix}\varepsilon_{t-1}^2 & \varepsilon_{t-1}\varepsilon_{t-2} & \cdots & \varepsilon_{t-1}\varepsilon_{t-p}\\ \varepsilon_{t-2}\varepsilon_{t-1} & \varepsilon_{t-2}^2 & \cdots & \varepsilon_{t-2}\varepsilon_{t-p}\\ \vdots & \vdots & \ddots & \vdots\\ \varepsilon_{t-p}\varepsilon_{t-1} & \varepsilon_{t-p}\varepsilon_{t-2} & \cdots & \varepsilon_{t-p}^2\end{bmatrix}.$$

Consider the main-diagonal terms, which are the sample variances of ε_{t-s}, for $1\leq s\leq p$. Under *Assumption 2.1*, using a similar argument of *Lemma 2.7*, we obtain

$$T^{-1}\sum_{t=1}^{T}\varepsilon_{t-s}^2\xrightarrow{p}\sigma_\varepsilon^2. \tag{4.15}$$

Now consider the off-diagonal terms $\sum_{t=1}^{T}\varepsilon_{t-s}\varepsilon_{t-j}$, for $j,s\in\{1,\ldots,p\}$ and $j\neq s$. Given the symmetric property of a variance-covariance matrix, we shall take $s<j$.

Observe that $\{\varepsilon_{t-s}\varepsilon_{t-j}\}$ is a martingale difference sequence, since

$$\mathbb{E}(\varepsilon_{t-s}\varepsilon_{t-j}|\mathscr{F}_{t-s-1})=\mathbb{E}(\varepsilon_{t-s}|\mathscr{F}_{t-s-1})\varepsilon_{t-j}=0.$$

We can then apply the martingale Laws of Large Numbers given in *Lemma 2.1* to $\varepsilon_{t-s}\varepsilon_{t-j}$ after checking the condition in equation (2.43).

Consider

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}(|\varepsilon_{t-s}\varepsilon_{t-j}|^2|\mathscr{F}_{t-s-1})}{t^2} = \sum_{t=1}^{\infty} \frac{\mathbb{E}(\varepsilon_{t-s}^2|\mathscr{F}_{t-s-1})\varepsilon_{t-j}^2}{t^2}.$$

Following from *Assumption 2.1(b)* that $\mathbb{E}(\varepsilon_{t-s}^2|\mathscr{F}_{t-s-1}) = \sigma_\varepsilon^2$, where σ_ε^2 is a finite constant, we get

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}(|\varepsilon_{t-s}\varepsilon_{t-j}|^2|\mathscr{F}_{t-s-1})}{t^2} = \sigma_\varepsilon^2 \sum_{t=1}^{\infty} \frac{\varepsilon_{t-j}^2}{t^2}.$$

Taking expectation of both sides gives

$$\mathbb{E}\{\sum_{t=1}^{\infty} \frac{\mathbb{E}(|\varepsilon_{t-s}\varepsilon_{t-j}|^2|\mathscr{F}_{t-s-1})}{t^2}\} = \sigma_\varepsilon^2 \mathbb{E}\{\sum_{t=1}^{\infty} \frac{\varepsilon_{t-j}^2}{t^2}\}.$$

Since ε_{t-j}^2/t^2 is non-negative, we can interchange the expectation operator and the infinite summation by the monotone convergence theorem to get

$$\mathbb{E}\{\sum_{t=1}^{\infty} \frac{\mathbb{E}(|\varepsilon_{t-s}\varepsilon_{t-j}|^2|\mathscr{F}_{t-s-1})}{t^2}\} = \sigma_\varepsilon^2 \sum_{t=1}^{\infty} \frac{\mathbb{E}(\varepsilon_{t-j}^2)}{t^2},$$

By the law of iterated expectations and again using *Assumption 2.1(b)*, the unconditional expectation of ε_{t-j} is also σ_ε^2 and we obtain

$$\mathbb{E}\{\sum_{t=1}^{\infty} \frac{\mathbb{E}(|\varepsilon_{t-s}\varepsilon_{t-j}|^2|\mathscr{F}_{t-s-1})}{t^2}\} = \sigma_\varepsilon^4 \sum_{t=1}^{\infty} \frac{1}{t^2} = \sigma_\varepsilon^4 \frac{\pi^2}{6} < \infty.$$

This implies that the condition specified in equation (2.43) is satisfied, so we can safely apply *Lemma 2.1* to $\varepsilon_{t-s}\varepsilon_{t-j}$ such that

$$T^{-1} \sum_{t=1}^{T} \varepsilon_{t-s}\varepsilon_{t-j} \xrightarrow{p} 0. \tag{4.16}$$

The proof is completed by combining the results in equations (4.15) and (4.16).

□

The next lemma establishes the difference between $\sum_{t=p+1}^{T} \varepsilon_t \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and $\sum_{t=p+1}^{T} \varepsilon_t \boldsymbol{\varepsilon}_{\boldsymbol{t-1}}$

is $\mathbf{o_p}(T^{\frac{1}{2}})$.

Lemma 4.4. *If Assumptions 2.1, 2.3 and 2.4 are satisfied, then*

$$\sum_{t=p+1}^{T} \varepsilon_t \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} = \sum_{t=p+1}^{T} \varepsilon_t \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} + \mathbf{o_p}(T^{\frac{1}{2}}).$$

Proof of Lemma 4.4.

Using the model equation and Notation 4.1, we have $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} = \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} - \boldsymbol{\nabla}_{\boldsymbol{t-1}}$. Therefore

$$\sum_{t=p+1}^{T} \varepsilon_t \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} = \sum_{t=p+1}^{T} \varepsilon_t \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} - \sum_{t=p+1}^{T} \varepsilon_t \boldsymbol{\nabla}_{\boldsymbol{t-1}}. \tag{4.17}$$

To show that the difference between $\sum_{t=p+1}^{T} \varepsilon_t \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}$ and $\sum_{t=p+1}^{T} \varepsilon_t \boldsymbol{\varepsilon}_{\boldsymbol{t-1}}$, it suffices to show the second term on the RHS of the above equation is of order $T^{\frac{1}{2}}$.

From Notation 4.1, $\boldsymbol{\nabla}_{\boldsymbol{t-1}} = (\nabla_{t-1}, \cdots, \nabla_{t-p})'$, where $\nabla_{t-s} = g(x_{t-s}; \widehat{\boldsymbol{\theta}}) - g(x_{t-s}; \boldsymbol{\theta_0})$ for $s \in \{1, \ldots, p\}$. To complete the proof, we will show that the elements of the vector $\sum_{t=p+1}^{T} \varepsilon_t \boldsymbol{\nabla}_{\boldsymbol{t-1}}$, $\sum_{t=p+1}^{T} \varepsilon_t \nabla_{t-s}$ for $s \in \{1, \ldots, p\}$, are $o_p(T^{\frac{1}{2}})$.

Consider the term $g(\cdot; \widehat{\boldsymbol{\theta}})$ in ∇_{t-s}. The Taylor expansion of $g(\cdot; \widehat{\boldsymbol{\theta}})$ with respect to $\widehat{\boldsymbol{\theta}}$ around $\boldsymbol{\theta_0}$ gives

$$g(x_{t-s}; \widehat{\boldsymbol{\theta}}) = g(x_{t-s}; \boldsymbol{\theta_0}) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \dot{g}_{\boldsymbol{\theta}}(x_{t-s}; \boldsymbol{\theta_0}) + R(x_{t-s}; \widehat{\boldsymbol{\theta}}),$$

where $R(x_{t-s}; \widehat{\boldsymbol{\theta}}) = g(x_{t-s}; \widehat{\boldsymbol{\theta}}) - g(x_{t-s}; \boldsymbol{\theta_0}) - (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \dot{g}_{\boldsymbol{\theta}}(x_{t-s}; \boldsymbol{\theta_0})$ denotes the higher order expansion terms by Lemma 2.3. Then ∇_{t-s} can be written as

$$\nabla_{t-s} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \dot{g}_{\boldsymbol{\theta}}(x_{t-s}; \boldsymbol{\theta_0}) + R(x_{t-s}; \widehat{\boldsymbol{\theta}}).$$

Plug this into $\sum_{t=p+1}^{T} \varepsilon_t \nabla_{t-s}$ and rearrange to get

$$\sum_{t=p+1}^{T} \varepsilon_t \nabla_{t-s} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' \sum_{t=p+1}^{T} \dot{g}_{\boldsymbol{\theta}}(x_{t-s}; \boldsymbol{\theta_0}) \varepsilon_t + \sum_{t=p+1}^{T} R(x_{t-s}; \widehat{\boldsymbol{\theta}}) \varepsilon_t. \tag{4.18}$$

- Consider the first term on the RHS. Pre-multiply the summation term by $(N'_{T,\boldsymbol{\theta_0}})^{-1}N'_{T,\boldsymbol{\theta_0}}$ to get

$$(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})'\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_{t-s};\boldsymbol{\theta_0})\varepsilon_t=\widehat{\boldsymbol{\phi}}'_T\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_{t-s};\boldsymbol{\theta_0})\varepsilon_t,$$

where $\widehat{\boldsymbol{\phi}}_T=N^{-1}_{T,\boldsymbol{\theta_0}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})$ by Notation 2.4.

Let $\mathbf{S_{\dot{g}\dot{g},t-s}}:=\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_{t-s};\boldsymbol{\theta_0})\dot{g}_{\boldsymbol{\theta}}(x_{t-s};\boldsymbol{\theta_0})'N_{T,\boldsymbol{\theta_0}}$. Next post-multiply $\widehat{\boldsymbol{\phi}}'_T$ by $\mathbf{S}^{\frac{1}{2}}_{\mathbf{\dot{g}\dot{g},t-s}}$ and its inverse to get

$$(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})'\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_{t-s};\boldsymbol{\theta_0})\varepsilon_t=\widehat{\boldsymbol{\phi}}'_T\mathbf{S}^{\frac{1}{2}}_{\mathbf{\dot{g}\dot{g},t-s}}\mathbf{S}^{-\frac{1}{2}}_{\mathbf{\dot{g}\dot{g},t-s}}\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_{t-s};\boldsymbol{\theta_0})\varepsilon_t.$$

Recognise that $\widehat{\boldsymbol{\phi}}_T$ is $\mathbf{o_p}(T^{\tau})$ under *Assumption 2.3*, where $0<\tau<\frac{1}{4}$. By *Assumption 2.4* , $\mathbf{S}^{\frac{1}{2}}_{\mathbf{\dot{g}\dot{g},t-s}}$ is $\mathbf{O_p}(1)$. Furthermore, under *Assumption 2.4*, we can invoke *Lemma 1* of Lai and Wei (1982) so that $\mathbf{S}^{-\mathbf{1/2}}_{\mathbf{\dot{g}\dot{g},t-s}}\sum_{t=p+1}^{T}N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_{t-s};\boldsymbol{\theta_0})\varepsilon_t$ is $\mathbf{O_p}(logT)$. This is, in turn, dominated by $\mathbf{O_p}(T^{\alpha}),\forall\alpha>0$. Throughout the proof, we shall assume $\alpha<\frac{1}{4}$. Combining these results then yields

$$(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})'\sum_{t=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_{t-s};\boldsymbol{\theta_0})\varepsilon_t=\mathbf{o_p}(T^{\tau})\cdot\mathbf{O_p}(1)\cdot\mathbf{O_p}(T^{\alpha})=o_p(T^{\tau+\alpha}). \qquad (4.19)$$

where $\tau+\alpha<\frac{1}{2}$.

- Now consider the second term on the RHS of equation (4.18). Using the Cauchy-Schwarz Inequality we obtain

$$|\sum_{t=p+1}^{T}R(x_{t-s};\widehat{\boldsymbol{\theta}})\varepsilon_t|\leq[\sum_{t=p+1}^{T}\varepsilon_t^2]^{\frac{1}{2}}\cdot\{\sum_{t=p+1}^{T}[R(x_{t-s};\widehat{\boldsymbol{\theta}})]^2\}^{\frac{1}{2}}.$$

Lemma 2.7 holds under *Assumption 2.1* which gives $\sum_{t=p+1}^{T}\varepsilon_t^2$ is $O_p(T)$. By *Assumptions 2.3* and *2.4*, we could apply *Lemma 2.3* to show that $\sum_{t=p+1}^{T}[R(x_{t-s};\widehat{\boldsymbol{\theta}})]^2$

is $o_p(1)$. Thus, we obtain

$$\left| \sum_{t=p+1}^{T} \varepsilon_t R(x_{t-s}; \widehat{\boldsymbol{\theta}}) \right| \leq [O_p(T)]^{\frac{1}{2}} \cdot [o_p(1)]^{\frac{1}{2}} = o_p(T^{\frac{1}{2}}). \tag{4.20}$$

Substitute equations (4.19) and (4.20) back into equation (4.18) to get

$$\sum_{t=p+1}^{T} \varepsilon_t \nabla_{t-s} = O_p(T^{2\alpha}) + o_p(T^{\frac{1}{2}}) = o_p(T^{\frac{1}{2}}).$$

□

Recall from *Notation 4.3* that $\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} = \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \boldsymbol{\varepsilon}'_{\mathbf{t-1}}$ and note that $N_{T,u} = T^{-\frac{1}{2}} \mathbf{I}_{\mathbf{d_u}}$. We proceed to illustrate that the covariance between $\dot{g}_{\widehat{u},t}$ and $\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}$ is asymptotically equivalent to that between $\dot{g}_{u,t}$ and $\boldsymbol{\varepsilon}_{\mathbf{t-1}}$.

Lemma 4.5. *By Assumptions 2.1, 2.3 and 2.4,*

$$\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\mathbf{t-1}} = \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} + \mathbf{o_p}(T^{\frac{1}{2}-\tau}).$$

Proof of Lemma 4.5.

The main idea of the proof is to show that the norm of the difference between $\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\mathbf{t-1}}$ and $\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon}$ is of order $T^{\frac{1}{2}-\tau}$.

Recall from Notation 4.1 that $\boldsymbol{\nabla}_{\mathbf{t-1}}$ denotes the $p \times 1$ vector whose elements are $\nabla_{t-s} = g(x_{t-s}; \widehat{\boldsymbol{\theta}}) - g(x_{t-s}; \boldsymbol{\theta_0})$ where $1 \leq s \leq p$. Using the model equation, we obtain $\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}} = \boldsymbol{\varepsilon}_{\mathbf{t-1}} - \boldsymbol{\nabla}_{\mathbf{t-1}}$. By Notation 4.2, $N'_{T,u} \dot{g}_{\widehat{u},t} = N'_{T,u} \dot{g}_{u,t} + \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}$. Substitute these two expressions into $\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\mathbf{t-1}}$ and rearrange to get

$$\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\mathbf{t-1}} = \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \boldsymbol{\varepsilon}'_{\mathbf{t-1}} - \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \boldsymbol{\nabla}'_{\mathbf{t-1}} + \sum_{t=p+1}^{T} \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}} \boldsymbol{\varepsilon}'_{\mathbf{t-1}} - \sum_{t=p+1}^{T} \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}} \boldsymbol{\nabla}'_{\mathbf{t-1}}.$$

Note that the first term on the RHS is exactly $\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon}$. Further rearranging gives

$$\left\| \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} - \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} \right\| = \left\| \sum_{t=p+1}^{T} \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}\boldsymbol{\varepsilon}'_{\boldsymbol{t-1}} - \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\boldsymbol{\nabla}'_{\boldsymbol{t-1}} - \sum_{t=p+1}^{T} \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}\boldsymbol{\nabla}'_{\boldsymbol{t-1}} \right\|.$$

By the triangle inequality and norm compatibility, we get

$$\left\| \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} - \mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} \right\| \leq \sum_{t=p+1}^{T} \left\| \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}} \right\| \|\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}\| + \sum_{t=p+1}^{T} \left\| N'_{T,u}\dot{g}_{u,t} \right\| \|\boldsymbol{\nabla}_{\boldsymbol{t-1}}\| + \sum_{t=p+1}^{T} \left\| \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}} \right\| \|\boldsymbol{\nabla}_{\boldsymbol{t-1}}\|. \quad (4.21)$$

To complete the proof, it suffices to show that the terms on the RHS are small. Apply the Cauchy-Schwarz inequality to each of the three terms to obtain

- $\sum_{t=p+1}^{T} ||\dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}|| ||\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|| \leq [\sum_{t=p+1}^{T} ||\dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}||^2]^{\frac{1}{2}} \cdot [\sum_{t=p+1}^{T} ||\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}||^2]^{\frac{1}{2}}$.

  Recall that $\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}$ is a $p \times 1$ vector of lag error terms. Given that p is finite, by *Assumption 2.1*, we obtain $\sum_{t=p+1}^{T} \varepsilon_{t-s}^2 = O_p(T)$ for $s \in \{1, \ldots, p\}$ using *Lemma 2.7*. Therefore, $\sum_{t=p+1}^{T} ||\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}||^2$ is of order T.

  Under *Assumptions 2.3* and *2.4*, *Lemma 2.5* establishes that $\sum_{t=p+1}^{T} ||\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}}||^2$ is $o_p(T^{-2\tau})$. Recall from Notation 4.2 that $\dot{\boldsymbol{\nabla}}_{\boldsymbol{t}} = (\dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}, \dot{\boldsymbol{\nabla}}_{\boldsymbol{v,t}})'$. It must follow that $\sum_{t=p+1}^{T} ||\dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}||^2$ is also of order $T^{-2\tau}$. Thus,

$$\sum_{t=p+1}^{T} \left\| \dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}} \right\| \|\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}\| \leq \left[o_p(T^{-2\tau}) \right]^{\frac{1}{2}} \cdot \left[O_p(T) \right]^{\frac{1}{2}} = o_p(T^{\frac{1}{2}-\tau}). \quad (4.22)$$

- $\sum_{t=p+1}^{T} ||N'_{T,u}\dot{g}_{u,t}|| ||\boldsymbol{\nabla}_{\boldsymbol{t-1}}|| \leq [\sum_{t=p+1}^{T} ||N'_{T,u}\dot{g}_{u,t}||^2]^{\frac{1}{2}} \cdot [\sum_{t=p+1}^{T} ||\boldsymbol{\nabla}_{\boldsymbol{t-1}}||^2]^{\frac{1}{2}}$.

  By Notation 4.1, $\boldsymbol{\nabla}_{\boldsymbol{t-1}}$ is a $p \times 1$ vector of ∇_{t-s} for $s \in \{1, \ldots, p\}$. Under *Assumptions 2.3* and *2.4*, *Lemma 2.4* holds, and we have $\sum_{t=1}^{T} \nabla_{t-s}^2$ are $o_p(T^{2\tau})$. Thus, $\sum_{t=p+1}^{T} ||\boldsymbol{\nabla}_{\boldsymbol{t-1}}||^2$ is also of order $T^{2\tau}$. From *Assumption 2.4*, $\sum_{t=p+1}^{T} ||N'_{T,u}\dot{g}_{u,t}||^2$ is

$O_p(1)$. Then

$$\sum_{t=p+1}^{T} ||N'_{T,u}\dot{g}_{u,t}||||\boldsymbol{\nabla_{t-1}}|| \leq \left[o_p(T^{2\tau})\right]^{\frac{1}{2}} \cdot \left[O_p(1)\right]^{\frac{1}{2}} = o_p(T^{\tau}). \tag{4.23}$$

- $\sum_{t=p+1}^{T} ||\dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}||||\boldsymbol{\nabla_{t-1}}|| \leq [\sum_{t=p+1}^{T} ||\dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}||^2]^{\frac{1}{2}} \cdot [\sum_{t=p+1}^{T} ||\boldsymbol{\nabla_{t-1}}||^2]^{\frac{1}{2}}$.

  By *Lemmas 2.4* and *2.5*,

$$\sum_{t=p+1}^{T} ||\dot{\boldsymbol{\nabla}}_{\boldsymbol{u,t}}||||\boldsymbol{\nabla_{t-1}}|| \leq \left[o_p(T^{-2\tau})\right]^{\frac{1}{2}} \cdot \left[o_p(T^{2\tau})\right]^{\frac{1}{2}} = o_p(1). \tag{4.24}$$

Combining the results from equations (4.22) to (4.24) completes the proof of *Lemma 4.5*.

□

Following a similar derivation as that for *Lemma 4.5*, we next demonstrate the sample covariance between $\dot{g}_{\widehat{u},t}$ and the contemporaneous error ε_t is asymptotically equivalent to its population counterpart.

Lemma 4.6. *By Assumptions 2.1, 2.3 and 2.4,*

$$\sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\varepsilon_t = \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\varepsilon_t + \mathbf{o_p}(T^{-3\tau}).$$

Proof of Lemma 4.6.

Recall that

$$N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} N_{T,u} & 0 \\ 0 & N_{T,v} \end{bmatrix}' \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix} = \begin{bmatrix} N'_{T,u}\dot{g}_{\widehat{u},t} \\ N'_{T,v}\dot{g}_{\widehat{v},t} \end{bmatrix}.$$

Therefore, $\sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{\widehat{u},t}\varepsilon_t$ is the first d_u elements of the $q \times 1$ vector $\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\varepsilon_t$. To prove Lemma 4.6, it suffices to show that the difference

between $\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\varepsilon_t$ and $\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_0)\varepsilon_t$ is small.

Taking a first-order Taylor expansion of $\dot{g}_{\boldsymbol{\theta}}(\cdot;\widehat{\boldsymbol{\theta}})$ with respect to $\widehat{\boldsymbol{\theta}}$ around $\boldsymbol{\theta}_0$ gives

$$N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_0) + N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_t^*$ is an intermediate point between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ that is dependent on x_t. Post-multiplying the $q\times q$-dimensional $\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ matrix by $N_{T,\boldsymbol{\theta}_0}N_{T,\boldsymbol{\theta}_0}^{-1}$ yields

$$N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) = N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_0) + N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0}\widehat{\boldsymbol{\phi}}_T,$$

where $\widehat{\boldsymbol{\phi}}_T = N_{T,\boldsymbol{\theta}_0}^{-1}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ according to Notation 2.4. Substitute the above equation for $N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})$ in $\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\varepsilon_t$ and rearrange to get

$$\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})\varepsilon_t - \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_0)\varepsilon_t = \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0}\varepsilon_t\widehat{\boldsymbol{\phi}}_T. \tag{4.25}$$

Recall from *Assumption 2.3* that $\widehat{\boldsymbol{\phi}}_T$ is of order T^{τ}. We now proceed to show that $\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0}\varepsilon_t$ on the RHS converges to zero as $T\to\infty$.

First, note that under *Assumption 2.4*, x_t is adapted to the filtration $\mathscr{F}_{t-1}$, whereas ε_t is $\mathscr{F}_t$-adapted by *Assumption 2.1*. Then, invoking the martingale CLT in Lemma 4.1 gives

$$\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0}\varepsilon_t \xrightarrow{D} N\{\mathbf{0}, \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0}N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0}\}.$$

To complete the proof, it remains to show that the variance vanishes asymptotically so that this normal distribution is a degenerate one that has probability mass at zero being one. By the triangle inequality and norm compatibility, we have

$$\left\| \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0}N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0} \right\| \leq \sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta}_0}\ddot{g}_{\boldsymbol{\theta}\boldsymbol{\theta}}(x_t;\boldsymbol{\theta}_t^*)N_{T,\boldsymbol{\theta}_0} \right\|^2.$$

Using Lemma 2.3a, which holds under *Assumption 2.4*, we get

$$\sum_{t=p+1}^{T} \left\| N'_{T,\boldsymbol{\theta_0}} \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*) N_{T,\boldsymbol{\theta_0}} \right\|^2 = \mathbf{o_p}(T^{-4\tau}).$$

This implies that $\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \ddot{g}_{\boldsymbol{\theta\theta}}(x_t;\boldsymbol{\theta}_t^*) N_{T,\boldsymbol{\theta_0}} \varepsilon_t$ is also of order $\mathbf{o_p}(T^{-4\tau})$. Substituting above results back into equation (4.25) yields

$$\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \varepsilon_t - \sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\boldsymbol{\theta_0}) \varepsilon_t = \mathbf{o_p}(T^{\tau}) \cdot \mathbf{o_p}(T^{-4\tau}) = \mathbf{o_p}(T^{-3\tau}).$$

□

Recall from *Assumption 4.2* that

$$\mathbf{H_{u,t}} = \begin{bmatrix} \boldsymbol{\varepsilon_{t-1}} \\ \dot{g}_{u,t} \end{bmatrix}.$$

The next lemma is the key distributional results to prove the main theorems.

Lemma 4.7. *If Assumptions 2.1, 2.4, 4.2 and 4.3 are satisfied, then*

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \mathbf{H_{u,t}} \xrightarrow{D} \mathbf{N}\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_\varepsilon^2 \begin{bmatrix} \sigma_\varepsilon^2 \mathbf{I_p} & \mathbf{A}' \\ \mathbf{A} & \mathbf{M_u} \end{bmatrix} \}.$$

Proof of Lemma 4.7.

Notice that

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \mathbf{H_{u,t}} = T^{-\frac{1}{2}} \sum_{t=1}^{T} \varepsilon_t \mathbf{H_{u,t}} - T^{-\frac{1}{2}} \sum_{t=1}^{p} \varepsilon_t \mathbf{H_{u,t}}.$$

First, consider the second term on the RHS of the above equation. Note that for each t, the random variable ε_t and the random vector $\mathbf{H_{u,t}}$ are both of order 1. Thus, by Slutsky's theorem, $\varepsilon_t \mathbf{H_{u,t}}$ must also be stochastically bounded. Given that p is fixed

and finite, $\sum_{t=1}^{p} \varepsilon_t \mathbf{H_{u,t}}$ is $\mathbf{O_p}(1)$, and

$$T^{-\frac{1}{2}} \sum_{t=1}^{p} \varepsilon_t \mathbf{H_{u,t}} = T^{-\frac{1}{2}} \cdot \mathbf{O_p}(1) = \mathbf{O_p}(T^{-\frac{1}{2}}).$$

To complete the proof, we need to show that $T^{-\frac{1}{2}} \sum_{t=1}^{T} \varepsilon_t \mathbf{H_{u,t}}$ converges in distribution.

By *Assumption 2.4*, $\{x_t\}$ is $\mathscr{F}_{t-1}$-adapted. Therefore, the vectors $\dot{g}_{u,t}$ and $\mathbf{H_{u,t}}$ are also $\mathscr{F}_{t-1}$-adapted. Under *Assumption 2.1*, $\{\varepsilon_t\}$ is a martingale difference sequence and the conditional expectation $\mathbb{E}(\varepsilon_t|\mathscr{F}_{t-1})$ is zero. Thus, the vector $\sum_{t=1}^{T} \varepsilon_t \mathbf{H_{u,t}}$ is a martingale. This could be checked by verifying that the summand is a martingale difference sequence:

$$\mathbb{E}(\varepsilon_t \mathbf{H_{u,t}}|\mathscr{F}_{t-1}) = \mathbb{E}(\varepsilon_t|\mathscr{F}_{t-1})\mathbf{H_{u,t}} = \mathbf{0}.$$

Once we have verified the two conditions in *Lemma 4.1*, an appeal to the martingale CLT will complete the proof of this lemma. Yet, one should notice that a stricter proof would require the application of a multivariate version of the CLT for martingale difference sequence [see Theorem 3.3 in Helland (1982)] or the Cramer-Wold device [see Theorem 25.5 of Davidson (1994)]. The Cramer-Wold device paves the way for the derivation of the distribution of a random vector, which follows from the uniqueness of the characteristic function of the vector. In our case, the Cramer-Wold device suggests that $T^{-\frac{1}{2}} \sum_{t=1}^{T} \varepsilon_t \mathbf{H_{u,t_t}}$ is $p + d_u$-variate normal if any projection of $T^{-\frac{1}{2}} \sum_{t=1}^{T} \varepsilon_t \mathbf{H_{u,t}}$ onto a one-dimensional subspace is itself normal.

1. We begin by checking that *Lemma 4.1(1)* is satisfied. Consider the conditional variance $\mathbb{E}[\varepsilon_t^2 \mathbf{H_{u,t}} \mathbf{H_{u,t}}'|\mathscr{F}_{t-1}]$, and we now show that the sample conditional variance, $T^{-1} \sum_{t=1}^{T} \mathbb{E}[\varepsilon_t^2 \mathbf{H_{u,t}} \mathbf{H_{u,t}}'|\mathscr{F}_{t-1}]$, converges in probability to the population counterpart. Recall from *Assumption 2.1(b)* that $\mathbb{E}[\varepsilon_t^2|\mathscr{F}_{t-1}] = \sigma_\varepsilon^2 < \infty$, and from *Assumption 2.4*

that $\mathbf{H_{u,t}}$ is $\mathscr{F}_{t-1}$-adapted. To verify the first condition, note that

$$T^{-1}\sum_{t=1}^{T}\mathbb{E}[\varepsilon_t^2\mathbf{H_{u,t}}\mathbf{H_{u,t}}'|\mathscr{F}_{t-1}] = T^{-1}\sum_{t=1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}'\mathbb{E}[\varepsilon_t^2|\mathscr{F}_{t-1}] = \sigma_\varepsilon^2 T^{-1}\sum_{t=1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}'. \tag{4.26}$$

Consider the sample variance-covariance matrix $T^{-1}\sum_{t=1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}'$ and recall that $\mathbf{H_{u,t}} = (\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}', \dot{g}_{u,t}')'$. It can be shown that

$$T^{-1}\sum_{t=1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}' = \begin{bmatrix} T^{-1}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}' & T^{-1}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}\dot{g}_{u,t}' \\ T^{-1}\sum_{t=1}^{T}\dot{g}_{u,t}\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}' & T^{-1}\sum_{t=1}^{T}\dot{g}_{u,t}\dot{g}_{u,t}' \end{bmatrix}.$$

The diagonal elements of the matrix $T^{-1}\sum_{t=1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}'$ are just the sample variances of $\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}$ and $\dot{g}_{u,t}$. Under *Assumption 2.1*, we can apply the results in Lemma 4.3 to get $T^{-1}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}' \xrightarrow{p} \sigma_\varepsilon^2\mathbf{I_p}$. By *Assumption 4.2(a)* and noting that $N_{T,u} = T^{-\frac{1}{2}}\mathbf{I_{d_u}}$, $T^{-1}\sum_{t=1}^{T}\dot{g}_{u,t}\dot{g}_{u,t}'$ converges in probability to $\mathbf{M_u}$. The off-diagonal elements are covariance terms, whose probability limit is $\mathbf{A}$ by *Assumptions 4.2(b)*. So,

$$T^{-1}\sum_{t=1}^{T}\mathbf{H_{u,t}}\mathbf{H_{u,t}}' \xrightarrow{p} \begin{bmatrix} \sigma_\varepsilon^2\mathbf{I_p} & \mathbf{A}' \\ \mathbf{A} & \mathbf{M_u} \end{bmatrix}. \tag{4.27}$$

Combining equations (4.26) and (4.27) gives

$$T^{-1}\sum_{t=1}^{T}\mathbb{E}[\varepsilon_t^2\mathbf{H_{u,t}}\mathbf{H_{u,t}}'|\mathscr{F}_{t-1}] \xrightarrow{p} \sigma_\varepsilon^2\begin{bmatrix} \sigma_\varepsilon^2\mathbf{I_p} & \mathbf{A}' \\ \mathbf{A} & \mathbf{M_u} \end{bmatrix}.$$

and the first condition of *Lemma 4.1* is satisfied.

2. It therefore remains to check *Lemma 4.1(2)* that $T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}\mathbb{E}\{||\varepsilon_t\mathbf{H_{u,t}}||^{2+\delta}|\mathscr{F}_{t-1}\}$

converges in probability to zero for some $\delta > 0$. Consider

$$T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}\mathbb{E}\{||\varepsilon_t\mathbf{H_{u,t}}||^{2+\delta}|\mathscr{F}_{t-1}\} = T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}\mathbb{E}\{|\varepsilon_t|^{2+\delta}||\mathbf{H_{u,t}}||^{2+\delta}|\mathscr{F}_{t-1}\}$$
$$= T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}||\mathbf{H_{u,t}}||^{2+\delta}\mathbb{E}\{|\varepsilon_t|^{2+\delta}|\mathscr{F}_{t-1}\},$$

where the second equality follows from the fact that $\mathbf{H_{u,t}}$ is $\mathscr{F}_{t-1}$-adapted. Using the condition in *Assumption 2.1(c)* that $\sup_t \mathbb{E}[|\varepsilon_t|^{\varphi}|\mathscr{F}_{t-1}] < \infty$ for some $\varphi > 2$, we obtain

$$T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}\mathbb{E}\{||\varepsilon_t\mathbf{H_{u,t}}||^{2+\delta}|\mathscr{F}_{t-1}\} \leq \{\sup_t \mathbb{E}[|\varepsilon_t|^{2+\delta}|\mathscr{F}_{t-1}]\}\cdot T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}||\mathbf{H_{u,t}}||^{2+\delta}.$$

Let $C_1 = \sup_t \mathbb{E}[|\varepsilon_t|^{2+\delta}|\mathscr{F}_{t-1}]$. Take the expectation of the above equation and we may deduce that

$$\mathbb{E}|T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}\mathbb{E}\{||\varepsilon_t\mathbf{H_{u,t}}||^{2+\delta}|\mathscr{F}_{t-1}\}| \leq C_1\cdot T^{-(1+\frac{\delta}{2})}\sum_{t=1}^{T}\mathbb{E}||\mathbf{H_{u,t}}||^{2+\delta}$$
$$\leq C_1\cdot T^{-\frac{\delta}{2}}\sup_t \mathbb{E}||\mathbf{H_{u,t}}||^{2+\delta}.$$

The verification of *Lemma 4.1(2)* is complete once we show that $\sup_t \mathbb{E}||\mathbf{H_{u,t}}||^{2+\delta}$ is bounded in probability for some $\delta > 0$, since C_1 is $O_p(1)$ by *Assumption 2.1(c)* and $T^{-\frac{\delta}{2}}$ vanishes as T goes to infinity.

Recall that $\mathbf{H_{u,t}} = (\boldsymbol{\varepsilon'_{t-1}}, \dot{g}'_{u,t})'$. According to *Assumptions 2.1(c)* and *4.3*, we have both $\sup_t \mathbb{E}|\varepsilon_t|^{\varphi}$ and $\sup_t \mathbb{E}\|\dot{g}_{u,t}\|^{\varphi}$ bounded in probability for some φ greater than 2. It must then follow that $\sup_t \mathbb{E}||\mathbf{H_{u,t}}||^{2+\delta}$ is bounded for some $\delta > 0$.

Having verified the conditions given in *Lemma 4.1*, we can now apply the martingale CLT to $T^{-\frac{1}{2}}\sum_{t=1}^{T}\varepsilon_t\mathbf{H_{u,t}}$ to obtain

$$T^{-\frac{1}{2}}\sum_{t=1}^{T}\varepsilon_t\mathbf{H_{u,t}} \xrightarrow{D} \mathbf{N}\{\begin{bmatrix}0\\0\end{bmatrix}, \sigma_\varepsilon^2\begin{bmatrix}\sigma_\varepsilon^2\mathbf{I_p} & \mathbf{A}'\\ \mathbf{A} & \mathbf{M_u}\end{bmatrix}\}.$$

□

4.C Asymptotics of $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$

Having established some auxiliary lemmas regarding the key components of $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$, we are ready to derive its limiting distribution.

Recall from Notation 2.5 that

$$Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] = \sum_{t=p+1}^{T} \varepsilon_t [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' \{ \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' \}^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] \varepsilon_t.$$

The following lemma concerns the middle matrix of $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$.

Lemma 4.8. *By Assumptions 2.1, 2.3, 2.4 and 4.2,*

$$T^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' = T^{-1} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]' + \mathbf{o_p}(T^{-\tau}).$$

Proof of Lemma 4.8.

We need to prove that the difference between $T^{-1}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]'$ and $T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]'$ is small. Using Notation 2.6 and expanding the terms to get

$$T^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' = T^{-1} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} - T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} [\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t}]^{-1} T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}.$$

Pre-and post-multiplying the matrix $[\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}]^{-1}$ by $N_{T,u}N_{T,u}^{-1}$ and $(N'_{T,u})^{-1}N'_{T,u}$, respectively, gives

$$T^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' = T^{-1} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} - T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \dot{g}'_{\widehat{u},t} N_{T,u} [\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u}]^{-1} T^{-\frac{1}{2}} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}. \quad (4.28)$$

1. Consider the first term on the RHS of equation (4.28), $\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}$. Using *Lemma 4.2* , which holds under *Assumptions 2.1*, *2.3* and *2.4*, we have

$$T^{-1} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} = T^{-1} \sum_{t=p+1}^{T} \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} \boldsymbol{\varepsilon}'_{\boldsymbol{t-1}} + \mathbf{o_p}(T^{\tau-\frac{1}{2}}). \tag{4.29}$$

2. Consider the middle matrix in the second term, $\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u}$. By *Assumptions 2.3* and *2.4*, we could use the results in *Lemma 2.8* to get

$$\sum_{t=p+1}^{T} N'_{T,\boldsymbol{\theta_0}} \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}) \dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})' N_{T,\boldsymbol{\theta_0}} = \mathbf{S_{\dot{g}\dot{g}}} \Big\{ \mathbf{I_q} + \mathbf{o_p}(T^{-\tau}) \Big\}.$$

From Notation 3.1, this implies

$$\begin{bmatrix} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u} & \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{v},t} N_{T,v} \\ \sum_{t=p+1}^{T} N'_{T,v} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{u},t} N_{T,u} & \sum_{t=p+1}^{T} N'_{T,v} \dot{g}_{\widehat{v},t} \dot{g}'_{\widehat{v},t} N_{T,v} \end{bmatrix} = \begin{bmatrix} \mathbf{S_{\dot{g}_u \dot{g}_u}} & \mathbf{S_{\dot{g}_u \dot{g}_v}} \\ \mathbf{S_{\dot{g}_v \dot{g}_u}} & \mathbf{S_{\dot{g}_v \dot{g}_v}} \end{bmatrix} \cdot \Big\{ \mathbf{I_q} + \mathbf{o_p}(T^{-\tau}) \Big\}.$$

Thus,

$$\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u} = \mathbf{S_{\dot{g}_u \dot{g}_u}} \{ \mathbf{I_{d_u}} + \mathbf{o_p}(T^{-\tau}) \}. \tag{4.30}$$

3. Consider the vector $\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}}$ in equation (4.28). By *Lemma 4.5*, which holds under *Assumptions 2.1*, *2.3* and *2.4*,

$$\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{t-1}} = \mathbf{S_{\dot{g}_u \varepsilon}} + \mathbf{o_p}(T^{\frac{1}{2}-\tau}), \tag{4.31}$$

where $\mathbf{S_{\dot{g}_u \varepsilon}} = \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \boldsymbol{\varepsilon}'_{\boldsymbol{t-1}}$ by Notation 4.3.

Substituting equations (4.29) to (4.31) back into equation (4.28) and rearranging gives

$$\begin{aligned} T^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}]' &= T^{-1} \sum_{t=p+1}^{T} \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} \boldsymbol{\varepsilon}'_{\boldsymbol{t-1}} + \mathbf{o_p}(T^{\tau-\frac{1}{2}}) \\ - \{ T^{-\frac{1}{2}} \mathbf{S_{\dot{g}_u \varepsilon}} + \mathbf{o_p}(T^{-\tau}) \}' \{ \mathbf{S_{\dot{g}_u \dot{g}_u}} [\mathbf{I_{d_u}} + \mathbf{o_p}(T^{-\tau})] \}^{-1} \{ T^{-\frac{1}{2}} \mathbf{S_{\dot{g}_u \varepsilon}} + \mathbf{o_p}(T^{-\tau}) \}. \end{aligned} \tag{4.32}$$

Recall from Notation 3.1 that $\mathbf{S_{\dot{g}_u\dot{g}_u}} = \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u}$. By *Assumption 4.2(a)*, the inverse of $\mathbf{S_{\dot{g}_u\dot{g}_u}}$ exists and is bounded, so $\mathbf{S^{-1}_{\dot{g}_u\dot{g}_u}}$ is $\mathbf{O_p}(1)$. From *Assumption 4.2(b)*, $T^{-\frac{1}{2}}\mathbf{S_{\dot{g}_u\varepsilon}}$ is also $\mathbf{O_p}(1)$. According to *Assumption 2.3*, we have $0 < \tau < \frac{1}{4}$. Substitute these results back into equation (4.32) and rearrange to get

$$T^{-1}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' = T^{-1}\sum_{t=p+1}^{T}\boldsymbol{\varepsilon_{t-1}\varepsilon'_{t-1}} - T^{-1}\mathbf{S'_{\dot{g}_u\varepsilon}}\mathbf{S^{-1}_{\dot{g}_u\dot{g}_u}}\mathbf{S_{\dot{g}_u\varepsilon}} + \mathbf{o_p}(T^{-\tau}). \tag{4.33}$$

Observe that the sum of the first two terms on the RHS is exactly $T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]'$ by Notation 2.6, giving

$$T^{-1}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' - T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]' = \mathbf{o_p}(T^{-\tau}).$$

□

The next lemma establishes the probability limit of the middle matrix of $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$.

Lemma 4.9. *By Assumptions 2.1, 4.1 and 4.2,*

$$\{T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]'\}^{-1} \xrightarrow{p} (\sigma_\varepsilon^2\mathbf{I_p} - \mathbf{A'M^{-1}_u A})^{-1}.$$

Proof of Lemma 4.9.

Using Notation 2.6,

$$T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]' = T^{-1}\sum_{t=p+1}^{T}\boldsymbol{\varepsilon_{t-1}\varepsilon'_{t-1}} - T^{-\frac{1}{2}}\mathbf{S'_{\dot{g}_u\varepsilon}}\mathbf{S^{-1}_{\dot{g}_u\dot{g}_u}}\mathbf{T}^{-\frac{1}{2}}\mathbf{S_{\dot{g}_u\varepsilon}},$$

where $\mathbf{S_{\dot{g}_u\varepsilon}} = \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\boldsymbol{\varepsilon'_{t-1}}$ and $\mathbf{S^{-1}_{\dot{g}_u\dot{g}_u}} = \{\sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u}\}^{-1}$.

Lemma 4.3 holds under *Assumption 2.1*, so $T^{-1}\sum_{t=p+1}^{T}\boldsymbol{\varepsilon_{t-1}\varepsilon'_{t-1}}$, which is the first term on the RHS, converges in probability to $\sigma_\varepsilon^2\mathbf{I_p}$.

Next, consider the second term on the RHS. By *Assumption 4.2*, the probability limits of $\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}^{-\mathbf{1}}$ and $T^{-\frac{1}{2}}\mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\boldsymbol{\varepsilon}}$ are $\mathbf{M}_\mathbf{u}^{-\mathbf{1}}$ and $\mathbf{A}'$, respectively. Then, by Slutsky's Theorem,

$$T^{-\frac{1}{2}}\mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\boldsymbol{\varepsilon}}\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}^{-\mathbf{1}}\mathbf{T}^{-\frac{1}{2}}\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\boldsymbol{\varepsilon}} \xrightarrow{p} \mathbf{A}'\mathbf{M}_\mathbf{u}^{-\mathbf{1}}\mathbf{A}.$$

Combining these results gives

$$T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]' \xrightarrow{p} \sigma_\varepsilon^2\mathbf{I_p} - \mathbf{A}'\mathbf{M}_\mathbf{u}^{-\mathbf{1}}\mathbf{A}.$$

Assumptions 4.1 and *4.2* ensures the invertibility of the matrix $\sigma_\varepsilon^2\mathbf{I_p} - \mathbf{A}'\mathbf{M}_\mathbf{u}^{-\mathbf{1}}\mathbf{A}$, leading to

$$\{T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]'\}^{-1} \xrightarrow{p} (\sigma_\varepsilon^2\mathbf{I_p} - \mathbf{A}'\mathbf{M}_\mathbf{u}^{-\mathbf{1}}\mathbf{A})^{-1}.$$

□

We next show that the vector component of $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$, $\sum_{t=p+1}^{T}\varepsilon_t[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]$, when normalised by $T^{-\frac{1}{2}}$, is asymptotically equivalent to $\sum_{t=p+1}^{T}\varepsilon_t[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]$.

Lemma 4.10. *If Assumptions 2.1, 2.3, 2.4, 4.2 and 4.3 are satisfied, then*

$$T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\varepsilon_t[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] = T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\varepsilon_t[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}] + \mathbf{o_p}(T^{-\tau}).$$

Proof of Lemma 4.10.

Using Notation 2.6, we can obtain

$$\begin{aligned}T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\varepsilon_t[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] = T^{-\frac{1}{2}}&\sum_{t=p+1}^{T}\varepsilon_t\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} \\ &- T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}\dot{g}'_{\widehat{u},t}\{\sum_{t=p+1}^{T}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}\}^{-1}\sum_{t=p+1}^{T}\dot{g}_{\widehat{u},t}\varepsilon_t. \quad (4.34)\end{aligned}$$

1. Consider the first term in equation (4.34). Under *Assumptions 2.1*, *2.3* and *2.4*,

Lemma 4.4 implies that

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}} = T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \boldsymbol{\varepsilon}_{\mathbf{t-1}} + \mathbf{o_p}(1). \tag{4.35}$$

2. We now show that the second term in equation (4.34) is asymptotically equivalent to $T^{-\frac{1}{2}} \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} \mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \varepsilon_t$. Pre- and post-multiplying the middle matrix, $\{\sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t}\}^{-1}$ by $N_{T,u} N_{T,u}^{-1}$ and its transpose, respectively, gives

$$\begin{aligned} T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}} \dot{g}'_{\widehat{u},t} \{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \}^{-1} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \varepsilon_t \\ = T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}} \dot{g}'_{\widehat{u},t} N_{T,u} \{ \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} N_{T,u} \}^{-1} \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \varepsilon_t. \end{aligned}$$

2.a) From *Lemma 2.8*, which holds under *Assumptions 2.3* and *2.4*, we have

$$\{\sum_{t=1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \dot{g}_{\widehat{u},t} N_{T,u}\}^{-1} = \{\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} \Big\{ \mathbf{I_{d_u}} + \mathbf{o_p}(T^{-\tau}) \Big\}\}^{-1} = \Big\{ \mathbf{I_{d_u}} + \mathbf{o_p}(T^{-\tau}) \Big\} \mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}. \tag{4.36}$$

2.b) Using *Lemma 4.5*, which is satisfied under *Assumptions 2.1*, *2.3* and *2.4*, we get

$$\sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}} \dot{g}'_{\widehat{u},t} N_{T,u} = \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} + \mathbf{o_p}(T^{\frac{1}{2}-\tau}). \tag{4.37}$$

2.c) Now consider the column vector, $\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \varepsilon_t$. Under *Assumptions 2.1*, *2.3* and *2.4*, we can apply *Lemma 4.6* to obtain

$$\sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{\widehat{u},t} \varepsilon_t = \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \varepsilon_t + \mathbf{o_p}(T^{-3\tau}). \tag{4.38}$$

Combining equations (4.36) to (4.38) yields

$$\begin{aligned} & T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}} \dot{g}'_{\widehat{u},t} \{ \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} \}^{-1} \sum_{t=p+1}^{T} \dot{g}_{\widehat{u},t} \varepsilon_t \\ = & \{ T^{-\frac{1}{2}} \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} + \mathbf{o_p}(T^{-\tau}) \} \Big\{ \mathbf{I_{d_u}} + \mathbf{o_p}(T^{-\tau}) \Big\} \mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} \{ \sum_{t=p+1}^{T} N'_{T,u} \dot{g}_{u,t} \varepsilon_t + \mathbf{o_p}(T^{-3\tau}) \}. \end{aligned}$$

Both $\mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}$ and $T^{-\frac{1}{2}}\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon}$ are $\mathbf{O_p}(1)$ following from *Assumptions 4.2*. Furthermore, under *Assumptions 2.1, 2.4, 4.2* and *4.3*, we could invoke the results in *Lemma 4.7* so that $T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\varepsilon_t\mathbf{H_{u,t}}$ follows a multivariate normal distribution, where $\mathbf{H_{u,t}} = (\boldsymbol{\varepsilon}'_{\mathbf{t-1}}, \dot{g}'_{u,t})'$. Using the property that any marginal distribution of a multivariate normal joint distribution is also multivariate normal and noting that $N_{T,u} = T^{-\frac{1}{2}}\mathbf{I_{d_u}}$, we get $\sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\varepsilon_t = \mathbf{O_p}(1)$.

It then follows that

$$T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}\dot{g}'_{\widehat{u},t}\{\sum_{t=p+1}^{T}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}\}^{-1}\sum_{t=p+1}^{T}\dot{g}_{\widehat{u},t}\varepsilon_t = \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon}\mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}\sum_{t=p+1}^{T}N'_{T,u}\dot{g}_{u,t}\varepsilon_t + \mathbf{o_p}(T^{-\tau}). \tag{4.39}$$

The proof is completed by substituting equations (4.35) and (4.39) back into equation (4.34) and recognising that

$$T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\varepsilon_t[\boldsymbol{\varepsilon}_{\mathbf{t-1}}|\dot{g}_{u,t}] = T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\varepsilon_t\boldsymbol{\varepsilon}_{\mathbf{t-1}} - \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon}\mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}\sum_{t=p+1}^{T}N'_{T,u}\dot{g}_{u,t}\varepsilon_t.$$

□

From Theorem 3.1, the asymptotic properties of the numerator of the test statistic, TR^2, depend entirely on $Q[\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}|\dot{g}_{\widehat{u},t}]$, whose limiting distribution is given in the following lemma.

Lemma 4.11. *If Assumptions 2.1, 2.3, 2.4, 4.1 to 4.3 are satisfied, then*

$$Q[\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}|\dot{g}_{\widehat{u},t}] = Q[\boldsymbol{\varepsilon}_{\mathbf{t-1}}|\dot{g}_{u,t}] + o_p(1) \xrightarrow{D} \sigma^2_\varepsilon\chi^2_p.$$

Proof of Lemma 4.11.

The proof falls naturally into two steps. The first step is to demonstrate that $Q[\boldsymbol{\varepsilon}_{\mathbf{t-1}}|\dot{g}_{u,t}]$ converges in distribution to $\sigma^2_\varepsilon\chi^2_p$. And the second step is to show that

$$Q[\widehat{\boldsymbol{\varepsilon}}_{\mathbf{t-1}}|\dot{g}_{\widehat{u},t}] = Q[\boldsymbol{\varepsilon}_{\mathbf{t-1}}|\dot{g}_{u,t}] + o_p(1), \tag{4.40}$$

so that $Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] \xrightarrow{D} \sigma_\varepsilon^2 \chi_p^2$.

1. Step 1

Consider the quantity $Q[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]$. By Notation 2.5,

$$Q[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}] = T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]' \{T^{-1} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]'\}^{-1} \cdot T^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]\varepsilon_t. \quad (4.41)$$

First, consider the middle matrix. Using *Lemma 4.9*, which holds under *Assumptions 2.1*, *4.1* and *4.2*, we have

$$\{T^{-1} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]'\}^{-1} \xrightarrow{p} (\sigma_\varepsilon^2 \mathbf{I_p} - \mathbf{A'M_u^{-1}A})^{-1}. \quad (4.42)$$

Now take the vector component $T^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]\varepsilon_t$. Following from Notation 2.6, we have

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \left[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}\right] = T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \boldsymbol{\varepsilon}_{\boldsymbol{t-1}} - T^{-\frac{1}{2}} \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} \mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \dot{g}_{u,t}\varepsilon_t,$$

where $\mathbf{S}_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} = \sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\boldsymbol{\varepsilon}'_{\boldsymbol{t-1}}$ by Notation 4.3, $\mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} = \{\sum_{t=p+1}^{T} N'_{T,u}\dot{g}_{u,t}\dot{g}'_{u,t}N_{T,u}\}^{-1}$ by Notation 2.3 and note that $N_{T,u} = T^{-\frac{1}{2}}\mathbf{I_{d_u}}$. Further rearranging gives

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \left[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}\right] = \left[\mathbf{I_p}, \quad -T^{-\frac{1}{2}} \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} \mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}\right] T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \mathbf{H_{u,t}}, \quad (4.43)$$

where $\mathbf{H_{u,t}} = (\boldsymbol{\varepsilon}'_{\boldsymbol{t-1}}, \dot{g}'_{u,t})'$.

We begin by analysing the $p \times (p + d_u)$ matrix $[\mathbf{I_p}, -T^{-\frac{1}{2}} \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} \mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}]$. By *Assumption 4.2*, $\mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}} \xrightarrow{p} \mathbf{M_u^{-1}}$ and $T^{-\frac{1}{2}} \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} \xrightarrow{p} \mathbf{A}'$. Therefore we have

$$\left[\mathbf{I_p}, \quad -T^{-\frac{1}{2}} \mathbf{S}'_{\dot{\mathbf{g}}_\mathbf{u}\varepsilon} \mathbf{S}^{-\mathbf{1}}_{\dot{\mathbf{g}}_\mathbf{u}\dot{\mathbf{g}}_\mathbf{u}}\right] \xrightarrow{p} \left[\mathbf{I_p}, \quad -\mathbf{A'M_u^{-1}}\right]. \quad (4.44)$$

Consider the remaining terms in equation (4.43). Using *Lemma 4.7*, which applies

under *Assumptions 2.1, 2.4, 4.2* and *4.3*, we obtain

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \mathbf{H_{u,t}} \xrightarrow{D} \mathbf{N}_{p+d_u}\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \sigma_\varepsilon^2 \begin{bmatrix} \sigma_\varepsilon^2 \mathbf{I_p} & \mathbf{A}' \\ \mathbf{A} & \mathbf{M_u} \end{bmatrix} \}. \tag{4.45}$$

Combine equations (4.44) and (4.45) together with Slutsky's Theorem to get

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \left[\boldsymbol{\varepsilon_{t-1}} | \dot{g}_{u,t} \right] \xrightarrow{D} \left[\mathbf{I_p}, \quad -\mathbf{A}'\mathbf{M_u^{-1}} \right] \mathbf{N}_{p+d_u}\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \sigma_\varepsilon^2 \begin{bmatrix} \sigma_\varepsilon^2 \mathbf{I_p} & \mathbf{A}' \\ \mathbf{A} & \mathbf{M_u} \end{bmatrix} \}.$$

Rearranging yields

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t \left[\boldsymbol{\varepsilon_{t-1}} | \dot{g}_{u,t} \right] \xrightarrow{D} N_p[\mathbf{0}, \sigma_\varepsilon^2(\sigma_\varepsilon^2 \mathbf{I_p} - \mathbf{A}'\mathbf{M_u^{-1}}\mathbf{A})]. \tag{4.46}$$

Substitute equations (4.42) and (4.46) back into equation (4.41), and invoke Slutsky's Theorem to obtain

$$Q[\boldsymbol{\varepsilon_{t-1}} | \dot{g}_{u,t}] \xrightarrow{D} N_p[\mathbf{0}, \sigma_\varepsilon^2(\sigma_\varepsilon^2 \mathbf{I_p} - \mathbf{A}'\mathbf{M_u^{-1}}\mathbf{A})]' \cdot (\sigma_\varepsilon^2 \mathbf{I_p} - \mathbf{A}'\mathbf{M_u^{-1}}\mathbf{A})^{-1} \cdot N_p[\mathbf{0}, \sigma_\varepsilon^2(\sigma_\varepsilon^2 \mathbf{I_p} - \mathbf{A}'\mathbf{M_u^{-1}}\mathbf{A})],$$

which could be further rearranged as

$$Q[\boldsymbol{\varepsilon_{t-1}} | \dot{g}_{u,t}] \xrightarrow{D} \{N_p\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I_p}\right)\}' \{N_p\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I_p}\right)\} = \sigma_\varepsilon^2 \chi_p^2.$$

2. Step 2

The task is now to demonstrate equation (4.40). From Notation 2.5

$$Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] = T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}]' \{T^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}]'\}^{-1} \cdot T^{-\frac{1}{2}} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}} | \dot{g}_{\widehat{u},t}] \varepsilon_t. \tag{4.47}$$

Lemma 4.10 applies under *Assumptions 2.1*, *2.3*, *2.4*, *4.2* and *4.3*, and we can obtain

$$T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] = T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}] + \mathbf{o_p}(T^{-\tau}). \tag{4.48}$$

Under *Assumptions 2.1*, *2.3*, *2.4* and *4.2*, *Lemma 4.8* shows that

$$T^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' = T^{-1} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]' + \mathbf{o_p}(T^{-\tau}).$$

In addition, under *Assumptions 2.1*, *4.1* and *4.2*, *Lemma 4.9* indicates that the inverse of $T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]'$ exists and is stochastically bounded. It then follows

$$T^{-1} \sum_{t=p+1}^{T} [\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}]' = T^{-1} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]'\{\mathbf{I_p} + \mathbf{o_p}(T^{-\tau})\}. \tag{4.49}$$

Substitute equations (4.48) and (4.49) back into equation (4.47) to get

$$\begin{aligned} Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] = \{T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}] + \mathbf{o_p}(T^{-\tau})\}'\{T^{-1} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]' \\ \cdot \{\mathbf{I_p} + \mathbf{o_p}(T^{-\tau})\}\}^{-1}\{T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}] + \mathbf{o_p}(T^{-\tau})\}. \end{aligned} \tag{4.50}$$

Since $\{T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]'\}^{-1}$ is $\mathbf{O_p}(1)$, and equation (4.46) in *Step 1* shows that $T^{-\frac{1}{2}}\sum_{t=1}^{T}\varepsilon_t[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]$ is also $\mathbf{O_p}(1)$. Rearranging equation (4.50) gives

$$\begin{aligned} Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\widehat{u},t}] = \{T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]\}'\{T^{-1} \sum_{t=p+1}^{T} [\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]'\}^{-1} \\ \cdot \{T^{-\frac{1}{2}} \sum_{t=p+1}^{T} \varepsilon_t[\boldsymbol{\varepsilon}_{\boldsymbol{t-1}}|\dot{g}_{u,t}]\} + \mathbf{o_p}(T^{-\tau}). \end{aligned}$$

Observing that the first term on the RHS

$$\{T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\varepsilon_t[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]\}'\{T^{-1}\sum_{t=p+1}^{T}[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}][\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]'\}^{-1}$$
$$\cdot\{T^{-\frac{1}{2}}\sum_{t=p+1}^{T}\varepsilon_t[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]\} = Q[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}]$$

gives

$$Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\hat{u},t}] = Q[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}] + \mathbf{o_p}(1) \xrightarrow{D} \sigma_\varepsilon^2\chi_p^2.$$

□

4.D Proof of Theorem 4.1

By Theorem 3.1, which applies under *Assumptions 2.1*, *2.3* to *2.5*, *3.1* to *3.3*, the test statistic takes the form

$$TR^2 = \frac{Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\hat{u},t}]}{T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2} + o_p(1),$$

Following from Lemma 4.11, which is satisfied under *Assumptions 2.1*, *2.3*, *2.4*, *4.1* to *4.3*, we have for the numerator

$$Q[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\hat{u},t}] = Q[\boldsymbol{\varepsilon_{t-1}}|\dot{g}_{u,t}] + o_p(1) \xrightarrow{D} \sigma_\varepsilon^2\chi_p^2.$$

Using Lemma 2.7, which holds under *Assumption 2.1*, we have for the denominator

$$T^{-1}\sum_{t=p+1}^{T}\varepsilon_t^2 \xrightarrow{p} \sigma_\varepsilon^2.$$

Since $\sigma_\varepsilon^2 > 0$ by *Assumption 2.1*, invoking Slutsky's Theorem yields

$$TR^2 \xrightarrow{D} \frac{\sigma_\varepsilon^2\chi_p^2}{\sigma_\varepsilon^2} = \chi_p^2.$$

□

4.E Equivalence of the Technical Assumptions — Assumption 2.5 and Assumption 3.3(a)

We now show that *Assumption 2.5* is implied by *Assumptions 2.4*, *3.2*, *3.3* and *4.4* jointly. The key idea is that *Assumption 2.5* is equivalent to *Assumption 3.3(a)* under *Assumptions 2.4*, *3.2*, *3.3(b)* and *4.4*.

Lemma 4.12. *If Assumptions 2.4, 3.2, 3.3 and 4.4 are satisfied, then*

$$\{\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})][\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]'\}^{-\frac{1}{2}}\sum_{t=p+1}^{T}[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]\varepsilon_t=\mathbf{O_p}(1).$$

Proof of Lemma 4.12.

By Notation 2.6,

$$[\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}|\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})]=\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}-\sum_{s=p+1}^{T}\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{s-1}}\dot{g}_{\boldsymbol{\theta}}(x_s;\widehat{\boldsymbol{\theta}})'[\sum_{s=p+1}^{T}\dot{g}_{\boldsymbol{\theta}}(x_s;\widehat{\boldsymbol{\theta}})\dot{g}_{\boldsymbol{\theta}}(x_s;\widehat{\boldsymbol{\theta}})']^{-1}\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}}).$$

We provide the proof here for the case when $p=1$.

Recall that $\dot{g}_{\boldsymbol{\theta}}(x_t;\widehat{\boldsymbol{\theta}})=(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})'$. The aim is to demonstrate

$$\frac{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']\varepsilon_t}{\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2\}^{\frac{1}{2}}}\approx\frac{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t}{\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{\frac{1}{2}}},\tag{4.51}$$

where the RHS is $O_p(1)$ by *Assumption 3.3(a)*.

We begin by evaluating the denominator and numerator on the LHS of the above equation separately.

1. Consider the denominator. By definition,

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2=\sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}^2-\sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}'\{\sum_{t=2}^{T}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}'\}^{-1}\sum_{t=2}^{T}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}\widehat{\varepsilon}_{t-1}.\tag{4.52}$$

We first show we can write $\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2$ as

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2=\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2-\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t}\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\cdot\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}], \quad (4.53)$$

and then illustrate that the second term on the RHS is small so the above equation is asymptotically equivalent to the following

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2=\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\{1+\mathbf{o_p}(1)\}.$$

(a) Let

$$\mathbf{L}=\begin{bmatrix}\mathbf{I_{d_u}} & \mathbf{0}\\ -\sum_{s=2}^{T}\dot{g}_{\widehat{v},s}\dot{g}'_{\widehat{u},s}(\sum_{s=2}^{T}\dot{g}_{\widehat{u},s}\dot{g}'_{\widehat{u},s})^{-1} & \mathbf{I_{d_v}}\end{bmatrix},$$

where the lower off-diagonal element, $\sum_{s=2}^{T}\dot{g}_{\widehat{v},s}\dot{g}'_{\widehat{u},s}(\sum_{s=2}^{T}\dot{g}_{\widehat{u},s}\dot{g}'_{\widehat{u},s})^{-1}$, is the estimate of the slope parameter from a simple regression of $\dot{g}_{\widehat{v},t}$ on $\dot{g}_{\widehat{u},t}$. So we have

$$\mathbf{L}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}=\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}\end{bmatrix}, \quad (4.54)$$

Pre- and post-multiplying the middle matrix in the second term of the RHS of equation (4.52) with $\mathbf{L}'[\mathbf{L}']^{-1}$ and its transpose, respectively, gives

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2=\sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}^2-\sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}'\mathbf{L}'\cdot\{\sum_{t=2}^{T}\mathbf{L}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}'\mathbf{L}'\}^{-1}\sum_{t=2}^{T}\mathbf{L}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}\widehat{\varepsilon}_{t-1}. \quad (4.55)$$

Observe that

$$\sum_{t=2}^{T} \mathbf{L} \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix} \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix}' \mathbf{L}' = \sum_{t=2}^{T} \begin{bmatrix} \dot{g}_{\widehat{u},t}, \dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t} \end{bmatrix} \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t} \end{bmatrix}',$$

which could be further rearranged as

$$\sum_{t=2}^{T} \mathbf{L} \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix} \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix}' \mathbf{L}' = \begin{bmatrix} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} & \sum_{t=2}^{T} \dot{g}_{\widehat{u},t} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}]' \\ \sum_{t=2}^{T} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}] \dot{g}'_{\widehat{u},t} & \sum_{t=2}^{T} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}] [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}]' \end{bmatrix}.$$

Recognise that the off-diagonal terms, $\sum_{t=2}^{T} \dot{g}_{\widehat{u},t} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}]'$, are zeros, since the residuals from the regression of $\dot{g}_{\widehat{v},t}$ on $\dot{g}_{\widehat{u},t}$ are orthogonal to $\dot{g}_{\widehat{u},t}$. Thus, the matrix becomes

$$\sum_{t=2}^{T} \mathbf{L} \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix} \begin{bmatrix} \dot{g}_{\widehat{u},t} \\ \dot{g}_{\widehat{v},t} \end{bmatrix}' \mathbf{L}' = \begin{bmatrix} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t} & \mathbf{0} \\ \mathbf{0} & \sum_{t=2}^{T} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}] [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}]' \end{bmatrix}. \quad (4.56)$$

Substituting equations (4.54) and (4.56) back into equation (4.55), and rearranging yields

$$\begin{aligned} \sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1} | (\dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})']^2 = \sum_{t=2}^{T} \widehat{\varepsilon}_{t-1}^2 - \sum_{t=2}^{T} \widehat{\varepsilon}_{t-1} \dot{g}'_{\widehat{u},t} (\sum_{t=2}^{T} \dot{g}_{\widehat{u},t} \dot{g}'_{\widehat{u},t})^{-1} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t} \widehat{\varepsilon}_{t-1} \\ - \sum_{t=2}^{T} \widehat{\varepsilon}_{t-1} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}]' \{ \sum_{t=2}^{T} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}] [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}]' \}^{-1} \sum_{t=2}^{T} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}] \widehat{\varepsilon}_{t-1}. \end{aligned}$$

Combining the first two terms on the RHS of the above equation gives $\sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1} | \dot{g}_{\widehat{u},t}]^2$. Notice that the vector $\sum_{t=2}^{T} [\dot{g}_{\widehat{v},t} | \dot{g}_{\widehat{u},t}] \widehat{\varepsilon}_{t-1}$ in the last term is equivalent to $\sum_{t=2}^{T} \dot{g}_{\widehat{v},t} [\widehat{\varepsilon}_{t-1} | \dot{g}_{\widehat{u},t}]$. Therefore we can rewrite the above equation

as

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2 = \sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2 - \sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t}\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}], \quad (4.57)$$

which is exactly the same as equation (4.53).

(b) We now show that the second term on the RHS of equation (4.57) vanishes asymptotically. Pre- and post-multiplying the middle matrix by the matrix $\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}$ and its transpose, respectively, gives

$$\begin{aligned}
&\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t}\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}] \\
&= \sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1} \\
&\cdot\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}}\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]. \quad (4.58)
\end{aligned}$$

Let

$$\mathscr{A}_{1,T} := \{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}}\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}],$$

and

$$\mathscr{A}_{2,T} := \{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}.$$

So we can restate equation (4.58) as

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t}\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}] = \mathscr{A}'_{1,T}\mathscr{A}_{2,T}\mathscr{A}_{1,T}, \quad (4.59)$$

and equation (4.57) can be rewritten as

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2 = \sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2 - \mathscr{A}'_{1,T}\mathscr{A}_{2,T}\mathscr{A}_{1,T}. \quad (4.60)$$

We now show that $\mathscr{A}'_{1,T}\mathscr{A}_{2,T}\mathscr{A}_{1,T}$ is small. First, consider the $\mathscr{A}_{2,T}$ term. Rearranging gives

$$\mathscr{A}_{2,T} = \{(\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t})^{-\frac{1}{2}} \sum_{t=2}^{T} [\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'(\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t})^{-\frac{1}{2}}\}^{-1}$$

Using Notation 2.6, we can expand the matrix $\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'$ to get

$$\mathscr{A}_{2,T} = \{(\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t})^{-\frac{1}{2}}[\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t} - \sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{u},t}(\sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t})^{-1} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{v},t}] \cdot (\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t})^{-\frac{1}{2}}\}^{-1}.$$

Multiplying out brackets and further simplifying yields

$$\mathscr{A}_{2,T} = \{\mathbf{I}_{\mathbf{d_v}} - \mathscr{B}'_T\mathscr{B}_T\}^{-1},$$

where $\mathscr{B}_T := (\sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t})^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{v},t}(\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t})^{-\frac{1}{2}}$. By *Assumption 3.2*, $\mathscr{B}_T$ is $\mathbf{O_p}(T^{\psi-\frac{1}{2}})$, where $0 \leq \psi < \frac{1}{2}$, and we can obtain

$$\mathscr{A}_{2,T} = \{\mathbf{I}_{\mathbf{d_v}} - \mathscr{B}'_T\mathscr{B}_T\}^{-1} = \{\mathbf{I}_{\mathbf{d_v}} + \mathbf{O_p}(T^{2\psi-1})\}^{-1} = \mathbf{I}_{\mathbf{d_v}} + \mathbf{O_p}(T^{2\psi-1}). \tag{4.61}$$

Substituting the above equation back into equation (4.59) gives

$$\mathscr{A}'_{1,T}\mathscr{A}_{2,T}\mathscr{A}_{1,T} = \mathscr{A}'_{1,T}\mathscr{A}_{1,T}\{1 + O_p(T^{2\psi-1})\}.$$

Next, pre- and post-multiply the RHS of the above equation by $\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{\frac{1}{2}}\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{-\frac{1}{2}}$ and its transpose, respectively, to get

$$\mathscr{A}'_{1,T}\mathscr{A}_{2,T}\mathscr{A}_{1,T} = \{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{\frac{1}{2}}\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{-\frac{1}{2}}\mathscr{A}'_{1,T}\mathscr{A}_{1,T} \cdot \{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{-\frac{1}{2}}\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{\frac{1}{2}}\{1 + O_p(T^{2\psi-1})\}. \tag{4.62}$$

Recall that

$$\mathscr{A}_{1,T} = \{\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}].$$

By *Assumption 3.3(b)*, we have $\mathscr{A}_{1,T}\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{-\frac{1}{2}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}})$. Plugging this result back into equation (4.62) and rearranging gives

$$\mathscr{A}'_{1,T}\mathscr{A}_{2,T}\mathscr{A}_{1,T} = \sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2 \cdot O_p(T^{2\psi-1}).$$

Substituting the above equation back into equation (4.60) and rearranging yields

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})']^2 = \sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\{1 + O_p(T^{2\psi-1})\}. \tag{4.63}$$

2. Now take the numerator in equation (4.51). By definition,

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})']\varepsilon_t = \sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}\varepsilon_t - \sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}'\{\sum_{t=2}^{T}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}'\}^{-1}\sum_{t=2}^{T}\begin{bmatrix}\dot{g}_{\widehat{u},t}\\ \dot{g}_{\widehat{v},t}\end{bmatrix}\varepsilon_t.$$

We first show that this is equivalently to

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})']\varepsilon_t = \sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t$$
$$- \sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]\varepsilon_t,$$

and then illustrate that the second term on the RHS is of order $T^{\psi-\frac{1}{2}} \cdot \log T$.

(a) Using the same type of manipulation as in the analysis of the denominator, we

can obtain

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']\varepsilon_t = \sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}\varepsilon_t - \sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}\dot{g}'_{\widehat{u},t}(\sum_{t=2}^{T}\dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t})^{-1}\sum_{t=2}^{T}\dot{g}_{\widehat{u},t}\varepsilon_t$$
$$-\sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]\varepsilon_t.$$

Noting that the first two terms on the RHS of the above equation are exactly$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t$, we can rewrite the above equation as

$$\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']\varepsilon_t = \sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t$$
$$-\sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]\varepsilon_t. \quad (4.64)$$

(b) Take the second term on the RHS of the above equation, and we now illustrate that this is of order $T^{\psi-\frac{1}{2}}\cdot\log T$. Pre- and post-multiplying the middle matrix by $\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}$ and its transpose, respectively, results in

$$\sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]\varepsilon_t$$
$$=\sum_{t=2}^{T}\widehat{\varepsilon}_{t-1}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}$$
$$\cdot\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}}\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]\varepsilon_t. \quad (4.65)$$

Recall that

$$\mathscr{A}_{1,T} = \{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}}\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}],$$

where $\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]$ is equivalent to $\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]\widehat{\varepsilon}_{t-1}$, and

$$\mathscr{A}_{2,T} = \{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}\{\sum_{t=2}^{T}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1}\{\sum_{t=2}^{T}\dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{\frac{1}{2}}.$$

Let

$$\mathscr{A}_{3,T} := \{\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}} \sum_{t=2}^{T} [\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]\varepsilon_t.$$

Thus we can rewrite equation (4.65) as

$$\sum_{t=2}^{T} \widehat{\varepsilon}_{t-1}[\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\{\sum_{t=2}^{T} [\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}][\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]'\}^{-1} \sum_{t=2}^{T} [\dot{g}_{\widehat{v},t}|\dot{g}_{\widehat{u},t}]\varepsilon_t = \mathscr{A}'_{1,T}\mathscr{A}_{2,T}\mathscr{A}_{3,T},$$

and equation (4.64) becomes

$$\sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})']\varepsilon_t = \sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t - \mathscr{A}'_{1,T}\mathscr{A}_{2,T}\mathscr{A}_{3,T}.$$

The goal is to show that the second term on the RHS is $O_p(T^{\psi-\frac{1}{2}} \cdot \log T)$.

Recall from equation (4.61) that $\mathscr{A}_{2,T} = \mathbf{I_{d_v}} + \mathbf{O_p}(T^{2\psi-1})$. Plug this result into the above equation and rearrange to get

$$\sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})']\varepsilon_t = \sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t - \mathscr{A}'_{1,T}\mathscr{A}_{3,T}[1 + O_p(T^{2\psi-1})]. \quad (4.66)$$

First consider $\mathscr{A}'_{1,T}$. By *Assumptions 3.3* and *4.4*, we have

$$\mathscr{A}'_{1,T} = \sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\dot{g}'_{\widehat{v},t}\{\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}} = \mathbf{O_p}(T^{\psi-\frac{1}{2}}). \quad (4.67)$$

Next, consider the term $\mathscr{A}_{3,T}$. By Notation 2.6, we have

$$\mathscr{A}_{3,T} = \{\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\varepsilon_t - \{\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{u},t}\{\sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}\}^{-1} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\varepsilon_t.$$

Recall that $\mathscr{B}_T := (\sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t})^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{v},t}(\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t})^{-\frac{1}{2}}$. We can re-

arrange the above equation as

$$\mathscr{A}_{3,T} = \{\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\varepsilon_t - \mathscr{B}'_T\{\sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}\}^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\varepsilon_t. \tag{4.68}$$

Consider the first term on the RHS of the above equation. Since $\dot{g}_{\widehat{v},t}$ is adapted to the filtration $\mathscr{F}_{t-1}$ by *Assumption 2.4*, we can apply *Lemma 1* of Lai and Wei (1982) to get

$$\{\sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\dot{g}'_{\widehat{v},t}\}^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{v},t}\varepsilon_t = \mathbf{O_p}(\log T). \tag{4.69}$$

Next consider the second term on the RHS of equation (4.68). Following from the *Assumption 3.2*, we have

$$\mathscr{B}_T = \mathbf{O_p}(T^{\psi-\frac{1}{2}}). \tag{4.70}$$

In addition, we could invoke *Lemma 1* of Lai and Wei (1982) to get

$$\{\sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\dot{g}'_{\widehat{u},t}\}^{-\frac{1}{2}} \sum_{t=2}^{T} \dot{g}_{\widehat{u},t}\varepsilon_t = \mathbf{O_p}(\log T), \tag{4.71}$$

since $\dot{g}_{\widehat{u},t}$ is adapted to the filtration $\mathscr{F}_{t-1}$ according to *Assumption 2.4*.

Plug the results in equations (4.69), (4.70) and (4.71) back into equation (4.68) to obtain

$$\mathscr{A}_{3,T} = \mathbf{O_p}(\log T) + \mathbf{O_p}(T^{\psi-\frac{1}{2}}) \cdot \mathbf{O_p}(\log T) = \mathbf{O_p}(\log T). \tag{4.72}$$

Substituting equations (4.67) and (4.72) back to equation (4.66) and rearranging yields

$$\sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t}, \dot{g}'_{\widehat{v},t})']\varepsilon_t = \sum_{t=2}^{T} [\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t + O_p(T^{\psi-\frac{1}{2}} \cdot \log T). \tag{4.73}$$

Recall from equation (4.51) we want to show that

$$\frac{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']\varepsilon_t}{\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2\}^{\frac{1}{2}}} \approx \frac{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t}{\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{\frac{1}{2}}}.$$

Substituting equations (4.63) and (4.73) for the denominator and numerator on the LHS, respectively, gives

$$\frac{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']\varepsilon_t}{\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|(\dot{g}'_{\widehat{u},t},\dot{g}'_{\widehat{v},t})']^2\}^{\frac{1}{2}}} = \frac{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]\varepsilon_t + \mathbf{O_p}(T^{\psi-\frac{1}{2}}\cdot\log T)}{\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2[1+O_p(T^{2\psi-1})]\}^{\frac{1}{2}}}. \tag{4.74}$$

Following from the *Assumption 4.4*, we have

$$\{\sum_{t=2}^{T}[\widehat{\varepsilon}_{t-1}|\dot{g}_{\widehat{u},t}]^2\}^{-\frac{1}{2}} = o_p(\frac{T^{\frac{1}{2}-\psi}}{\log T}).$$

Plugging the above equation into equation (4.74) completes the proof. □

4.F Proof of Theorem 4.2

In the derivation Theorem 4.1, *Assumption 2.5* is needed to invoke Theorem 3.1. Then provided that this assumption is under control, we can follow the recipe of Appendix 4.D to derive the limiting distribution of the test statistic. From Lemma 4.12, *Assumption 2.5* always holds under *Assumptions 2.4*, *3.2*, *3.3* and *4.4*.

□

Part II

Least Trimmed Squares

CHAPTER 5

Introduction for Part II

It is well known that the classical ordinary least squares (OLS) estimator is sensitive to outliers, and this motivated the research on robust regression methods. An estimator that is robust against possible outliers was proposed by Rousseeuw (1984) and the term 'least trimmed squares' (LTS) was coined. The purpose of Part II is twofold. The first purpose is to study a Jarque and Bera (1980, 1987) type test for normality of non-outlying observations after running the LTS estimation (Chapter 6). The second purpose is to derive some limit properties for information based and cumulant based model selection criteria in the context of LTS regressions (Chapter 7).

This introduction chapter proceeds as follows. Section 5.1 provides an empirical illustration on how outliers can pose a problem in the classical OLS regression. Section 5.2 gives a brief discussion of robust regressions. Section 5.3 describes the model assumption of this part and reviews literature on normality tests. Section 5.4 introduces the model selection problem in the LTS context.

5.1 The Sensitivity of OLS to Outliers

The OLS estimator is widely used in regression analysis because of both the ease of its computation and tradition. However, as illustrated in Figure 5.1, unusual observations can have a large impact on the estimates. The data are taken from Rousseeuw and Leroy (1987), which contain 47 observations corresponding to the 47 stars in the

direction of Cygnus. The regressor is the log of effective temperature at the surface of the star, and the dependent variable is the log of its light intensity. Since there is only one explanatory variable, the structure is obvious on inspection. We can see that there are 41 observations which plausibly lie on the same regression line, two relatively close outliers and a cluster of four substantial bad leverage points, which are called the giant stars.

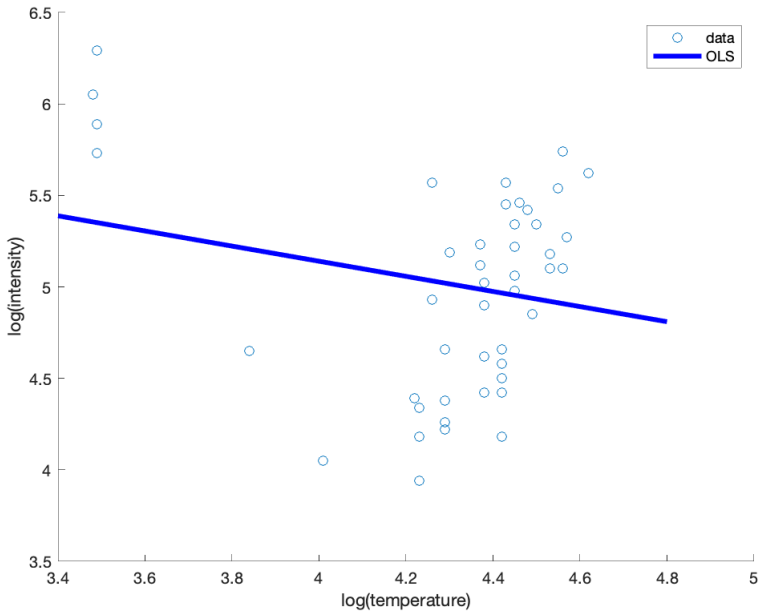


Fig. 5.1 Star Data with OLS Fit

While the main sequence presents a positive linear relationship, the four giants represent a different population that has low temperature but high output of light. These four bad leverage points influence the OLS regression line due to the poor estimation of parameters, and the line is attracted towards the giants with a negative slope.

Outlying observations both in the direction of the dependent variable and regressors can have a strong adverse effect on the OLS estimates, and they might remain unnoticed, especially when higher dimensional data are analysed. Therefore statistical methods that could accommodate or detect outlying observations have been developed, and one of them is the LTS estimator that we consider in this part.

5.2 Robust Regressions

In this thesis, we consider data points with high influence as outliers. The term 'influence' refers to a measure of the change in the estimated coefficients when a data point is removed. Therefore, the outliers under consideration are those that, if trimming them, there will be a substantial change in the value of the OLS estimators.

Once outliers are identified, one either delete or accommodate them. Robust statistical methods seek to accommodate outliers and try to fit the model imposed by the majority of the data. They aim to find a 'robust' fit, which is similar to the fit we would have found in the absence of the outliers. Therefore, the focus is on achieving robust estimators that are less influenced by the outliers and insensitive to small deviations from statistical assumptions.

There are two important concepts related to the field of robust regression, namely the breakdown point [Hampel (1968, 1971)] and the influence function [Hampel (1974)]. The breakdown point measures how resistant an estimation method is to the presence of outliers. It is the percentage of contaminants that is allowed in the data before rendering the estimator arbitrarily bad, and this could be as high as 50%. Therefore, the higher the breakdown point, the more robust the estimation method. The influence function measures the effect of a single outlier on the estimator. Ideally, we would like estimators to have bounded influence functions, which refers to the ability of the estimator to limit its reaction to contamination. The OLS estimator, which is known to be not robust, has no bounded influence and zero breakdown point.

Consider the standard regression model

$$y_i = x_i'\beta + \sigma\varepsilon_i, \quad i = 1, \ldots, n,$$

where $\beta \in \mathbb{R}^k$ and $\sigma \in \mathbb{R}_{>0}$. Let $r_i(\beta) = y_i - x_i'\beta$. We here give a brief review of some commonly used robust statistical methods.

In 1887, Edgeworth studied the Least Absolute Value (LAV) estimator, also known

as Least Absolute Deviations (LAD), which was originally introduced by Boscovich in 1757. The LAV estimators are obtained by minimising $\sum_{i=1}^{n} |r_i(\beta)|$. More recently, Huber (1964) introduced the M-estimators which attempt to minimise $\sum_{i=1}^{n} \rho[r_i(\beta)/\widehat{\sigma}]$ for some objective function $\rho(\cdot)$ and $\widehat{\sigma}$ is an estimator of the scale σ. Rousseeuw (1984) proposed the Least Median of Squares (LMS) estimator, which minimises the median, rather than the sum, of $r_i^2(\beta)$. In the same paper, he also introduced the LTS estimator. Suppose it is known that there are $h \leq n$ 'good' observations and $n-h$ outliers. Then we could estimate the model using the LTS method as follows. For a given value of β, we compute $r_{(i)}^2(\beta)$ for all i and sort them in increasing order: $r_{(1)}^2(\beta) \leq r_{(2)}^2(\beta) \leq \cdots \leq r_{(n)}^2(\beta)$. The sum of the h smallest $r_{(i)}^2(\beta)$, $\sum_{i=1}^{h} r_{(i)}^2(\beta)$, is then computed. This procedure is repeated for all possible values of β, and the one that returns the smallest sum is the LTS estimator.

Among the aforementioned robust estimation methods, the LAV estimator has bounded influence but low efficiency and zero break down point in general. The breakdown point and boundedness of the M-estimator depend both on the specification of the objective function and on the scale estimate, $\widehat{\sigma}$. An M-estimator with a non-decreasing score function, which is the derivative of the objective function $\rho(\cdot)$, in general has a breakdown point of $1/n$ which also tends to zero as $n \to \infty$. In addition, if the choice of the scale estimator, $\widehat{\sigma}$, is less resistant to outliers, this will in turn affects the robustness of the M-estimator. In contrast, both the LMS and LTS estimators have high breakdown points and bounded influence functions. These two estimators can have a breakdown point up to 50%. However, there is no closed form solution for the LMS estimator in general regression models. Therefore, we restrict our attention to the LTS estimator in this thesis, because it has close form solutions and is scale-invariant. Econometricians care about valid inferences. Since there exists a closed form solution for the LTS estimator, we can use the model to learn about the true data generation process, which makes it more appealing than the LMS estimator.

5.3 The Error Term Distribution and the Normality Test

LTS was originally proposed by Rousseeuw (1984) within the i.i.d. ε-contaminated errors framework but without reference to a likelihood function. He followed the ε-contaminated model used by Huber (1964), where errors are i.i.d. draws from a distribution that mixes a normal and an unknown contamination distribution. The errors are normal with probability $1-\varepsilon$ and otherwise sampled from the contamination distribution. Berenguer-Rico et al. (2019) proposed a different setup and derived a model where the LTS estimator is the maximum likelihood estimator (MLE). Specifically, they derived a semi-parametric model where the errors are not i.i.d., but only the h 'good' ones are i.i.d. normally distributed. We follow their specification for the error terms in this thesis, and this is formally defined in the next chapter. Using an extension of the traditional likelihood, where pairwise comparisons of probability measures are implemented, they showed that the LTS estimators are MLEs in their model.

It is important to assess the normality assumption on the 'good' errors. This assumption forms the ground for the equivalence of MLE and LTS estimators. If this normality assumption is violated, interpretation and inference may not be reliable or valid. Before proceeding with any relevant statistical procedures, this assumption should be checked. In Chapter 6, we consider the normality test based jointly on the third and fourth standardised moments — skewness and kurtosis.

The concept of cumulants was introduced by Thiele (1889), where it was termed 'half-invariants'. Based on Thiele's work, Pearson (1894) was among the first ones that initiated the modern theory of normality tests. He recognised that departures from normality could be detected using the third and fourth standardised moments of a distribution. Fisher (1930) formalised the idea of using cumulants based tests to test for normality. Under the null hypothesis of normality, Pearson (1930) derived the first four moments of the sampling distribution of skewness and kurtosis, using Fisher's result.

One of the more recent tests combining both standardised moments was studied by D'Agostino and Pearson (1973), in which the test statistic is the sum of normal approximations of skewness and kurtosis. The normality test studied in this thesis is based on the famous Jarque and Bera (1980, 1987) (hereafter JB) test. The JB test statistic was originally suggested by Bowman and Shenton (1975), which does not require transformations of the skewness and kurtosis measures. Suppose $h = n$, so there are no outliers, and we estimate the model using the standard OLS. Their test statistic is defined as

$$\frac{n}{6}(\widehat{\kappa}_3^2 + \frac{\widehat{\kappa}_4^2}{4}),$$

where $\widehat{\kappa}_3$ and $\widehat{\kappa}_4$ are the sample skewness and sample excess kurtosis, respectively. This test statistic is asymptotically chi-squared distributed with two degrees of freedom. This is because skewness and excess kurtosis are normally distributed and independent of each other asymptotically. Bowman and Shenton (1975) derived the asymptotic distribution but didn't study the power or finite sample behaviour of the test statistic. Using the same statistic, Jarque and Bera showed that this is also the Lagrangian Multiplier test statistic, and therefore has maximum local asymptotic power, against alternative distributions in the Pearson family. We shall show that, when $h < n$ is known and the model is estimated using the LTS method, the same asymptotic theory still applies for location-scale models.

5.4 The Least Trimmed Squares Method and Model Selection

In Figure 5.2, we have included 2 LTS fits to the stars data discussed above. The red line depicts the LTS regression with the four giants treated as outliers. The green line is the LTS regression with the two intermediate points also treated as outliers, so six observations have been trimmed. Both LTS lines fit predominantly to the main sequence with positive slopes and are far away from the four giant stars. Thus these bad leverage points will have large residuals from these two lines. For hard trimming

with six outliers of 47 observations, the line is the steepest. This figure also reflects that the LTS estimates are sensitive to the choice of h.

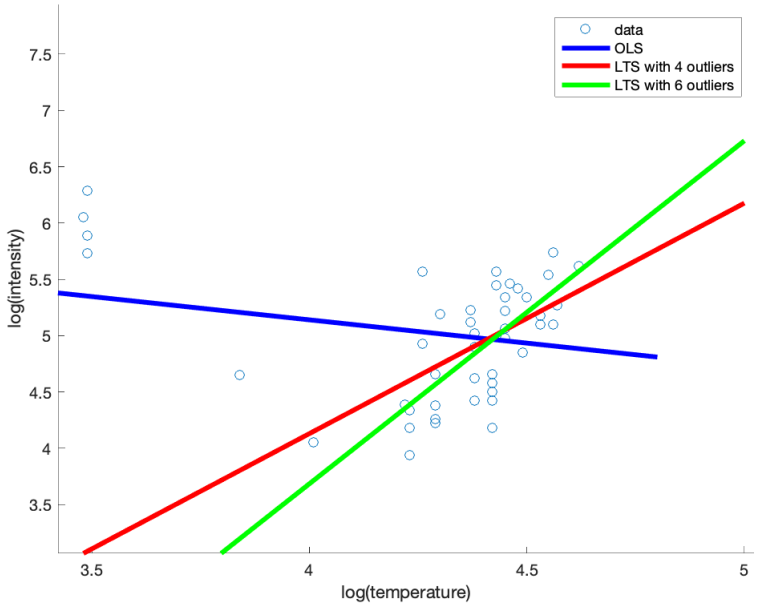


Fig. 5.2 Star Data with OLS and LTS Fits

If the number of outliers is known, then that number of cases should be trimmed. In practice, h is generally unknown and needs to be chosen prior to the LTS estimation. This amounts to the problem of selecting the optimal sample size of good observations from a set of candidate values denoted by l, where $l = 2, \ldots, n$. The goal is to find the set of good observations that not only minimises the residual sum of squares (RSS) but is also of the optimal size. This is a model selection problem specific to the LTS estimation. In the LTS context, we have a range of models to choose from depending on the values of l.

In the literature, stepwise procedures together with index plots, which allow data points to be added or deleted at each step, are commonly employed to identify outliers. This method, however, may not yield a consistent selection of the true model due to masking and swamping. Here, we consider two alternative types of model selection criteria. One is the information criteria (IC) and the other is the cumulant based normality test statistic. We establish the conditions, under which these two methods

yield consistent estimation of the true model in Chapter 7. This thesis aims to fill the gap and to provide some theoretical justifications for consistent model selection using these two types of criteria.

CHAPTER 6

Normality Test for LTS Regressions

6.1 Introduction

In classical theory, the error terms, ε_i's, are assumed to be i.i.d. $N(0,1)$-distributed. As discussed in the previous chapter, the OLS estimators are very sensitive to the presence of outliers and violations to the normality assumption may result in invalid inferences. Suppose it is known that there are $h \leq n$ 'good' observations and $n-h$ outliers. Rousseeuw (1984) introduced a robust estimator, the LTS estimator, which is obtained by finding the h subset of observations with the smallest OLS objective function. More recently, Berenguer-Rico et al. (2019) derived a model where the LTS estimators are MLEs under a different distributional assumption on the error terms. In this chapter, following their setup, we attempt to test for the normality of non-outlying observations using residuals from the LTS regression. We shall study the omnibus test for normality which is based on skewness and kurtosis in a location-scale model, taking the number of 'good' observations, h, as known.

The organisation of this chapter is as follows. The location-scale model is formally defined in Section 6.2. Section 6.3 is concerned with testing the normality assumption of the 'good' error terms. In Section 6.4, we present some numerical examples related to the normality test to evaluate the finite sample performance of the asymptotic results. Some concluding remarks are stated in Section 6.5.

6.2 The LTS Location-Scale Model

In this section, we consider the model developed in Berenguer-Rico et al. (2019). We focus on the location-scale case, for which the asymptotic theory is developed. Regression models are more challenging than the location-scale, as the outliers can be associated with the error distribution and/or with the regressors. Nonetheless, we expect the result formulated in this chapter to generalise to a regression model.

Consider the following location-scale model for observations y_1 to y_n,

$$y_i = \mu + \sigma\varepsilon_i, \qquad i = 1,\ldots,n. \tag{6.1}$$

The parameters μ and σ are parameters for location and scale, respectively. Let $h \leq n$ be given, and ζ be a set with h elements from $1,\ldots,n$. Here, ζ forms the set of indices for non-outlying error terms, so-called 'good' errors. We assume that the h 'good' errors are i.i.d. standard normally distributed. The outlying errors are independent when conditioning on the good ones, but not necessarily identically distributed, as stated formally in the following assumption.

Assumption 6.1 (LTS Model Disturbances)**.** *For $i \in \zeta$, let ε_i be i.i.d.-$N(0,1)$ distributed. For $j \notin \zeta$, let ξ_j be independent with distribution functions $G_j(z)$ for $z \in \mathbb{R}$ where G_j are continuous at 0. The 'outlier' errors are then defined as*

$$\varepsilon_j = (\max_{i\in\zeta} \varepsilon_i + \xi_j)1_{(\xi_j>0)} + (\min_{i\in\zeta} \varepsilon_i + \xi_j)1_{(\xi_j<0)}. \tag{6.2}$$

In *Assumption 6.1*, we allow the contamination distributions, G_j, to be observation specific, provided that they are continuous at zero, so the probability of being zero is zero. Then, depending on the sign of ξ_j, the outlying error, ε_j, is either the sum of the largest 'good' error and ξ_j if $\xi_j > 0$ or the sum of the smallest 'good' error and ξ_j if $\xi_j < 0$. Under equation (6.2), there is no overlap in the supports of the 'good' and 'outlier' errors.

Taking h as known, the parameters in equation (6.1) are estimated using the LTS approach. That is, by finding the 'optimal' set of observations that returns the smallest RSS. Given the specific structure of the location-scale model and *Assumption 6.1*, the estimation of ζ, the set of indices for 'good' observations, can be simplified to the estimation of a parameter that we will now define.

Let $\varepsilon_{(i)}$ denote the error terms in increasing order, so $\varepsilon_{(1)} \leq \varepsilon_{(2)} \leq \cdots \leq \varepsilon_{(n)}$. Note that in a location-scale model, the ordering of the observations, y, must be the same as the ordering of the error terms, ε. In addition, since the 'outlier' errors lie strictly outside the realised range of the 'good' ones, it turns out that the 'optimal' set of observations must have consecutive order statistics. Then, depending on the contamination distribution, G_j, some 'outliers' will randomly be smaller than the smallest 'good' error and thus located to the left of the latter.

Let δ be the count of left 'outliers' so that

$$\delta = \sum_{j \notin \zeta} 1_{(\xi_j < 0)} = \sum_{j \notin \zeta} 1_{(\varepsilon_j < \min_{i \in \zeta} \varepsilon_i)}.$$

The ordered disturbances can be rewritten as

$$\underbrace{\varepsilon_{(1)} \leq \varepsilon_{(2)} \leq \cdots \leq \varepsilon_{(\delta)}}_{\delta \text{ left 'outliers'}} < \underbrace{\varepsilon_{(\delta+1)} < \varepsilon_{(\delta+2)} < \cdots < \varepsilon_{(\delta+h)}}_{h \text{ 'good'}} < \underbrace{\varepsilon_{(\delta+h+1)} \leq \varepsilon_{(\delta+h+2)} \leq \cdots \leq \varepsilon_{(n)}}_{\bar{n}=n-h-\delta \text{ right 'outliers'}}.$$

The optimal h-subset then consists of the error terms of the form $\varepsilon_{(\delta+1)} < \cdots < \varepsilon_{(\delta+h)}$. Given the assumed normality of the 'good' observations, the probability that they take any specific value will be zero, so the inequality signs are strict between these observations, whereas there is no such restriction for the 'outlier' errors.

Using δ, the set of indices for 'good' observations, ζ, can be defined as follows:

$$\zeta = \{(i) | \delta + 1 \leq i \leq \delta + h\}.$$

So ζ is a collection of the ordered indices, where i is an integer that takes value

between $\delta+1$ and $\delta+h$. We hence replace the problem of finding the optimal set with the problem of finding an optimal δ.

The estimator for δ, assuming h is known (as indicated by the subscript h), is therefore

$$\widehat{\delta}_h := \underset{0\leq\delta\leq n-h}{\arg\min} \sum_{i=\delta+1}^{\delta+h} \{y_{(i)} - \widehat{\mu}_\delta\}^2 \qquad \text{where} \quad \widehat{\mu}_\delta = h^{-1} \sum_{i=\delta+1}^{\delta+h} y_{(i)}. \tag{6.3}$$

So the LTS estimator of δ is the one that minimises the RSS among all subsets of the sample with h observations. Accordingly, the optimal estimated ζ, denoted as $\widehat{\zeta}_h$, is

$$\widehat{\zeta}_h = \{(i) | \widehat{\delta}_h + 1 \leq i \leq \widehat{\delta}_h + h\},$$

and the estimated set of 'good' residuals is $\{\widehat{\varepsilon}_{(i)}\}_{(i)\in\widehat{\zeta}_h}$. The LTS estimators for μ and σ are then

$$\widehat{\mu}_{LTS} = h^{-1} \sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h} y_{(i)} = \bar{y}_{LTS}, \tag{6.4}$$

and

$$\widehat{\sigma}_{LTS} = \{h^{-1} \sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h} (y_{(i)} - \bar{y}_{LTS})^2\}^{1/2}. \tag{6.5}$$

Thus, $\widehat{\mu}_{LTS}$ and $\widehat{\sigma}_{LTS}$ are the sample mean and the sample standard deviation of the optimal set. Using *Assumption 6.1*, Berenguer-Rico et al. (2019) show that these LTS estimators are also MLEs in the LTS model.

Note that, the LTS estimator of μ in equation (6.4) is the same as that proposed by Rousseeuw (1984). On the other hand, the LTS estimator of σ in equation (6.5) suggested by Berenguer-Rico et al. (2019) differs from that studied by Rousseeuw. Specifically, the estimator of σ differs between that considered in Rousseeuw's framework and that studied in this thesis by a consistency factor: $\widehat{\sigma}^2_{Rousseeuw} = \varsigma\widehat{\sigma}^2_{LTS}$, where ς is the consistency factor. Recall from Section 5.3 in Chapter 5 that Rousseeuw discussed his LTS estimators within an ε-contaminated framework, and thus the consis-

tency factor ς would depend both on the specification of the contamination distribution and on the normal distribution in his model.

Rousseeuw considered the ε-contaminated framework in his LTS model when deriving his LTS estimators of μ and σ. However, in developing the consistency and asymptotic normality of μ, he only considered the situation where all the errors are normal without contamination. The asymptotic variance of μ also has a nonstandard expression in his framework. Furthermore, his LTS estimator of σ, which takes a different form as the one in equation (6.5), is consistent upon adjusting by a consistency factor. In contrast, Berenguer-Rico et al. (2019) show that the asymptotic theory of these two estimators in equations (6.4) and (6.5) is the same as of the standard OLS estimators, provided that *Assumption 6.1* is satisfied.

6.3 The Normality Test

In this section, we consider a cumulant based normality test for the estimated 'good' residuals where the true h is assumed to be known.

Suppose we have run the LTS estimation with the correct h given, and obtained the standardised residuals,

$$\widehat{\varepsilon}_{(i)} = \frac{y_{(i)} - \widehat{\mu}_{LTS}}{\widehat{\sigma}_{LTS}},$$

where $(i) \in \widehat{\zeta}_h$.

We suggest assessing the normality of the residuals via evaluating the skewness and kurtosis. If the errors are not normally distributed, the sample skewness should indicate a departure from a symmetric distribution and the sample excess kurtosis, which measures whether the data is peaked or flat relative to a normal distribution, will be different from zero.

The normality test is based on the following two statistics,

$$\widehat{\kappa}_{3,h} = \frac{h^{-1}\sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h}\widehat{\varepsilon}_{(i)}^3}{[h^{-1}\sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h}\widehat{\varepsilon}_{(i)}^2]^{3/2}} = \frac{S_{3,h}}{S_{2,h}^{3/2}}$$

and

$$\widehat{\kappa}_{4,h} = \frac{h^{-1}\sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h}\widehat{\varepsilon}_{(i)}^4}{[h^{-1}\sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h}\widehat{\varepsilon}_{(i)}^2]^2} - 3 = \frac{S_{4,h} - 3S_{2,h}^2}{S_{2,h}^2},$$

where $\widehat{\kappa}_{3,h}$ and $\widehat{\kappa}_{4,h}$ are the sample skewness and sample excess kurtosis of $\varepsilon_{(i)}$ ($i \in \zeta$), respectively, and $S_{p,h} = h^{-1}\sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h}\widehat{\varepsilon}_{(i)}^p$ for $p \in \{2,3,4\}$.

Under the null hypothesis that the 'good' errors are normally distributed, we show that

$$\sqrt{h}\begin{bmatrix} \frac{\widehat{\kappa}_{3,h}}{\sqrt{6}} \\ \frac{\widehat{\kappa}_{4,h}}{\sqrt{24}} \end{bmatrix} \xrightarrow{D} N_2(\mathbf{0}, \mathbf{I_2}),$$

where $\mathbf{0} = [0 \quad 0]'$, and $\mathbf{I_2}$ is a 2×2 identity matrix.

The main results are illustrated in the sketch here, with the formal proof given in Appendix 6.A. The analysis is twofold. First, we show that conditioning on the true δ, the correct set of 'good' observations is chosen so that

$$\sqrt{h}\begin{bmatrix} \frac{\widehat{\kappa}_{3,h\cdot\delta}}{\sqrt{6}} \\ \frac{\widehat{\kappa}_{4,h\cdot\delta}}{\sqrt{24}} \end{bmatrix} \xrightarrow{D} N_2(\mathbf{0}, \mathbf{I_2}),$$

where $\widehat{\kappa}_{3,h\cdot\delta}$ and $\widehat{\kappa}_{4,h\cdot\delta}$ denote the sample skewness and excess kurtosis derived under known δ. We then prove that when the estimation error for the number of left 'outliers', $\widehat{\delta}_h - \delta$, is small, both $h^{1/2}(\widehat{\kappa}_{3,h} - \widehat{\kappa}_{3,h\cdot\delta})$ and $h^{1/2}(\widehat{\kappa}_{4,h} - \widehat{\kappa}_{4,h\cdot\delta})$ are $o_p(1)$, conditional on δ. It has been shown in Berenguer-Rico et al. (2019, Theorem 5.2) that, when less than half of the observations are 'outliers', $\widehat{\delta}_h$ is a consistent estimator for δ.

To facilitate the analysis, it is sufficient to consider

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\widehat{\varepsilon}_{(i)}^{p} \quad \text{and} \quad S_{p,h} = h^{-1}\sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h}\widehat{\varepsilon}_{(i)}^{p}$$

for $p \in \{2,3,4\}$.

The key result to establish the asymptotic bivariate normal distribution is given in the following theorem.

Theorem 6.1. *If Assumption 6.1 is satisfied, then*

$$\sqrt{h}\begin{bmatrix} S_{3,h} \\ S_{4,h}-3S_{2,h}^2 \end{bmatrix} \xrightarrow{D} N\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}\}.$$

6.4 Simulation — The Normality Test in the LTS Location-Scale Model

In this section, we investigate the finite sample exact critical values of the normality test proposed in Section 6.3 by means of simulation studies.

Table 6.1 DGP for ε — Normality Test

DGP	ε
LS.1	$\varepsilon_i \sim i.i.d.N(0,1) \quad i=1,\ldots,n$
LS.2	For $i \in \zeta$: $\varepsilon_i \sim i.i.d.N(0,1)$ For $j \notin \zeta$: $\varepsilon_j = (\max_{i\in\zeta}\varepsilon_i + 1 + \xi_j)1_{(\xi_j>0)} + (\min_{i\in\zeta}\varepsilon_i - 1 + \xi_j)1_{(\xi_j<0)}$, where $\xi_j \sim i.i.d.U[-5,5]$
LS.3	For $i \in \zeta$: $\varepsilon_i \sim i.i.d.N(0,1)$ For $j \notin \zeta$: $\varepsilon_j = \min_{i\in\zeta}\varepsilon_i - 1 + \xi_j$, where $\xi_j \sim i.i.d.U[-5,-2]$
LS.4	For $i \in \zeta$: $\varepsilon_i \sim i.i.d.N(0,1)$ For $j \notin \zeta$: $\varepsilon_j = \max_{i\in\zeta}\varepsilon_i + 1 + \xi_j$, where $\xi_j \sim i.i.d.U[2,5]$
LS.5	For $i \in \zeta$: $\varepsilon_i \sim i.i.d.N(0,1)$ For $j \notin \zeta$: $\varepsilon_j = \max_{i\in\zeta}\varepsilon_i + 1 + \xi_j$, where $\xi_j \sim i.i.d.\chi_3^2$
LS.6	For $i \in \zeta$: $\varepsilon_i \sim i.i.d.N(0,1)$ For $j \notin \zeta$: $\varepsilon_j = (\max_{i\in\zeta}\varepsilon_i + 1 + \xi_j)1_{(\xi_j>0)} + (\min_{i\in\zeta}\varepsilon_i - 1 + \xi_j)1_{(\xi_j<0)}$, where $\xi_j \sim i.i.d.\chi_3^2 - 3$

We consider the location-scale model

$$y_i = \mu + \sigma\varepsilon_i, \qquad i=1,\ldots,n, \tag{6.6}$$

where the true μ is 0 and the true σ is 1, and examine six different DGPs for the error terms, which are given in Table 6.1. The model with error terms generated using LS.1 has no outliers and is therefore estimated by OLS. For DGPs LS.2 to LS.6, there are three different types of outliers, which locate to the left (LS.3), to the right (LS.4 and LS.5), or to both the left and right (LS.2 and LS.6) of the 'good' errors. For these DGPs, we fix $h = 0.8 \times n$ and apply the following algorithm to obtain the LTS estimators.

The Algorithm

1. For a given sample of size n, sort the data ($\varepsilon_i \to \varepsilon_{(i)}$).
2. Calculate the partial sums $\sum_{i=1}^{s} \varepsilon_{(i)}$ and $\sum_{i=1}^{s} \varepsilon_{(i)}^2$, for $s = 1, \ldots, n$.
3. Find $\widehat{\delta}_h$ by

$$\widehat{\delta}_h = \underset{0 \le \delta \le n-h}{\arg\min}\, RSS = \underset{0 \le \delta \le n-h}{\arg\min}\, \{ \sum_{i=1}^{\delta+h} \varepsilon_{(i)}^2 - \sum_{i=1}^{\delta} \varepsilon_{(i)}^2 - \frac{1}{h} [\sum_{i=1}^{\delta+h} \varepsilon_{(i)} - \sum_{i=1}^{\delta} \varepsilon_{(i)}]^2 \}$$

and compute $\widehat{\mu}_{LTS}$ and $\widehat{\sigma}_{LTS}$.

The test statistics for skewness and excess kurtosis in the m-th replication are

$$W_{\kappa_3,m} := \sqrt{h} \frac{\widehat{\kappa}_{3,h}}{\sqrt{6}} \quad \text{and} \quad W_{\kappa_4,m} := \sqrt{h} \frac{\widehat{\kappa}_{4,h}}{\sqrt{24}},$$

respectively. Following from Theorem 6.1 and using the property that joint normality implies marginal normality, both $W_{\kappa_3,m}$ and $W_{\kappa_4,m}$ converge in distribution to a standard normal as $n \to \infty$.

We compare the asymptotic critical values with the exact ones calculated by simulation for sample sizes $n = 50$, 100, 200, 400, 800, 1600, 10000 and 100000, and varying levels (denoted as α) for $\alpha = 0.5$, 0.05 and 0.025. The corresponding asymptotic critical values are 0, 1.6449 and 1.9600. For a given sample of size n, we sort the test statistics in ascending order to get the exact critical values

$$\widehat{q}_{\kappa_3,n} := W_{\kappa_3,int\{M \times (1-\alpha)\}} \quad \text{and} \quad \widehat{q}_{\kappa_4,n} := W_{\kappa_4,int\{M \times (1-\alpha)\}}.$$

The total number of replications, M, is 10000. In each replication, we first generate $h = 0.8 \times max(n)$ 'good' errors. Then, for each smaller sample of size n in the same replication, the first $h = 0.8 \times n$ ones are used for analysis. So the results reported here are based on common random numbers.

The exact critical values for the test are given below in Table 6.2. The corresponding Monte Carlo standard errors are given in Table 6.5 in the Appendix. The Monte Carlo standard errors are calculated using

$$se = \sqrt{\frac{\alpha(1-\alpha)}{M}} \varphi^{-1}(\widehat{q}_n),$$

where $\varphi(\cdot)$ denotes the standard normal density function.

Table 6.2 Exact Critical Values in Location Scale Models

n	Skewness			Excess Kurtosis			Skewness			Excess Kurtosis		
α	0.5	0.05	0.025	0.5	0.05	0.025	0.5	0.05	0.025	0.5	0.05	0.025
DGP	LS.1						LS.2					
50	0.0097	1.5530	1.8770	-0.3241	1.4771	2.0319	0.0062	1.5495	1.8749	-0.3435	1.3884	1.9252
100	0.0065	1.6349	1.9512	-0.2510	1.6009	2.0915	-0.0009	1.6083	1.9297	-0.2892	1.5610	2.1269
200	0.0033	1.6385	1.9274	-0.2101	1.6583	2.1633	0.0034	1.6248	1.9585	-0.2227	1.6200	2.1681
400	-0.0115	1.6391	1.9754	-0.1555	1.7046	2.1574	0.0055	1.6576	1.9937	-0.1608	1.6873	2.1544
800	0.0070	1.6397	1.9753	-0.1212	1.7311	2.1251	0.0120	1.6645	1.9667	-0.1256	1.6408	2.0336
1600	0.0173	1.6531	1.9706	-0.0742	1.7132	2.1375	0.0268	1.6684	1.9961	-0.0920	1.6611	2.0878
10000	-0.0044	1.6765	1.9659	-0.0217	1.6461	1.9486	-0.0043	1.6741	1.9751	-0.0367	1.6565	2.0303
100000	0.0026	1.6262	1.9273	-0.0365	1.6675	1.9897	0.0066	1.6191	1.9242	-0.0200	1.6479	1.9699
DGP	LS.3						LS.4					
50	0.0062	1.5495	1.8749	-0.3435	1.3884	1.9252	0.0065	1.5498	1.8758	-0.3435	1.3884	1.9252
100	-0.0009	1.6083	1.9297	-0.2892	1.5610	2.1269	-0.0002	1.6087	1.9303	-0.2892	1.5610	2.1269
200	0.0034	1.6248	1.9585	-0.2227	1.6200	2.1681	0.0034	1.6248	1.9585	-0.2227	1.6200	2.1681
400	0.0055	1.6576	1.9937	-0.1608	1.6873	2.1544	0.0055	1.6576	1.9937	-0.1608	1.6873	2.1544
800	0.0120	1.6645	1.9667	-0.1256	1.6408	2.0336	0.0120	1.6645	1.9667	-0.1256	1.6408	2.0336
1600	0.0268	1.6684	1.9961	-0.0920	1.6611	2.0878	0.0268	1.6684	1.9961	-0.0920	1.6611	2.0878
10000	-0.0043	1.6741	1.9751	-0.0367	1.6565	2.0303	-0.0043	1.6741	1.9751	-0.0367	1.6565	2.0303
100000	0.0066	1.6191	1.9242	-0.0200	1.6479	1.9699	0.0066	1.6191	1.9242	-0.0200	1.6479	1.9699
DGP	LS.5						LS.6					
50	0.0298	1.5453	1.9027	-0.3635	1.2807	1.7697	-0.0147	1.4822	1.8394	-0.3628	1.3575	1.9663
100	0.0421	1.6057	1.9214	-0.2888	1.4753	1.9639	-0.0040	1.5714	1.8855	-0.2850	1.5388	2.0681
200	0.0550	1.6560	1.9877	-0.2261	1.6051	2.0831	0.0186	1.6242	1.9514	-0.2207	1.6745	2.1822
400	0.0447	1.6454	1.9693	-0.1819	1.6779	2.1026	0.0105	1.6143	1.9415	-0.1788	1.7062	2.1890
800	0.0447	1.6585	1.9761	-0.1482	1.6739	2.1005	0.0192	1.6457	1.9608	-0.1444	1.7020	2.1348
1600	0.0102	1.6769	1.9562	-0.0946	1.7165	2.0984	-0.0105	1.6600	1.9398	-0.0888	1.7235	2.1135
10000	-0.0048	1.6416	1.9330	-0.0164	1.6576	2.0188	-0.0152	1.6301	1.9249	-0.0157	1.6642	2.0218
100000	0.0078	1.6330	1.9518	-0.0300	1.6703	1.9789	0.0035	1.6330	1.9512	-0.0300	1.6718	1.9789

Note: $W_{0.5} = 0$, $W_{0.05} = 1.64$, $W_{0.025} = 1.96$

From Table 6.2, we can see that, overall, the approximation by the standard normal distribution works quite well, indicating that the speed of convergence is rather fast. The results are consistent across different types of DGPs for the error terms. This suggests that, when the support of 'good' and 'bad' error terms are not overlapping, the test is robust to the location of the 'outlying' error terms — no 'outliers' (LS.1), left 'outliers only (LS.3), right 'outliers' only (LS.4 and LS.5), or both left and right

'outliers' (LS.2 and LS.6).

However, the pattern of convergence is not so clear. This might be due to the poor small sample performance of the cumulant based normality test; see Bowman and Shenton (1975), D'Agostino (1986), Thadewald and Buning (2007) and Doornik and Hansen (2008) for details. The main problem is that while asymptotically the sample skewness and kurtosis are independent, this may not hold in finite samples. In particular, as pointed out by Bowman and Shenton (1975), the finite sample distribution of the kurtosis could be far away from a normal even when the sample size is moderately large. This is reflected in Table 6.2, where the exact critical values for the skewness test statistic are closer to the asymptotic values than the ones for the excess kurtosis test statistic.

We also compare critical values of the test statistics in models where we vary distances between the 'good' and 'bad' observations. In particular, the outlying errors, ε_j ($j \notin \zeta$) are generated as follows

$$\varepsilon_j = (\max_{i\in\zeta} \varepsilon_i + d_1\xi_j + d_2)1_{(\xi_j>0)} + (\min_{i\in\zeta} \varepsilon_i + d_1\xi_j - d_2)1_{(\xi_j<0)}$$

where ξ_j are either i.i.d. $U[-5,5]$-distributed as in LS.2 or i.i.d. χ^2_3-distributed as in LS.5. We consider $d_1 = 0.5$, 1, 2, and 3 as well as $d_2 = 0$, 1 and 2. The results are summarised in Tables 6.3 and 6.4 (Monte Carlo standard errors are reported in Tables 6.6 and 6.7 in the Appendix).

In both tables, we do not observe a clear pattern of convergence for a given combination of d_1 and d_2, which is similar to our finding above. In addition, even when the outliers and 'good' errors become further apart, as one increases the magnitude of d_1 and/or d_2, the approximation does not become more precise.

There is, however, one important difference between the results reported in these two tables. Even if the 'bad' observations are sufficiently close to the 'good' ones ($d_1 = 0.5$), the distribution of the test statistic still resembles a normal distribution

when the contamination distribution is a symmetric one ($U[-5,5]$), as one can see in Table 6.3. However, Table 6.4 shows that the test statistic is very sensitive to outlying values at right contamination or when ξ_j follows a skewed distribution, in which case the test statistic detects departure from normality, even though there is no overlap in the realised ranges of the 'good' and 'bad' observations.

Table 6.3 Exact Critical Values for LS.2 with Varying Distances

n	Skewness			Excess Kurtosis			Skewness			Excess Kurtosis		
	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975
						$d_2 = 0$						
		$d_1 = 0.5$						$d_1 = 1$				
50	0.0091	1.5498	1.8766	-0.3437	1.3901	1.9296	0.0013	1.5022	1.8252	-0.3563	1.3374	1.8911
100	-0.0009	1.6083	1.9297	-0.2892	1.5610	2.1269	0.0098	1.5764	1.8848	-0.2831	1.5553	2.0584
200	0.0034	1.6248	1.9585	-0.2227	1.6200	2.1681	0.0091	1.5776	1.9198	-0.2374	1.6576	2.1543
400	0.0055	1.6576	1.9937	-0.1608	1.6873	2.1544	-0.0022	1.6078	1.9268	-0.1747	1.6547	2.1116
800	0.0120	1.6645	1.9667	-0.1256	1.6408	2.0336	0.0051	1.6522	1.9314	-0.1264	1.6643	2.0962
1600	0.0268	1.6684	1.9961	-0.0920	1.6611	2.0878	-0.0005	1.6440	1.9405	-0.1092	1.6780	2.0513
10000	-0.0043	1.6741	1.9751	-0.0367	1.6565	2.0303	-0.0028	1.6608	1.9623	-0.0395	1.6715	1.9820
100000	0.0066	1.6191	1.9242	-0.0200	1.6479	1.9699	-0.0242	1.6190	1.9295	-0.0399	1.6610	1.9935
		$d_1 = 2$						$d_1 = 3$				
50	0.0032	1.5197	1.8362	-0.3538	1.3602	1.9664	-0.0044	1.4946	1.8087	-0.3430	1.3016	1.7793
100	0.0060	1.5829	1.9009	-0.2939	1.4941	2.0033	-0.0051	1.5719	1.8975	-0.2798	1.4795	2.0140
200	-0.0246	1.6221	1.9749	-0.2350	1.6673	2.1710	-0.0005	1.6353	1.9391	-0.2241	1.6264	2.1431
400	-0.0216	1.6243	1.9820	-0.1887	1.6962	2.2392	0.0043	1.6438	1.9822	-0.1817	1.6875	2.1407
800	-0.0230	1.6363	1.9395	-0.1372	1.7388	2.1934	-0.0037	1.6538	1.9936	-0.1327	1.6729	2.1311
1600	-0.0279	1.6308	1.9355	-0.0963	1.7310	2.0991	0.0111	1.6519	1.9833	-0.1091	1.6989	2.0605
10000	-0.0205	1.6295	1.9114	-0.0303	1.7169	2.0583	-0.0184	1.6531	1.9248	-0.0364	1.6960	2.0767
100000	-0.0034	1.6516	1.9549	0.0030	1.6640	2.0182	-0.0111	1.6587	1.9435	-0.0249	1.6453	1.9720
						$d_2 = 1$						
		$d_1 = 0.5$						$d_1 = 1$				
50	-0.0026	1.4941	1.8243	-0.3437	1.3510	1.8822	0.0017	1.5476	1.8573	-0.3384	1.3245	1.8611
100	0.0154	1.5982	1.9127	-0.2858	1.4758	1.9687	-0.0104	1.5901	1.9286	-0.2768	1.5452	2.0494
200	0.0102	1.6299	1.9300	-0.2390	1.5732	2.0791	-0.0162	1.6793	2.0153	-0.2193	1.6815	2.2014
400	0.0127	1.6483	1.9661	-0.1670	1.6892	2.1461	-0.0237	1.6229	1.9433	-0.1609	1.6868	2.1433
800	0.0195	1.6399	1.9759	-0.1370	1.7136	2.1661	0.0074	1.6657	1.9500	-0.1216	1.6965	2.1439
1600	0.0051	1.6454	1.9602	-0.1086	1.7378	2.1244	-0.0025	1.6199	1.9379	-0.0900	1.6891	2.0897
10000	-0.0110	1.6637	1.9842	-0.0546	1.6735	2.0540	-0.0182	1.6305	1.9463	-0.0338	1.6411	1.9618
100000	0.0010	1.6362	1.9664	-0.0109	1.6717	1.9853	-0.0334	1.6341	1.9273	-0.0409	1.6541	2.0055
		$d_1 = 2$						$d_1 = 3$				
50	-0.0123	1.5327	1.8400	-0.3435	1.3857	1.9381	0.0058	1.5054	1.8505	-0.3491	1.3246	1.7978
100	-0.0036	1.5714	1.9229	-0.2802	1.5104	2.0610	0.0112	1.5842	1.9554	-0.2859	1.4907	1.9978
200	-0.0112	1.5967	1.9356	-0.2309	1.6314	2.1243	0.0074	1.5598	1.9005	-0.2333	1.5939	2.0324
400	0.0105	1.6512	1.9205	-0.1708	1.6893	2.1579	0.0084	1.6132	1.9084	-0.1849	1.6358	2.0992
800	0.0091	1.6651	2.0009	-0.1322	1.6925	2.1295	-0.0018	1.6101	1.8896	-0.1503	1.6722	2.0621
1600	-0.0022	1.6423	1.9727	-0.0836	1.6895	2.0763	0.0000	1.6061	1.9366	-0.1060	1.6590	2.0403
10000	-0.0045	1.6508	1.9922	-0.0390	1.6304	1.9871	-0.0014	1.6703	1.9853	-0.0292	1.6474	1.9843
100000	-0.0001	1.6708	2.0012	-0.0073	1.6462	1.9650	-0.0024	1.6620	1.9557	-0.0166	1.6342	1.9442
						$d_2 = 2$						
		$d_1 = 0.5$						$d_1 = 1$				
50	0.0127	1.5169	1.8692	-0.3421	1.3665	1.9555	0.0194	1.5725	1.8845	-0.3507	1.3492	1.8576
100	0.0045	1.5839	1.9253	-0.2805	1.5964	2.1508	0.0138	1.5823	1.9265	-0.2819	1.5670	2.0903
200	-0.0093	1.6123	1.9298	-0.2172	1.6287	2.1424	-0.0129	1.6182	1.9428	-0.2314	1.6245	2.1623
400	0.0058	1.6257	1.9471	-0.1618	1.6807	2.1272	-0.0135	1.6378	1.9351	-0.2002	1.6833	2.1663
800	0.0121	1.6316	1.9264	-0.1212	1.6967	2.0940	-0.0137	1.6609	1.9670	-0.1438	1.6979	2.1443
1600	0.0060	1.6381	1.9676	-0.0863	1.7032	2.0780	0.0009	1.6573	1.9558	-0.1133	1.6627	2.0876
10000	0.0034	1.6764	1.9738	-0.0527	1.6586	2.0011	0.0097	1.6346	1.9279	-0.0428	1.6491	2.0061
100000	0.0233	1.6757	2.0093	-0.0171	1.6317	2.0038	0.0052	1.6539	1.9804	-0.0267	1.6620	1.9841
		$d_1 = 2$						$d_1 = 3$				
50	-0.0032	1.5040	1.8066	-0.3584	1.3664	1.8448	0.0077	1.5210	1.8401	-0.3555	1.3152	1.8143
100	-0.0040	1.5110	1.8256	-0.2819	1.5762	2.0711	0.0216	1.5891	1.9334	-0.2891	1.5033	2.0169
200	-0.0174	1.6083	1.9267	-0.2314	1.6298	2.1518	0.0172	1.6089	1.9380	-0.2399	1.6376	2.1397
400	-0.0208	1.6362	1.9205	-0.1823	1.6388	2.1138	-0.0066	1.6298	1.9345	-0.1894	1.7033	2.1415
800	-0.0031	1.6579	1.9671	-0.1138	1.7533	2.1252	0.0026	1.6305	1.9475	-0.1262	1.7080	2.1311
1600	0.0052	1.6499	1.9906	-0.0868	1.7190	2.1222	-0.0008	1.6206	1.9473	-0.0940	1.6866	2.0928
10000	-0.0025	1.6393	1.9614	-0.0367	1.6584	1.9913	0.0208	1.6435	1.9660	-0.0480	1.6607	2.0389
100000	-0.0052	1.6159	1.9122	0.0040	1.6572	1.9600	0.0002	1.6204	1.9409	-0.0003	1.6476	1.9806

Table 6.4 Exact Critical Values for LS.5 with Varying Distances

n	Skewness			Excess Kurtosis			Skewness			Excess Kurtosis		
	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975
						$d_2 = 0$						
	$d_1 = 0.5$						$d_1 = 1$					
50	0.5130	1.7600	2.0647	-0.8229	0.5404	1.0561	0.3195	1.6319	1.9240	-0.4608	1.0643	1.5377
100	0.8267	2.0513	2.3194	-1.0343	0.5721	1.0835	0.3699	1.7303	2.0284	-0.4252	1.1560	1.6200
200	1.2198	2.4501	2.6956	-1.2828	0.4891	0.9672	0.4413	1.8278	2.1470	-0.3826	1.3032	1.7448
400	1.7649	2.9428	3.1669	-1.6204	0.1829	0.6709	0.4916	1.8726	2.1568	-0.3465	1.4105	1.7675
800	2.4607	3.6218	3.8373	-2.0542	-0.1884	0.3118	0.4955	1.9558	2.2312	-0.2930	1.4520	1.8622
1600	3.3983	4.5563	4.7782	-2.6072	-0.5683	-0.0532	0.4737	1.9747	2.2663	-0.2193	1.5066	1.8778
10000	7.7125	9.0543	9.2849	-4.9844	-2.5365	-2.0044	0.3407	1.9078	2.1937	-0.1357	1.5641	1.9311
100000	20.3670	23.3228	23.7129	-11.1466	-6.7552	-5.8237	0.1698	1.7942	2.1176	-0.0641	1.6169	1.9568
	$d_1 = 2$						$d_1 = 3$					
50	0.0150	1.5164	1.8475	-0.3561	1.3104	1.8374	0.0025	1.5043	1.8080	-0.3435	1.3030	1.8898
100	0.0171	1.5683	1.9198	-0.2871	1.5167	1.9892	-0.0159	1.5747	1.8775	-0.2907	1.5168	2.0528
200	0.0071	1.6261	1.9522	-0.2288	1.5980	2.1284	-0.0139	1.5844	1.9174	-0.2308	1.5755	2.0826
400	0.0118	1.6245	1.9562	-0.1662	1.6695	2.1215	0.0010	1.5915	1.9068	-0.1784	1.6753	2.2038
800	0.0083	1.6631	1.9387	-0.1159	1.7147	2.1169	-0.0098	1.5907	1.9100	-0.1417	1.6730	2.1225
1600	0.0069	1.6696	1.9811	-0.1101	1.7149	2.1439	-0.0044	1.6261	1.9553	-0.1025	1.7319	2.1426
10000	0.0167	1.6718	1.9884	-0.0477	1.6843	2.0270	0.0060	1.6800	2.0167	-0.0474	1.7007	2.0402
100000	0.0124	1.6615	1.9874	-0.0329	1.6604	2.0140	0.0241	1.6593	2.0012	-0.0182	1.6275	2.0005
						$d_2 = 1$						
	$d_1 = 0.5$						$d_1 = 1$					
50	0.3395	1.6556	1.9527	-0.4726	0.9591	1.3817	0.0423	1.5608	1.8574	-0.3448	1.2836	1.7688
100	0.4926	1.8106	2.0866	-0.4897	1.0486	1.5317	0.0306	1.5848	1.9094	-0.2764	1.4598	1.9344
200	0.6742	1.9738	2.2071	-0.5265	1.1256	1.5681	0.0416	1.6052	1.9509	-0.2318	1.5846	2.0560
400	0.8860	2.1976	2.4600	-0.5614	1.1485	1.5639	0.0291	1.6199	1.9290	-0.1777	1.5738	2.0780
800	1.1412	2.4342	2.7237	-0.6200	1.1214	1.5664	0.0077	1.6098	1.8822	-0.1445	1.6329	2.0354
1600	1.4134	2.8091	3.0465	-0.6746	1.0487	1.4432	0.0005	1.6359	1.9505	-0.1178	1.6618	2.0647
10000	2.4240	4.0045	4.2782	-0.9389	0.8271	1.1817	0.0083	1.6579	1.9879	-0.0551	1.6704	1.9949
100000	4.4789	6.8057	7.1981	-1.4936	0.3862	0.7328	0.0027	1.6764	1.9829	-0.0154	1.6468	1.9823
	$d_1 = 2$						$d_1 = 3$					
50	0.0069	1.5284	1.8656	-0.3507	1.3711	1.9085	-0.0007	1.5152	1.8705	-0.3488	1.3581	1.8950
100	-0.0039	1.5618	1.8809	-0.3022	1.5173	2.0967	0.0070	1.5812	1.9168	-0.2995	1.5150	2.0080
200	0.0004	1.5806	1.9134	-0.2515	1.6048	2.1142	-0.0054	1.6232	1.9086	-0.2486	1.6812	2.2242
400	-0.0139	1.6020	1.9078	-0.1960	1.7127	2.2144	0.0081	1.5952	1.9337	-0.1727	1.6599	2.1134
800	-0.0054	1.6263	1.9495	-0.1329	1.7123	2.1062	0.0122	1.6482	1.9772	-0.1483	1.7134	2.1288
1600	0.0055	1.6454	1.9529	-0.1112	1.7001	2.0764	0.0148	1.6665	1.9852	-0.1093	1.6712	2.0329
10000	0.0147	1.6581	1.9584	-0.0355	1.6847	2.0254	0.0203	1.6717	1.9626	-0.0618	1.6507	1.9810
100000	-0.0013	1.6811	2.0061	-0.0186	1.6380	2.0200	0.0108	1.6012	1.9149	-0.0526	1.6527	1.9839
						$d_2 = 2$						
	$d_1 = 0.5$						$d_1 = 1$					
50	0.0272	1.5511	1.9180	-0.3632	1.3012	1.8120	0.0115	1.5210	1.8477	-0.3428	1.3698	1.8525
100	0.0575	1.6135	1.9249	-0.2913	1.4633	1.9337	0.0016	1.5542	1.8866	-0.2724	1.4784	2.0339
200	0.1016	1.7136	2.0117	-0.2303	1.5338	2.0293	0.0154	1.6335	1.9846	-0.2080	1.6396	2.1154
400	0.1224	1.7013	2.0051	-0.1955	1.5970	2.0176	0.0005	1.6364	1.9856	-0.1711	1.6191	2.0892
800	0.1430	1.7397	2.0439	-0.1641	1.5948	2.0178	0.0167	1.6580	1.9751	-0.1156	1.6524	2.0624
1600	0.1420	1.7564	2.0403	-0.1281	1.6415	2.0162	0.0129	1.6704	1.9748	-0.0641	1.6830	2.0701
10000	0.1903	1.7950	2.1093	-0.0589	1.5695	1.9364	0.0074	1.6446	1.9340	-0.0424	1.6737	2.0514
100000	0.2573	1.9091	2.1905	-0.0966	1.5914	1.9087	-0.0072	1.6175	1.9862	-0.0177	1.6498	2.0063
	$d_1 = 2$						$d_1 = 3$					
50	0.0103	1.5028	1.8259	-0.3561	1.3212	1.8439	0.0025	1.5043	1.8080	-0.3435	1.3030	1.8898
100	0.0134	1.5642	1.9184	-0.2871	1.5182	1.9905	-0.0159	1.5747	1.8775	-0.2907	1.5168	2.0528
200	0.0067	1.6246	1.9505	-0.2288	1.5980	2.1284	-0.0139	1.5844	1.9174	-0.2308	1.5755	2.0826
400	0.0118	1.6245	1.9562	-0.1662	1.6695	2.1215	0.0010	1.5915	1.9068	-0.1784	1.6753	2.2038
800	0.0083	1.6631	1.9387	-0.1159	1.7147	2.1169	-0.0098	1.5907	1.9100	-0.1417	1.6730	2.1225
1600	0.0069	1.6696	1.9811	-0.1101	1.7149	2.1439	-0.0044	1.6261	1.9553	-0.1025	1.7319	2.1426
10000	0.0167	1.6718	1.9884	-0.0477	1.6843	2.0270	0.0060	1.6800	2.0167	-0.0474	1.7007	2.0402
100000	0.0124	1.6615	1.9874	-0.0329	1.6604	2.0140	0.0241	1.6593	2.0012	-0.0182	1.6275	2.0005

6.5 Conclusion

This chapter draws attention to the LTS estimation within the LTS location-scale model and shows that standard asymptotic theories for the skewness and kurtosis based normality test apply when the true h is known. The limiting distribution of the test statistic is derived under this new model. We compare the finite sample exact critical values of the test statistic with the asymptotic ones through a series of Monte Carlo experiments. Overall, the finite sample distribution of the test statistic is closely approximated by its asymptotic distribution.

There are several issues that remain unresolved or have triggered questions for future research. In this chapter, asymptotic results are formulated in the context of a location-scale model. We hope to extend these theoretical results to models with stochastic regressors in the near future. A formal proof is left for future work. The second issue involves the assumption of known h when deriving the asymptotic distribution of the normality test statistic. In practice, this condition is stronger than needed since h is usually unknown and needs to be estimated. Provided that the proportion of 'good' errors can be consistently estimated, it is possible to relax this condition and derive the same limiting results at the next stage. Another direction of interest is to examine small sample corrections for the normality test statistic. Since skewness and kurtosis measures are affected by sample sizes, literature on tests based on these shape parameters indicates a potential need for a weighting coefficient in the test statistic.

Appendix

6.A Proof of Theorem 6.1

In this section, we give a proof to Theorem 6.1 taking the true h as known. We shall suppress the subscript h in $\widehat{\delta}_h$ to simplify the notations.

6.A.1 Limiting Distribution Conditional on δ

Notice that conditional on δ, we have

$$S_{p,h\cdot\delta} = h^{-1} \sum_{i=\delta+1}^{\delta+h} \widehat{\varepsilon}_{(i)}^{p}.$$

Therefore, provided that δ is being conditioned on, the correct set of 'good' observations is identified. Recall that ζ is the collection of ordered indices for the set of 'good' observations,

$$\zeta = \{(i) | \delta + 1 \leq i \leq \delta + h\}.$$

The following lemma is an intermediate result on $S_{p,h\cdot\delta}$'s.

Lemma 6.1. *For $(i) \in \zeta$, let*

$$\widehat{\varepsilon}_{(i)} = \varepsilon_{(i)} - \overline{\varepsilon}_{\delta},$$

where $\overline{\varepsilon}_{\delta} = h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}$.

If Assumption 6.1 is satisfied, then conditional on δ,

$$S_{p,h\cdot\delta} = h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{p} + O_p(h^{-\frac{1}{2}}),$$

for $p \in \{2,3,4\}$.

Proof of Lemma 6.1.

For $p \in \{2,3,4\}$, we can apply the binomial rule to obtain

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\widehat{\varepsilon}_{(i)}^{p} = h^{-1}\sum_{i=\delta+1}^{\delta+h}(\varepsilon_{(i)}-\overline{\varepsilon}_{\delta})^{p} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\sum_{q=0}^{p}\binom{p}{q}(-1)^{q}\overline{\varepsilon}_{\delta}^{q}\varepsilon_{(i)}^{p-q}.$$

In the above summation term, when $q=0$, we have $h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^{p}$, whereas when $q=p$, the summand is $(-1)^{p}\overline{\varepsilon}_{\delta}^{p}$. Therefore, the above summation term can be divided into the sum of three terms as follows,

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^{p} + h^{-1}\sum_{i=\delta+1}^{\delta+h}\sum_{q=1}^{p-1}\binom{p}{q}(-1)^{q}\overline{\varepsilon}_{\delta}^{q}\varepsilon_{(i)}^{p-q} + (-1)^{p}\overline{\varepsilon}_{\delta}^{p}.$$

Since the first summation symbol in the middle term, $\sum_{i=\delta+1}^{\delta+h}$, only controls the term $\varepsilon_{(i)}^{p-q}$, we can further rearrange the above equation as

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^{p} + \sum_{q=1}^{p-1}\binom{p}{q}(-1)^{q}\overline{\varepsilon}_{\delta}^{q}h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^{p-q} + (-1)^{p}\overline{\varepsilon}_{\delta}^{p}. \tag{6.7}$$

Let $m_p := \mathbb{E}(\varepsilon_{(1)}^{p})$, and note that we have $m_0 := \mathbb{E}(\varepsilon_{(1)}^{0}) = 1$. Add and subtract $\sum_{q=0}^{p}\binom{p}{q}(-1)^{q}\overline{\varepsilon}_{\delta}^{q}m_{p-q}$ to the right-hand-side of the above equation, and rearrange to get

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^{p} + \sum_{q=1}^{p-1}\binom{p}{q}(-1)^{q}\overline{\varepsilon}_{\delta}^{q}h^{-1}\sum_{i=\delta+1}^{\delta+h}(\varepsilon_{(i)}^{p-q}-m_{p-q}) + \sum_{q=1}^{p}\binom{p}{q}(-1)^{q}\overline{\varepsilon}_{\delta}^{q}m_{p-q}.$$

Further rearranging gives

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^{p} + \sum_{q=1}^{p-1}\binom{p}{q}(-1)^q h^{-\frac{q+1}{2}}(\sqrt{h}\overline{\varepsilon}_\delta)^q h^{-\frac{1}{2}}\sum_{i=\delta+1}^{\delta+h}(\varepsilon_{(i)}^{p-q} - m_{p-q}) + \sum_{q=1}^{p}\binom{p}{q}(-1)^q h^{-\frac{q}{2}}(\sqrt{h}\overline{\varepsilon}_\delta)^q m_{p-q}. \quad (6.8)$$

Under *Assumption 6.1*, we have a sequence of i.i.d. $N(0,1)$-distributed random variables, so m_p exists and is finite. Given the good errors are i.i.d., it must follow that $\varepsilon_{(i)}^2$ and $\varepsilon_{(i)}^3$ are also i.i.d.. Invoking the Central Limit Theorem gives $h^{-\frac{1}{2}}\sum_{i=\delta+1}^{\delta+h}(\varepsilon_{(i)}^{p-q} - m_{p-q}) = O_p(1)$ and $\sqrt{h}\overline{\varepsilon}_\delta = O_p(1)$. Thus, the second and third terms in the above equation are $O_p(h^{-1})$ and $O_p(h^{-\frac{1}{2}})$, respectively. Plugging these results back into the equation to get

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^{p} + O_p(h^{-\frac{1}{2}}).$$

□

Lemma 6.1a. *If Assumption 6.1 is satisfied, then*

$$\sqrt{h}S_{3,h\cdot\delta} = h^{-\frac{1}{2}}\sum_{i=\delta+1}^{\delta+h}(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}) + O_p(h^{-\frac{1}{2}}),$$

and

$$\sqrt{h}S_{4,h\cdot\delta} = h^{-\frac{1}{2}}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^4 + O_p(h^{-\frac{1}{2}}).$$

Proof of Lemma 6.1a.

Recall from equation (6.8) that

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^{p} + \sum_{q=1}^{p-1}\binom{p}{q}(-1)^q h^{-\frac{q+1}{2}}(\sqrt{h}\overline{\varepsilon}_\delta)^q h^{-\frac{1}{2}}\sum_{i=\delta+1}^{\delta+h}(\varepsilon_{(i)}^{p-q} - m_{p-q}) + \sum_{q=1}^{p}\binom{p}{q}(-1)^q h^{-\frac{q}{2}}(\sqrt{h}\overline{\varepsilon}_\delta)^q m_{p-q}.$$

Multiplying both sides of the above equation by $\sqrt{h}$ gives

$$\sqrt{h}S_{p,h\cdot\delta} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{p} + \sum_{q=1}^{p-1} \binom{p}{q} (-1)^{q} h^{-\frac{q}{2}} (\sqrt{h}\overline{\varepsilon}_{\delta})^{q} h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} (\varepsilon_{(i)}^{p-q} - m_{p-q}) + \sum_{q=1}^{p} \binom{p}{q} (-1)^{q} h^{-\frac{q-1}{2}} (\sqrt{h}\overline{\varepsilon}_{\delta})^{q} m_{p-q}.$$

Following from *Assumption 6.1* and the Central Limit Theorem, the second term in the above equation is $O_p(h^{-\frac{1}{2}})$. Therefore, we can rewrite the above equation as

$$\sqrt{h}S_{p,h\cdot\delta} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{p} + \sum_{q=1}^{p} \binom{p}{q} (-1)^{q} h^{-\frac{q-1}{2}} (\sqrt{h}\overline{\varepsilon}_{\delta})^{q} m_{p-q} + O_p(h^{-\frac{1}{2}}).$$

Again invoking *Assumption 6.1* and the Central Limit Theorem, the middle term is of order 1 for $q = 1$ and $\frac{1}{\sqrt{h}}$ for $q > 1$. We could thus further rewrite the above equation as

$$\sqrt{h}S_{p,h\cdot\delta} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{p} - \binom{p}{1} m_{p-1} \sqrt{h}\overline{\varepsilon}_{\delta} + O_p(h^{-\frac{1}{2}}).$$

When $p = 3$, we have

$$\sqrt{h}S_{3,h\cdot\delta} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{3} - 3m_2 \sqrt{h}\overline{\varepsilon}_{\delta} + O_p(h^{-\frac{1}{2}}).$$

According to *Assumption 6.1*, m_2 is 1, and thus the above equation simplifies to

$$\sqrt{h}S_{3,h\cdot\delta} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} (\varepsilon_{(i)}^{3} - 3\varepsilon_{(i)}) + O_p(h^{-\frac{1}{2}}).$$

When $p = 4$, we have

$$\sqrt{h}S_{4,h\cdot\delta} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{4} - 4m_3 \sqrt{h}\overline{\varepsilon}_{\delta} + O_p(h^{-\frac{1}{2}}).$$

By *Assumption 6.1*, all odd moments equal zero, and therefore $m_3 = 0$, giving

$$\sqrt{h}S_{4,h\cdot\delta} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^4 + O_p(h^{-\frac{1}{2}}).$$

□

Recall that

$$\widehat{\kappa}_{3,h\cdot\delta} = \frac{S_{3,h\cdot\delta}}{S_{2,h\cdot\delta}^{3/2}} \quad \text{and} \quad \widehat{\kappa}_{4,h\cdot\delta} = \frac{S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2}{S_{2,h\cdot\delta}^2}.$$

The following result is well-known, and we shall adapt it here to the context of this chapter.

Lemma 6.2. *If Assumption 6.1 is satisfied, then*

$$\sqrt{h}\begin{bmatrix} S_{3,h\cdot\delta} \\ S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2 \end{bmatrix} \xrightarrow{D} N\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}\}.$$

Proof of Lemma 6.2.

First, consider $S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2$. Note that $(S_{2,h\cdot\delta} - 1)^2 = S_{2,h\cdot\delta}^2 - 2S_{2,h\cdot\delta} + 1$, and we have

$$\sqrt{h}(S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2) = \sqrt{h}S_{4,h\cdot\delta} - 3\sqrt{h}[(S_{2,h\cdot\delta} - 1)^2 + 2S_{2,h\cdot\delta} - 1].$$

Rearranging gives

$$\sqrt{h}(S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2)] = \sqrt{h}(S_{4,h\cdot\delta} - 6S_{2,h\cdot\delta} + 3) - 3\sqrt{h}(S_{2,h\cdot\delta} - 1)^2.$$

We could then apply Lemmas 6.1 and 6.1a, both of which are satisfied under *Assump-*

tion 6.1, to get

$$\sqrt{h}(S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2) = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} (\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3) + O_p(h^{-\frac{1}{2}}) \\ - 3\sqrt{h}[h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^2 - 1 + O_p(h^{-\frac{1}{2}})]^2. \quad (6.9)$$

Take the last term on the RHS of the above equation and rearrange to obtain

$$\sqrt{h}[h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^2 - 1 + O_p(h^{-\frac{1}{2}})]^2 = [h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} (\varepsilon_{(i)}^2 - 1) + O_p(1)] \\ \cdot [h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^2 - 1 + O_p(h^{-\frac{1}{2}})].$$

By the Law of Large Numbers and *Assumption 6.1*, $h^{-1}\sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^2$ converges in probability to 1. Thus $h^{-\frac{1}{2}}\sum_{i=\delta+1}^{\delta+h}(\varepsilon_{(i)}^2 - 1)$ is $O_p(1)$ and $h^{-1}\sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^2 - 1 = o_p(1)$. Combining these results gives

$$\sqrt{h}[h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^2 - 1 + O_p(h^{-\frac{1}{2}})]^2 = O_p(1) \times [o_p(1) + O_p(h^{-\frac{1}{2}})] = o_p(1).$$

Substitute the above equation back into equation (6.9) to obtain

$$\sqrt{h}(S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2) = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} (\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3) + o_p(1). \quad (6.10)$$

In addition, Lemma 6.1a also gives

$$\sqrt{h}S_{3,h\cdot\delta} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} (\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}) + O_p(h^{-\frac{1}{2}}).$$

Stacking $\sqrt{h}S_{3,h\cdot\delta}$ and $\sqrt{h}(S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2)$ into a vector gives

$$\sqrt{h} \begin{bmatrix} S_{3,h\cdot\delta} \\ S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2 \end{bmatrix} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} \begin{bmatrix} \varepsilon_{(i)}^3 - 3\varepsilon_{(i)} \\ \varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3 \end{bmatrix} + \mathbf{o_p}(1). \quad (6.11)$$

where $\mathbf{o_p}(1)$ is a 2×1 vector.

Since the 'good' errors are i.i.d., it follows that $\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}$ and $\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3$ are both i.i.d., with mean 0. Given that both the summands are concerned with the same observation, it follows that $\{ \begin{bmatrix} \varepsilon_{(i)}^3 - 3\varepsilon_{(i)} \\ \varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3 \end{bmatrix} \}$ is a sequence of i.i.d. random vectors, with mean vector $\mathbf{0}$. Applying the multivariate Central Limit Theorem gives

$$\sqrt{h} \begin{bmatrix} S_{3,h\cdot\delta} \\ S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2 \end{bmatrix} = h^{-\frac{1}{2}} \sum_{i=\delta+1}^{\delta+h} \begin{bmatrix} \varepsilon_{(i)}^3 - 3\varepsilon_{(i)} \\ \varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3 \end{bmatrix} + \mathbf{o_p}(1) \xrightarrow{D} N\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Omega\}, \tag{6.12}$$

where the variance covariance matrix

$$\Omega = \begin{bmatrix} Var(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}) & Cov(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}, \varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3) \\ Cov(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}, \varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3) & Var(\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3) \end{bmatrix}.$$

Under *Assumption 6.1*, $\varepsilon_{(i)}$ is i.i.d. $N(0,1)$-distributed, so we have for the diagonal elements of Ω,

$$Var(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}) = \mathbb{E}[(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)})^2] = \mathbb{E}(\varepsilon_{(i)}^6 - 6\varepsilon_{(i)}^4 + 9\varepsilon_{(i)}^2) = 15 - 6 \times 3 + 9 = 6,$$

and

$$\begin{aligned} Var(\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3) = \mathbb{E}[(\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3)^2] &= \mathbb{E}(\varepsilon_{(i)}^8 - 12\varepsilon_{(i)}^6 + 42\varepsilon_{(i)}^4 - 36\varepsilon_{(i)}^2 + 9) \\ &= 105 - 12 \times 15 + 42 \times 3 - 36 + 9 = 24. \end{aligned}$$

Recall that both $\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}$ and $\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3$ are mean 0. Since odd moments of a standard normal distribution are all equal to zero, the off-diagonal terms of Ω

$$\begin{aligned} Cov(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)}, \varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3) &= \mathbb{E}[(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)})(\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3)] \\ &\quad - \mathbb{E}[(\varepsilon_{(i)}^3 - 3\varepsilon_{(i)})]\mathbb{E}[(\varepsilon_{(i)}^4 - 6\varepsilon_{(i)}^2 + 3)] = 0. \end{aligned}$$

Thus, combining above results and plugging back into equation (6.12) yields

$$\sqrt{h}\begin{bmatrix} S_{3,h\cdot\delta} \\ S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2 \end{bmatrix} \xrightarrow{D} N\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}\}.$$

□

Lemma 6.2a. *If Assumption 6.1 is satisfied, then*

$$\sqrt{h}\begin{bmatrix} \frac{\widehat{\kappa}_{3,\delta}}{\sqrt{6}} \\ \frac{\widehat{\kappa}_{4,h\cdot\delta}}{\sqrt{24}} \end{bmatrix} \xrightarrow{D} N_2(\mathbf{0}, \mathbf{I_2})$$

where $\mathbf{0} = [0 \quad 0]'$*, and* $\mathbf{I_2}$ *is a* 2×2 *identity matrix.*

Proof of Lemma 6.2a.

Recall that

$$\sqrt{h}\frac{\widehat{\kappa}_{3,\delta}}{\sqrt{6}} = \frac{1}{S_{2,h\cdot\delta}^{3/2}} \times \frac{\sqrt{h}S_{3,h\cdot\delta}}{\sqrt{6}}$$

and

$$\sqrt{h}\frac{\widehat{\kappa}_{4,h\cdot\delta}}{\sqrt{24}} = \frac{1}{S_{2,h\cdot\delta}^{2}} \times \frac{\sqrt{h}(S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2)}{\sqrt{24}}.$$

Under *Assumption 6.1*, Lemma 6.2 is satisfied and we have

$$\sqrt{h}\begin{bmatrix} S_{3,h\cdot\delta} \\ S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2 \end{bmatrix} \xrightarrow{D} N\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}\}.$$

By the Continuous Mapping Theorem, we immediately have

$$\sqrt{h}\begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}^{-1/2}\begin{bmatrix} S_{3,h\cdot\delta} \\ S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2 \end{bmatrix} = \sqrt{h}\begin{bmatrix} \frac{S_{3,h\cdot\delta}}{\sqrt{6}} \\ \frac{S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2}{\sqrt{24}} \end{bmatrix} \xrightarrow{D} N_2(\mathbf{0}, \mathbf{I_2}). \tag{6.13}$$

Using Lemma 6.1, which holds by *Assumption 6.1*, and invoking the Law of Large

Numbers gives

$$S_{2,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^2 + O_p(h^{-\frac{1}{2}}) \xrightarrow{p} 1. \tag{6.14}$$

The proof is completed by combining equations (6.13) and (6.14) using the Slutsky's Theorem together with the Continuous Mapping Theorem. □

6.A.2 Vanishing Estimation Error

It remains to show that, $S_{p,h}$ and $S_{p,h\cdot\delta}$ (for $p \in \{2,3,4\}$) are very close to each other, provided that $\widehat{\delta}$ is a consistent estimator for δ under known h. Recall

$$S_{p,h\cdot\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\widehat{\varepsilon}_{(i)}^p = h^{-1}\sum_{i=\delta+1}^{\delta+h}(\varepsilon_{(i)} - \overline{\varepsilon}_{\delta})^p$$

and

$$S_{p,h} = h^{-1}\sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h}\widehat{\varepsilon}_{(i)}^p = h^{-1}\sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h}(\varepsilon_{(i)} - \overline{\varepsilon}_{\widehat{\delta}})^p,$$

where $\overline{\varepsilon}_{\delta} = h^{-1}\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}$ and $\overline{\varepsilon}_{\widehat{\delta}} = h^{-1}\sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h}\varepsilon_{(i)}$, respectively.

We begin by showing that the difference between the set of estimated 'good' errors, $\sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h}\varepsilon_{(i)}^p$, and the set of true 'good' errors, $\sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^p$, is of order $h^{\frac{1}{2}}$.

Lemma 6.3. *If Assumption 6.1 is satisfied, then conditional on δ,*

$$\sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h}\varepsilon_{(i)}^p - \sum_{i=\delta+1}^{\delta+h}\varepsilon_{(i)}^p = o_p(h^{\frac{1}{2}})$$

for $p \in \{1,2,3,4\}$.

Proof of Lemma 6.3.

Provided that *Assumption 6.1* is satisfied, we can apply *Theorem 5.2* in Berenguer-Rico et al. (2019) so that

$$\widehat{\delta} = \delta + o_p(h^{\frac{1}{2}-w}) \tag{6.15}$$

for some $w > 0$. Furthermore, as illustrated in their *Example 5.1*, the 'good' error

terms satisfy

$$\varepsilon_{(i)} = O_p[(2logh)^{\frac{1}{2}}]. \tag{6.16}$$

Without loss of generality, we consider the case where $\widehat{\delta} > \delta$. Let

$$D_{n,p} = \sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h} \varepsilon_{(i)}^p - \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^p.$$

Rearranging $D_{n,p}$ gives

$$D_{n,p} = (\sum_{i=\widehat{\delta}+1}^{\delta+h} \varepsilon_{(i)}^p + \sum_{i=\delta+h+1}^{\widehat{\delta}+h} \varepsilon_{(i)}^p) - (\sum_{i=\delta+1}^{\widehat{\delta}} \varepsilon_{(i)}^p + \sum_{i=\widehat{\delta}+1}^{\delta+h} \varepsilon_{(i)}^p) = \sum_{i=\delta+h+1}^{\widehat{\delta}+h} \varepsilon_{(i)}^p - \sum_{i=\delta+1}^{\widehat{\delta}} \varepsilon_{(i)}^p.$$

Note that the summands of $\sum_{i=\delta+h+1}^{\widehat{\delta}+h} \varepsilon_{(i)}^p$ are the right outliers, which satisfy

$$\varepsilon_{(i)} = \varepsilon_{(\delta+h+j)} = \varepsilon_{(\delta+h)} + \xi_{(j)} 1_{(\xi_{(j)}>0)},$$

for $i \in [\delta+h+1, \widehat{\delta}+h]$, $j \in [1, \widehat{\delta}-\delta]$, and $\varepsilon_{(\delta+h)}$ is the largest 'good' error term. Let $\tilde{\xi}_{(j)} = \xi_{(j)} 1_{(\xi_{(j)}>0)}$, and therefore we have $\varepsilon_{(\delta+h+j)} = \varepsilon_{(\delta+h)} + \tilde{\xi}_{(j)}$. We can then rewrite the expression for $D_{n,p}$ as

$$D_{n,p} = \sum_{j=1}^{\widehat{\delta}-\delta} (\varepsilon_{(\delta+h)} + \tilde{\xi}_{(j)})^p - \sum_{i=\delta+1}^{\widehat{\delta}} \varepsilon_{(i)}^p. \tag{6.17}$$

Take the second summation term on the RHS and change the index to j where $j = i-\delta$, so $\sum_{i=\delta+1}^{\widehat{\delta}} \varepsilon_{(i)}^p = \sum_{j=1}^{\widehat{\delta}-\delta} \varepsilon_{(\delta+j)}^p$. Then we can further rearrange equation (6.17) to obtain

$$D_{n,p} = \sum_{j=1}^{\widehat{\delta}-\delta} [(\varepsilon_{(\delta+h)} + \tilde{\xi}_{(j)})^p - \varepsilon_{(\delta+j)}^p].$$

Recall that $\tilde{\xi}_{(j)}$ is positive. Thus,

$$D_{n,p} \leq \sum_{j=1}^{\widehat{\delta}-\delta} [(|\varepsilon_{(\delta+h)}| + \tilde{\xi}_{(j)})^p - \varepsilon_{(\delta+j)}^p]. \tag{6.18}$$

Notice that the function $f(x) = x^k$ where $x \geq 0$ is convex for $k \geq 1$. We can invoke Jensen's inequality to get the result $(a+b)^k \leq 2^{k-1}(a^k + b^k)$, for $a, b \geq 0$. It follows that $(|\varepsilon_{(\delta+h)}| + \tilde{\xi}_{(j)})^p \leq 2^{p-1}(|\varepsilon_{(\delta+h)}|^p + \tilde{\xi}^p_{(j)})$. Substituting this inequality back into equation (6.18) yields

$$D_{n,p} \leq 2^{p-1} \sum_{j=1}^{\widehat{\delta}-\delta} (|\varepsilon_{(\delta+h)}|^p + \tilde{\xi}^p_{(j)} - \varepsilon^p_{(\delta+j)}). \tag{6.19}$$

When p is even, that is when $p \in \{2,4\}$, the last term inside the brackets, $\varepsilon^p_{(\delta+j)}$, is non-negative. The above inequality can thus be further simplified to

$$D_{n,p} \leq 2^{p-1} \sum_{j=1}^{\widehat{\delta}-\delta} (|\varepsilon_{(\delta+h)}|^p + \tilde{\xi}^p_{(j)}).$$

Notice that the summands $\tilde{\xi}_{(j)}$, where $j \in [1, \widehat{\delta} - \delta]$, must be no greater than $\tilde{\xi}_{(\widehat{\delta}-\delta)}$, and thus

$$D_{n,p} \leq 2^{p-1}(\widehat{\delta} - \delta)(|\varepsilon_{(\delta+h)}|^p + \tilde{\xi}^p_{(\widehat{\delta}-\delta)}).$$

For finite p, 2^{p-1} is bounded and therefore is $O_p(1)$. Recall from equation (6.15) that $\widehat{\delta} = \delta + o_p(h^{\frac{1}{2}-w})$. So the difference between $\widehat{\delta} - \delta = o_p(h^{\frac{1}{2}-w})$ for some $w > 0$. In addition, using the result in equation (6.16), we have $\varepsilon^p_{(\delta+h)} = O_p[(2logh)^{\frac{p}{2}}]$. Note that $log(h)$ is dominated by h^η, for any $\eta > 0$, by L'Hôpital's rule. Then, $O_p[(2logh)^{\frac{p}{2}}]$ is dominated by $o_p(h^w)$ for some $w > 0$. Furthermore, *Lemma A.1* in Berenguer-Rico et al. (2019) shows that $\tilde{\xi}_{(\widehat{\delta}-\delta)}$ is $O_p(1)$. In consequence,

$$D_{n,p} \leq O_p(1)o_p(h^{\frac{1}{2}-w})[o_p(h^w) + O_p(1)] = o_p(h^{\frac{1}{2}}).$$

It remains to evaluate the situation where p is odd, so that $p \in \{1,3\}$. Re-consider equation (6.19) and notice that the $\varepsilon^p_{(\delta+j)} \geq \varepsilon^p_{(\delta+1)}$. This gives

$$D_{n,p} \leq 4 \sum_{j=1}^{\widehat{\delta}-\delta} (|\varepsilon_{(\delta+h)}|^p + \tilde{\xi}^p_{(j)} - \varepsilon^p_{(\delta+1)}).$$

By the same type of manipulation as above, we can further rewrite the inequality as

$$D_{n,p} \leq 4(\widehat{\delta} - \delta)(|\varepsilon_{(\delta+h)}|^p + \tilde{\xi}^p_{(\widehat{\delta}-\delta)} - \varepsilon^p_{(\delta+1)}).$$

Applying the above results, $\widehat{\delta} - \delta = o_p(h^{\frac{1}{2}-w})$, $\varepsilon^p_{(i)} = o_p(h^w)$ (for $i \in \zeta$), and $\tilde{\xi}_{(\widehat{\delta}-\delta)} = O_p(1)$ yields

$$D_{n,p} \leq 4o_p(h^{\frac{1}{2}-w})[o_p(h^w) + O_p(1) + o_p(h^w)] = o_p(h^{\frac{1}{2}}).$$

□

We now proceed to show the difference between $S_{p,h}$ and $S_{p,h\cdot\delta}$ (for $p \in \{2,3,4\}$) is of order $h^{-\frac{1}{2}}$. Recall that

$$S_{p,h\cdot\delta} = h^{-1} \sum_{i=\delta+1}^{\delta+h} (\varepsilon_{(i)} - \overline{\varepsilon}_\delta)^p$$

and

$$S_{p,h} = h^{-1} \sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h} (\varepsilon_{(i)} - \overline{\varepsilon}_{\widehat{\delta}})^p,$$

where $\overline{\varepsilon}_\delta = h^{-1}\sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}$ and $\overline{\varepsilon}_{\widehat{\delta}} = h^{-1}\sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h} \varepsilon_{(i)}$, respectively.

Lemma 6.4. *If Assumption 6.1 is satisfied, then conditional on δ,*

$$S_{p,h} = S_{p,h\cdot\delta} + o_p(h^{-\frac{1}{2}}),$$

for $p \in \{2,3,4\}$.

Proof of Lemma 6.4.

Using the binomial rule and equation (6.7), we have

$$S_{p,h} = h^{-1} \sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h} (\varepsilon_{(i)} - \overline{\varepsilon}_{\widehat{\delta}})^p = h^{-1} \sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h} \varepsilon_{(i)}^p + \sum_{q=1}^{p-1} \binom{p}{q} (-1)^q \overline{\varepsilon}_{\widehat{\delta}}^q h^{-1} \sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h} \varepsilon_{(i)}^{p-q} + (-1)^p \overline{\varepsilon}_{\widehat{\delta}}^p.$$

Under *Assumption 6.1*, we can apply Lemma 6.3 to get

$$h^{-1} \sum_{i=\widehat{\delta}+1}^{\widehat{\delta}+h} \varepsilon_{(i)}^p = h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^p + o_p(h^{-\frac{1}{2}}),$$

for $p \in \{1,2,3,4\}$. So $\overline{\varepsilon}_{\widehat{\delta}} = \overline{\varepsilon}_{\delta} + o_p(h^{-\frac{1}{2}})$. Plugging these results back into $S_{p,h}$ gives

$$S_{p,h} = h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^p + o_p(h^{-\frac{1}{2}}) + \sum_{q=1}^{p-1} \binom{p}{q} (-1)^q [\overline{\varepsilon}_{\delta} + o_p(h^{-\frac{1}{2}})]^q [h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{p-q} + o_p(h^{-\frac{1}{2}})] + (-1)^p [\overline{\varepsilon}_{\delta} + o_p(h^{-\frac{1}{2}})]^p.$$

Under *Assumption 6.1*, we have a sequence of i.i.d. $N(0,1)$-distributed random variables, so all moments of $\varepsilon_{(i)}$ exist and are finite (for $(i) \in \zeta$). Thus, by the LLN, both $\overline{\varepsilon}_{\delta}$ and $h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{p-q}$ are $O_p(1)$. We can then further rearrange the above equation to get

$$\begin{aligned} S_{p,h} &= h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^p + \sum_{q=1}^{p-1} \binom{p}{q} (-1)^q \overline{\varepsilon}_{\delta}^q h^{-1} \sum_{i=\delta+1}^{\delta+h} \varepsilon_{(i)}^{p-q} + (-1)^p \overline{\varepsilon}_{\delta}^p + o_p(h^{-\frac{1}{2}}) \\ &= h^{-1} \sum_{i=\delta+1}^{\delta+h} (\varepsilon_{(i)} - \overline{\varepsilon}_{\delta})^p + o_p(h^{-\frac{1}{2}}) \\ &= S_{p,h \cdot \delta} + o_p(h^{-\frac{1}{2}}). \end{aligned}$$

□

6.A.3 Limiting Distribution Under Estimated δ

Having established above lemmas, we are now ready to prove Theorem 6.1.

Proof of Theorem 6.1.

Under *Assumption 6.1*, we can apply Lemma 6.4 to get

$$S_{p,h} = S_{p,h\cdot\delta} + o_p(h^{-\frac{1}{2}}),$$

for $p \in \{2,3,4\}$. Multiplying both sides of the above equation by $\sqrt{h}$ gives

$$\sqrt{h}S_{p,h} = \sqrt{h}S_{p,h\cdot\delta} + o_p(1).$$

Thus,

$$\sqrt{h}\begin{bmatrix} S_{3,h} \\ S_{4,h} - 3S_{2,h}^2 \end{bmatrix} = \sqrt{h}\begin{bmatrix} S_{3,h\cdot\delta} \\ S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2 \end{bmatrix} + \mathbf{o_p}(1).$$

Recall from Lemma 6.2 that

$$\sqrt{h}\begin{bmatrix} S_{3,h\cdot\delta} \\ S_{4,h\cdot\delta} - 3S_{2,h\cdot\delta}^2 \end{bmatrix} \xrightarrow{D} N\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}\}.$$

It then follows

$$\sqrt{h}\begin{bmatrix} S_{3,h} \\ S_{4,h} - 3S_{2,h}^2 \end{bmatrix} \xrightarrow{D} N\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}\}.$$

□

6.B Monte Carlo Standard Errors for Simulations

Table 6.5 Monte Carlo Standard Errors for Table 6.2

n	Skewness			Excess Kurtosis			Skewness			Excess Kurtosis		
	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975
DGP	LS.2						LS.3					
50	0.0125	0.0182	0.0228	0.0132	0.0163	0.0308	0.0125	0.0181	0.0227	0.0133	0.0143	0.0250
100	0.0125	0.0208	0.0263	0.0129	0.0197	0.0349	0.0125	0.0199	0.0252	0.0131	0.0185	0.0376
200	0.0125	0.0209	0.0251	0.0128	0.0216	0.0406	0.0125	0.0205	0.0266	0.0128	0.0203	0.0411
400	0.0125	0.0209	0.0275	0.0127	0.0234	0.0401	0.0125	0.0216	0.0286	0.0127	0.0227	0.0399
800	0.0125	0.0210	0.0275	0.0126	0.0244	0.0374	0.0125	0.0218	0.0271	0.0126	0.0210	0.0309
1600	0.0125	0.0214	0.0273	0.0126	0.0237	0.0384	0.0125	0.0220	0.0287	0.0126	0.0217	0.0346
10000	0.0125	0.0223	0.0270	0.0125	0.0212	0.0261	0.0125	0.0222	0.0275	0.0125	0.0215	0.0307
100000	0.0125	0.0205	0.0251	0.0125	0.0219	0.0283	0.0125	0.0203	0.0249	0.0125	0.0212	0.0272
DGP	LS.4						LS.5					
50	0.0125	0.0181	0.0227	0.0133	0.0143	0.0250	0.0125	0.0182	0.0227	0.0133	0.0143	0.0250
100	0.0125	0.0199	0.0252	0.0131	0.0185	0.0376	0.0125	0.0199	0.0252	0.0131	0.0185	0.0376
200	0.0125	0.0205	0.0266	0.0128	0.0203	0.0411	0.0125	0.0205	0.0266	0.0128	0.0203	0.0411
400	0.0125	0.0216	0.0286	0.0127	0.0227	0.0399	0.0125	0.0216	0.0286	0.0127	0.0227	0.0399
800	0.0125	0.0218	0.0271	0.0126	0.0210	0.0309	0.0125	0.0218	0.0271	0.0126	0.0210	0.0309
1600	0.0125	0.0220	0.0287	0.0126	0.0217	0.0346	0.0125	0.0220	0.0287	0.0126	0.0217	0.0346
10000	0.0125	0.0222	0.0275	0.0125	0.0215	0.0307	0.0125	0.0222	0.0275	0.0125	0.0215	0.0307
100000	0.0125	0.0203	0.0249	0.0125	0.0212	0.0272	0.0125	0.0203	0.0249	0.0125	0.0212	0.0272
DGP	LS.6						LS.7					
50	0.0125	0.0180	0.0239	0.0134	0.0124	0.0187	0.0125	0.0164	0.0212	0.0134	0.0137	0.0270
100	0.0125	0.0198	0.0248	0.0131	0.0162	0.0269	0.0125	0.0188	0.0231	0.0131	0.0178	0.0332
200	0.0126	0.0215	0.0282	0.0129	0.0198	0.0343	0.0125	0.0204	0.0263	0.0128	0.0222	0.0423
400	0.0125	0.0212	0.0272	0.0127	0.0223	0.0357	0.0125	0.0201	0.0258	0.0127	0.0234	0.0430
800	0.0125	0.0216	0.0276	0.0127	0.0222	0.0355	0.0125	0.0212	0.0268	0.0127	0.0233	0.0382
1600	0.0125	0.0223	0.0265	0.0126	0.0238	0.0354	0.0125	0.0217	0.0257	0.0126	0.0241	0.0365
10000	0.0125	0.0210	0.0253	0.0125	0.0216	0.0300	0.0125	0.0206	0.0250	0.0125	0.0218	0.0302
100000	0.0125	0.0207	0.0263	0.0125	0.0220	0.0277	0.0125	0.0207	0.0263	0.0125	0.0221	0.0277

Table 6.6 Monte Carlo Standard Errors for Table 6.3

n	Skewness			Excess Kurtosis			Skewness			Excess Kurtosis		
	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975
						$d_2 = 0$						
		$d_1 = 0.5$						$d_1 = 1$				
50	0.0125	0.0182	0.0228	0.0133	0.0144	0.0252	0.0125	0.0169	0.0207	0.0134	0.0134	0.0234
100	0.0125	0.0199	0.0252	0.0131	0.0185	0.0376	0.0125	0.0189	0.0231	0.0130	0.0183	0.0326
200	0.0125	0.0205	0.0266	0.0128	0.0203	0.0411	0.0125	0.0190	0.0247	0.0129	0.0216	0.0398
400	0.0125	0.0216	0.0286	0.0127	0.0227	0.0399	0.0125	0.0199	0.0250	0.0127	0.0215	0.0364
800	0.0125	0.0218	0.0271	0.0126	0.0210	0.0309	0.0125	0.0214	0.0253	0.0126	0.0218	0.0352
1600	0.0125	0.0220	0.0287	0.0126	0.0217	0.0346	0.0125	0.0211	0.0257	0.0126	0.0223	0.0321
10000	0.0125	0.0222	0.0275	0.0125	0.0215	0.0307	0.0125	0.0217	0.0268	0.0125	0.0221	0.0279
100000	0.0125	0.0203	0.0249	0.0125	0.0212	0.0272	0.0125	0.0203	0.0252	0.0125	0.0217	0.0285
		$d_1 = 2$						$d_1 = 3$				
50	0.0125	0.0173	0.0211	0.0133	0.0138	0.0271	0.0125	0.0167	0.0201	0.0133	0.0127	0.0191
100	0.0125	0.0191	0.0238	0.0131	0.0167	0.0291	0.0125	0.0188	0.0237	0.0130	0.0163	0.0297
200	0.0125	0.0204	0.0275	0.0129	0.0219	0.0413	0.0125	0.0208	0.0256	0.0129	0.0205	0.0389
400	0.0125	0.0204	0.0279	0.0128	0.0230	0.0480	0.0125	0.0211	0.0279	0.0127	0.0227	0.0387
800	0.0125	0.0208	0.0257	0.0127	0.0248	0.0434	0.0125	0.0214	0.0286	0.0126	0.0221	0.0379
1600	0.0125	0.0207	0.0255	0.0126	0.0244	0.0354	0.0125	0.0214	0.0280	0.0126	0.0231	0.0327
10000	0.0125	0.0206	0.0243	0.0125	0.0239	0.0326	0.0125	0.0214	0.0250	0.0125	0.0230	0.0338
100000	0.0125	0.0214	0.0265	0.0125	0.0218	0.0300	0.0125	0.0216	0.0259	0.0125	0.0211	0.0274
						$d_2 = 1$						
		$d_1 = 0.5$						$d_1 = 1$				
50	0.0125	0.0167	0.0207	0.0133	0.0136	0.0230	0.0125	0.0181	0.0220	0.0133	0.0131	0.0221
100	0.0125	0.0196	0.0244	0.0131	0.0162	0.0272	0.0125	0.0193	0.0251	0.0130	0.0180	0.0320
200	0.0125	0.0206	0.0252	0.0129	0.0188	0.0340	0.0125	0.0224	0.0298	0.0128	0.0225	0.0441
400	0.0125	0.0213	0.0270	0.0127	0.0228	0.0391	0.0125	0.0204	0.0259	0.0127	0.0227	0.0389
800	0.0125	0.0210	0.0276	0.0127	0.0237	0.0409	0.0125	0.0219	0.0262	0.0126	0.0230	0.0390
1600	0.0125	0.0211	0.0267	0.0126	0.0247	0.0374	0.0125	0.0203	0.0256	0.0126	0.0228	0.0347
10000	0.0125	0.0218	0.0280	0.0126	0.0222	0.0323	0.0125	0.0206	0.0260	0.0125	0.0210	0.0268
100000	0.0125	0.0208	0.0271	0.0125	0.0221	0.0281	0.0125	0.0208	0.0251	0.0125	0.0215	0.0292
		$d_1 = 2$						$d_1 = 3$				
50	0.0125	0.0177	0.0213	0.0133	0.0143	0.0256	0.0125	0.0170	0.0217	0.0133	0.0131	0.0197
100	0.0125	0.0188	0.0249	0.0130	0.0171	0.0327	0.0125	0.0192	0.0265	0.0131	0.0166	0.0288
200	0.0125	0.0195	0.0255	0.0129	0.0207	0.0374	0.0125	0.0184	0.0238	0.0129	0.0195	0.0309
400	0.0125	0.0214	0.0247	0.0127	0.0228	0.0402	0.0125	0.0201	0.0242	0.0127	0.0208	0.0354
800	0.0125	0.0219	0.0290	0.0126	0.0229	0.0378	0.0125	0.0200	0.0233	0.0127	0.0221	0.0328
1600	0.0125	0.0210	0.0274	0.0126	0.0228	0.0338	0.0125	0.0198	0.0255	0.0126	0.0216	0.0314
10000	0.0125	0.0213	0.0285	0.0125	0.0206	0.0282	0.0125	0.0220	0.0281	0.0125	0.0212	0.0280
100000	0.0125	0.0221	0.0290	0.0125	0.0212	0.0270	0.0125	0.0217	0.0265	0.0125	0.0208	0.0259
						$d_2 = 2$						
		$d_1 = 0.5$						$d_1 = 1$				
50	0.0125	0.0173	0.0225	0.0133	0.0139	0.0265	0.0125	0.0188	0.0231	0.0133	0.0136	0.0220
100	0.0125	0.0192	0.0250	0.0130	0.0195	0.0395	0.0125	0.0191	0.0250	0.0130	0.0186	0.0348
200	0.0125	0.0200	0.0252	0.0128	0.0206	0.0388	0.0125	0.0202	0.0258	0.0129	0.0204	0.0405
400	0.0125	0.0205	0.0260	0.0127	0.0224	0.0376	0.0125	0.0209	0.0254	0.0128	0.0225	0.0409
800	0.0125	0.0207	0.0250	0.0126	0.0230	0.0351	0.0125	0.0217	0.0271	0.0127	0.0231	0.0390
1600	0.0125	0.0209	0.0271	0.0126	0.0233	0.0339	0.0125	0.0216	0.0265	0.0126	0.0218	0.0346
10000	0.0125	0.0223	0.0274	0.0126	0.0216	0.0290	0.0125	0.0208	0.0251	0.0125	0.0213	0.0293
100000	0.0125	0.0222	0.0295	0.0125	0.0207	0.0291	0.0125	0.0214	0.0278	0.0125	0.0217	0.0280
		$d_1 = 2$						$d_1 = 3$				
50	0.0125	0.0169	0.0200	0.0134	0.0139	0.0215	0.0125	0.0174	0.0213	0.0134	0.0130	0.0203
100	0.0125	0.0171	0.0207	0.0130	0.0189	0.0334	0.0125	0.0193	0.0254	0.0131	0.0169	0.0299
200	0.0125	0.0199	0.0250	0.0129	0.0206	0.0396	0.0125	0.0199	0.0256	0.0129	0.0209	0.0386
400	0.0125	0.0208	0.0247	0.0127	0.0209	0.0365	0.0125	0.0206	0.0254	0.0128	0.0233	0.0388
800	0.0125	0.0216	0.0271	0.0126	0.0254	0.0374	0.0125	0.0206	0.0261	0.0126	0.0235	0.0379
1600	0.0125	0.0213	0.0284	0.0126	0.0239	0.0372	0.0125	0.0203	0.0261	0.0126	0.0227	0.0350
10000	0.0125	0.0209	0.0268	0.0125	0.0216	0.0284	0.0125	0.0211	0.0270	0.0125	0.0217	0.0313
100000	0.0125	0.0202	0.0244	0.0125	0.0216	0.0267	0.0125	0.0203	0.0257	0.0125	0.0212	0.0278

Table 6.7 Monte Carlo Standard Errors for Table 6.4

n	Skewness			Excess Kurtosis			Skewness			Excess Kurtosis		
	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975	0.5	0.95	0.975
					$d_2 = 0$							
	$d_1 = 0.5$						$d_1 = 1$					
50	0.0143	0.0257	0.0330	0.0176	0.0063	0.0068	0.0132	0.0207	0.0249	0.0139	0.0096	0.0128
100	0.0176	0.0448	0.0576	0.0214	0.0064	0.0070	0.0134	0.0244	0.0306	0.0137	0.0107	0.0145
200	0.0264	0.1099	0.1481	0.0285	0.0062	0.0062	0.0138	0.0290	0.0392	0.0135	0.0128	0.0179
400	0.0595	0.4148	0.5893	0.0466	0.0056	0.0049	0.0141	0.0315	0.0401	0.0133	0.0148	0.0187
800	0.2588	3.8532	6.1667	0.1034	0.0056	0.0041	0.0142	0.0370	0.0472	0.0131	0.0157	0.0222
1600	4.0347	175.9801	354.9799	0.3750	0.0064	0.0039	0.0140	0.0384	0.0510	0.0128	0.0170	0.0228
10000	1.034E+11	3.462E+15	2.054E+16	3111.0199	0.1363	0.0292	0.0133	0.0337	0.0434	0.0126	0.0186	0.0253
100000	1.493E+88	7.159E+115	4.948E+119	1.196E+25	4.429E+7	9.061E+4	0.0127	0.0273	0.0368	0.0126	0.0202	0.0265
	$d_1 = 2$						$d_1 = 3$					
50	0.0125	0.0172	0.0216	0.0134	0.0129	0.0212	0.0125	0.0169	0.0201	0.0133	0.0128	0.0233
100	0.0125	0.0187	0.0247	0.0131	0.0173	0.0283	0.0125	0.0189	0.0228	0.0131	0.0173	0.0322
200	0.0125	0.0205	0.0263	0.0129	0.0196	0.0377	0.0125	0.0192	0.0246	0.0129	0.0189	0.0342
400	0.0125	0.0204	0.0265	0.0127	0.0220	0.0371	0.0125	0.0194	0.0241	0.0127	0.0222	0.0444
800	0.0125	0.0218	0.0256	0.0126	0.0238	0.0368	0.0125	0.0194	0.0242	0.0127	0.0221	0.0372
1600	0.0125	0.0220	0.0279	0.0126	0.0238	0.0390	0.0125	0.0205	0.0265	0.0126	0.0245	0.0389
10000	0.0125	0.0221	0.0283	0.0125	0.0226	0.0305	0.0125	0.0224	0.0299	0.0125	0.0232	0.0314
100000	0.0125	0.0217	0.0282	0.0125	0.0217	0.0297	0.0125	0.0216	0.0290	0.0125	0.0205	0.0289
					$d_2 = 1$							
	$d_1 = 0.5$						$d_1 = 1$					
50	0.0133	0.0215	0.0263	0.0140	0.0087	0.0102	0.0125	0.0185	0.0220	0.0133	0.0125	0.0187
100	0.0141	0.0281	0.0345	0.0141	0.0095	0.0126	0.0125	0.0192	0.0242	0.0130	0.0159	0.0254
200	0.0157	0.0383	0.0447	0.0144	0.0103	0.0134	0.0125	0.0198	0.0262	0.0129	0.0192	0.0324
400	0.0186	0.0611	0.0807	0.0147	0.0106	0.0133	0.0125	0.0203	0.0252	0.0127	0.0188	0.0339
800	0.0240	0.1057	0.1598	0.0152	0.0102	0.0133	0.0125	0.0200	0.0230	0.0127	0.0207	0.0311
1600	0.0340	0.2825	0.4054	0.0157	0.0095	0.0111	0.0125	0.0208	0.0262	0.0126	0.0217	0.0330
10000	0.2366	16.5820	36.9016	0.0195	0.0077	0.0079	0.0125	0.0216	0.0282	0.0126	0.0220	0.0286
100000	284.5966	6.241E+7	6.975E+8	0.0382	0.0059	0.0051	0.0125	0.0223	0.0279	0.0125	0.0212	0.0279
	$d_1 = 2$						$d_1 = 3$					
50	0.0125	0.0176	0.0223	0.0133	0.0140	0.0242	0.0125	0.0172	0.0225	0.0133	0.0137	0.0236
100	0.0125	0.0185	0.0230	0.0131	0.0173	0.0353	0.0125	0.0191	0.0246	0.0131	0.0172	0.0294
200	0.0125	0.0191	0.0244	0.0129	0.0198	0.0366	0.0125	0.0204	0.0242	0.0129	0.0224	0.0464
400	0.0125	0.0197	0.0241	0.0128	0.0237	0.0454	0.0125	0.0195	0.0254	0.0127	0.0217	0.0365
800	0.0125	0.0205	0.0262	0.0126	0.0237	0.0360	0.0125	0.0212	0.0276	0.0127	0.0237	0.0377
1600	0.0125	0.0211	0.0263	0.0126	0.0232	0.0338	0.0125	0.0219	0.0281	0.0126	0.0221	0.0309
10000	0.0125	0.0216	0.0266	0.0125	0.0226	0.0304	0.0125	0.0221	0.0269	0.0126	0.0213	0.0278
100000	0.0125	0.0224	0.0293	0.0125	0.0209	0.0301	0.0125	0.0197	0.0245	0.0126	0.0214	0.0280
					$d_2 = 2$							
	$d_1 = 0.5$						$d_1 = 1$					
50	0.0125	0.0182	0.0246	0.0134	0.0127	0.0202	0.0125	0.0174	0.0216	0.0133	0.0140	0.0218
100	0.0126	0.0201	0.0250	0.0131	0.0159	0.0254	0.0125	0.0183	0.0232	0.0130	0.0163	0.0310
200	0.0126	0.0237	0.0296	0.0129	0.0177	0.0307	0.0125	0.0207	0.0280	0.0128	0.0209	0.0367
400	0.0126	0.0232	0.0292	0.0128	0.0196	0.0300	0.0125	0.0208	0.0281	0.0127	0.0203	0.0347
800	0.0127	0.0248	0.0316	0.0127	0.0195	0.0300	0.0125	0.0216	0.0275	0.0126	0.0214	0.0328
1600	0.0127	0.0255	0.0314	0.0126	0.0210	0.0299	0.0125	0.0220	0.0275	0.0126	0.0225	0.0334
10000	0.0128	0.0274	0.0362	0.0126	0.0187	0.0255	0.0125	0.0211	0.0254	0.0125	0.0222	0.0321
100000	0.0130	0.0338	0.0431	0.0126	0.0194	0.0242	0.0125	0.0202	0.0281	0.0125	0.0213	0.0293
	$d_1 = 2$						$d_1 = 3$					
50	0.0125	0.0169	0.0207	0.0134	0.0131	0.0214	0.0125	0.0169	0.0201	0.0133	0.0128	0.0233
100	0.0125	0.0186	0.0246	0.0131	0.0173	0.0284	0.0125	0.0189	0.0228	0.0131	0.0173	0.0322
200	0.0125	0.0204	0.0262	0.0129	0.0196	0.0377	0.0125	0.0192	0.0246	0.0129	0.0189	0.0342
400	0.0125	0.0204	0.0265	0.0127	0.0220	0.0371	0.0125	0.0194	0.0241	0.0127	0.0222	0.0444
800	0.0125	0.0218	0.0256	0.0126	0.0238	0.0368	0.0125	0.0194	0.0242	0.0127	0.0221	0.0372
1600	0.0125	0.0220	0.0279	0.0126	0.0238	0.0390	0.0125	0.0205	0.0265	0.0126	0.0245	0.0389
10000	0.0125	0.0221	0.0283	0.0125	0.0226	0.0305	0.0125	0.0224	0.0299	0.0125	0.0232	0.0314
100000	0.0125	0.0217	0.0282	0.0125	0.0217	0.0297	0.0125	0.0216	0.0290	0.0125	0.0205	0.0289

CHAPTER 7

Model Selection in LTS Regressions

7.1 Introduction

In Chapter 6, we studied a Jarque-Bera type test to assess the normality assumption on the 'good' errors in the LTS model. In this chapter, we shall discuss the model selection problem specific to LTS regressions and this relates to the selection of h, the number of 'good' observations. The two selection criteria — one is information based and the other is cumulant based — considered here build heavily on the normality assumption. Provided that the normality specification of the 'good' errors is correct, these two selection methods may consistently select the true model (with some additional assumptions). We examine some limit properties of these two model selection methods and derive conditions under which they give consistent results.

The plan of this chapter is set out as follows. The general LTS model is formally defined in Section 7.2 . In Section 7.3, we develop some asymptotic results using the simple location-scale model. We first give conditions for consistent model selection using the information criteria (IC) method in Section 7.3.1. Inconsistency conditions will also be shown. We present the consistency theorem using the cumulant based normality test statistic in Section 7.3.2. Section 7.4 compares the ability of consistent model selection using these two methods in location-scale models as well as regression models through Monte-Carlo simulation studies. Some concluding remarks are given in Section 7.5.

7.2 The General LTS Model

In this section, we give a formal description of the LTS model proposed by Berenguer-Rico et al. (2019) and the LTS estimator discussed in the previous chapters.

Consider the linear regression model for a sample $\{y_i, x_i\}_{i=1}^n$ with a dependent variable y_i and a vector of k regressors x_i

$$y_i = x_i'\beta + \sigma\varepsilon_i, \quad i = 1, \ldots, n,$$

where $\beta \in \mathbb{R}^k$ is the slope coefficient vector and $\sigma \in \mathbb{R}_{>0}$ is the scale parameter. Similar to Chapter 6, ζ forms the set of indices for non-outlying errors, which contains $h \leq n$ elements. The error terms follow the same specification as that given in *Assumption 6.1*. So for $i \in \zeta$, the 'good' errors, ε_i, are $N(0,1)$-distributed and mutually independent. Conditioning on the 'good' errors, the outliers, ε_j (for $j \notin \zeta$), are independent and may come from different contamination distributions, G_j. These G_j distributions could be dependent on the regressors, so this model allows for vertical and/or leverage outliers. The parameters of this model are therefore β, σ, ζ, which is any h-subset of $1, \ldots, n$, and the G_j's, which are arbitrary distributions assumed continuous at zero. A key characteristic of this model specification is that the ranges of realised values of the 'good' and 'bad' error terms are well separated.

Taking h as known, the LTS estimator is formally defined as

$$\widehat{\beta}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} r_{(i)}^2(\beta), \tag{7.1}$$

where, for a given value of β, $r_{(i)}^2(\beta)$ is the i-th ordered statistic among $r_1^2(\beta), \ldots, r_n^2(\beta)$ with $r_i(\beta) = y_i - x_i'\beta$. The minimisation of the objective function can also be viewed as a process in which every time one chooses a subsample of h observations and finds some β that minimises the RSS for the selected subsample. We obtain $\binom{n}{h}$ values of β's if we repeat this for every possible subsample. The one associated with the

smallest value of the objective function is the LTS estimate. Let $\widehat{\zeta}_h$ denote the LTS estimator of ζ. Therefore, β is the OLS estimator in that set of observations whose indices belong to the set $\widehat{\zeta}_h$. The LTS estimate of the scale parameter is then

$$\widehat{\sigma}_{LTS} = \{h^{-1} \sum_{i \in \widehat{\zeta}_h} (y_i - x_i' \widehat{\beta}_{LTS})^2\}^{1/2}. \tag{7.2}$$

Under *Assumption 6.1*, Berenguer-Rico et al. (2019) show that $\widehat{\beta}_{LTS}$ and $\widehat{\sigma}_{LTS}$ are equivalent to the maximum likelihood estimates for β and σ, respectively, in the LTS model. Analogous to the discussion in Chapter 6, while the LTS estimator of β in equation (7.1) is the same as that in Rousseeuw (1984), the LTS estimator of σ in equation (7.2) is not.

7.3 The Location-Scale Model and the Model Selection Problem

In this section, we consider the same location-scale model as that studied in Chapter 6. We shall extend the methods discussed in this section to regression models involving regressors later in the simulations. We expect that the same asymptotic results apply to general regression models, yet formal proof is left for future work.

Consider the following location scale model for observations y_1 to y_n,

$$y_i = \mu + \sigma \varepsilon_i, \qquad i = 1, \ldots, n. \tag{7.3}$$

As discussed in the previous section, assuming h as known, we obtain the LTS estimates for the parameters in equation (7.3) by deriving $\widehat{\zeta}_h$. Following from the discussion in Chapter 6, this in turn is equivalent to estimating the number of left 'outliers', $\delta = \sum_{j \notin \zeta} 1_{(\varepsilon_j < \min_{i \in \zeta} \varepsilon_i)}$.

Recall that $\varepsilon_{(i)}$ denote the error terms in increasing order. If h is given, the estima-

tor for δ is

$$\widehat{\delta}_h := \underset{0\leq\delta\leq n-h}{\arg\min} \sum_{i=\delta+1}^{\delta+h} \{y_{(i)} - \widehat{\mu}_\delta\}^2 \qquad \text{where} \quad \widehat{\mu}_\delta = h^{-1} \sum_{i=\delta+1}^{\delta+h} y_{(i)}. \tag{7.4}$$

Accordingly, the LTS estimators for μ and σ are

$$\widehat{\mu}_{LTS} = h^{-1} \sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h} y_{(i)} = \bar{y}_{LTS}, \tag{7.5}$$

and

$$\widehat{\sigma}_{LTS} = \{h^{-1} \sum_{i=\widehat{\delta}_h+1}^{\widehat{\delta}_h+h} (y_{(i)} - \bar{y}_{LTS})^2\}^{1/2}. \tag{7.6}$$

The LTS method detects the set of outliers taking the number of good observations, h, as given. However, h is usually unknown in reality and thus needs to be estimated. Recall from Section 5.4 in Chapter 5 that we have a range of models to select from depending on the values of l. Here, l represents the possible sizes of the set of good observations and we have $2 \leq l \leq n$. The aim is to find the model with the lowest value of the selected criterion. The estimator of the true h, $\widehat{h}$, is thus defined as

$$\widehat{h} := \underset{l\in[\underline{l},\bar{l}]}{\arg\min}\, \Phi(l), \tag{7.7}$$

where $\Phi(\cdot)$ is some predefined criterion function, $\underline{l}$ and $\bar{l}$ are the lower and upper bounds of the subset of values of l satisfying $2 \leq \underline{l} < \bar{l} \leq n$. While in theory we could choose $\underline{l} = 2$ and $\bar{l} = n$, in empirical analysis we might focus on a sub-range of $[2,n]$, i.e., $[0.5 \times n, n]$.

Note that in the robust literature, one usually assumes $\underline{l} \geq \frac{n}{2}$, and therefore we cannot have more outliers than 'good' observations in the data. This assumption is usually made with respect to the ε-contaminated framework. Indeed, if more than half of the data are drawn from the contaminated distribution and less than half of the data are drawn from the normal distribution, we could flip our thinking as to which observa-

tions are 'abnormal' and maintain the assumption of less than half of the data being outliers. So we could treat the 'contaminated' distribution that generates more than $\frac{n}{2}$ outliers as the 'good' distribution, and accordingly treat the normal distribution as the contaminated distribution which generates less than $\frac{n}{2}$ 'bad' observations. Swapping the labels is possible in the ε-contaminated framework, because these errors are assumed to be i.i.d..

This renaming, however, is not feasible in the LTS framework studied in this thesis. By Assumption 6.1, the error terms are not i.i.d, but only the 'good' ones are. There is a dependence structure between the outlying errors and the 'good' errors considered here, since the former is generated conditioning on the latter. While it is infeasible to swap labels of the 'good' and 'bad' errors, it is possible to allow for $\underline{l} < \frac{n}{2}$ in our model. In order for this to work, we would need to impose some regularity assumptions on the contamination distributions G_j's [see Assumption 5.2 in Berenguer-Rico et al. (2019) for details]. Berenguer-Rico et al. (2019) showed that using Assumption 5.2 in their paper and Assumption 6.1 in this thesis, the standard asymptotic theory for $\widehat{\mu}_{LTS}$ and $\widehat{\sigma}_{LTS}$ still applies when more than half of the data are outliers.

Ideally, we hope that $\widehat{h}$ converges in probability to the true h. However, since h increases with the sample size, it diverges as n increases, and therefore there is no probability limit. Instead, we consider an alternative quantity — the fraction of 'good' observations, which is a constant because h and n both grow at the same rate. Let $\gamma_l = l/n$, and accordingly $\gamma_h = h/n$, so γ_h is the true fraction of 'good' errors. Using equation (7.7), we could define its estimator $\widehat{\gamma}$ as

$$\widehat{\gamma} := \frac{1}{n} \underset{l \in [\underline{l}, \bar{l}]}{\arg\min}\, \Phi(l) = \frac{\widehat{h}}{n}. \tag{7.8}$$

We suggest two criterion methods $\Phi(\cdot)$ so that

$$plim_{n\to\infty} \widehat{\gamma} = \gamma_h.$$

Therefore, the criteria $\Phi(\cdot)$ yield a consistent estimate of γ_h if for all $\eta > 0$, we have both

$$\mathbb{P}\{\Phi(l) > \Phi(h) | \gamma_h - \gamma_l > \eta\} \to 1 \text{ as } n \to \infty$$

and

$$\mathbb{P}\{\Phi(l) > \Phi(h) | \gamma_l - \gamma_h > \eta\} \to 1 \text{ as } n \to \infty$$

Let

$$\mathbb{S}_{l,h} := \Phi(l) - \Phi(h). \tag{7.9}$$

Equivalently, we can write the above conditions as $inf_l \mathbb{S}_{l,h} > 0$ $w.p.1$ for $|\gamma_l - \gamma_h| > \eta$.

In what follows, we first consider consistent estimation of the true γ using the IC method, and later discuss an alternative selection strategy that utilises the cumulant based normality test statistic.

7.3.1 The Information Criteria Method

A general form of the IC can be written as the sum of a loss term and a penalty term:

$$\Phi_{IC}(l) = log(\widehat{\sigma}_l^2) + \frac{n-l}{n} f(n), \tag{7.10}$$

where $\widehat{\sigma}_l$ is the LTS estimator of σ for a given value of l,

$$\widehat{\sigma}_l = \{h^{-1} \sum_{i \in \widehat{\zeta}_l} (y_{(i)} - h^{-1} \sum_{i \in \widehat{\zeta}_l} y_{(i)})^2\}^{1/2}, \tag{7.11}$$

similar to that defined above in equation (7.6). Here, $\widehat{\zeta}_l$ is the estimator of ζ if the selected sub-sample is of size l. With the assumed normality of the 'good' errors, this coincides with the MLE. Here $n-l$ is the number of outliers, which is the same as the number of the contamination distributions, G_j. Thus, $n-l$ is approximately equal to the number of parameters (although strictly speaking, the number of parameters is $n-l+3$, taking μ, σ, and δ into account).

The penalty function, $f(n)$, is a known, non-stochastic function of the sample size, n. If $f(n) = 2$, this gives the classic Akaike Information Criterion (AIC) (Akaike (1973)). The well-known Bayesian Information Criterion (BIC) introduced by Schwartz (1978) uses the penalty $f(n) = log(n)$. Another common choice for the penalty is the $2loglog(n)$ suggested by Hannan and Quinn (1979).

The question here is how this penalty should be chosen such that it yields a consistent selection of the true model, the model with size h. The main problem is that the dimension of the parameter space, $n-l$, is no longer a constant, unlike in the standard model selection problem. So the number of candidate models changes subject to the sample size, which is unbounded. As $n \to \infty$, we will be selecting the true model from an infinite number of models. We need a penalty that approaches infinity but does not diverge faster than some constant multiple of $loglog(n)$, as established in the following theorem.

Theorem 7.1. *If Assumption 6.1 is satisfied and the second moment of ξ_j is finite, then the estimate*

$$\widehat{\gamma} := \frac{1}{n} \arg\min_{l \in [\underline{l}, \bar{l}]} \Phi_{IC}(l)$$

converges in probability to γ_h, provided that

1. *$f(n) \to \infty$ as $n \to \infty$, and*
2. *$f(n) \leq c\{loglog(n)\}$, where $0 < c < \frac{1}{\gamma_{\bar{l}} - \gamma_{\underline{l}}}$.*

The $loglog(n)$ rate in the above theorem comes from the extreme value theorem for normal distributions [see, for example, Fisher and Tippett (1928) and Leadbetter et al. (1983)]. Under Assumption 6.1, the largest 'good' error grows at approximately $\sqrt{2log(n)}$ rate. This implies that $\widehat{\sigma}_l^2$ is of order $2log(n)$, and thus $log(\widehat{\sigma}_l^2)$ is approximately of order $loglog(n)$.

The upper bound for the choice of the constant c, depends on both the interval $[\underline{l}, \bar{l}]$ and how precise we would like to be with this rate. A sketch of the proof is given

in Appendix 7.A, where we find the smallest upper bound for c being $\frac{1}{\gamma_{\bar{l}}-\gamma_{\underline{l}}}$, however, a formal proof is left for future work. If $\bar{l}=n$ and the true γ_h were known, then the smallest upper bound would be $\frac{1}{1-\gamma_h}$. Berenguer-Rico et al. (2019) found the smallest upper bound to be 2, if $[\underline{l},\bar{l}]=[0.5n,n]$.

The rate at which the penalty approaches infinity is what determines the consistency of $\widehat{\gamma}$. In particular, we require $\frac{f(n)(\gamma_{\bar{l}}-\gamma_{\underline{l}})}{loglog(n)} \to 0$ as $n \to \infty$. If this requirement fails to hold, we will not be able to consistently estimate γ_h, as stated in the next theorem (with proof in Appendix 7.B).

Theorem 7.2. *If Assumption 6.1 is satisfied and the second moment of ξ_j is finite, then the estimator,*

$$\widehat{\gamma} := \frac{1}{n} \arg\min_{l\in[\underline{l},\bar{l}]} \Phi_{IC}(l),$$

is inconsistent for γ_h, if $f(n) > c\{loglog(n)\}$, where $c > \frac{1}{\gamma_{\bar{l}}-\gamma_{\underline{l}}}$.

Under the same assumptions as in Theorem 7.1, Theorem 7.2 states that if the penalty function, $f(n)$, is such that

$$\frac{f(n)(\gamma_{\bar{l}}-\gamma_{\underline{l}})}{loglog(n)} \to \infty \text{ as } n \to \infty,$$

$\Phi_{IC}(l)$ will always choose a larger model and we will not be able to consistently estimate γ_h.

Therefore, the choice of $f(n)$ is crucial to the consistent estimation of γ_h. Here, we discuss a few examples of $f(n)$ to illustrate the above two theorems.

Table 7.1 Comparison of Different $f(n)$ Penalties

$f(n)$	Consistency	Direction of Inconsistency
2	Inconsistent	Lower Bound
$log(n)$	Inconsistent	Upper Bound
$loglog(n)$	Consistent	–

The first penalty in Table 7.1 corresponds to the one used in the AIC. We can see that with $f(n)=2$, the resulting estimator of γ_h is inconsistent towards the lower bound

of the values of γ under analysis. This is because the requirement in Theorem 7.1 that $f(n)$ needs to diverge to infinity as n increases is violated.

In contrast, when $f(n) = log(n)$, which is the penalty used in the BIC, $\widehat{\gamma}$ is inconsistent towards the upper bound of γ, since it diverges too quickly. While BIC is asymptotically consistent in many model selection scenarios, it is inconsistent in the LTS model selection problem. In a standard model selection where one selects among a finite set of nested models, BIC consistently selects the true model [Stone (1979)] as the penalty penalises an over-parametrised model even more as the dataset size n increases, and therefore pick up the correct model for large samples. However, in the LTS case, not only the set of candidate models grows with n but also the complexity of the model increases with n. Thus, the penalisation coefficient, which depends on the dimension of the parameter space, $n-l$, is no longer a finite constant. Theorem 7.2 shows that BIC tends to favour bigger models than the true one in the problem considered here, as the $log(n)$ penalty is too harsh. This also implies that, any faster divergence rate than $log(n)$ is not favoured, and we require $c\{loglog(n)\} < log(n)$ for consistent model selection.

If we consider the last penalty in Table 7.1, $loglog(n)$, then the corresponding IC will yield consistent estimation, as the $f(n)$ function converge to infinity at a slower rate, and thus weaken the effect of the penalty as $n \to \infty$.

7.3.2 The Cumulant Based Normality Test Statistic Method

Employing a similar logic to the normality test discussed in Chapter 6, we construct a criterion function based on the cumulant based normality test statistic to select the correct LTS model.

Let

$$S_{p,l} = l^{-1} \sum_{i=\widehat{\delta}_l+1}^{\widehat{\delta}_l+l} \widehat{\varepsilon}_{(i)}^{p}, \quad \text{for} \quad p \in \{2,3,4\},$$

where $\widehat{\delta}_l$ is the LTS estimator of δ if the set of optimal observations is of size l. Then,

for an LTS model with size l, the criterion function is based on the sample skewness and excess kurtosis of the estimated 'good' errors,

$$\widehat{\kappa}_{3,l} = \frac{S_{3,l}}{S_{2,l}^{3/2}} \quad \text{and} \quad \widehat{\kappa}_{4,l} = \frac{S_{4,l} - 3S_{2,l}^2}{S_{2,l}^2},$$

respectively.

The resulting criterion function is

$$\Phi_{T_h}(l) = l\frac{\widehat{\kappa}_{3,l}^2}{6} + l\frac{\widehat{\kappa}_{4,l}^2}{24}. \tag{7.12}$$

Unlike the IC method, whose consistency depends upon the chosen penalty, the Φ_{T_h} is a consistent selection tool under the LTS model, as given in the following theorem.

Theorem 7.3. *If Assumption 6.1 is satisfied, then the estimator*

$$\widehat{\gamma} := \frac{1}{n} \operatorname*{arg\,min}_{l \in [\underline{l}, \bar{l}]} \Phi_{T_h}(l)$$

is consistent for γ_h.

Intuitively, if $l = h$, then from equation (7.9), we have

$$\mathbb{S}_{l,h} = \Phi_{T_h}(l) - \Phi_{T_h}(h) = 0.$$

When $l < h$, using the results in Berenguer-Rico et al. (2019), we can show that $\widehat{\zeta}_l$ is asymptotically a 'subset' of the 'good' observations. The observations in $\widehat{\zeta}_l$, instead of following a complete normal distribution, follow a truncated normal. However, the normalisations of 6 and 24 in equation (7.12) assume that the selected observations are drawn from a complete normal distribution. Berenguer-Rico and Nielsen (2017) showed that when the normal distribution is truncated, the usual normality test has the power to reject the null hypothesis. This is because the truncated normal distribution belongs to models under the alternative hypothesis. This implies that, when $l < h$, $\mathbb{S}_{l,h}$

must be strictly greater than zero, since the criterion function is based on the complete normality assumption.

On the other hand, if $l > h$, asymptotically $\widehat{\zeta}_l$ will pick up all the 'good' observations as well as some bad ones. Given that some outliers are included, the cumulant based normality test criterion will also have the power to reject the null hypothesis that the selected sample follows a normal distribution, and therefore $\mathbb{S}_{l,h} > 0$.

A sketch of proof for Theorem 7.3 is provided in Appendix 7.C, but the formal proof is left for future work.

7.4 Simulation — Consistent Model Selection

In this section, we carry out various simulation studies to evaluate the finite sample properties of $\widehat{h}$ and $\widehat{\gamma}$. We are interested in the model selection ability of the IC using different penalties as well as the cumulant based normality test statistic criterion. Our goal is to check how different criteria select the optimal model from all the candidate models as the sample size increases. Throughout, we shall fix γ_h at 0.8. We start with a simple location-scale model, and later extend the study to regression models.

7.4.1 A Location-Scale Model

Let ζ be a set with h elements, where $h = 0.8 \times n$. Here, we consider an example with right outliers only, so all the outliers lie strictly to the right of the realised range of the good errors. The DGPs for the error terms are described in Table 7.2.

Table 7.2 DGP for ε — Model Selection

DGP	ε
LS	For $i \in \zeta$: $\varepsilon_i \sim i.i.d.N(0,1)$ For $j \notin \zeta$: $\varepsilon_j = \max_{i\in\zeta} \varepsilon_i + 1 + 3\xi_j$, where $\xi_j \sim i.i.d.\chi^2_3$

We estimate the location-scale model

$$y_i = \mu + \sigma\varepsilon_i, \qquad i = 1,\ldots,n, \tag{7.13}$$

where the true μ is 0 and the true σ is 1. For $l \in [0.5n, n]$, we estimate each model by LTS and estimate h by

$$\widehat{h} = \underset{l \in [0.5n,n]}{\arg\min}\, \Phi(l), \tag{7.14}$$

where $\Phi(l)$ could be $\Phi_{T_h}(l)$ or $\Phi_{IC}(l)$. So, we restrict the analysis to $\gamma_l \in [0.5, 1]$ and $\widehat{\gamma} = \widehat{h}/n$.

The Algorithm

1. For a given sample of size n, sort the data ($\varepsilon_i \to \varepsilon_{(i)}$).
2. Calculate the partial sums $\sum_{i=1}^{s} \varepsilon_{(i)}$ and $\sum_{i=1}^{s} \varepsilon_{(i)}^2$, for $s = 1, \ldots, n$.
3. For a given $l \in [0.5n, n]$, find $\widehat{\delta}_l$ by

$$\widehat{\delta}_l = \underset{0 \le \delta \le n-l}{\arg\min}\, RSS = \underset{0 \le \delta \le n-l}{\arg\min} \{ \sum_{i=1}^{\delta+l} \varepsilon_{(i)}^2 - \sum_{i=1}^{\delta} \varepsilon_{(i)}^2 - \frac{1}{l} [\sum_{i=1}^{\delta+l} \varepsilon_{(i)} - \sum_{i=1}^{\delta} \varepsilon_{(i)}]^2 \}$$

and compute the values of various $\Phi(l)$.

4. Apply the criterion function in equation (7.14) to select the optimal model. That is, find $\widehat{h}$ by finding the value of l that minimises $\Phi(l)$ over the range from 50% to 100% of n.

Computational Issues

We consider multiple (n, M) combinations, where n is the sample size (ranging from 50 to 800000), and M is the number of replications (M=1, 100, or 1000). The number of replications is small in this study due to the limit of computational power. In step 3 of the algorithm discussed above, for each value of $l \in [0.5n, n]$, we need to store $\binom{n}{l}$ candidate values of δ_l in the background. In order to produce $\widehat{h}$, we need to store a total of $\sum_{l=0.5n}^{n} \binom{l}{n}$ for a given sample size. This number increases nonlinearly with n, resulting in an 'out of power' error. While the small replication number affects the precision of the estimated results, it provides useful implications of the upper and lower bounds of the penalty functions that we are interested in.

Results

We examine two types of comparisons. First, we compare the performance of

$\widehat{h}$ obtained by $\Phi_{T_h}(l)$ as well as $\Phi_{IC}(l)$ using different penalty functions. Table 7.3 summarises the model selection results of these criteria. The first column is the sample size, and the second column is the value of the true h for a given n. We consider the cumulant based normality test statistic (T_h), and the ICs using penalties given by $f(n) = loglog(n)$, $log(n)$, $log(n^2)$, and $log(n^3)$. For each criterion, two statistics, $\widehat{h}_{bias}$ and $\widehat{h}_{sd}$, are reported, which are the Monte Carlo average and standard deviation of the estimation error $\widehat{h} - h$. While we report here $\widehat{h}_{bias}$ and $\widehat{h}_{sd}$, note that we could consistently estimate γ_h, but not h, according to the discussions in Section 7.3.

We observe that, in general, the cumulant based test statistic criterion works better than IC. It can be seen that the finite sample bias with IC using $f(n)$ increasing to infinity more rapidly blows up more quickly than those with 'slower' $f(n)$. While the penalties $log(n^2)$ and $log(n^3)$ work quite well in relatively small samples, in samples where n is greater than 10000, they tend to select overfitted models. For the IC with more liberal penalties on the model size, such as $log(n)$ and $loglog(n)$, there is considerable under-fitting even when the sample size is very large. Table 7.3 suggests that while theoretically $loglog(n)$ is a better choice than $log(n)$, since the latter would result in inconsistent estimates of γ_h, in practice the former can be too conservative, as shown in Table 7.4. Table 7.4 reports the values of $log(n)$ and $loglog(n)$ for different sample sizes. We can see that even when $n = 800000$, $log(n)$ is smaller than 15 and $loglog(n)$ barely exceeds 2.5, suggesting that the rate at which the logarithm function diverges is rather slow in reality. Therefore, the $loglog(n)$ penalty may not be much better than $log(n)$ for the DGP considered. Generally, $\Phi_{T_h}(l)$ is more robust than the $\Phi_{IC}(l)$ as the sample size varies.

In Table 7.5, the estimated bias and empirical standard errors of $\widehat{\gamma}$ are presented. Among all methods, there is a common tendency that the estimated values obtained using $\Phi_{T_h}(l)$ and IC with the $log(n)$-penalty are approaching the true value (and meanwhile mean square errors are decreasing as the sample size increases). In contrast, the $loglog(n)$-penalty grows at such a slow rate that we could hardly see any improvement

Table 7.3 Simulated Bias and Standard Deviation for $\widehat{h}$ — Comparison 1

		T_h		$loglog(n)$		$log(n)$		$log(n^2)$		$log(n^3)$	
n	h	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$
M=1000											
50	40	-3.8990	4.9994	-15.0000	0.0000	-12.8820	4.2301	-0.6730	1.2797	0.3170	1.3944
100	80	-2.4170	5.7755	-30.0000	0.0000	-16.5520	11.1141	-0.4750	0.8149	0.9270	3.0471
200	160	-0.9800	2.4065	-60.0000	0.0000	-11.9910	11.8967	-0.4170	0.6853	5.3420	11.7049
400	320	-0.8110	1.7100	-120.0000	0.0000	-10.4930	5.2716	-0.4290	0.6598	52.2310	32.0237
800	640	-0.9000	1.9328	-240.0000	0.0000	-12.9100	5.0612	-0.4260	0.6834	149.9950	4.7983
1600	1280	-0.8880	2.0673	-480.0000	0.0000	-16.4670	5.3059	-0.4440	0.6598	306.2800	4.9138
M=100											
10000	8000	-0.9700	2.4056	-3000.0000	0.0000	-37.9700	6.7726	1540.4100	701.5800	1962.6700	7.9570
M=1											
100000	80000	-2.0000	NA	-30000.0000	NA	-138.0000	NA	19413.0000	NA	19842.0000	NA
500000	400000	2.0000	NA	-150000.0000	NA	-298.0000	NA	98175.0000	NA	99552.0000	NA
800000	640000	-6.0000	NA	-240000.0000	NA	-378.0000	NA	157348.0000	NA	159365.0000	NA

Table 7.4 Logarithm Values

n	$log(n)$	$loglog(n)$
50	3.9120	1.3641
100	4.6052	1.5272
200	5.2983	1.6674
400	5.9915	1.7903
800	6.6846	1.8998
1600	7.3778	1.9985
10000	9.2103	2.2203
100000	11.5129	2.4435
500000	13.1224	2.5743
800000	13.5924	2.6095

as n increases, as supported by the results in Tables 7.3 and 7.4. In addition, neither the $log(n^2)$ or $log(n^3)$-penalty yields a satisfactory outcome, since the bias becomes more positive as n shoots up. In general, the estimate using $\Phi_{T_h}(l)$ is more consistent and becomes more accurate compared with $\Phi_{IC}(l)$ estimates as n grows. The performance of IC depends heavily on the chosen penalty as well as the sample size.

Table 7.5 Simulated Bias and Standard Deviation for $\widehat{\gamma}$ — Comparison 1

		T_h		$loglog(n)$		$log(n)$		$log(n^2)$		$log(n^3)$	
n	h	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}sd$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$
M=1000											
50	40	-0.0780	0.1000	-0.3000	0.0000	-0.2576	0.0846	-0.0135	0.0256	0.0063	0.0279
100	80	-0.0242	0.0578	-0.3000	0.0000	-0.1655	0.1111	-0.0048	0.0081	0.0093	0.0305
200	160	-0.0049	0.0120	-0.3000	0.0000	-0.0600	0.0595	-0.0021	0.0034	0.0267	0.0585
400	320	-0.0020	0.0043	-0.3000	0.0000	-0.0262	0.0132	-0.0011	0.0016	0.1306	0.0801
800	640	-0.0011	0.0024	-0.3000	0.0000	-0.0161	0.0063	-0.0005	0.0009	0.1875	0.0060
1600	1280	-0.0006	0.0013	-0.3000	0.0000	-0.0103	0.0033	-0.0003	0.0004	0.1914	0.0031
M=100											
10000	8000	-0.0001	0.0002	-0.3000	0.0000	-0.0038	0.0007	0.1540	0.0702	0.1963	0.0008
M=1											
100000	80000	0.0000	NA	-0.3000	NA	-0.0014	NA	0.1941	NA	0.1984	NA
500000	400000	0.0000	NA	-0.3000	NA	-0.0006	NA	0.1964	NA	0.1991	NA
800000	640000	0.0000	NA	-0.3000	NA	-0.0005	NA	0.1967	NA	0.1992	NA

Secondly, we assess the ability to consistently select the correct model using different $c\{loglog(n)\}$ rate penalties. We examine several possible values of the constant c to provide some clues for how to select the c that yields consistent estimation of γ_h in Theorem 7.1. We consider $c = 0.5$, 1, 2, 4, 8, and 16.

Table 7.6 and Table 7.7 report the simulation results for $\widehat{h}$ and $\widehat{\gamma}$, respectively. We observe that the more liberal choices of c (0.5 and 1) suffer from considerable under-fitting, whereas the more aggressive choices of c (8 and 16) lead to overfitted models. As the sample size increases, there is no improvement in $c = 1$ over 0.5, confirming that the rate at which $loglog(n)$ diverges is too slow in practice. When $c = 8$, $c\{loglog(n)\}$ performs similar to the $log(n^2)$ in the above comparison, which diverges too quickly as the sample size increases, indicating that c should be smaller than 8 for this particular DGP. This is in accordance with the theoretical findings in Section 7.3.1. Theorem 7.1 establishes that, for consistent estimation, the upper bound for c needs to be smaller than $\frac{1}{1-\gamma_h}$. Hence, with $\gamma_h = 0.8$, the upper bound should be smaller than 5. Therefore, any c that is greater than 5 would result in inconsistent estimation. The simulated results suggest that $c = 2$ and 4 seem to be the appropriate choices here, with the latter performing slightly better than the former, which provide further evidence that c should be smaller than 5.

Table 7.6 Simulated Bias and Standard Deviation for $\widehat{h}$ — Comparison 2

		$0.5loglog(n)$		$loglog(n)$		$2loglog(n)$		$4loglog(n)$		$8loglog(n)$		$16loglog(n)$	
n	h	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$
M=1000													
50	40	-15.0000	0.0000	-15.0000	0.0000	-14.9920	0.1482	-4.3000	4.8732	0.1450	1.1495	9.0790	1.7233
100	80	-30.0000	0.0000	-30.0000	0.0000	-29.9720	0.5191	-3.1290	3.8677	0.1140	1.1800	19.2120	1.1456
200	160	-60.0000	0.0000	-60.0000	0.0000	-59.9470	0.9127	-3.4100	2.6459	0.0060	1.0388	38.9360	1.2375
400	320	-120.0000	0.0000	-120.0000	0.0000	-119.8500	3.0051	-4.8850	3.0101	-0.1050	0.5461	78.4080	1.4218
800	640	-240.0000	0.0000	-240.0000	0.0000	-239.8660	4.2375	-7.4620	3.5523	-0.1560	0.5347	157.5230	1.7874
1600	1280	-480.0000	0.0000	-480.0000	0.0000	-478.7380	18.3584	-11.6600	4.2140	-0.2380	0.5307	315.9820	2.2634
M=100													
10000	8000	-3000.0000	0.0000	-3000.0000	0.0000	-1498.2300	817.5273	-44.7500	6.9912	90.9100	401.4985	1984.3300	4.1318
M=1													
100000	80000	-30000.0000	NA	-30000.0000	NA	-6210.0000	NA	-288.0000	NA	18984.0000	NA	19903.0000	NA
500000	400000	-150000.0000	NA	-150000.0000	NA	-23931.0000	NA	-1103.0000	NA	95918.0000	NA	99619.0000	NA
800000	640000	-240000.0000	NA	-240000.0000	NA	-36196.0000	NA	-1639.0000	NA	153628.0000	NA	159422.0000	NA

Table 7.7 Simulated Bias and Standard Deviation for $\widehat{\gamma}$ — Comparison 2

		$0.5loglog(n)$		$loglog(n)$		$2loglog(n)$		$4loglog(n)$		$8loglog(n)$		$16loglog(n)$	
n	h	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$
M=1000													
50	40	-0.3000	0.0000	-0.3000	0.0000	-0.2998	0.0030	-0.0860	0.0975	0.0029	0.0230	0.1816	0.0345
100	80	-0.3000	0.0000	-0.3000	0.0000	-0.2997	0.0052	-0.0313	0.0387	0.0011	0.0118	0.1921	0.0115
200	160	-0.3000	0.0000	-0.3000	0.0000	-0.2997	0.0046	-0.0171	0.0132	0.0000	0.0052	0.1947	0.0062
400	320	-0.3000	0.0000	-0.3000	0.0000	-0.2996	0.0075	-0.0122	0.0075	-0.0003	0.0014	0.1960	0.0036
800	640	-0.3000	0.0000	-0.3000	0.0000	-0.2998	0.0053	-0.0093	0.0044	-0.0002	0.0007	0.1969	0.0022
1600	1280	-0.3000	0.0000	-0.3000	0.0000	-0.2992	0.0115	-0.0073	0.0026	-0.0001	0.0003	0.1975	0.0014
M=100													
10000	8000	-0.3000	0.0000	-0.3000	0.0000	-0.1498	0.0818	-0.0045	0.0007	0.0091	0.0401	0.1984	0.0004
M=1													
100000	80000	-0.3000	NA	-0.3000	NA	-0.0621	NA	-0.0029	NA	0.1898	NA	0.1990	NA
500000	400000	-0.3000	NA	-0.3000	NA	-0.0479	NA	-0.0022	NA	0.1918	NA	0.1992	NA
800000	640000	-0.3000	NA	-0.3000	NA	-0.0452	NA	-0.0020	NA	0.1920	NA	0.1993	NA

7.4.2 A Linear Regression Model

To show the effectiveness of the estimation methods, we consider a more general case by modifying the model in equation (7.13):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the true β_0 and β_1 are 0 and 1, respectively. The DGP for the error terms is the same as that described in Table 7.2. The regressors are generated as follows:

- For $i \in \zeta$: $x_i \sim i.i.d.U[-5,5]$.
- For $j \notin \zeta$: $x_j = \max_{i \in \zeta} x_i + 3 + v_j$, where $v_j \sim i.i.d.\chi^2_3$.

In this way, we have constructed a model with bad leverage points. The model is estimated using the *ltsregres* function in the 'MATLAB Library for Robust Analysis' toolbox with $\widehat{h}$ obtained by minimising the objective functions, $\Phi(l)$, and $\widehat{\gamma} = \widehat{h}/n$. With the location-scale model, the derivation of the LTS estimators follows a simpler algorithm, and therefore, the computation burden is smaller. Nonetheless, it still soaks up computer memories. Concerning the complexity of the model selection problem and the computation burden with the presence of regressors, instead of searching for the optimal l among a continuous interval, we evaluate $\Phi(l)$ at certain percentiles of n. Specifically, we consider $\gamma_l \in \Gamma$, where

$$\Gamma = \{0.55, 0.6, 0.75, 0.76, 0.77, 0.78, 0.79, 0.8, 0.81, 0.82, 0.83, 0.84, 0.85, 0.9, 0.95, 1\}.$$

Given the discontinuous nature of Γ, we examine a more extensive set of (n, M) combinations here. Tables 7.8 and 7.9 summarise the finite sample properties of $\widehat{h}$ and $\widehat{\gamma}$ obtained using $\Phi_{IC}(l)$ with different penalty functions as well as $\Phi_{T_h}(l)$. Note that some of the standard errors being 0.000 does not mean that the variances are exactly zero, but that we are evaluating among a discrete range of values.

Table 7.8 Simulated Bias and Standard Deviation for $\widehat{h}$ — Regression

		T_h		$loglog(n)$		$log(n)$		$log(n^2)$		$log(n^3)$	
n	h	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$	$\widehat{h}_{bias}$	$\widehat{h}_{sd}$
m=1000											
100	80	0.0310	4.9863	-25.0000	0.0000	-16.0850	9.9758	1.2200	4.0898	11.5830	5.8377
300	240	-0.3240	4.1637	-75.0000	0.0000	-9.8430	5.5848	11.5350	17.8283	46.5000	5.1114
500	400	-0.9350	3.0348	-125.0000	0.0000	-11.4250	4.9415	36.6150	34.3969	79.4000	9.5253
800	640	-1.2640	3.2119	-200.0000	0.0000	-12.9920	5.3610	103.0960	38.1802	131.1200	17.9295
1000	800	-1.1700	3.8400	-250.0000	0.0000	-14.6400	5.6482	140.4500	34.1682	167.2000	23.7639
2000	1600	-0.9800	4.3195	-500.0000	0.0000	-20.3800	3.0108	300.0000	0.0000	380.8000	39.4070
3000	2400	-0.3300	3.1306	-750.0000	0.0000	-29.8800	1.8945	450.0000	0.0000	596.7000	22.0135
4000	3200	-0.1200	2.1887	-1000.0000	0.0000	-39.5600	4.1742	600.0000	0.0000	799.6000	8.9398
5000	4000	-0.1000	2.2349	-1250.0000	0.0000	-48.3000	9.0660	750.0000	0.0000	1000.0000	0.0000
8000	6400	0.0000	0.0000	-2000.0000	0.0000	-66.3200	30.1358	1200.0000	0.0000	1600.0000	0.0000
10000	8000	0.0000	0.0000	-2500.0000	0.0000	-65.1000	47.6892	1500.0000	0.0000	2000.0000	0.0000
15000	12000	0.0000	0.0000	-3750.0000	0.0000	-15.9000	46.1987	2254.5000	57.9492	3000.0000	0.0000
20000	16000	0.0000	0.0000	-5000.0000	0.0000	-1.6000	17.8258	3039.0000	193.6918	4000.0000	0.0000
25000	20000	0.0000	0.0000	-6250.0000	0.0000	0.0000	0.0000	3973.7500	479.4302	5000.0000	0.0000
30000	24000	0.0000	0.0000	-7500.0000	0.0000	0.0000	0.0000	5119.5000	738.9288	6000.0000	0.0000
35000	28000	0.0000	0.0000	-8750.0000	0.0000	0.0000	0.0000	6464.5000	806.8555	7000.0000	0.0000
m=10											
50000	40000	0.0000	0.0000	-12500.0000	0.0000	0.0000	0.0000	10000.0000	0.0000	10000.0000	0.0000
100000	80000	0.0000	0.0000	-25000.0000	0.0000	0.0000	0.0000	20000.0000	0.0000	20000.0000	0.0000
m=1											
1000000	800000	0.0000	NA	-250000.0000	NA	150000.0000	NA	200000.0000	NA	200000.0000	NA

Table 7.9 Simulated Bias and Standard Deviation for $\widehat{\gamma}$ — Regression

		T_h		$loglog(n)$		$log(n)$		$log(n^2)$		$log(n^3)$	
n	h	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}sd$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$	$\widehat{\gamma}_{bias}$	$\widehat{\gamma}_{sd}$
m=1000											
100	80	0.0003	0.0499	-0.2500	0.0000	-0.1609	0.0998	0.0122	0.0409	0.1158	0.0584
300	240	-0.0011	0.0139	-0.2500	0.0000	-0.0328	0.0186	0.0385	0.0594	0.1550	0.0170
500	400	-0.0019	0.0061	-0.2500	0.0000	-0.0229	0.0099	0.0732	0.0688	0.1588	0.0191
800	640	-0.0016	0.0040	-0.2500	0.0000	-0.0162	0.0067	0.1289	0.0477	0.1639	0.0224
1000	800	-0.0012	0.0038	-0.2500	0.0000	-0.0146	0.0056	0.1405	0.0342	0.1672	0.0238
2000	1600	-0.0005	0.0022	-0.2500	0.0000	-0.0102	0.0015	0.1500	0.0000	0.1904	0.0197
3000	2400	-0.0001	0.0010	-0.2500	0.0000	-0.0100	0.0006	0.1500	0.0000	0.1989	0.0073
4000	3200	0.0000	0.0005	-0.2500	0.0000	-0.0099	0.0010	0.1500	0.0000	0.1999	0.0022
5000	4000	0.0000	0.0004	-0.2500	0.0000	-0.0097	0.0018	0.1500	0.0000	0.2000	0.0000
8000	6400	0.0000	0.0000	-0.2500	0.0000	-0.0083	0.0038	0.1500	0.0000	0.2000	0.0000
10000	8000	0.0000	0.0000	-0.2500	0.0000	-0.0065	0.0048	0.1500	0.0000	0.2000	0.0000
15000	12000	0.0000	0.0000	-0.2500	0.0000	-0.0011	0.0031	0.1503	0.0039	0.2000	0.0000
20000	16000	0.0000	0.0000	-0.2500	0.0000	-0.0001	0.0009	0.1520	0.0097	0.2000	0.0000
25000	20000	0.0000	0.0000	-0.2500	0.0000	0.0000	0.0000	0.1590	0.0192	0.2000	0.0000
30000	24000	0.0000	0.0000	-0.2500	0.0000	0.0000	0.0000	0.1707	0.0246	0.2000	0.0000
35000	28000	0.0000	0.0000	-0.2500	0.0000	0.0000	0.0000	0.1847	0.0231	0.2000	0.0000
m=100											
50000	40000	0.0000	0.0000	-0.2500	0.0000	0.0000	0.0000	0.2000	0.0000	0.2000	0.0000
100000	80000	0.0000	0.0000	-0.2500	0.0000	0.0000	0.0000	0.2000	0.0000	0.2000	0.0000
m=1											
1000000	800000	0.0000	NA	-0.2500	NA	0.1500	NA	0.2000	NA	0.2000	NA

The results are similar to those for the location-scale model reported in the previous subsection. Firstly, both the $log(n^2)$ and $log(n^3)$ penalties lead to overfitted models in this DGP, and the bias becomes more significant as the sample size increases. On the contrary, the criterion with $loglog(n)$ penalty is too liberal in practice, and the estimated bias for $\widehat{\gamma}$ is fixed at -0.25, which means the smallest possible model is always selected. One can see that $\widehat{\gamma}$ obtained using the $log(n)$-penalty approaches the true value as n increases to 100000. However, the bias becomes positive as n continues to rise to 1000000, suggesting that there is a turning point where this criterion starts to select an overfitted model. This agrees with the results in Theorem 7.2 that $log(n)$ penalty is too aggressive to produce consistent estimates.

Overall, $\Phi_{T_h}(l)$ performs uniformly better than $\Phi_{IC}(l)$, regardless of the choice of $f(n)$

7.5 Conclusion

In this chapter, we provide some theoretical justifications for consistent model selection in the location-scale LTS model, which is about pinning down the true proportion of outliers in the dataset. We show that, under the LTS model in Assumption 6.1, the cumulant based normality test statistic method consistently selects the true model. Consistency of IC holds provided the penalty term satisfies the weak requirements that $f(n) \to \infty$ and $f(n)/c\{loglog(n)\} \to 0$ as $n \to \infty$ for some constant c, under an additional assumption on the second moment of the contamination distribution. The rate condition implies that there will be a breakdown in consistency if the penalty increases too fast relative to $loglog(n)$. Simulations confirm the model selection consistency and that the cumulant based normality test statistic method works better than IC.

Note that the results in this chapter built heavily on Assumption 6.1, and more specifically, the normality assumption on the 'good' errors. When the normality assumption is slightly violated, the IC approach might still be valid since the normal distribution based maximum likelihood procedure may still yield consistent results. However, it may not be legitimate to use the cumulant based approach to choose the optimal sample size, since the normalisations in the cumulant based normality test statistic are based on the normality assumption. With slight deviation from normality, it may require adjusting with consistency factors to deliver consistent model selection. If the methods for testing the normality assumption indicate a significant departure, we recommend one modelling a more specific error structure given the data and using alternative approaches to estimate the model. For example, one could begin with an LTS estimator with a small pre-defined value of l and use the Forward Search algorithm [Atkinson et al. (2010)] to find the optimal sample size of 'good' observations.

If it is known that the i.i.d. 'good' errors follow a uniform distribution, one could use the LMS estimation, rather than the LTS method, to estimate the optimal model.

This chapter leaves several directions open for future work. First, asymptotic results are formulated in the context of a location-scale model. Simulation studies suggest that it is possible to extend these theoretical results to models with stochastic regressors. A formal proof is left for future research. Second, information criteria with different penalty terms can be further investigated as the rate at which the $loglog(n)$ penalty diverges is quite slow in reality. Finally, the inference based on the model selected by these two selection methods have not been examined and discussed, which could be included in future work.

Appendix

7.A Proof of Theorem 7.1

Sketch of Proof for Theorem 7.1.

Recall that $\gamma_h = h/n$ and $\gamma_l = l/n$. Then equation (7.10) can be rewritten as

$$\Phi_{IC}(l) = log(\widehat{\sigma}_l^2) + (1-\gamma_l)f(n),$$

and therefore

$$\mathbb{S}_{l,h} = \Phi(l) - \Phi(h) = log(\frac{\widehat{\sigma}_l^2}{\widehat{\sigma}_h^2}) - (\gamma_l - \gamma_h)f(n). \tag{7.15}$$

The key is to show that $\inf_l \mathbb{S}_{l,h} > 0$ *w.p.*1 when $|\gamma_l - \gamma_h| > \eta$, for some $\eta > 0$.

When $l < h$ such that $\gamma_l - \gamma_h < 0$, provided that $f(n) \to +\infty$ as $n \to \infty$, the second term of the above equation diverges. So long as the first term does not diverge to negative infinity, the infimum of $\mathbb{S}_{l,h}$ is strictly positive with probability one. To demonstrate this, we can first use Lemma A.4 in Berenguer-Rico et al. (2019) to show that asymptotically $\widehat{\zeta}_l$ is strictly a subset of ζ. If l selects a subset of the true 'good' observations asymptotically, then $\widehat{\zeta}_l$ returns a truncated normal distribution. Using the results in Butler (1982), we have $\widehat{\sigma}_l^2 \xrightarrow{p} \sigma_l^2 < \sigma^2$. Since $\widehat{\sigma}_h^2 \xrightarrow{p} \sigma^2$, it follows that $log(\widehat{\sigma}_l^2/\widehat{\sigma}_h^2) \xrightarrow{p} log(\sigma_l^2/\sigma^2)$, where $-\infty < log(\sigma_l^2/\sigma^2) < 0$ if l is not too small such that σ_l^2 is not too close to zero.

If $l > h$ such that $\gamma_l - \gamma_h > \eta$, we expect that ζ is asymptotically a subset of $\widehat{\zeta}_l$

using Lemma A.8 in Berenguer-Rico et al. (2019). Thus, for some $\iota > 0$, we have

$$\mathbb{P}\{|\frac{\#(\widehat{\zeta}_l \cap \zeta)}{h} - 1| > \iota\} \to 0 \quad \text{as} \quad n \to \infty. \tag{7.16}$$

Consequently, $\widehat{\zeta}_l$ not only contains all the 'good' observations asymptotically but also some outliers, and we have

$$\widehat{\sigma}_l^2 = \frac{1}{l}\sum_{i\in\zeta} \varepsilon_{(i)}^2 - \frac{1}{l}\sum_{\substack{i\in\zeta \\ \text{but } i\notin\widehat{\zeta}_l}} \varepsilon_{(i)}^2 + \frac{1}{l}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)} \varepsilon_{(i)}^2 - (\frac{1}{l}\sum_{i\in\widehat{\zeta}_l} \varepsilon_{(i)})^2, \tag{7.17}$$

where ζ^c is the complement of the set ζ. From equation (7.16), there are ι 'good' observations that do not belong to $\widehat{\zeta}_l$ in finite samples. Using the Lemma A.7 in Berenguer-Rico et al. (2019), the second term in the above equation converges in probability to zero as $\iota \to 0$, and therefore is of order $o_p(1)$. Since the first term in equation (7.17) is non-negative, it follows that

$$\widehat{\sigma}_l^2 \geq \frac{1}{l}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)} \varepsilon_{(i)}^2 - (\frac{1}{l}\sum_{i\in\widehat{\zeta}_l} \varepsilon_{(i)})^2.$$

Under Assumption 6.1, we have $\max_{i\in\zeta} \varepsilon_{(i)} = O_p[(2log(h))^{1/2}]$. The outliers must grow (at least) at the same rate as the largest 'good' observation, since there is no overlap in their supports, giving $\frac{1}{(l-h)}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)} \varepsilon_{(i)}^2 = O_p[2log(h)]$. Take the last term in the above equation, and notice that

$$\frac{1}{l}\sum_{i\in\widehat{\zeta}_l} \varepsilon_{(i)} = \frac{1}{l}\sum_{i\in\zeta} \varepsilon_{(i)} + \frac{1}{l}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)} \varepsilon_{(i)},$$

where the first term on the RHS converges in probability to zero. Now consider the second term, and suppose that all the outlying observations are positive so there is no left 'outliers'. Then for $i \in (\widehat{\zeta}_l \cap \zeta^c)$, $\varepsilon_{(i)} = O_p[(2log(h))^{1/2}] + \xi_i$, where $\xi_i > 0$. Under the finite second moment assumption, ξ_i must be bounded in probability and

therefore is of order $O_p(1)$. Thus, the second term in the above equation is of order $\frac{l-h}{l}\sqrt{2log(h)}$ by Assumption 6.1, giving

$$\widehat{\sigma}_l^2 \geq O_p[\frac{l-h}{l}2log(h) - (\frac{l-h}{l})^2 2log(h)] = O_p[\frac{l-h}{l}(1-\frac{l-h}{l})2log(h)].$$

Note that $O_p[log(h)] = O_p[log(\gamma_h n)] = O_p[log(n)]$, which implies that $\widehat{\sigma}_l^2$ is at least of order $\frac{l-h}{l}(1-\frac{l-h}{l})2log(n)$. Observe that $\widehat{\sigma}_h^2 = O_p(1)$, which gives

$$log(\widehat{\sigma}_l^2/\widehat{\sigma}_h^2) \geq log\{\frac{l-h}{l}(1-\frac{l-h}{l})2log(n)\} \approx loglog(n).$$

Combining above results yields

$$\mathbb{S}_{l,h} \geq loglog(n) - (\gamma_l - \gamma_h)f(n).$$

In order to consistently estimate γ_h, we require the infimum of $\mathbb{S}_{l,h}$ to be greater than 0 $w.p.1$. Therefore a sufficient condition is $loglog(n) > (\gamma_l - \gamma_h)f(n)$, which implies the penalty function should satisfy

$$f(n) < \frac{loglog(n)}{\gamma_l - \gamma_h}.$$

This inequality should hold for any $\gamma_l - \gamma_h > \eta$ satisfying $\underline{l} \leq h < l \leq \bar{l}$. Observe that $\gamma_l - \gamma_h \leq \gamma_{\bar{l}} - \gamma_{\underline{l}}$, so the lower bound for $(\gamma_l - \gamma_h)^{-1}$ is $(\gamma_{\bar{l}} - \gamma_{\underline{l}})^{-1}$. This in turn requires

$$f(n) < \frac{loglog(n)}{\gamma_{\bar{l}} - \gamma_{\underline{l}}},$$

which is equivalent to that $c < \frac{1}{\gamma_{\bar{l}} - \gamma_{\underline{l}}}$. □

7.B Proof of Theorem 7.2

Sketch of Proof for Theorem 7.2.

The goal is to show that when $f(n) > \frac{loglog(n)}{\gamma_l - \gamma_h}$, $\Phi_{IC}(l)$ gives inconsistent estimate of γ_h. From the proof of Theorem 7.1, it suffices to consider the case where $l > h$. Recall equation (7.17) that

$$\widehat{\sigma}_l^2 = \frac{1}{l}\sum_{i\in\zeta}\varepsilon_{(i)}^2 - \frac{1}{l}\sum_{\substack{i\in\zeta \\ \text{but } i\notin\widehat{\zeta}_l}}\varepsilon_{(i)}^2 + \frac{1}{l}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)}\varepsilon_{(i)}^2 - (\frac{1}{l}\sum_{i\in\widehat{\zeta}_l}\varepsilon_{(i)})^2.$$

Since both the second and the last terms are non-negative, it follows that

$$\widehat{\sigma}_l^2 \leq \frac{1}{l}\sum_{i\in\zeta}\varepsilon_{(i)}^2 + \frac{1}{l}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)}\varepsilon_{(i)}^2. \tag{7.18}$$

The first term on the RHS is of order $\frac{h}{l}$ which is equivalent to $\frac{\gamma_h}{\gamma_l}$. Now take the second term, and again we shall consider the case where we have only right 'outliers'. So for $i \in (\widehat{\zeta}_l \cap \zeta^c)$, $\varepsilon_{(i)} = O_p[(2log(h))^{1/2}] + \xi_i$, where $\xi_i > 0$. Recall that $O_p[log(h)] = O_p[log(n)]$, so

$$\begin{aligned}\frac{1}{l}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)}\varepsilon_{(i)}^2 = \frac{1}{l}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)}\{O_p[(2log(n))^{1/2}] + \xi_i\}^2 \\ \leq O_p[\frac{l-h}{l}2log(n)] + \frac{1}{l}\sum_{i\in(\widehat{\zeta}_l\cap\zeta^c)}\xi_i^2.\end{aligned}$$

where the inequality follows from the fact that $(a+b)^2 \leq a^2+b^2$. Under the assumption that ξ_i has finite second moment, the last term in the above equation is of order $\frac{l-h}{l}$.

Substituting these results back into equation (7.18) gives $\widehat{\sigma}_l^2 \leq O_p[\frac{l-h}{l}2log(n)]$. Since $\widehat{\sigma}_h^2$ is of order 1, we then have

$$\mathbb{S}_{l,h} \leq loglog(n) - (\gamma_l - \gamma_h)f(n).$$

In order to consistently estimate γ_h, we require the infimum of $\mathbb{S}_{l,h}$ to be greater than

0 $w.p.1$. Therefore, if

$$f(n) > \frac{loglog(n)}{\gamma_l - \gamma_h},$$

$\Phi_{IC}(l)$ will always choose an l that is greater than h, and we cannot consistently estimate γ. Again, this inequality inequality needs to hold for any $\gamma_l - \gamma_h > \eta$ where $\underline{l} \leq h < l \leq \bar{l}$. In the proof for Theorem 7.1, we find the lower bound for $(\gamma_l - \gamma_h)^{-1}$ being $(\gamma_{\bar{l}} - \gamma_{\underline{l}})^{-1}$. So if

$$f(n) > \frac{loglog(n)}{\gamma_{\bar{l}} - \gamma_{\underline{l}}},$$

the information criterion method will not deliver consistent estimate of γ_h. □

7.C Proof of Theorem 7.3

Sketch of Proof for Theorem 7.3.

Following a similar argument as the proof for Theorem 7.1, we show that the cumulant based normality test method is consistent. Note that

$$\mathbb{S}_{l,h} = l\frac{\widehat{\kappa}_{3,l}^2}{6} + l\frac{\widehat{\kappa}_{4,l}^2}{24} - (h\frac{\widehat{\kappa}_{3,h}^2}{6} + h\frac{\widehat{\kappa}_{4,h}^2}{24}).$$

If $l < h$ and asymptotically $\widehat{\zeta}_l \subset \zeta$, ε_i $(i \in \widehat{\zeta}_l)$ follows a truncated normal and we can show that $\inf_l \mathbb{S}_{l,h} := \Phi(l) - \Phi(h) > 0\, w.p.1$ using the results in Berenguer-Rico et al. (2019).

If $l > h$ and asymptotically $\zeta \subset \widehat{\zeta}_l$, ε_i $(i \in \widehat{\zeta}_l)$, the sum of the first two terms in $\mathbb{S}_{l,h}$ is dominated by the second term, which is of order $O_p(2log(h))^4$, since the largest 'good' observation is of order $O_p(2log(h))^{1/2}$. The sum of the terms inside the brackets is $O_p(1)$, and thus $\mathbb{S}_{l,h}$ diverges to $+\infty$ as $n \to \infty$. □

REFERENCES

Aitchison, J. and Silvey, S. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, *29*(3), 813–828.

Aitchison, J. and Silvey, S. (1960). Maximum-likelihood estimation procedures and associated tests of significance. *Journal of the Royal Statistical Society*, *22*(1), 154–171.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki (Eds.), *Proceedings of the Second International Symposium of Information Theory* (pp. 267–281). Akademia Kiado: Budapest.

Ali, M. (1984). An approximation to the null distribution and power of the Durbin-Watson statistic. *Biometrika*, *71*(2), 253–261.

Ali, M. (1987). Durbin–Watson and generalized Durbin–Watson tests for autocorrelations and randomness. *Journal of Business and Economic Statistics*, *5*(2), 195–203.

Andrews, D. and Ploberger, W. (1996). Testing for serial correlation against an ARMA $(1,1)$ process. *Journal of the American Statistical Association*, *91*(435), 1331–1342.

Atkinson, A., Riani, M., and Cerioli, A. (2010). The forward search: theory and data analysis. *Journal of the Korean Statistical Society*, *2*(39), 117–134.

Belsley, D. (1997). A small-sample correction for testing for gth-order serial correlation with artificial regressions. *Computational Economics*, *10*(3), 197–229.

Belsley, D. (2000). A small-sample correction for testing for joint serial correlation with artificial regressions. *Computational Economics*, *16*(1), 5–45.

Berenguer-Rico, V., Johansen, S., and Nielsen, B. (2019). Models where the least trimmed squares and least median of squares estimators are maximum likelihood. *Oxford Department of Economics Discussion Paper 879*.

Berenguer-Rico, V. and Nielsen, B. (2017). Marked and weighted empirical processes of residuals with applications to robust regressions. *IDEAS Working Paper Series from RePEc*.

Berenguer-Rico, V. and Nielsen, B. (2020). Cumulated sum of squares statistics for non-linear and non-stationary regressions. *Econometric Theory*, *36*(1), 1–47.

Bowman, K. and Shenton, L. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, *62*(2), 243–250.

Box, G. and Pierce, D. (1970). Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *Journal of the American Statistical Association*, *65*(332), 1509–1526.

Breusch, T. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, *17*(31), 334–355.

Brown, B. (1971). Martingale central limit theorems. *The Annals of Mathematical Statistics*, *42*(1), 59–66.

Brüggemann, R., Lütkepohl, H., and Saikkonen, P. (2006). Residual autocorrelation testing for vector error correction models. *Journal of Econometrics*, *134*(2), 579–604.

Butler, R. (1982). Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. *The Annals of Statistics*, *10*(1), 197–204.

Cao, C., Lin, J., and Shi, J. (2014). Diagnostics on nonlinear model with scale mixtures of skew-normal and first-order autoregressive errors. *Statistics*, *48*(5), 1033–1047.

Chabot-Hallé, D. and Duchesne, P. (2008). Diagnostic checking of multivariate nonlinear time series models with martingale difference errors. *Statistics and Probability Letters*, *78*(8), 997–1005.

Chan, N. and Wei, C. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *The Annals of Statistics*, *16*(1), 367–401.

Chen, Y. (2008). A unified approach to standardized-residuals-based correlation tests for GARCH-type models. *Journal of Applied Econometrics*, *23*(1), 111–133.

Chow, Y. S. (1965). Local convergence of martingales and the Law of Large Numbers. *The Annals of Mathematical Statistics*, *36*(2), 552–558.

Cochrane, D. and Orcutt, G. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32–61.

D'Agostino, R. (1986). Tests for the normal distribution. In R. D'Agostino and M. Stephens (Eds.), *Goodness of Fit Techniques* (pp. 367–419). Marcel Dekker: New York.

D'Agostino, R. and Pearson, E. (1973). Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika*, *60*(3), 613–622.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. Oxford: Oxford University Press.

Davidson, R. and MacKinnon, J. (1990). Specification tests based on artificial regressions. *Journal of the American Statistical Association*, *85*(409), 220–227.

Davidson, R. and MacKinnon, J. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

Davies, N., Triggs, C., and Newbold, P. (1977). Significance levels of the Box-Pierce portmanteau statistic in finite samples. *Biometrika*, *64*(3), 517–522.

Dickey, D. and Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427–431.

Doornik, J. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, *70*(Supp.), 927–939.

Doornik, J. and Hendry, F. (2013). *PcGive 14 Volume I: Empirical Econometric Modelling* (2013 edition. ed.). London: Timberlake Consultants.

Dougherty, C. (2011). *Introduction to Econometrics* (4 ed.). Oxford: Oxford University Press.

Dufour, J.-M. and King, M. (1991). Optimal invariant tests for the autocorrelation coefficient in linear regressions with stationary or non-stationary AR(1) errors. *Journal of Econometrics*, *47*(1), 115–143.

Durbin, J. (1970). Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables. *Econometrica*, *38*(3), 410–421.

Durbin, J. and Watson, G. (1950). Testing for serial correlation in least squares regression I. *Biometrika*, *37*(34), 409–428.

Durbin, J. and Watson, G. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, *38*(12), 159–178.

Durbin, J. and Watson, G. (1971). Testing for serial correlation in least squares regression III. *Biometrika*, *58*(1), 1–19.

Edgerton, D. and Shukur, G. (1999). Testing autocorrelation in a system perspective testing autocorrelation. *Econometric Reviews*, *18*(4), 343–386.

Farebrother, R. (1980). The Durbin-Watson test for serial correlation when there is no intercept in the regression. *Econometrica*, *48*(6), 1535–1563.

Fisher, R. (1930). The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character (1905-1934)*, *130*(812), 16–28.

Fisher, R. and Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*(2), 180–190.

Fisher, T. and Gallagher, C. (2012). New weighted Portmanteau statistics for time series goodness of fit testing. *Journal of the American Statistical Association*, *107*(498), 777–787.

Flett, T. (1958). A mean value theorem. *The Mathematical Gazette*, *42*(339), 38–39.

Godfrey, L. (1978a). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, *46*(6), 1293–1301.

Godfrey, L. (1978b). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, *46*(6), 1303–1310.

Godfrey, L. (2007). Alternative approaches to implementing Lagrange multiplier tests for serial correlation in dynamic regression models. *Computational Statistics and Data Analysis*, *51*(7), 3282–3295.

Greene, W. (2012). *Econometric Analysis* (7 ed.). Harlow: Pearson.

Hall, P. and Heyde, C. (1980). *Martingale limit theory and its applications*. New York: Academic Press.

Hampel, F. (1968). Contributions to the theory of robust estimation.

Hampel, F. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, *42*(6), 1887–1896.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393.

Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 190–195.

Harrison, M. (1972). On testing for serial correlation in regression when the bounds test is inconclusive. *Economic and Social Review*, *4*(1), 41–57.

Hayashi, F. (2000). *Econometrics*. Princeton, NJ.; Oxford: Princeton University Press.

Helland, I. (1982). Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics*, *9*(2), 79–94.

Hendry, D. and Nielsen, B. (2007). *Econometric modeling: A likelihood approach*. Oxford: Princeton University Press.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Inder, B. (1986). An approximation to the null distribution of the Durbin-Watson statistic in models containing lagged dependent variables. *Econometric Theory*, *2*(3), 413–428.

Inder, B. (1990). A new test for autocorrelation in the disturbances of the dynamic linear regression model. *International Economic Review*, *31*(2), 341–354.

Jarque, C. and Bera, A. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259.

Jarque, C. and Bera, A. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, *55*(2), 163–172.

Jeong, K.-J. (1985). A new approximation of the critical point of the Durbin-Watson test for serial correlation. *Econometrica*, *53*(2), 477–482.

King, M. (1985). A point optimal test for autoregressive disturbances. *Journal of Econometrics*, *27*(1), 21–37.

Kyriazidou, E. (1998). Testing for serial correlation in multivariate regression models. *Journal of Econometrics*, *86*(2), 193–220.

Lai, T. and Wei, C. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, *10*(1), 154–166.

Leadbetter, M., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer series in statistics. New York: Springer-Verlag.

Li, W. (1992). On the asymptotic standard errors of residual autocorrelations in nonlinear time series modelling. *Biometrika*, *79*(2), 435–437.

Ling, S., Zhu, K., and Yee, C. (2013). Diagnostic checking for non-stationary ARMA models with an application to financial data. *North American Journal of Economics and Finance*, *26*, 624–639.

Ljung, G. and Box, G. (1978). On a measure of lack of fit in time-series models. *Biometrika*, *65*(2), 297–303.

MacKinnon, J. (1992). Model specification tests and artificial regressions. *Journal of Economic Literature*, *30*(1), 102–146.

Monti, A. (1994). A proposal for a residual autocorrelation test in linear models. *Biometrika*, *81*(4), 776–780.

Nankervis, J. and Savin, N. (2010). Testing for serial correlation: Generalized Andrews-Ploberger tests. *Journal of Business and Economic Statistics*, *28*(2), 246–255.

Nerlove, M. and Wallis, K. (1966). Use of the Durbin-Watson statistic in inappropriate situations. *Econometrica*, *34*(1), 235–238.

Nielsen, B. (2006a). Correlograms for non-stationary autoregressions. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *68*, 707–720.

Nielsen, B. (2006b). Order determination in general vector autoregressions. In H.-C. Ho, C.-K. Ing, and T. Lai (Eds.), *Time Series and Related Topics: In Memory of Ching-Zong Wei*, volume 52 (pp. 93–112). IMS Lecture Notes and Monograph Series.

Pearson, E. (1930). A further development of tests for normality. *Biometrika*, *22*(1/2), 239–249.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, *185*, 71–110.

Phillips, P. (1987). Time series regression with a unit root. *Econometrica*, *55*(2), 277–301.

Phillips, P. (2007). Regression with slowly varying regressors and nonlinear trends. *Econometric Theory*, *23*(4), 557–614.

Phillips, P. and Solo, V. (1992). Asymptotics for linear processes. *The Annals of Statistics*, *20*(2), 971–1001.

Poskitt, D. and Tremayne, A. (1980). Testing the specification of a fitted autoregressive moving average model. *Biometrika*, *67*(2), 359–363.

Poskitt, D. and Tremayne, A. (1981). An approach to testing linear time series models. *The Annal of Statistics*, *9*(5), 974–986.

Pötscher, B. (1985). The behaviour of the Lagrangian multiplier test in testing the orders of an ARMA-model. *Metrika*, *32*(1), 129–150.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871–880.

Rousseeuw, P. and Leroy, A. (1987). *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. New York.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Schwarz, H. and Waldvogel, J. (1989). *Numerical analysis: A comprehensive introduction*. Chichester: Wiley.

Shin, D. and Lee, J. (1996). Distribution of residual autocorrelations in non-stationary autoregressive processes. *Journal of Time Series Analysis*, *17*(1), 105–109.

Silvapulle, P. and King, M. (1993). Nonnested testing for autocorrelation in the linear regression model. *Journal of Econometrics*, *58*(3), 295–314.

Silvey, S. (1959). The Lagrangian multiplier test. *The Annals of Mathematical Statistics*, *30*(2), 389–407.

Srianantha kumar, S. and King, M. (2006). A new approximate point optimal test of a composite null hypothesis. *Journal of Econometrics*, *130*(1), 101–122.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 276–278.

Thadewald, T. and Buning, H. (2007). Jarque-Bera test and its competitors for testing normality — a power comparison. *Journal of Applied Statistics*, *34*(1), 87–105.

Theil, H. and Shonkwiler, J. (1986). Monte Carlo tests of autocorrelation. *Economics Letters*, *20*(2), 157–160.

Thiele, T. (1889). *Forelæsninger over almindelig Iagttagelseslære: Sandsynlighedsregning og mindste Kvadraters Methode*. København: Reitzel.

Vinod, H. (1973). Generalisation of the Durbin-Watson statistic for higher order autoregressive processes. *Communications in Statistics*, *2*(2), 115–144.

Wallis, K. (1972). Testing for fourth order autocorrelation in quarterly regression equations. *Econometrica*, *40*(4), 617–636.

Wang, Q. and Zhu, K. (2018). On a measure of lack of fit in nonlinear cointegrating regression with endogeneity. *Statistica Sinica*.

White, K. (1992). The Durbin-Watson test for autocorrelation in nonlinear models. *Review of Economics and Statistics*, *74*(2), 370–373.

Wooldridge, J. (2013). *Introductory Econometrics: A Modern Approach* (5 ed.). Mason, OH: South-Western Cengage Learning.

Wu, C. (1981). Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, *3*(9), 501–513.