Essays on Asymptotics of Outlier Detection Algorithms with Applications to Economics

Xiyu Jiao
Department of Economics and Mansfield College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Economics

Trinity, 2019

Abstract

In this thesis, we study a "heuristic approach" that are frequently used for outlier robustness analysis in either the classical or instrumental variables regression. In applied economics, it is a frequent concern whether a tiny set of atypical observations may have invalidated the key empirical findings. To check the robustness of the conclusion especially with respect to outliers, the heuristic approach is to first run least squares regression and remove observations with residuals beyond a chosen cut-off value. Then, re-run regression with selected observations and compare the updated estimate with the original one relative to their standard errors. This procedure can be iterated until the robust result is obtained. The leading purpose of this thesis is to develop asymptotic theory that formally justifies this simple robust procedure. The argument involves a theory of a new class of the weighted and marked empirical processes of residuals. Asymptotics are derived under the null hypothesis that there is no data contamination.

Chapter 1 The asymptotic theory is established in the linear regression, where the cut-off is chosen for the heuristic approach can be varying with sample size and iteration step. We show the tightness and convergence of the algorithm. The Poisson exceedance theory is established for the gauge (proportion of falsely discovered outliers). The theory is valid for stationary, deterministic trends, or unit roots regressors.

Chapter 2 Considering the time series regression, we propose two sets of testing procedures for the null hypothesis of no outliers based on the heuristic approach: first, a set of tests on the proportion and count of detected outliers; second, a set of scaling tests. Tests on the proportion and count have power against the number of outliers, while scaling tests have power against both the number and magnitude of outliers. Empirical applications including Mankiw, Romer, and Weil (1992) and Pretis (2019) show versatility of the tests.

Chapter 3 This chapter extends the asymptotic theory of the heuristic approach to the IVs regression. With the established weak convergence, we provide the adjusted inference robust to outliers. Then, we compare the efficiency of the robust estimator relative to the ordinary one. At last, a Hausman type test is constructed to perform outlier robustness checks, enabling us to compare between two estimates (the robust and ordinary) in a formal statistical way. An empirical application to Acemoglu et al. (2017) shows the utility of the proposed theory.

To my family Yuxia Ma, Yujie Ma, and Yong Jiao

Acknowledgements

The DPhil has been a highly non-stationary process with many outliers, I am thankful to all those who helped me detect all these outliers and pulled me back to the track to steer my thesis towards completion.

My deepest thanks go to my advisor, Bent Nielsen, who has been a constant source of support and guidance throughout my DPhil. When I chose to study economics at Oxford six years ago, it was Bent who chose to guide and help me first as my departmental supervisor and later as my MPhil and DPhil thesis advisor. With contagious passion and enthusiasm, he led me into the world of econometric theory; helped me build up foundations in probability and statistics; taught me how to rigorously establish asymptotic theory for econometric methods; and made me strive to be a better researcher and teacher. Every time I did not know how to prove a theorem, Bent has always been available to discuss my research and to find a way to solve the problem. Without his never-failing trust, unfailing support and constant inspiration, this thesis would not have been written.

I am extremely grateful to James Duffy who has been very generous with his time to educate me; to read my papers; and to provide advice for the future research. Together with Bent, James helped me read through two books "Convergence of Probability Measures" and "Weak Convergence and Empirical Processes", which essentially equip me with sufficient knowledge in asymptotic convergence and empirical processes to analyze the econometric problem in this thesis. James has also been a very kind friend, allowing me to use his departmental office where I have finished the most part of my job market paper. I thank James for his kindness and encouragement.

I am also very grateful to David Hendry, who provided ideas for my research and pointed me to many interesting empirical applications. Before I started my MPhil/DPhil at Oxford, it was David who recommended me to study asymptotics with Bent, so without him the path of my academic career may have been a very different one.

A special thanks has to be given to Qiwei Yao, who was the first to introduce me the modern statistics (through his lectures based on the textbook "All of Statistics") and to encourage me to pursue the PhD degree. Without his recommendation letter, I would not have the chance to study the DPhil at Oxford. Throughout my DPhil, Qiwei constantly encourages and motivates me by sharing his study and research experience in

statistics, though he does not have any official responsibilities for my supervision. His input at all levels of my research deserves a special mention.

It has been a great honor to be part of the academic family in the econometrics group. I have benefited a lot at the department from attending lectures, seminars, and conferences, and I am indebted to Johannes Abeler (for guiding through the job market), Steve Bond (for pointing to an empirical application that was later included in my job market paper), Jenifer Castle (for very supportive comments), Anders Kock (for encouraging me to meet and discuss with external visitors presenting at Oxford), Sophocles Mavroeidis (for many fruitful discussions), and Felix Pretis (for proposing an outlier testing idea that eventually became a co-authored paper).

I must acknowledge my college advisors Vanessa Berenguer-Rico and Eileen Tipoe at Mansfield, who not only offer academic guidance for my research, but also provide pastoral support. Due to both of them, I have been taken care very well within the college.

I am also thankful to my Oxford peer econometricians Liang Chen, Zoe Fannon, Olga Gdula, Yingyu Guo, Jonas Harnau, Katharine Lauderdale, Andrew Martinez, Matthias Qian, Wen Xu, and Qianzi Zeng, all of whom have contributed in various ways to my research work.

In the wider world of econometrics and statistics, I would like to thank Jesús Gonzalo, Emmanuel Guerre, Soren Jøhansen, Hira Koul, Takamitsu Kurita, Leandro Magnusson, Taisuke Otsu, Eric Renault, Yuya Sasaki, Winfried Stute, Qiying Wang, and Martin Weidner for their detailed comments and very enjoyable conversations.

I would like to thank the funding from Nuffield College (the Research Assistantship) and Department of Economics (the Department's Doctoral Studentship Scheme, the Final-Year Departmental Doctoral Bursary, and the Graduate Teaching Assistantship), which have supported my study and research, otherwise my DPhil would have been impossible. The doctoral studentship also provided me a fantastic opportunity to teach the first year MPhil econometrics classes at the department, which has been an extremely enjoyable experience in Oxford. Teaching have been great fun and deepened my understanding of econometrics, so I am thankful to all the students in my classes.

In academia in China, I wanted to express my gratitude to He Guo (for sharing his career experience and constantly giving life advice), Yao Ouyang (for inviting me to Hunan Normal University and helping me on the job market), and Dongtao Zou (for recommending me to study abroad and writing the strong reference letter), though I am

not their official student.

My friends balanced my life in Oxford and made it enjoyable. I would like to thank Bowen Dai, Benran Huang, Niels Buus Lassen, Song Li, Ang Liu, Kai Liu, Jialong Lu, Xuxin Mao, Meng Miao, Xinjie Wang, Xuan Wang, Wenchuan Wu, Ji Yan, Pengfei Zhang, and all my friends in the Oxford Chin Woo Basketball Team for their advice, support, and friendship.

Lastly, but most importantly, the greatest appreciation goes to my family Yuxia Ma, Yujie Ma, and Yong Jiao. Without their unconditional love and support, I would not have reached this point. My gratitude to them is beyond words. A very special thank you to Shuang Liu, who made me the time during my DPhil the best time of my life. Shuang balanced my life from exceptional hard work in her own caring that made it much more enjoyable. We together successfully navigated the highs and lows of our academic and personal life. My hope is that we will also successfully navigate beyond our DPhil studies.

Contents

Introduction **1**

Motivation 2

Background 5

Our contribution 6

1 Asymptotic Analysis of Iterated 1-step Huber-skip M-estimators with Varying Cut-offs **11**

1.1 Introduction 11

1.2 Model and outlier detection algorithms 14

1.2.1 Model 14

1.2.2 The iterated 1-step Huber-skip M-estimator algorithm 15

1.3 The main results 16

1.3.1 Assumptions 16

1.3.2 Properties of the iterated estimators 17

1.3.3 Properties of the gauge 19

1.4 Weighted and marked empirical process 20

1.4.1 Assumptions 21

1.4.2 A new empirical process result 22

1.4.3 A result for the two-sided empirical process 22

1.5 Conclusion 23

1.6 Appendix A: A metric on $\mathbb{R}$ and some inequalities 24

1.7 Appendix B: Proofs of the empirical process results 26

1.8 Appendix C: Proofs of the main results 30

1.8.1 An auxillary result for product moments 30

1.8.2 Proofs of Theorem 1.3.1, 1.3.2 and 1.3.3 32

1.8.3 Proof of Theorem 1.3.4 35

2 Testing the Presence of Outliers in Regression Models **37**
2.1 Introduction . . . 37
2.1.1 Related literature . . . 39
2.2 Proportion and count tests . . . 40
2.2.1 Gaussian approximation to the gauge . . . 41
2.2.2 Outlier proportion test: A standard normal test . . . 44
2.2.3 Outlier count test: A Poisson test . . . 45
2.3 Scaling tests . . . 46
2.3.1 Scaling sum test . . . 47
2.3.2 Scaling supremum test . . . 48
2.3.3 Scaling global test . . . 49
2.3.4 Implementation and use of the tests . . . 52
2.4 Simulation study . . . 53
2.4.1 Size under the null . . . 53
2.4.2 Power under alternatives . . . 56
2.4.3 Recommendations based on simulation results . . . 61
2.5 Applications . . . 61
2.5.1 Economic growth in a cross-sectional model . . . 61
2.5.2 The effects of a carbon tax on transport emissions in a difference-in-differences panel model . . . 63
2.5.3 Re-visiting studies using impulse indicator saturation . . . 66
2.6 Conclusion . . . 68
2.7 Appendix: Proofs . . . 70

3 A Simple Robust Procedure in Instrumental Variables Regression **75**
3.1 Introduction . . . 76
3.2 Model and a robust procedure . . . 82
3.2.1 Model . . . 82
3.2.2 Robust statistical algorithms to detect outliers . . . 84
3.3 The main results . . . 87
3.3.1 Assumptions . . . 87
3.3.2 Asymptotics of Algorithm 3.2.1 . . . 89
3.3.3 Weak convergence of Algorithm 3.2.2 and 3.2.3 . . . 93
3.3.4 Efficiency comparison . . . 96
3.3.5 Valid inference on structural parameters β . . . 99

3.3.6 A new Hausman type test of outlier robustness 102
3.4 Empirical application to Acemoglu et al. (2017) 107
3.5 Weighted and marked empirical process 116
3.5.1 Assumptions 117
3.5.2 Asymptotic results for empirical process 117
3.5.3 A result for the two-sided empirical process 119
3.6 Conclusion 121
3.7 Appendix A: A metric on $\mathbb{R}$ and some inequalities 121
3.8 Appendix B: Proofs of the empirical process results 125
3.8.1 Uniform $n^{1/2}$-convergence of empirical process 125
3.8.2 Linearization of compensator 130
3.8.3 Tightness of empirical process 133
3.8.4 Uniform n-convergence of empirical process 138
3.8.5 Results for empirical processes of absolute residuals 140
3.9 Appendix C: Proofs of the main results 141
3.9.1 Auxillary results for product moments 141
3.9.2 Results for Algorithm 3.2.1 144
3.9.3 Results for Algorithm 3.2.2 and 3.2.3 152
3.9.4 Results for a new type of Hausman test 155

Conclusion **156**

References **158**

Introduction

In this thesis, we study a class of outlier detection algorithms that are frequently used in either classical or instrumental variables regressions by many empirical economists. These algorithms can detect outlying observations and provide robust estimators of population parameters. In empirical research in economics, there is a frequent concern whether a tiny set of atypical observations may invalidate the key empirical results. For example, Toda and Walsh (2015) report that the estimate of the risk aversion coefficient is significantly changed by removing 200 observations classified as outliers from the full sample with the size 400000. Thus, all valid empirical findings should be resistant to a change in a small portion of data; also argued by Young (2017). This requires an econometric sample selection method that can perform outlier robustness analysis.

To carry out outlier analysis in classical or instrumental variables regressions, the common practice is to first run least squares regression estimates and remove observations with residuals beyond a chosen cut-off value. Then, rerun regression with selected observations and compare the updated estimate with the original one relative to their standard errors. This procedure can be iterated until the robust result is obtained. In the thesis, we develop asymptotic theory that formally justifies this heuristic approach.

Formalizing the asymptotic theory of the above heuristic approach is important for outlier robustness checks, since it is being widely applied in the empirical economics literature. For instance, De Long and Summers (1994) investigate the effect of equipment investment on economic growth and evaluate whether the positive impact is due to some outlying observations using the heuristic approach. Similarly, Acemoglu, Johnson, and Robinson (2012) and Acemoglu et al. (2017) use the procedure to examine the robustness of their conclusion to outliers when they estimate the institutional impact on economic development.

The established theory will give a guidance on better implementing the algorithm given the cut-off value, to identify outliers, to obtain consistent and robust estimators, and to make valid inference on structural parameters. In addition, the asymptotic theory

contributes to setting the cut-off as the tuning parameter in the algorithm to target the average frequency of false positives. Lastly, with the weak convergence results, we construct a test of the null of no outliers examining the overall presence of outliers and a test of outlier robustness examining whether the parameter of interest changes its value drastically before and after removing observations classified as outliers.

The argument involves a new class of weighted and marked empirical processes of residuals, thus we develop its uniform and weak law sufficient to formalize such the heuristic practice as the formal statistical investigation of outlier robustness analysis. To show the utility of our theory, we re-visit the study by Mankiw, Romer, and Weil (1992) and Pretis (2019), who estimate the augmented Solow growth model and the effect of a carbon tax on CO2 emissions respectively. We also re-examine outlier analysis in Acemoglu et al. (2017) estimating the effect of democracy on GDP per capita.

This thesis is conceptually split into two parts. In the first part, which contains the first two chapters, we consider the time series regression equation, which includes stationary, or deterministically or stochastically trending regressors. Extending an existing theory of the described outlier detection algorithm and its rate of falsely discovered outliers in the first chapter, we propose two sets of tests for the overall presence of outliers in the second chapter. In the second part, the third chapter generalizes all the asymptotic results to the instrumental variables regression equation.

Motivation

Outliers are data points that differ significantly from other observations, and they are ubiquitous in economic data due to several different reasons. Possible causes of outliers might be human errors in data collection, recording, or entry; measurement or experimental errors; changes in system behaviour; fraudulent behaviour; and interventions to an economy[1]. Statistically, a frequent cause of outliers is a mixture of two distributions, i.e. data are drawn from a population to be contaminated by some arbitrary distributions, although an outlier can just arise by random chance due to inherent variability of data. Finally, incorrect assumptions about the distribution of the data can also lead to the presence of outliers.

Undetected outliers give rise to three statistical problems. First, it is well known that estimation of population parameters is significantly distorted, see the detailed discussion

[1] see Engle, Hendry, and Richard (1983) and Bazinas and Nielsen (2015).

in Rousseeuw and Leroy (1987). Second, outliers greatly affect the performance of statistical inference. Berenguer-Rico and Nielsen (2018) and Berenguer-Rico and Wilms (2018) show that a small number of outliers substantially influence the outcome of specification checks testing whether residuals are i.i.d. normal. Third, as described in Clements and Hendry (1998), appearance of outliers at or near the forecast origin leads to forecast failure.

Economists are particularly concerned about whether their crucial empirical findings are driven by a tiny set of outliers. In the empirical economics literature, a recurrent critique is to question whether a paper examines the robustness of its conclusion, especially to a change in a small portion of data. Auerbach, Hassett, and Oliner (1994) question De Long and Summers (1991) whether the estimated positive capital effect (on economic growth) may have been driven by outliers. De Long and Summers (1994) agree that they face the sample-selection issue and their findings are particularly sensitive to one outlier. On estimating the institutional impact on economic growth, Acemoglu, Johnson, and Robinson (2001) receive a lot of comments on outlier robustness of their conclusion. For instance, Albouy (2012) ask whether the validity of instruments by mortality rates is ruined by some outlying observations. In response to the critique, Acemoglu, Johnson, and Robinson (2012) perform outlier analysis to check their empirical results.

To solve the problems caused by outliers, the statistics literature[2] designs some robust estimators of population parameters in the classical regression. Young (2018) argues the two stage least squares in the instrumental variables regression is more sensitive to outliers than the ordinary least squares in the classical regression by performing a comprehensive survey of 1400 IVs regressions in 32 papers published in journals of the American Economic Association. Thus, two types of robust methods (quantile and rank regression) are proposed and subsequently extended to the instrumental variables regression (see Honoré and Hu, 1994). However, it is well known that the quantile approach is not robust to bad leverage points where the joint distribution of dependent and independent variables shift, see Rousseeuw and Leroy (1987), while the rank regression has been used much less often to perform instrumental variables regressions in economics (see Honoré and Powell, 1994).

Even though there are existing robust methods in statistics, empirical economists still prefer applying a so called "heuristic approach" to check the robustness of their analysis to outliers. The heuristic approach is to first run regression with the full sample and

[2]see the thorough treatment in Rousseeuw and Leroy (1987) and Maronna, Martin, and Yohai (2006).

to classify outliers by examining whether residuals are larger than the chosen cut-off. Then, re-run regression with the selected non-outlying observations and compare the updated estimator with the original one relative to their standard errors. For example, on estimating the effect of democracy on economic growth, Acemoglu et al. (2017) apply this heuristic approach to argue the robustness of their conclusion to outliers. They choose the cut-off value 1.96, and first run OLS and its updated robust estimator (see columns (1) and (2) in Table 0.0.1). Due to the endogeneity problem that selection into democracy is affected by its own country's GDP, they secondly instrument democracy by the regional waves of democratization to obtain 2SLS and its updated robust estimator (see columns (3) and (4) in Table 0.0.1). They end up with the robustness conclusion simply by comparing the outlier removal with the full sample estimate (but without applying a formal statistical test). Notice that this one-step procedure can also be iterated until convergence.

	OLS		2SLS	
GDP	(1)	(2)	(3)	(4)
Democracy	0.787	0.556	1.149	0.941
	(0.228)	(0.150)	(0.607)	(0.393)
GDP 1st lag	1.238	1.226	1.238	1.224
	(0.013)	(0.009)	(0.013)	(0.009)
GDP 2nd lag	−0.207	−0.198	−0.205	−0.194
	(0.020)	(0.014)	(0.020)	(0.014)
GDP 3rd lag	−0.026	−0.027	−0.029	−0.032
	(0.019)	(0.013)	(0.019)	(0.013)
GDP 4th lag	−0.043	−0.030	−0.040	−0.027
	(0.012)	(0.008)	(0.012)	(0.008)
Sample size	6336	6044	6309	6019

Table 0.0.1: Acemoglu et al. (2017) estimate the effect of democracy on log GDP per capita in the panel setup. They include four lags of GDP as regressors to control the dynamics of the GDP process and add dummies for each country and year to control the fixed effects. (1) represents the full sample OLS and (2) the outlier removal OLS. (3) represents the full sample 2SLS instrumenting democracy by regional waves of democratization, and (4) the outlier removal 2SLS. Outliers are classified in the classical regression in (2) (IVs regression in (4)) if the standardized (structural) residuals are beyond the cut-off 1.96.

Establishing asymptotic theory of such a simple robust procedure is significantly im-

portant. For instance, without a formal hypothesis test, we cannot draw any robustness conclusion rooted in the statistical reasoning to evaluate the distance between the full sample estimate (1) and outlier removal estimate (2) (or between (3) and (4)). However, to the best of our knowledge, this heuristic approach lacks formal justification, and it is technically demanding to build up the probability tools (empirical processes) to obtain the joint distribution of the baseline and outlier removal estimators necessary to formalize the outlier robustness check by statistical tests. The leading purpose of this thesis is thus to derive asymptotic theory of this simple robust procedure.

Background

In this section, we formally define the described "heuristic approach" in either the classical or IVs regression. Then, introduce the concept of gauge, which is a proportion of falsely discovered outliers, to set the cut-off c as the tuning parameter in such a class of outlier detection algorithms.

We first consider a classical linear regression equation for the data set $\{(y_i, x_i)\}_{i=1}^n$

$$y_i = x_i'\beta + u_i, \quad i = 1, 2, \ldots, n,$$

where u_i/σ are i.i.d. and independent of the regressors x_i with the known density f. Regressors x_i can be stationary or deterministically or stochastically trending. Outliers are pairs of observations that do not conform with the model. Given an initial estimator $(\widetilde{\beta}, \widetilde{\sigma}^2)$ and the cut-off c, an observation i is non-outlying if

$$v_i = 1_{(|y_i - x_i'\widetilde{\beta}| \leq \widetilde{\sigma} c)}.$$

The choice of the cut-off c is related to the known reference density f. For selected non-outlying observations, we re-run a least squares to get the updated estimator

$$\widehat{\beta} = (\sum_{i=1}^n x_i x_i' v_i)^{-1} (\sum_{i=1}^n x_i y_i v_i), \qquad \widehat{\sigma}^2 = \varsigma^{-2} (\sum_{i=1}^n v_i)^{-1} \{\sum_{i=1}^n (y_i - x_i'\widehat{\beta})^2 v_i\}.$$

In the instrumental variables regression where regressors x_i are uncorrelated with structural errors u_i, we instrument x_i by z_i so that the data set $\{(x_i, z_i)\}_{i=1}^n$ satisfies the projection equation

$$x_i' = z_i'\Pi + r_i', \quad i = 1, 2, \ldots, n.$$

We assume the instruments z_i are independent of the structural errors u_i and the projection errors r_i. We further suppose $\operatorname{rank} \Pi = d_x$ and $\operatorname{rank} \mathsf{E} z_i z_i' = d_z$. Outliers are also

detected through v_i if structural residuals are beyond the chosen cut-off. The updated 2SLS is then

$$\widehat{\beta} = (\widehat{\Pi}' \sum_{i=1}^{n} z_i z_i' v_i \widehat{\Pi})^{-1} (\widehat{\Pi}' \sum_{i=1}^{n} z_i y_i v_i), \qquad \widehat{\sigma}^2 = \varsigma^{-2} (\sum_{i=1}^{n} v_i)^{-1} \{ \sum_{i=1}^{n} (y_i - x_i' \widehat{\beta})^2 v_i \},$$

where the location parameter in projection is estimated by

$$\widehat{\Pi} = (\sum_{i=1}^{n} z_i z_i' v_i)^{-1} (\sum_{i=1}^{n} z_i x_i' v_i).$$

Notice that ς^2 is the bias correction factor required for consistently estimating the variance, and it will be defined explicitly later in the main chapters. Also note that the above one-step procedure can be iterated in practice. The central contribution will be to derive asymptotic theory of such a class of robust estimators and to show convergence of the iterated algorithm.

Outlier detection algorithms have a positive probability to find outliers even when, in fact, the data generation process has no outliers. The key question that arises is whether the outliers that it discovers are indeed outliers in the true DGP. Only if the rate of falsely discovered outliers is small, it is very likely that the identified observations are true outliers. Thus, the choice of cut-off c as the tuning parameter of the algorithm is related to this concept, which can be formalized as follows. The algorithms assign stochastic indicators v_i to all observations so that $v_i = 0$ when observation i is declared as an outlier, otherwise $v_i = 1$. Under the null of no outliers, the empirical gauge is

$$\widehat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} (1 - v_i).$$

This thesis will be to give formal justification to pick the tuning parameter c indirectly from the gauge by deriving its weak limits.

Our contribution

The leading purpose is to establish asymptotic theory of the defined heuristic approach for outlier robustness analysis in either classical or IVs regressions. The theory includes uniform consistency and weak limits of the iterated estimators as well as convergence of the algorithms.

With the derived weak convergence result of the algorithm, we will propose a consistent updated estimator of σ^2 and provide a valid inference on β based on the robust

estimator (2) and (4) in Table 0.0.1. After removing outliers, the standard method will underestimate σ^2 using the sub-sample of non-outlying observations due to the sample truncation, though the updated β estimator is consistent. Moreover, without the correct limiting distribution of the outlier removal estimator of β, the heuristic procedure uses the usual 2SLS asymptotics to perform inference with the selected non-outlying observations. Thus, the standard errors are erroneously calculated in columns (2) and (4), so the standard method for inference is invalid. To fix the problem, we introduce the bias correction factor ς^2 to adjust the updated σ^2 estimator and derive the weak limit of the updated β estimator, which together allow us to perform valid inference on structural parameters β in columns (2) and (4).

The performance of the algorithms can be evaluated by a gauge, which is a measure of type I error and can give us an indirect way of choosing the cut-off c. However, due to a lack of distributional results on the gauge, there was no obvious cut-off selection rule. This thesis contributes to showing normal limit or Poisson limit of the gauge when the cut-off is either fixed or goes to infinity under the null of no outliers.

This thesis also constructs a formal test of outlier robustness, which can be used to check whether the empirical results are driven by outliers, i.e. whether the robust estimate (2) or (4) is different from the full sample one (1) or (3). We suggest using the Durbin-Hausman-Wu test statistics to measure the distance between the outlier removal and the full sample estimators, and then show that it converges to a chi-square distribution.

Asymptotic analysis of robust methods in the iterated one-step frame is by no means new. One-step (Huber, L, and M) estimators have been considered by Bickel (1975), Ruppert and Carroll (1980), and Welsh and Ronchetti (2002). The idea of iterating one-step estimators can be found in Dollinger and Staudte (1991) and Cavaliere and Georgiev (2013). The heuristic approach that we are interested in is mimics the Huber-skip M-estimators. For the iterated one-step Huber-skip M and L-estimators, asymptotic theory has been recently established by Berenguer-Rico and Nielsen (2018), Johansen and Nielsen (2009, 2013, 2016a, 2016b, 2018).

In the first chapter, we build an asymptotic theory which can explore how variation in the cut-off affects the algorithm. In particular, we prove the tightness and fixed point theorems for the iterated robust estimators with the varying cut-off. Moreover, this chapter demonstrates an asymptotic Poisson distribution to the gauge in a situation where the cut-off increases with the sample size while the number of iterations also

increases.

The presence of outlying observations in a regression equation can be indicative of model misspecification. For instance, estimating a linear model, when the true DGP is non-linear, may result in a high number of outliers, consequently, it is important to check for possible contamination. In the second chapter, using the distributional results of the gauge, we propose two sets of tests for the overall presence of outliers: first, a set of tests on whether the observed proportion of outliers deviate from their expected value. Second, a set of scaling tests on whether the number of outliers decreases proportionally with the increasing cut-off. Simulation studies show the first type test has power against the number of outliers present, while the second type has power against both outlier magnitude and number.

Motivated by many IVs empirical papers, the third chapter extends the asymptotic theory to the IVs model. Statistics is the tradeoff between efficiency and robustness: the estimator produced by the heuristic approach is robust to outliers, however loses efficiency in terms of having the higher asymptotic variance under the null of no outliers, as compared to the full sample estimator. This is because the robust estimator has a certain probability to falsely remove observations even when there are no outliers in the data. We then draw the efficiency comparison plot against both cut-offs and iteration steps. Moreover, we construct a statistical test of outlier robustness based on the Hausman type test statistics comparing between the ordinary and robust estimators. It is then useful to check whether (2) or (4) is different from (1) or (3) in Table 0.0.1, indicating whether the key empirics is robust to the sample selection.

Our empirical applications show the versatility of the proposed theory. In a cross-sectional model of economic growth based on Mankiw, Romer, and Weil (1992), we find little evidence of significant outlying countries. In a panel difference-in-difference model of the impact of a carbon tax on CO2 emissions we identify significant outliers in the un-taxed control group, the presence of which impacts the estimated effect of the tax.

This thesis also re-explores the outlier robustness analysis in Table 0.0.1 in Acemoglu et al. (2017). We first adjust standard errors in columns (2) and (4), and then conduct a formal Hausman type test of outlier robustness to statistically check whether the robust estimate (2) or (4) is equal to (1) or (3). For most coefficients, the tests are rejected at 5% significance level, indicating the outlier removal estimator is distinct from the full sample one at least in magnitutde. The adjusted inference based on robust estimates cannot reject the hypothesis that the coefficient on democracy equals zero, thus the result of

outlier analysis does not robustly support that democracy has a strictly positive effect on GDP per capita, though there is little evidence for the view that democracy is a constraint for economic development.

The argument involves a theory for a new class of weighted and marked empirical processes of residuals, which can be defined from the generalized empirical distribution function

$$\widehat{\mathsf{G}}_{u,n}^{w,p}(a,b,c) = \frac{1}{n}\sum_{i=1}^{n} w_{in} u_i^p v_i^{a,b,c}.$$

In the regression, rewrite $v_i = 1_{(|u_i - x_{in}' b| \leq \sigma c + n^{-1/2} ac)}$, where a,b represent normalized estimation errors, but are chosen deterministically, which suffices if arguments are uniform in a,b. Estimation errors $\widetilde{a} = n^{1/2}(\widetilde{\sigma} - \sigma)$, $\widetilde{b} = N^{-1}(\widetilde{\beta} - \beta)$ and $x_{in} = N' x_i$. The normalization matrix N depends on the stochastic property of x_i, say $N = n^{-1/2} I_{d_x}$ if $\{x_i\}_{i=1}^n$ is stationary. The weights w_{in} are either of $1, n^{1/2} x_{in}, n x_{in} x_{in}'$ and p as either of $0, 1, 2$. Similarly, in the IVs regression $v_i^{a,b,c} = 1_{(|u_i - z_{in}' \Pi b - n^{-1/2} r_i' b| \leq \sigma c + n^{-1/2} ac)}$ while w_{in} are combinations of the normalized intruments z_{in} instead. We derive asymptotic expansions that are uniform in a, b, c and allow for a near $n^{1/4}$ inefficiency in the estimation uncertainties a, b.

The empirical process literature dates back to Kolmogorov (1933) and Smirnov (1939) who proposed a type of goodness of fit tests that check whether the distribution under the null is well specified by comparing it with the empirical distribution function. To build asymptotics of Kolmogorov-Smirnov type test statistics, Doob (1949) first demonstrated weak convergence of empirical distribution function, and then Donsker (1952) established the empirical process central limit theorem to close a gap in Doob's proof. Because there exists the measurability issue in $D[0,1]$ space, the classical method applies the Skorokhod metric instead of uniform topology to avoid non-measurability[3]. Meanwhile, many researchers still equip $D[0,1]$ space with uniform metric but use the Hoffmann-Jørgensen[4] definition of weak convergence instead. With the new concept of weak convergence, they extend the classical theory to the empirical process indexed by VC class of sets or functions using the entropy and bracketing argument[5]. However, their work has not yet been extended to deal with the weights w_{in} and marks u_i^p appearing in our context.

[3] see details in Billingsley (1968).

[4] a sequence of random elements X_n weakly converges to the limiting random element X if and only if $\mathsf{E}^* f(X_n) \to \mathsf{E} f(X)$ as $n \to \infty$ for any continuous and bounded function f where E^* denotes the outer expectation.

[5] see the summary in van der Vaart and Wellner (1996).

Following the classical idea, Koul and Ossiander (1994) used the entropy argument to first build up the theory for the weighted empirical process in the autoregressive model[6]. Recently, empirical processes including both weights and marks have been analyzed by a series of papers, see Engler and Nielsen (2009), Johansen and Nielsen (2009, 2016a), Berenguer-Rico and Nielsen (2018), and Berenguer-Rico, Johansen, and Nielsen (2019a, b). This thesis further extends their work and show uniform convergence for the empirical process constructed by $n^{1/2}\widehat{\mathsf{G}}_{u,n}^{w,p}(a,b,c)$. We also provide the linearization for the compensator of $n^{1/2}\widehat{\mathsf{G}}_{u,n}^{w,p}(a,b,c)$.

[6]also see Koul (2002) and Koul and Ling (2006).

Chapter 1

Asymptotic Analysis of Iterated 1-step Huber-skip M-estimators with Varying Cut-offs

ABSTRACT[1]. We consider outlier detection algorithms for time series regression based on iterated 1-step Huber-skip M-estimators. This chapter analyses the role of varying cut-offs in such algorithms. The argument involves an asymptotic theory for a new class of weighted and marked empirical processes allowing for estimation errors of the scale and the regression coefficient.

1.1 Introduction

We consider outlier detection methods that are based on iterated 1-step Huber-skip M-estimators for linear regression models with regressors that are stationary or deterministically or stochastically trending. Each 1-step estimator relies on a cut-off value when classifying observations as outliers or not. In this chapter we allow the cut-off value to vary with sample size and iteration step. To analyze the iterated 1-step estimator asymptotically, we generalize some recent results for residual empirical processes, which allow for variation in location, scale and quantile. The model is a linear regression

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1.1.1}$$

[1]This chapter is co-authored with Bent Nielsen, Department of Economics and Nuffield College, University of Oxford.

where ε_i/σ are independent of $\mathcal{F}_{i-1} = \sigma(x_1, \ldots, x_i, \varepsilon_1, \ldots, \varepsilon_{i-1})$ with the common reference density f. Outliers are pairs of observations that do not conform with the model.

The iterated 1-step Huber-skip M-estimator mimics the Huber (1964) skip estimator. An initial estimator $(\widetilde{\beta}, \widetilde{\sigma}^2)$ is used to decide which observations are outlying through

$$v_i = 1_{(|y_i - x_i'\widetilde{\beta}| \leq \widetilde{\sigma} c)}, \tag{1.1.2}$$

where the choice of the cut-off c is related to the known reference density f. For those observations that are not outlying, we run a least squares regression and get the one-step Huber-skip estimator

$$\widehat{\beta} = (\sum_{i=1}^{n} x_i x_i' v_i)^{-1} (\sum_{i=1}^{n} x_i y_i v_i), \tag{1.1.3}$$

$$\widehat{\sigma}^2 = \varsigma^{-2} (\sum_{i=1}^{n} v_i)^{-1} \{\sum_{i=1}^{n} (y_i - x_i'\widehat{\beta})^2 v_i\}, \tag{1.1.4}$$

where ς^2 is the consistency factor as in (1.2.2). This step can be iterated. The iteration may be initiated by a robust estimator. More simply we get the Robustified Least Squares and the Impulse Indicator Saturation starting with the full or split sample least squares. The latter algorithm was introduced in the empirical work of US food expenditure by Hendry (1999), see also Hendry and Doornik (2014).

Outlier detection algorithms have a positive probability to find outliers even when, in fact, the data generation process has no outliers. We evaluate the performance of such algorithms by the concept of a gauge, which is the expected retention rate of falsely discovered outliers. This is a measure of type I error and it gives us an indirect way of choosing the cut-off c. It is defined as follows. The algorithms assign stochastic indicators v_i to all observations such as in (1.1.2) so that $v_i = 0$ when observation i is declared as an outlier, otherwise $v_i = 1$. When the model has no contamination, the sample and population gauge are

$$\widehat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} (1 - v_i), \qquad \mathsf{E}\widehat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}(1 - v_i). \tag{1.1.5}$$

Hoover and Perez (1999) originally introduced the idea of a gauge in a simulation study of general-to-specific variable selection algorithms. The concept of a gauge was formally proposed by Hendry and Santos (2010) as the expected retention rate of irrelevant regressors in the context of model selection algorithms. Doornik (2009) and Hendry and Doornik (2014) presented a comprehensive simulation study on the gauge for the model

selection algorithm Autometrics. Johansen and Nielsen (2016b) gave an asymptotic analysis for the gauge of some outlier detection algorithms.

One-step estimators have been considered before, see Bickel (1975), Ruppert and Carroll (1980). The 1-step Huber-skip estimator was studied by Welsh and Ronchetti (2002). Asymptotic distribution theory was made by Hendry, Johansen and Santos (2008) for the location model and by Johansen and Nielsen (2009) for the time series regression. Iteration was investigated in Johansen and Nielsen (2013). Johansen and Nielsen (2016b) established an asymptotic expansion for the sample gauge. All these asymptotic analyses are restricted to the situation where the cut-off and the number of iterations are not both increasing.

The purpose of this chapter is to build an asymptotic theory which can explore how variation in the cut-off affects the iterated 1-step Huber-skip M-estimator. In particular, we prove the tightness and fixed point theorems for the iterated 1-step M-estimator with the varying cut-off. Moreover, this chapter demonstrates an asymptotic Poisson distribution to the gauge in a situation where the cut-off increases with the sample size while the number of iterations also increases.

The argument involves a theory for a new class of weighted and marked empirical processes. This is defined from the generalized empirical distribution function

$$\widehat{\mathsf{F}}_n^{g,p}(a,b,c) = \frac{1}{n}\sum_{i=1}^{n} g_{in}\varepsilon_i^p 1_{(\varepsilon_i \leq \sigma c + n^{-1/2}ac + x_{in}'b)}, \tag{1.1.6}$$

where the weights g_{in} are combinations of the normalized $\mathcal{F}_{i-1}$ measurable regressors x_{in} and ε_i^p are the $\mathcal{F}_i$ adapted marks, while a, b represent the normalized estimation errors for σ, β. We derive asymptotic expansions that are uniform in a, b, c and allow for a near $n^{1/4}$ inefficiency in the estimation uncertainties a, b. This generalizes results by Koul and Ossiander (1994) and Koul (2002) who did not allow unbounded marks and divergence of the normalized estimation errors.

The outline of this chapter is the following. We first review the model and iterated 1-step Huber-skip M-estimator algorithm in §1.2. Then, this chapter presents the main results in §1.3. §1.4 provides theory for a new class of the weighted and marked empirical process with proofs in Appendix §1.6, 1.7. Proofs of the main theorems in §1.3 follow in Appendix §1.8.

1.2 Model and outlier detection algorithms

The regression equation with some notations is described first. It is widely known that least squares estimators are sensitive to outliers. We review the iterated 1-step Huber-skip M-estimators including the Robustified Least Squares and the Impulse Indicator Saturation.

1.2.1 Model

Suppose we have data (y_i, x_i), $i = 1, 2, \ldots, n$, where y_i is univariate and x_i is multivariate with dimension $\dim x$. Assume the data satisfies the regression equation

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

This setting can represent both classical regression and time series models. Moreover, regressors x_i can be a deterministic or stochastic trend. Innovations ε_i are independent of the filtration $\mathcal{F}_{i-1}$ generated by $(x_1, \ldots, x_i, \varepsilon_1, \ldots, \varepsilon_{i-1})$, and are identically distributed with scale σ so that ε_i/σ has the known density f and distribution function $\mathsf{F}(c) = \mathsf{P}(\varepsilon_i/\sigma \leq c)$. In practice, f, F will often be assumed to be standard normal or at least symmetric. Outlier detection algorithms use absolute residuals and then calculate robust least squares estimators from the non-outlying sample. This implicitly assumes symmetry, while non-symmetry leads to bias forms. We assume symmetry when analyzing the iterated 1-step Huber-skip M-estimator algorithm in §1.3, but not for the general empirical process results in §1.4.

Denote the absolute error $|\varepsilon_i|/\sigma$ has a density g and a cumulative distribution function $\mathsf{G}(c) = \mathsf{P}(|\varepsilon_i|/\sigma \leq c)$ for $c > 0$. With a symmetry assumption, $\mathsf{G}(c) = 2\mathsf{F}(c) - 1$ and $\mathsf{g}(c) = 2\mathsf{f}(c)$. Define $\psi = \mathsf{G}(c)$ so the probability of exceeding the cut-off c is $\gamma = 1 - \psi$. Suppose the k-th moment of the density f exists, then introduce

$$\tau_k = \int_{-\infty}^{\infty} u^k \mathsf{f}(u) du, \qquad \tau_k^c = \int_{-c}^{c} u^k \mathsf{f}(u) du. \tag{1.2.1}$$

Thus $\tau_0^c = \psi$, $\tau_2 = 1$ while $\tau_k = \tau_k^c = 0$ for odd k when assuming symmetry. Define the conditional variance of ε_i/σ given $(|\varepsilon_i|/\sigma \leq c)$ as

$$\varsigma_c^2 = \frac{\tau_2^c}{\psi} = \frac{\int_{-c}^{c} u^2 \mathsf{f}(u) du}{\mathsf{P}(|\varepsilon_i| \leq \sigma c)}. \tag{1.2.2}$$

This will be used as a bias correction factor for the variance estimate computed from the selected non-outlying sample. For a standard normal reference distribution, we have $\tau_2^c = \psi - 2c\mathsf{f}(c)$, $\tau_4^c = 3\psi - 2c(c^2 + 3)\mathsf{f}(c)$ and $\tau_4 = 3$.

1.2.2 The iterated 1-step Huber-skip M-estimator algorithm

We first define the iterated 1-step Huber-skip M-estimator algorithm. Specific examples include the Robustified Least Squares and the Impulse Indicator Saturation.

Algorithm 1.2.1. ***Iterated 1-step Huber-skip M-estimator***. *Choose a cut-off* $c > 0$.

1. Choose initial estimators $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ *and let* $m = 0$.

2. Define indicator variables for selecting non-outlying observations

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \leq \widehat{\sigma}_c^{(m)} c)}. \tag{1.2.3}$$

3. Compute least squares estimators

$$\widehat{\beta}_c^{(m+1)} = (\sum_{i=1}^n x_i x_i' v_{i,c}^{(m)})^{-1} (\sum_{i=1}^n x_i y_i v_{i,c}^{(m)}), \tag{1.2.4}$$

$$(\widehat{\sigma}_c^{(m+1)})^2 = \varsigma_c^{-2} (\sum_{i=1}^n v_{i,c}^{(m)})^{-1} \{\sum_{i=1}^n (y_i - x_i'\widehat{\beta}_c^{(m+1)})^2 v_{i,c}^{(m)}\}. \tag{1.2.5}$$

4. Let $m = m + 1$ *and repeat 2 and 3.*

In §1.3 we show how to choose the cut-off c indirectly from the gauge defined in (1.1.5). The algorithm could start with a robust estimator, while the Robustified Least Squares is initiated using the full sample least squares. The latter is not robust with respect to high leverage points in cross section data. Leverage points seem to be less of a problem in time series models when lagged variables are included as regressors.

Another example is the Impulse Indicator Saturation which was initially proposed by Hendry (1999) in his empirical work. The algorithm was studied comprehensively by Hendry, Johansen and Santos (2008), see also Hendry and Doornik (2014). The idea is to divide full sample into two sub-samples and use regression estimates calculated from each sub-sample to detect outliers in the other sub-sample.

Algorithm 1.2.2. ***Impulse Indicator Saturation***. *Choose a cut-off* $c > 0$.

1.1. Split full sample into two sets $\mathcal{I}_j$, $j = 1, 2$ *of* n_j *observations where* $\sum_{j=1}^2 n_j = n$.

1.2. Calculate least squares estimators based upon each sub-sample $\mathcal{I}_j$ *for* $j = 1, 2$

$$\widehat{\beta}_j = (\sum_{i \in \mathcal{I}_j} x_i x_i')^{-1} (\sum_{i \in \mathcal{I}_j} x_i y_i), \qquad \widehat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} (y_i - x_i'\widehat{\beta}_j)^2. \tag{1.2.6}$$

1.3. Define the initial indicator variables for selecting non-outlying observations

$$v_{i,c}^{(-1)} = 1_{(i \in \mathcal{I}_1)} 1_{(|y_i - x_i'\widehat{\beta}_2| \leq \widehat{\sigma}_2 c)} + 1_{(i \in \mathcal{I}_2)} 1_{(|y_i - x_i'\widehat{\beta}_1| \leq \widehat{\sigma}_1 c)}. \tag{1.2.7}$$

1.4. Compute $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ *using* (1.2.4), (1.2.5) *with* $m = -1$, *and then let* $m = 0$.
2. Follow the step 2,3,4 in Algorithm 1.2.1.

The Impulse Indicator Saturation is possibly more robust than the Robustified Least Squares when we have prior knowledge that outliers are located in a particular subset of the whole sample. Doornik (2009) argued the choice of the initial sets $\mathcal{I}_1$ and $\mathcal{I}_2$ should be iterated since the location of contaminated observations is unknown in most practical situations.

1.3 The main results

We start by listing the assumptions. Then follows the new tightness and fixed point result for the iterated estimator defined in Algorithm 1.2.1. Finally the gauge of the iterated estimator is analysed. The result is uniform in the cut-off value, which generalises Johansen and Nielsen (2009, 2013) who set the threshold fixed. This allows us to analyse the gauge of the iterated estimator when the cut-off value is drifting.

1.3.1 Assumptions

We list the sufficient assumptions for asymptotic theory of iterated 1-step Huber-skip M-estimators. These assumptions are somewhat stronger than they need to be. In section §1.4 on the one-sided empirical process, we will introduce some weaker assumptions. For instance, we will then abandon the symmetry assumption of f.

Innovations ε_i and regressors x_i must satisfy some moment conditions so as to carry out asymptotic analysis. Regressors x_i can be temporally dependent and trending deterministically or stochastically. Note that N is a normalisation matrix for x_i such that $N = n^{-1/2} I_{\dim x}$ for stationarity and $N = n^{-1} I_{\dim x}$ for random walks. Recall the dimension of the regressor x_i is $\dim x$ and denote $I_{\dim x}$ as the identity matrix with the dimension $\dim x$.

Assumption 1.3.1. *Let* $\mathcal{F}_i$ *be an increasing sequence of* σ*-fields so* ε_{i-1} *and* x_i *are* $\mathcal{F}_{i-1}$ *measurable and* ε_i *is independent of* $\mathcal{F}_{i-1}$. *Let* ε_i/σ *have a symmetric, continuously differentiable density* f *which is positive on* $(-\infty, \infty)$. *For some values of* κ, η *such that* $0 \leq \kappa < \eta \leq 1/4$, *choose* $r \geq 2$ *so*

$$2^{r-1} \geq 1 + (1/4 + \kappa - \eta)(1 + \dim x). \tag{1.3.1}$$

Let $q = 1 + 2^{r+1}$. Denote $c_0 > 0$ as a finite number. Suppose

(i) the density f *satisfies*

(a) $c^q \mathsf{f}(c)$, $|c^{q+1}\dot{\mathsf{f}}(c)|$ *are decreasing for large* c*;*

(b) $\mathsf{f}(c - n^{-1/4}A)/\mathsf{f}(c) = \mathrm{O}(1)$ *for some* $A > 0$ *and* $c = \mathrm{o}(n^{1/4})$ *as* $n \to \infty$*;*

(c) $\mathsf{f}(c)/[c\{1 - \mathsf{F}(c)\}] = \mathrm{O}(1)$ *for* $c \to \infty$*;*

(ii) the regressors x_i *satisfy*

(a) $\Sigma_n = \sum_{i=1}^n N' x_i x_i' N \xrightarrow{\mathsf{P}} \Sigma \overset{a.s.}{>} 0$*;*

(b) $\max_{1 \le i \le n} |n^{1/2-\kappa} N' x_i| = \mathrm{O}_{\mathsf{P}}(1)$*;*

(c) $n^{-1}\mathsf{E} \sum_{i=1}^n |n^{1/2} N' x_i|^q = \mathrm{O}(1)$*;*

(iii) the initial estimator $(\widetilde{\beta}, \widetilde{\sigma}^2)$ *satisfies*

(a) $N^{-1}(\widetilde{\beta} - \beta) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$*;*

(b) $n^{1/2}(\widetilde{\sigma}^2 - \sigma^2) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$.

There is a trade-off between κ, η, the dimension $\dim x$ and the required number of moments r, see Johansen and Nielsen (2016a, Remark 3.1). Conditions (i), (ii) are satisfied in a range of situations, see Johansen and Nielsen (2016a, 2016b). In particular, the condition (i) is satisfied by the normal and t distribution. The condition (ii) is satisfied by stationary, random walk and deterministically trending regressors. Condition (iii) allows the standardised estimation errors to diverge at a rate of $n^{1/4-\eta}$ rather than being bounded in probability. In particular, $\eta = 1/4$ can be chosen for estimators with standard convergence rates.

1.3.2 Properties of the iterated estimators

We show the asymptotic theory for the iterated 1-step Huber-skip M-estimator with the varying cut-off. The first result is an 1-step stochastic expansion of the updated estimator in terms of the original estimator, a kernel and a small remainder term.

Theorem 1.3.1. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 1.2.1. Suppose Assumption 1.3.1*(ia, ii) *holds, and that* $N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ *are* $\mathrm{O}_{\mathsf{P}}(1)$*. Then uniformly in* $c \in [c_0, \infty)$ *and as* $n \to \infty$

$$N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = \frac{2c\mathsf{f}(c)}{\psi} N^{-1}(\widehat{\beta}_c^{(m)} - \beta) + (\psi\Sigma)^{-1} \sum_{i=1}^n N' x_i \varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)} + \mathrm{o}_{\mathsf{P}}(1),$$

$$n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = \frac{c(c^2 - \varsigma_c^2)\mathsf{f}(c)}{\tau_2^c} n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) + \frac{1}{2\sigma\tau_2^c} n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \varsigma_c^2\sigma^2) 1_{(|\varepsilon_i| \le \sigma c)} + \mathrm{o}_{\mathsf{P}}(1).$$

Theorem 1.3.1 shows that the updated estimation error for β depends on the previous estimation error for β, but not on the estimation uncertainty for σ. The estimation error for σ has a similar property. This is a consequence of symmetry imposed on the density f. More complex situations can also be analyzed where the reference distribution f is non-symmetric and the cut-off c is chosen in a matching way, see Johansen and Nielsen (2009). The proof uses the empirical process theory in §1.4.

The next result shows that the iterated estimator is tight in iteration $m \in [0, \infty)$ and in the cut-off value $c \in [c_0, \infty)$. This builds on Johansen and Nielsen (2013).

Theorem 1.3.2. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 1.2.1. Suppose Assumption 1.3.1(ia, ii, iii) holds with $\eta = 1/4$. Then as $n \to \infty$*

$$\sup_{0 \le m < \infty} \sup_{c_0 \le c < \infty} |N^{-1}(\widehat{\beta}_c^{(m)} - \beta)| + |n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)| = \mathrm{O}_{\mathsf{P}}(1).$$

Assumption 1.3.1(iii) with $\eta = 1/4$ corresponds to a standard convergence rate for the initial estimator. Theorem 1.3.1 provides the 1-step relationship between the updated and the original estimator. Since $\sup_{c_0 \le c < \infty} \max\{|2c\mathsf{f}(c)/\psi|, |c(c^2 - \varsigma_c^2)\mathsf{f}(c)/\tau_2^c|\} < 1$ implied by Assumption 1.3.1(ia), see Johansen and Nielsen (2013, Theorem 3.5), a geometric argument and mathematical induction are used to show the tightness result.

The fixed point result can now be shown. Initially the tight estimator is assumed available. These are iterated through the one-step equation presented in Theorem 1.3.1.

Theorem 1.3.3. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 1.2.1. Suppose Assumption 1.3.1(ia, ii, iii) holds with $\eta = 1/4$. Then for all $\epsilon, \delta > 0$ a pair $m_0, n_0 > 0$ exists, so for all $m > m_0$ and $n > n_0$*

$$\mathsf{P}\{\sup_{c_0 \le c < \infty} |N^{-1}(\widehat{\beta}_c^{(m)} - \widehat{\beta}_c^*)| + |n^{1/2}(\widehat{\sigma}_c^{(m)} - \widehat{\sigma}_c^*)| > \delta\} < \epsilon,$$

where

$$N^{-1}(\widehat{\beta}_c^* - \beta) = \frac{1}{\psi - 2c\mathsf{f}(c)} \Sigma^{-1} \sum_{i=1}^{n} N' x_i \varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)},$$

$$n^{1/2}(\widehat{\sigma}_c^* - \sigma) = \frac{1}{2\sigma\{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)\}} n^{-1/2} \sum_{i=1}^{n} (\varepsilon_i^2 - \varsigma_c^2 \sigma^2) 1_{(|\varepsilon_i| \le \sigma c)}.$$

Based on Theorem 1.3.2, if the initial estimator is bounded in a large compact set with large probability, then any iterated estimator takes values in the same compact set no matter what value of the cut-off c is chosen in the interval $[c_0, \infty)$. The proof of Theorem 1.3.3 is to further argue the deviation between the m-fold iterated estimator

and the fixed point is the sum of two terms vanishing exponentially and in probability respectively when m and n are sufficiently large.

The iterated 1-step Huber-skip M-estimator can be seen as a special case of iteratively reweighted least squares with binary weights. Dollinger and Staudte (1991) applied an influence function argument to demonstrate convergence of iteratively reweighted least squares with smooth weights. Even if the spirit is similar, our proof is different due to binary weights. The idea of iterating one-step estimator can also be found in an analysis of the first order autoregression with infinite variance by Cavaliere and Georgiev (2013).

1.3.3 Properties of the gauge

Johansen and Nielsen (2016b) proved the Poisson approximation to the gauge for the finite step Huber-skip M-estimator. But the iterated result was not established, since they did not have the empirical process theory which investigates the varying quantile c and estimation errors for β and σ. This chapter shows the Poisson approximation to the gauge for the iterated 1-step Huber-skip M-estimator.

A Poisson exceedence theory arises in the scenario where the cut-off value c is set to allow the fixed number λ of outliers regardless of the sample size n. For some $\lambda > 0$, the cut-off value c_n is set so as to let

$$n\mathsf{P}(|\varepsilon_i| > \sigma c_n) = \lambda. \tag{1.3.2}$$

Notice that $c_n \to \infty$ as $n \to \infty$. Define $v_{i,c_n}^{(m)}$, $\widehat{\beta}_{c_n}^{(m+1)}$, $(\widehat{\sigma}_{c_n}^{(m+1)})^2$ by replacing c by c_n in expressions (1.2.3), (1.2.4), (1.2.5). The corresponding sample gauge is

$$\widehat{\gamma}_{c_n}^{(m)} = \frac{1}{n}\sum_{i=1}^{n}(1 - v_{i,c_n}^{(m)}) = \frac{1}{n}\sum_{i=1}^{n} 1_{(|y_i - x_i'\widehat{\beta}_{c_n}^{(m)}| > \widehat{\sigma}_{c_n}^{(m)} c_n)}. \tag{1.3.3}$$

Theorem 1.3.2, 1.3.3 shows that any iterated estimator is tight, so lower and upper bounds can be found for the indicators appearing in the gauge. By exploring these bounds, the following Poisson limit theorem arises.

Theorem 1.3.4. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 1.2.1. Let c_n be defined from (1.3.2). Suppose Assumption 1.3.1 holds with $\eta = 1/4$. Then for all $0 \leq m < \infty$ and as $n \to \infty$, the sample gauge in (1.3.3) satisfies*

$$n\widehat{\gamma}_{c_n}^{(m)} \xrightarrow{\mathsf{D}} \text{Poisson}(\lambda).$$

			x					
λ	c_{100}	c_{200}	0	1	2	3	4	5
5	1.960	2.241	0.01	0.04	0.12	0.27	0.44	0.62
1	2.576	2.807	0.37	0.74	0.92	0.98	1.00	
0.5	2.807	3.023	0.61	0.91	0.98	1.00		
0.25	3.023	3.227	0.78	0.97	1.00			
0.1	3.291	3.481	0.90	1.00				

Table 1.3.1: The probability of detecting at most x outliers approximated by a Poisson distribution for a given λ, and the cut-off $c_n = \Phi^{-1}\{1 - \lambda/(2n)\}$ for $n = 100, 200$.

Table 1.3.1 assumes that ε_i/σ follows a standard normal distribution. For a given λ, the cut-off in (1.3.2) satisfies $c_n = \Phi^{-1}\{1 - \lambda/(2n)\}$. Cut-off values are shown for $n = 100, 200$. The Poisson approximation gives the probability of finding at most x outliers. There is an increase from 62% to 90% for the probability of detecting at most $x = \lambda$ outliers as λ declines from 5 to 0.1. The reason is due to the left skewness of the Poisson distribution. In particular, we focus on the case where $\lambda = 1$ and $n = 100$. The cut-off is $c_n = 2.58$ and the probability to find at most 1, 2 outliers are 0.74, 0.92. This means it regularly finds 2 outliers when there are none.

The Robustified Least Squares and Impulse Indicator Saturation are special versions of iterated 1-step Huber-skip M-estimators with different starting points. Their initial points do not depend on the cut-off, and thus satisfy the tightness property. Therefore, Theorems 1.3.1, 1.3.2, 1.3.3 and 1.3.4 apply for these algorithms.

1.4 Weighted and marked empirical process

Consider the weighted and marked empirical distribution function

$$\widehat{\mathsf{F}}_n^{g,p}(a, b, c) = \frac{1}{n}\sum_{i=1}^{n} g_{in}\varepsilon_i^p 1_{(\varepsilon_i \le \sigma c + n^{-1/2}ac + x_{in}'b)},$$

with $\mathcal{F}_{i-1}$ adapted weights g_{in} and $\mathcal{F}_i$ measurable marks ε_i^p. Let $a \in \mathbb{R}$, $b \in \mathbb{R}^{\dim x}$ represent estimation errors $\widetilde{a} = n^{1/2}(\widetilde{\sigma} - \sigma)$, $\widetilde{b} = N^{-1}(\widetilde{\beta} - \beta)$, while $c \in \mathbb{R}$ is the quantile. Define normalized regressors $x_{in} = N'x_i$ so that $\sum_{i=1}^n x_{in}x_{in}'$ converges. For example, $N = n^{-1/2}I_{\dim x}$ if $\{x_i\}_{i=1}^n$ is stationary, while $N = n^{-1}I_{\dim x}$ for a random walk. Our interest focuses on weights g_{in} given as either of 1, $n^{1/2}N'x_i$, $nN'x_ix_i'N$ and p as either

of 0, 1, 2. To form the empirical process, introduce the compensator

$$\overline{\mathsf{F}}_n^{g,p}(a,b,c) = \frac{1}{n}\sum_{i=1}^{n} g_{in}\mathsf{E}_{i-1}\varepsilon_i^p 1_{(\varepsilon_i \leq \sigma c + n^{-1/2}ac + x_{in}'b)}, \tag{1.4.1}$$

where $\mathsf{E}_{i-1}(\cdot) = \mathsf{E}(\cdot|\mathcal{F}_{i-1})$. Note that $\overline{\mathsf{F}}_n^{1,0}(0,0,c) = \mathsf{F}(c) = \mathsf{P}(\varepsilon_i \leq \sigma c)$. The empirical process is

$$\mathbb{F}_n^{g,p}(a,b,c) = n^{1/2}\{\widehat{\mathsf{F}}_n^{g,p}(a,b,c) - \overline{\mathsf{F}}_n^{g,p}(a,b,c)\}. \tag{1.4.2}$$

We will show convergence that is uniform in a, b, c for the above process. This generalizes results by Koul and Ossiander (1994), who had no marks and no variation a in scale, by Johansen and Nielsen (2009), who had no variation in quantile c, and by Johansen and Nielsen (2016a), who had no variation a in scale.

In the following we first present the assumptions and then the results follow.

1.4.1 Assumptions

In this section, the density f is not necessarily symmetric. The conditions listed below are weaker than Assumption 1.3.1.

Assumption 1.4.1. *Let $\mathcal{F}_i$ be an increasing sequence of σ-fields so ε_{i-1}, x_i and g_{in} are $\mathcal{F}_{i-1}$ measurable and ε_i is independent of $\mathcal{F}_{i-1}$. Let ε_i/σ have a continuously differentiable density f which is positive on $(-\infty,\infty)$. Let p, r, η, κ be given so $p, r \in \mathbb{N}_0$ and $0 \leq \kappa < \eta \leq 1/4$. Suppose*

(i) the density f satisfies

(a) moments: $\int_{-\infty}^{\infty} |u|^{2^r p}\mathsf{f}(u)du < \infty$;

(b) boundedness: $\sup_{c\in\mathbb{R}}\{(1+|c|^{2^r p+1})\mathsf{f}(c) + (1+|c|^{2^r p+2})|\dot{\mathsf{f}}(c)|\} < \infty$;

(c) smoothness: a $C_{\mathrm{H}} \in \mathbb{N}$ exists so that for all $c > 0$

$$\frac{\sup_{u\geq c}(1+u^{2^r p})\mathsf{f}(u)}{\inf_{0\leq u\leq c}(1+u^{2^r p})\mathsf{f}(u)} \leq C_{\mathrm{H}}, \qquad \frac{\sup_{u\leq -c}(1+u^{2^r p})\mathsf{f}(u)}{\inf_{-c\leq u\leq 0}(1+u^{2^r p})\mathsf{f}(u)} \leq C_{\mathrm{H}};$$

(ii) the regressors x_i satisfy $\max_{1\leq i\leq n}|n^{1/2-\kappa}N'x_i| = \mathrm{O}_{\mathsf{P}}(1)$;

(iii) the weights g_{in} satisfy

(a) $n^{-1}\mathsf{E}\sum_{i=1}^n |g_{in}|^{2^r}(1+|n^{1/2}N'x_i|) = \mathrm{O}(1)$;

(b) $n^{-1}\sum_{i=1}^n |g_{in}|(1+|n^{1/2}N'x_i|^2) = \mathrm{O}_{\mathsf{P}}(1)$.

Remark 1.4.1. *Assumption 1.3.1(ia, iib, iic) implies Assumption 1.4.1 with $r \geq 2$ satisfying (1.3.1) when g_{in} is either of 1, $n^{1/2}N'x_i$, $nN'x_ix_i'N$ and p is either of 0, 1, 2. Details are given in Lemma 1.7.2 in the appendix.*

1.4.2 A new empirical process result

We present two asymptotic results. The first theorem shows that the estimation error for the scale and regression parameters is negligible uniformly in the quantile.

Theorem 1.4.1. *Let $c_\psi = \mathsf{F}^{-1}(\psi)$. Suppose Assumption 1.4.1 holds with $r \geq 2$ satisfying (1.3.1). Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B} |\mathbb{F}_n^{g,p}(a,b,c_\psi) - \mathbb{F}_n^{g,p}(0,0,c_\psi)| = \mathsf{o}_\mathsf{P}(1).$$

The proof involves a chaining argument. For this, we apply an iterated martingale inequality, see Lemma 1.6.3, to explore the tail behaviour of the maximum of a family of martingales. We can, however, split this argument into two parts. First, we keep a fixed and consider variation in b, c. This has already been done in Theorem 4.1 of Johansen and Nielsen (2016a). Secondly, we keep b fixed and consider variation in a, c. This proof is new, and the details of its chaining argument are slightly different.

The second result provides a linearization of the compensator.

Theorem 1.4.2. *Let $c_\psi = \mathsf{F}^{-1}(\psi)$. Suppose Assumption 1.4.1(ib, iiib) holds with $r = 0$. Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B} |n^{1/2}\{\overline{\mathsf{F}}_n^{g,p}(a,b,c_\psi) - \overline{\mathsf{F}}_n^{g,p}(0,0,c_\psi)\} - \sigma^{p-1}c_\psi^p \mathsf{f}(c_\psi) n^{-1/2}\sum_{i=1}^n g_{in}(n^{-1/2}ac_\psi + x_{in}'b)| = \mathrm{O}_\mathsf{P}(n^{-2\eta}).$$

Finally, Johansen and Nielsen (2016a, Theorem 4.4) showed the process $\mathbb{F}_n^{g,p}(0,0,c_\psi)$ is tight, see tightness in Billingsley (1968).

1.4.3 A result for the two-sided empirical process

The 1-step Huber-skip M-estimator involves indicators depending on the absolute value of the residuals. We therefore present some results for a class of two-sided weighted and marked empirical processes.

Define the weighted and marked absolute empirical distribution function

$$\widehat{\mathsf{G}}_n^{g,p}(a,b,c) = \frac{1}{n}\sum_{i=1}^n g_{in}\varepsilon_i^p 1_{(|\varepsilon_i - x_{in}'b| \leq \sigma c + n^{-1/2}ac)}. \tag{1.4.3}$$

We suppose a such that $\sigma + n^{-1/2}a > 0$, in which case it suffices to consider $c \geq 0$. This restriction on a is satisfied when choosing a as $\widetilde{a} = n^{1/2}(\widetilde{\sigma} - \sigma)$ so that $\sigma + n^{-1/2}\widetilde{a} = \widetilde{\sigma} > 0$. Introduce the compensator of $\widehat{\mathsf{G}}_n^{g,p}(a,b,c)$

$$\overline{\mathsf{G}}_n^{g,p}(a,b,c) = \frac{1}{n}\sum_{i=1}^{n} g_{in}\mathsf{E}_{i-1}\varepsilon_i^p 1_{(|\varepsilon_i - x_{in}'b| \leq \sigma c + n^{-1/2}ac)}. \tag{1.4.4}$$

Note that $\overline{\mathsf{G}}_n^{1,0}(0,0,c) = \mathsf{G}(c) = \mathsf{P}(|\varepsilon_i| \leq \sigma c)$. Then the absolute empirical process is

$$\mathbb{G}_n^{g,p}(a,b,c) = n^{1/2}\{\widehat{\mathsf{G}}_n^{g,p}(a,b,c) - \overline{\mathsf{G}}_n^{g,p}(a,b,c)\}. \tag{1.4.5}$$

We can now derive asymptotic theory for the absolute empirical process from Theorems 1.4.1, 1.4.2. These results are presented under more restrictive Assumption 1.3.1, where the innovation distribution is symmetric, see Remark 1.4.1 and Lemma 1.7.2. In this section, we only consider g_{in} chosen as 1, $n^{1/2}N'x_i$, $nN'x_ix_i'N$ and p as 0, 1, 2.

Theorem 1.4.3. *Let $c_\psi = \mathsf{G}^{-1}(\psi)$. Suppose Assumption 1.3.1(ia, iib, iic) holds. Then for all $B > 0$ and as $n \to \infty$*

$$\sup_{0 \leq \psi \leq 1} \sup_{|a|,|b| \leq n^{1/4-\eta}B} |\mathbb{G}_n^{g,p}(a,b,c_\psi) - \mathbb{G}_n^{g,p}(0,0,c_\psi)| = \mathrm{o}_{\mathsf{P}}(1).$$

Theorem 1.4.4. *Let $c_\psi = \mathsf{G}^{-1}(\psi)$. Suppose Assumption 1.3.1(ia, iic) holds. Then for all $B > 0$ and as $n \to \infty$*

$$\sup_{0 \leq \psi \leq 1} \sup_{|a|,|b| \leq n^{1/4-\eta}B} |n^{1/2}\{\overline{\mathsf{G}}_n^{g,p}(a,b,c_\psi) - \overline{\mathsf{G}}_n^{g,p}(0,0,c_\psi)\}$$
$$- 2\sigma^{p-1}c_\psi^p \mathsf{f}(c_\psi)n^{-1/2}\sum_{i=1}^{n} g_{in}\{1_{(p \text{ is even})}n^{-1/2}ac_\psi + 1_{(p \text{ is odd})}x_{in}'b\}| = \mathrm{O}_{\mathsf{P}}(n^{-2\eta}).$$

1.5 Conclusion

This chapter contributes to the asymptotic theory of iterated 1-step Huber-skip M-estimators. The results are derived under the null hypothesis that there are no outliers in the model. It is well known that the first order asymptotic approximation is fragile in some small finite sample situations. Therefore, it would be of interest to carry out simulation studies to evaluate the finite sample performance of the results in this chapter. Likewise it would be of interest to extend the result to situations where outliers are actually present in the data generating process. Scenario possibly contain single outliers, clusters of outliers, level shifts, symmetric or non-symmetric outliers. In such situations,

we would analyze the potency, which is the retention rate for relevant outliers. Moreover, it would be possible to compare the potency of two distinct outlier detection algorithms with the same gauge.

1.6 Appendix A: A metric on $\mathbb{R}$ and some inequalities

The asymptotic theory uses a chaining argument. This involves a partitioning of the quantile axis using a metric on $\mathbb{R}$, which is presented first. Then follows some preliminary inequalities including an iterated exponential martingale inequality.

Define the function

$$J_{i,p}(x,y) = (\frac{\varepsilon_i}{\sigma})^p \{1_{(\varepsilon_i/\sigma \leq y)} - 1_{(\varepsilon_i/\sigma \leq x)}\}. \tag{1.6.1}$$

Our interest focus on $J_{i,p}(x,y)$ of order 2^r with $r \in \mathbb{N}$. Note that $u^{2^r p}$ is non-negative since $2^r p$ is even for $p \in \mathbb{N}_0$ and $r \in \mathbb{N}$. Introduce a positive and increasing function

$$\mathsf{H}_r(x) = \int_{-\infty}^{x} (1 + u^{2^r p}) \mathsf{f}(u) du. \tag{1.6.2}$$

The derivative of this function is $\dot{\mathsf{H}}_r(x) = (1 + x^{2^r p})\mathsf{f}(x)$. Then, denote the constant

$$H_r = \mathsf{H}_r(\infty) = \int_{-\infty}^{\infty} (1 + u^{2^r p}) \mathsf{f}(u) du, \tag{1.6.3}$$

which is finite by Assumption 1.4.1(ia). Selection of the specific $r \in \mathbb{N}$ will be more clear in proofs of the empirical process results. The intuition of $\mathsf{H}_r(x)$ is obtained through setting $p = 0$ so that $\mathsf{H}_r(x) = 2\mathsf{F}(x)$, $\dot{\mathsf{H}}_r(x) = 2\mathsf{f}(x)$ and $H_r = 2$. Therefore, $\mathsf{H}_r(x)$ is the generalization of the distribution $\mathsf{F}(x) \sim \varepsilon_i/\sigma$. For $x \leq y$ and $0 \leq s \leq r$, then

$$0 \leq |\mathsf{E}\{J_{i,p}(x,y)^{2^s}\}| \leq \mathsf{E}\{|J_{i,p}(x,y)|^{2^s}\} \leq \mathsf{H}_r(y) - \mathsf{H}_r(x), \tag{1.6.4}$$

since $|u^p| < |u^q| + 1$ for $q \geq p \geq 0$. Let $|\mathsf{H}_r(x) - \mathsf{H}_r(y)|$ be the H_r-distance for $x, y \in \mathbb{R}$.

In the context of chaining, partition the range of $\mathsf{H}_r(c)$ into K intervals of equal size H_r/K. In other words, partition the support into K intervals by endpoints

$$-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = \infty, \tag{1.6.5}$$

with $c_{-k} = c_0$ for $k \in \mathbb{N}$ so that for $1 \leq k \leq K$

$$\mathsf{H}_r(c_k) - \mathsf{H}_r(c_{k-1}) = \frac{H_r}{K}. \tag{1.6.6}$$

We first present two preliminary inequalities.

Lemma 1.6.1. *If* $|\tilde{c} - c| \leq |Ac + B|$ *and* $|A| \leq 1/2$*, then*

$$|c| \leq \frac{|\tilde{c}| + |B|}{1 - |A|}, \qquad (Ac + B)^2 \leq 16(A^2\tilde{c}^2 + B^2).$$

Proof of Lemma 1.6.1. *First inequality.* Since $|Ac+B| \leq |A||c|+|B|$, the assumption implies $c - |A||c| - |B| \leq \tilde{c} \leq c + |A||c| + |B|$. Suppose $c \geq 0$, then the lower inequality gives $c(1-|A|) - |B| \leq \tilde{c}$ so that $c \leq (\tilde{c}+|B|)/(1-|A|)$. Suppose $c < 0$, then the upper inequality gives $\tilde{c} \leq c(1-|A|) + |B|$ so that $(\tilde{c} - |B|)/(1-|A|) \leq c$. Combine to get $|c| \leq \max\{|(\tilde{c}+|B|)/(1-|A|)|, |(\tilde{c}-|B|)/(1-|A|)|\} \leq (|\tilde{c}| + |B|)/(1-|A|)$.

Second inequality. The first inequality in the lemma, $(x+y)^2 \leq 2(x^2+y^2)$ and $|A| \leq 1/2$ imply $c^2 \leq 8(\tilde{c}^2 + B^2)$ and $(Ac+B)^2 \leq 2(A^2c^2 + B^2)$. Combine to get $(Ac+B)^2 \leq 2(8A^2\tilde{c}^2 + 8A^2B^2 + B^2) \leq 16(A^2\tilde{c}^2 + B^2)$. ■

The following lemma concerns the H_r-distance of multiplicative shifts.

Lemma 1.6.2. *Let* $c_\psi = \mathsf{F}^{-1}(\psi)$*. Suppose Assumption 1.4.1*(ia, ib) *holds. For any* $B > 0$*, there exists* $C > 0$ *so*

$$\sup_{0 \leq \psi \leq 1} \sup_{|a| \leq n^{1/4-\eta}B} |\mathsf{H}_r\{c_\psi(1 + n^{-1/2}\frac{a}{\sigma})\} - \mathsf{H}_r(c_\psi)| \leq Cn^{-1/4-\eta}.$$

Proof of Lemma 1.6.2. Denote $\mathcal{H} = |\mathsf{H}_r\{c_\psi(1 + n^{-1/2}a/\sigma)\} - \mathsf{H}_r(c_\psi)|$. Apply the first order mean value theorem at the point c_ψ to get $\mathcal{H} = |\sigma^{-1}n^{-1/2}a||c_\psi||\dot{\mathsf{H}}_r(\tilde{c}_\psi)|$, where $|\tilde{c}_\psi - c_\psi| \leq |\sigma^{-1}n^{-1/2}ac_\psi|$ and $\dot{\mathsf{H}}_r(\tilde{c}_\psi) = (1 + \tilde{c}_\psi^{2r_p})\mathsf{f}(\tilde{c}_\psi)$.

There exists n_0, so for any $n > n_0$ then $|\sigma^{-1}n^{-1/2}a| \leq 1/2$ uniformly in $|a| \leq n^{1/4-\eta}B$. First, for $n > n_0$, the first inequality in Lemma 1.6.1 shows $|c_\psi| \leq |\tilde{c}_\psi|/(1 - |\sigma^{-1}n^{-1/2}a|)$ so that $|c_\psi| \leq 2|\tilde{c}_\psi|$. It follows

$$\mathcal{H} \leq \sigma^{-1}n^{-1/2}n^{1/4-\eta}B2|\tilde{c}_\psi||\dot{\mathsf{H}}_r(\tilde{c}_\psi)| \leq 2\sigma^{-1}B \sup_{c \in \mathbb{R}} |c||\dot{\mathsf{H}}_r(c)|n^{-1/4-\eta}.$$

Thus $\mathcal{H} \leq Cn^{-1/4-\eta}$ by Assumption 1.4.1(ib) that $\sup_{c\in\mathbb{R}} |c||\dot{\mathsf{H}}_r(c)| < \infty$.

Second, consider $n \leq n_0$. Note $\mathsf{H}_r(x) \leq \mathsf{H}_r(\infty) = H_r$ for any x so that the triangle inequality shows $\mathcal{H} \leq 2H_r$. With $0 < \eta \leq 1/4$, it follows

$$\mathcal{H} \leq 2H_r n^{1/4+\eta}n^{-1/4-\eta} \leq 2H_r n_0^{1/4+\eta}n^{-1/4-\eta} = Cn^{-1/4-\eta},$$

where $C = 2H_r n_0^{1/4+\eta}$ is finite since $H_r < \infty$ by Assumption 1.4.1(ia). ■

The chaining argument involves the tail behaviour of the maximum of a family of martingales which can be controlled using the following iterated martingale inequality taken from Johansen and Nielsen (2016a). It builds on an exponential martingale inequality derived by Bercu and Touati (2008, Theorem 2.1).

Lemma 1.6.3. *(Johansen and Nielsen, 2016a, Theorem 5.2) For l so $1 \leq l \leq L$, let $z_{l,i}$ be $\mathcal{F}_i$ adapted satisfying $\mathrm{E} z_{l,i}^{2^{\bar{r}}} < \infty$ for some $\bar{r} \in \mathbb{N}$. Let $D_r = \max_{1 \leq l \leq L} \sum_{i=1}^{n} \mathsf{E}_{i-1} z_{l,i}^{2^r}$ for $1 \leq r \leq \bar{r}$. Suppose, for some $\varsigma \geq 0$, $\lambda > 0$, that $L = \mathrm{O}(n^{\lambda})$ and $\mathrm{E} D_r = \mathrm{O}(n^{\varsigma})$ for $r \leq \bar{r}$. If $\upsilon > 0$ is chosen such that*

(i) $\varsigma < 2\upsilon$;

(ii) $\varsigma + \lambda < \upsilon 2^{\bar{r}}$;

then, for all $\kappa > 0$ and as $n \to \infty$

$$\lim_{n \to \infty} \mathsf{P}\{\max_{1 \leq l \leq L} |\sum_{i=1}^{n} (z_{l,i} - \mathsf{E}_{i-1} z_{l,i})| > \kappa n^{\upsilon}\} = 0.$$

1.7 Appendix B: Proofs of the empirical process results

We start by proving Theorem 1.4.1 when $b = 0$. Subsequently, we combine this with Theorem 4.1 in Johansen and Nielsen (2016a) to prove Theorem 1.4.1 for general b.

Lemma 1.7.1. *Let $c_{\psi} = \mathsf{F}^{-1}(\psi)$. Suppose Assumption 1.4.1(ia, ib, iii) holds with $r = 2$. Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0 \leq \psi \leq 1} \sup_{|a| \leq n^{1/4-\eta} B} |\mathbb{F}_n^{g,p}(a, 0, c_{\psi}) - \mathbb{F}_n^{g,p}(0, 0, c_{\psi})| = \mathrm{o}_{\mathsf{P}}(1).$$

Proof of Lemma 1.7.1. Let $c_{\psi^\dagger} = c_{\psi}(1 + n^{-1/2} a/\sigma)$ so $\mathbb{F}_n^{g,p}(a, 0, c_{\psi}) = \mathbb{F}_n^{g,p}(0, 0, c_{\psi^\dagger})$. Note $c_{\psi^\dagger}$ can be greater or less than c_{ψ}, since a such that $|a| \leq n^{1/4-\eta} B$ and c_{ψ} can be either positive or negative. Assume $c_{\psi} < c_{\psi^\dagger}$ without loss of generality. Denote $R(c_{\psi}, c_{\psi^\dagger}) = \mathbb{F}_n^{g,p}(0, 0, c_{\psi^\dagger}) - \mathbb{F}_n^{g,p}(0, 0, c_{\psi})$. The aim is to prove $\mathcal{R}_n = \mathrm{o}_{\mathsf{P}}(1)$ for $n \to \infty$ where $\mathcal{R}_n = \sup_{0 \leq \psi \leq 1} \sup_{|a| \leq n^{1/4-\eta} B} |R(c_{\psi}, c_{\psi^\dagger})|$.

1. *Partition the support.* For $\delta, n > 0$ partition the range of quantiles c as laid out in (1.6.5) with $K = \mathrm{int}(H_r n^{1/2}/\delta)$, since $H_r < \infty$ for $r = 2$ by Assumption 1.4.1(ia).

2. *Assign c_{ψ} and $c_{\psi^\dagger}$ to the partitioned support.* For each ψ and $\psi^\dagger$ there exist $k \leq k^\dagger$ and grid points so that $c_{k-1} < c_{\psi} \leq c_k$ and $c_{k^\dagger - 1} < c_{\psi^\dagger} \leq c_{k^\dagger}$.

3. *Apply chaining.* Relate c_{ψ} to the nearest right grid point c_k and $c_{\psi^\dagger}$ to the nearest left grid point $c_{k^\dagger - 1}$. Add and subtract $\mathbb{F}_n^{g,p}(0, 0, c_k)$ and $\mathbb{F}_n^{g,p}(0, 0, c_{k^\dagger - 1})$ to $R(c_{\psi}, c_{\psi^\dagger})$. The triangle inequality gives $|R(c_{\psi}, c_{\psi^\dagger})| \leq |R(c_{\psi}, c_k)| + |R(c_k, c_{k^\dagger - 1})| + |R(c_{k^\dagger - 1}, c_{\psi^\dagger})|$. Note that if c_{ψ}, $c_{\psi^\dagger}$ are in the same interval, then $|R(c_k, c_{k^\dagger - 1})| = |R(c_{k-1}, c_k)|$. If c_{ψ},

$c_{\psi^\dagger}$ are in the neighbouring intervals, then $|R(c_k, c_{k^\dagger-1})| = 0$. Apply chaining to obtain $\mathcal{R}_n \leq \mathcal{R}_{n,1} + \mathcal{R}_{n,2} + \mathcal{R}_{n,3} + \mathcal{R}_{n,4}$, where

$$\mathcal{R}_{n,1} = \max_{1\leq k<k^\dagger-1<K} |R(c_k, c_{k^\dagger-1})|, \qquad \mathcal{R}_{n,2} = \max_{1\leq k\leq K} |R(c_{k-1}, c_k)|,$$
$$\mathcal{R}_{n,3} = \max_{1\leq k\leq K} \sup_{c_{k-1}<c_\psi\leq c_k} |R(c_\psi, c_k)|, \qquad \mathcal{R}_{n,4} = \max_{1\leq k^\dagger\leq K} \sup_{c_{k^\dagger-1}<c_{\psi^\dagger}\leq c_{k^\dagger}} |R(c_{k^\dagger-1}, c_{\psi^\dagger})|.$$

Thus, it suffices to show $\mathcal{R}_{n,j} = \mathrm{o_P}(1)$ for $j = 1, 2, 3, 4$ as $n \to \infty$.

4. *The term* $\mathcal{R}_{n,1}$ *is* $\mathrm{o_P}(1)$. Use Lemma 1.6.3 with $\upsilon = 1/2$. Let g_{in} have coordinates $g^*_{in} = \sigma^p g_{in}$. Recall the notation $J_{i,p}(x, y)$ in (1.6.1). Write the coordinates of $R(c_k, c_{k^\dagger-1})$ as $n^{-1/2}\sum_{i=1}^n (z_{l,i} - \mathsf{E}_{i-1} z_{l,i})$ with $z_{l,i} = g^*_{in} J_{i,p}(c_k, c_{k^\dagger-1})$, where l represents the indices $k, k^\dagger$ with $L \leq K^2$. Two conditions of Lemma 1.6.3 need to be verified.

The parameter λ. The set of indices l has the size $L = \mathrm{O}(n^\lambda)$ where $\lambda = 1$, since $L \leq K^2$ and $K = \mathrm{O}(n^{1/2})$.

The parameter ς. Consider $1 \leq r \leq \bar{r} = 2$. By construction of partition and assignment in steps 1, 2, then $c_\psi \leq c_k < c_{k^\dagger-1} < c_{\psi^\dagger}$. Thus,

$$\mathsf{E}_{i-1} J^{2^r}_{i,p}(c_k, c_{k^\dagger-1}) \leq \mathsf{H}_r(c_{k^\dagger-1}) - \mathsf{H}_r(c_k) \leq \mathsf{H}_r(c_{\psi^\dagger}) - \mathsf{H}_r(c_\psi) \leq C n^{-1/4-\eta},$$

by Lemma 1.6.2 using Assumption 1.4.1(ia, ib) for some finite $C > 0$. We then find

$$D_r = \max_{1\leq l\leq L} \sum_{i=1}^n \mathsf{E}_{i-1} z^{2^r}_{l,i} = \max_{1\leq k<k^\dagger-1<K} \sum_{i=1}^n g^{*2^r}_{in} \mathsf{E}_{i-1} J^{2^r}_{i,p}(c_k, c_{k^\dagger-1}) \leq C n^{-1/4-\eta} \sum_{i=1}^n g^{*2^r}_{in}.$$

Assumption 1.4.1$(iiia)$ shows $\mathsf{E} n^{-1}\sum_{i=1}^n g^{*2^r}_{in} = \mathrm{O}(1)$. Thus, with $\varsigma = 3/4 - \eta$, we have $\mathrm{E}D_r = \mathrm{O}(n^\varsigma)$.

Condition (i) is that $\varsigma < 2\upsilon$. This holds since $\eta > 0$ so $\varsigma = 3/4 - \eta < 1 = 2\upsilon$.

Condition (ii) is that $\varsigma + \lambda < \upsilon 2^{\bar{r}}$ where $\bar{r} = 2$. This is satisfied since $\eta > 0$ so $\varsigma + \lambda = 7/4 - \eta < 2 = \upsilon 2^{\bar{r}}$.

5. *The term* $\mathcal{R}_{n,2}$ *is* $\mathrm{o_P}(1)$. Use Lemma 1.6.3 with $\upsilon = 1/2$ and $z_{l,i} = g^*_{in} J_{i,p}(c_{k-1}, c_k)$, where index $l = k$ has the size $L = K$. Two conditions of Lemma 1.6.3 need to be shown.

The parameter λ. The size $L = \mathrm{O}(n^\lambda)$ where $\lambda = 1/2$, since $L = K$ and $K = \mathrm{O}(n^{1/2})$.

The parameter ς. Consider $1 \leq r \leq \bar{r} = 2$. The equality (1.6.6) shows

$$\mathsf{E}_{i-1} J^{2^r}_{i,p}(c_{k-1}, c_k) \leq \mathsf{H}_r(c_k) - \mathsf{H}_r(c_{k-1}) = \frac{H_r}{K} = \mathrm{O}(n^{-1/2}).$$

Then, we find

$$D_r = \max_{1\leq l\leq L} \sum_{i=1}^n \mathsf{E}_{i-1} z^{2^r}_{l,i} = \max_{1\leq k\leq K} \sum_{i=1}^n g^{*2^r}_{in} \mathsf{E}_{i-1} J^{2^r}_{i,p}(c_{k-1}, c_k) = \mathrm{O}(n^{-1/2}) \sum_{i=1}^n g^{*2^r}_{in}.$$

It follows that $\mathrm{E}D_r = \mathrm{O}(n^{\varsigma})$ where $\varsigma = 1/2$ by Assumption 1.4.1(*iiia*).

Condition (*i*) holds, since $\varsigma = 1/2 < 1 = 2\upsilon$.

Condition (*ii*) holds, since $\varsigma + \lambda = 1 < 2 = \upsilon 2^{\bar{r}}$.

6. *Decompose the term* $\mathcal{R}_{n,3}$. Apply the triangle and Jensen's inequality to obtain,

$$|R(c_\psi, c_k)| \leq n^{-1/2} \sum_{i=1}^{n} |g_{in}^*| \{ |J_{i,p}(c_\psi, c_k)| + \mathsf{E}_{i-1} |J_{i,p}(c_\psi, c_k)| \}.$$

For $c_{k-1} < c_\psi \leq c_k$ where $1 \leq k \leq K$, we have $|J_{i,p}(c_\psi, c_k)| \leq |J_{i,p}(c_{k-1}, c_k)|$. Then,

$$\mathcal{R}_{n,3} \leq \max_{1 \leq k \leq K} n^{-1/2} \sum_{i=1}^{n} |g_{in}^*| \{ |J_{i,p}(c_{k-1}, c_k)| + \mathsf{E}_{i-1} |J_{i,p}(c_{k-1}, c_k)| \}.$$

Therefore, it can be argued that $\mathcal{R}_{n,3} \leq \widetilde{\mathcal{R}}_{n,3} + 2\overline{\mathcal{R}}_{n,3}$, where

$$\widetilde{\mathcal{R}}_{n,3} = \max_{1 \leq k \leq K} n^{-1/2} \sum_{i=1}^{n} |g_{in}^*| \{ |J_{i,p}(c_{k-1}, c_k)| - \mathsf{E}_{i-1} |J_{i,p}(c_{k-1}, c_k)| \},$$

$$\overline{\mathcal{R}}_{n,3} = \max_{1 \leq k \leq K} n^{-1/2} \sum_{i=1}^{n} |g_{in}^*| \mathsf{E}_{i-1} |J_{i,p}(c_{k-1}, c_k)|.$$

Thus, it suffices to show $\widetilde{\mathcal{R}}_{n,3}$ and $\overline{\mathcal{R}}_{n,3}$ are $\mathrm{o}_{\mathsf{P}}(1)$ as $n \to \infty$.

7. *The term* $\widetilde{\mathcal{R}}_{n,3}$ *is* $\mathrm{o}_{\mathsf{P}}(1)$. Argue along the lines of step 5 to show $\widetilde{\mathcal{R}}_{n,3} = \mathrm{o}_{\mathsf{P}}(1)$.

8. *Bounding the term* $\overline{\mathcal{R}}_{n,3}$. Use the equality (1.6.6) and $K = \mathrm{O}(H_r n^{1/2}/\delta)$ to get

$$\mathsf{E}_{i-1} |J_{i,p}(c_{k-1}, c_k)| \leq \mathsf{H}_r(c_k) - \mathsf{H}_r(c_{k-1}) = \frac{H_r}{K} = \mathrm{O}(n^{-1/2}\delta).$$

We then find $\overline{\mathcal{R}}_{n,3} = \mathrm{O}(n^{-1/2}\delta) n^{-1/2} \sum_{i=1}^{n} |g_{in}^*| = \mathrm{O}_{\mathsf{P}}(\delta)$ by Assumption 1.4.1(*iiib*) that $n^{-1} \sum_{i=1}^{n} |g_{in}^*| = \mathrm{O}_{\mathsf{P}}(1)$. Thus, choosing δ sufficiently small, $\overline{\mathcal{R}}_{n,3}$ is small in probability.

9. *The term* $\mathcal{R}_{n,4}$ *is* $\mathrm{o}_{\mathsf{P}}(1)$. This is similar as to show $\mathcal{R}_{n,3} = \mathrm{o}_{\mathsf{P}}(1)$. Thus the same argument can be made through steps $6, 7, 8$. ■

Proof of Theorem 1.4.1. The term of interest is $\mathcal{W} = \mathbb{F}_n^{g,p}(a, b, c_\psi) - \mathbb{F}_n^{g,p}(0, 0, c_\psi)$. Denote $c_{\psi^\dagger} = c_\psi(1 + n^{-1/2}a/\sigma)$. Notice that $\mathbb{F}_n^{g,p}(a, b, c_\psi) = \mathbb{F}_n^{g,p}(0, b, c_{\psi^\dagger})$ so it follows that $\mathcal{W} = \mathbb{F}_n^{g,p}(0, b, c_{\psi^\dagger}) - \mathbb{F}_n^{g,p}(0, 0, c_\psi)$. Add and subtract $\mathbb{F}_n^{g,p}(a, 0, c_\psi) = \mathbb{F}_n^{g,p}(0, 0, c_{\psi^\dagger})$ and apply the triangle inequality to get

$$|\mathcal{W}| \leq |\mathbb{F}_n^{g,p}(0, b, c_{\psi^\dagger}) - \mathbb{F}_n^{g,p}(0, 0, c_{\psi^\dagger})| + |\mathbb{F}_n^{g,p}(a, 0, c_\psi) - \mathbb{F}_n^{g,p}(0, 0, c_\psi)|.$$

Thus, the problem reduces to showing

$$\sup_{0 \leq \psi^\dagger \leq 1} \sup_{|b| \leq n^{1/4-\eta} B} |\mathbb{F}_n^{g,p}(0, b, c_{\psi^\dagger}) - \mathbb{F}_n^{g,p}(0, 0, c_{\psi^\dagger})| = \mathrm{o}_{\mathsf{P}}(1), \tag{1.7.1}$$

$$\sup_{0 \leq \psi \leq 1} \sup_{|a| \leq n^{1/4-\eta} B} |\mathbb{F}_n^{g,p}(a, 0, c_\psi) - \mathbb{F}_n^{g,p}(0, 0, c_\psi)| = \mathrm{o}_{\mathsf{P}}(1). \tag{1.7.2}$$

Then (1.7.1) was considered by Theorem 4.1 in Johansen and Nielsen (2016a) using Assumption 1.4.1$(i, ii, iiia)$ with $r \geq 2$ such that (1.3.1) holds. Further, (1.7.2) follows from Lemma 1.7.1, which requires Assumption 1.4.1(ia, ib, iii) with $r = 2$. ■

Proof of Theorem 1.4.2. The term of interest is

$$D_n(a,b,c_\psi) = n^{1/2}\{\overline{\mathsf{F}}_n^{g,p}(a,b,c_\psi) - \overline{\mathsf{F}}_n^{g,p}(0,0,c_\psi)\} \\ - \sigma^{p-1}c_\psi^p \mathsf{f}(c_\psi) n^{-1/2} \sum_{i=1}^n g_{in}(n^{-1/2}ac_\psi + x_{in}'b).$$

Let $w_i^{a,b,c_\psi} = 1_{(\varepsilon_i \leq \sigma c_\psi + n^{-1/2}ac_\psi + x_{in}'b)} - 1_{(\varepsilon_i \leq \sigma c_\psi)}$, $h_i(a,b,c_\psi) = (n^{-1/2}ac_\psi + x_{in}'b)/\sigma$ and $s(c) = c^p\mathsf{f}(c)$. Define $S_i(a,b,c_\psi) = \mathsf{E}_{i-1}\varepsilon_i^p w_i^{a,b,c_\psi} - \sigma^p h_i(a,b,c_\psi)s(c_\psi)$ then we have $D_n(a,b,c_\psi) = n^{-1/2}\sum_{i=1}^n g_{in}S_i(a,b,c_\psi)$. Write $S_i(a,b,c_\psi)$ as an integral and apply the second order Taylor expansion at c_ψ to obtain

$$S_i(a,b,c_\psi) = \sigma^p\{\int_{c_\psi}^{c_\psi + h_i(a,b,c_\psi)} s(u)du - h_i(a,b,c_\psi)s(c_\psi)\} = \frac{\sigma^p}{2}h_i^2(a,b,c_\psi)\dot{s}(\tilde{c}_\psi),$$

where $|\tilde{c}_\psi - c_\psi| \leq |h_i(a,b,c_\psi)|$. There exists $n_0 > 0$ so for any $n > n_0$ we have $|\sigma^{-1}n^{-1/2}a| \leq 1/2$. We then apply the second inequality in Lemma 1.6.1 to obtain $h_i^2(a,b,c_\psi) \leq 16\{n^{-1}a^2\tilde{c}_\psi^2 + (x_{in}'b)^2\}/\sigma^2$. Exploit bounds $|a|, |b| \leq n^{1/4-\eta}B$ and the inequality $x^2 + y^2 \leq (1+x^2)(1+y^2)$ to get

$$|S_i(a,b,c_\psi)| = \mathrm{O}(n^{-1/2-2\eta})(1 + |n^{1/2}x_{in}|^2)(1 + \tilde{c}_\psi^2)|\dot{s}(\tilde{c}_\psi)|.$$

Since $(1 + \tilde{c}_\psi^2)|\dot{s}(\tilde{c}_\psi)| \leq \sup_{c\in\mathbb{R}}(1+c^2)|\dot{s}(c)| < \infty$ by Assumption 1.4.1(ib) with $r = 0$, then $|S_i(a,b,c_\psi)| = \mathrm{O}(n^{-1/2-2\eta})(1 + |n^{1/2}x_{in}|^2)$ uniformly in ψ, a, b. Triangle inequality gives

$$|D_n(a,b,c_\psi)| \leq n^{-1/2}\sum_{i=1}^n |g_{in}||S_i(a,b,c_\psi)| = \mathrm{O}(n^{-2\eta})n^{-1}\sum_{i=1}^n |g_{in}|(1 + |n^{1/2}x_{in}|^2).$$

By Assumption 1.4.1$(iiib)$, this term is of order $\mathrm{O}_\mathsf{P}(n^{-2\eta})$ uniformly in ψ, a, b. ■

The absolute empirical process results are given under more restrictive Assumption 1.3.1, so the next lemma concerns the relationship between Assumption 1.3.1 and 1.4.1.

Lemma 1.7.2. *Suppose g_{in} is either of 1, $n^{1/2}N'x_i$, $nN'x_ix_i'N$ and p is either of 0, 1, 2. Then Assumption 1.3.1(ia, iib, iic) implies Assumption 1.4.1 with $r \geq 2$ satisfying (1.3.1).*

Proof of Lemma 1.7.2. Assumption 1.3.1(ia) shows Assumption 1.4.1(ia, ic), while Assumption 1.4.1(ib) further needs continuous differentiability of f, see discussion in Johansen and Nielsen (2016a, Remark 4.1(c)). Assumption 1.3.1(iib) is the same as Assumption 1.4.1(ii). Assumption 1.3.1(iic) implies Assumption 1.4.1($iiia$) and ($iiic$) by the Markov inequality. ■

Proof of Theorem 1.4.3. The term of interest is $\mathcal{G} = \mathbb{G}_n^{g,p}(a, b, c_\psi) - \mathbb{G}_n^{g,p}(0, 0, c_\psi)$. Our focus is on the absolute quantile $c_\psi = \mathsf{G}^{-1}(\psi) > 0$ rather than the one-sided quantile $c_{\psi^*} = \mathsf{F}^{-1}(\psi^*) \in \mathbb{R}$. Note $|\varepsilon_i|/\sigma \sim \mathsf{G}$ and $\varepsilon_i/\sigma \sim \mathsf{F}$. Since

$$1_{(|\varepsilon_i - x'_{in}b| \leq \sigma c + n^{-1/2}ac)} = 1_{(\varepsilon_i \leq \sigma c + n^{-1/2}ac + x'_{in}b)} - 1_{(\varepsilon_i \leq -\sigma c - n^{-1/2}ac + x'_{in}b)}$$

and by (1.4.2), (1.4.5), we have $\mathbb{G}_n^{g,p}(a, b, c) = \mathbb{F}_n^{g,p}(a, b, c) - \lim_{c^\dagger \downarrow c} \mathbb{F}_n^{g,p}(a, b, -c^\dagger)$ for any $c > 0$. By this and the triangle inequality, then for any $c_\psi = \mathsf{G}^{-1}(\psi) > 0$,

$$|\mathcal{G}| \leq |\mathbb{F}_n^{g,p}(a, b, c_\psi) - \mathbb{F}_n^{g,p}(0, 0, c_\psi)| + \lim_{c_\psi^\dagger \downarrow c_\psi} |\mathbb{F}_n^{g,p}(a, b, -c_\psi^\dagger) - \mathbb{F}_n^{g,p}(0, 0, -c_\psi^\dagger)|.$$

These vanish uniformly in ψ, a, b by Theorem 1.4.1 using Assumption 1.4.1 with $r \geq 2$ such that (1.3.1) holds. Lemma 1.7.2 shows that Assumption 1.3.1(ia, iib, iic) suffices. ■

Proof of Theorem 1.4.4. Argue as in the proof of Theorem 1.4.3 but using Theorem 1.4.2 instead of Theorem 1.4.1. Due to the symmetry of f, the correction term is then

$$\sigma^{p-1} c_\psi^p \mathsf{f}(c_\psi) n^{-1/2} \sum_{i=1}^n g_{in}[\{1 + (-1)^p\} n^{-1/2} a c_\psi + \{1 - (-1)^p\} x'_{in} b].$$

This reduces as desired. ■

1.8 Appendix C: Proofs of the main results

We first present an anxillary result for asymptotic expansions of product moments. Then, the tightness and fixed point result are shown for the iterated estimators. At last, we provide the proof of the Poisson exceedence theory for the gauge.

1.8.1 An auxillary result for product moments

The 1-step Huber-skip M-estimators are least squares estimators for selected observations. The following result describes the asymptotic behaviour of the corresponding product moments. For this purpose introduce the indicators

$$v_i^{a,b,c} = 1_{(|\varepsilon_i - x'_{in}b| \leq \sigma c + n^{-1/2}ac)}. \tag{1.8.1}$$

Lemma 1.8.1. *Suppose Assumption 1.3.1(ia, ii) holds. Then we have expansions*

$$\begin{aligned}
n^{-1/2}\sum_{i=1}^{n} v_i^{a,b,c} &= n^{-1/2}\sum_{i=1}^{n} 1_{(|\varepsilon_i|\le\sigma c)} + 2\mathsf{f}(c)\frac{ac}{\sigma} + R_v(a,b,c),\\
n^{-1/2}\sum_{i=1}^{n} \varepsilon_i^2 v_i^{a,b,c} &= n^{-1/2}\sum_{i=1}^{n} \varepsilon_i^2 1_{(|\varepsilon_i|\le\sigma c)} + 2\sigma^2c^2\mathsf{f}(c)\frac{ac}{\sigma} + R_{v\varepsilon\varepsilon}(a,b,c),\\
\sum_{i=1}^{n} N'x_i\varepsilon_i v_i^{a,b,c} &= \sum_{i=1}^{n} N'x_i\varepsilon_i 1_{(|\varepsilon_i|\le\sigma c)} + 2c\mathsf{f}(c)\Sigma b + R_{vx\varepsilon}(a,b,c),\\
n^{1/2}\sum_{i=1}^{n} N'x_ix_i'Nv_i^{a,b,c} &= n^{1/2}\sum_{i=1}^{n} N'x_ix_i'N1_{(|\varepsilon_i|\le\sigma c)} + 2\mathsf{f}(c)\Sigma\frac{ac}{\sigma} + R_{vxx}(a,b,c).
\end{aligned}$$

Let $R(a,b,c) = |R_v(a,b,c)| + |R_{v\varepsilon\varepsilon}(a,b,c)| + |R_{vx\varepsilon}(a,b,c)| + |R_{vxx}(a,b,c)|$. *Then for any* $B > 0$ *and as* $n \to \infty$

$$\sup_{0<c<\infty}\sup_{|a|,|b|\le n^{1/4-\eta}B} |R(a,b,c)| = \mathsf{o}_\mathsf{P}(1).$$

Remark 1.8.1. *The first and fourth item in Lemma 1.8.1 adjusted by the rate* $n^{-1/2}$ *have the expansion*

$$n^{-1}\sum_{i=1}^{n} v_i^{a,b,c} = \psi + R_v'(a,b,c), \qquad \sum_{i=1}^{n} N'x_ix_i'Nv_i^{a,b,c} = \psi\Sigma + R_{vxx}'(a,b,c),$$

where any $B > 0$ *and as* $n \to \infty$

$$\sup_{0<c<\infty}\sup_{|a|,|b|\le n^{1/4-\eta}B} |R_v'(a,b,c)| + |R_{vxx}'(a,b,c)| = \mathsf{o}_\mathsf{P}(1).$$

Indeed, for the first expansion, the law of large numbers shows $n^{-1}\sum_{i=1}^{n} 1_{(|\varepsilon_i|\le\sigma c)} = \psi + \mathsf{o}_\mathsf{P}(1)$, while $\sup_{c\in\mathbb{R}} |c|\mathsf{f}(c) < \infty$ by Assumption 1.3.1(*ia*) and $n^{-1/2}a$ vanishes. For the second expansion, decompose

$$\sum_{i=1}^{n} N'x_ix_i'N1_{(|\varepsilon_i|\le\sigma c)} = \sum_{i=1}^{n} N'x_ix_i'N\{1_{(|\varepsilon_i|\le\sigma c)} - \psi\} + \sum_{i=1}^{n} N'x_ix_i'N\psi.$$

The first item vanishes by the Chebyshev inequality and Assumption 1.3.1(*iia*, *iic*), while the second converges to $\psi\Sigma$.

Proof of Lemma 1.8.1. The general class of empirical processes is

$$\mathcal{M}_n = n^{-1/2}\sum_{i=1}^{n} g_{in}\varepsilon_i^p v_i^{a,b,c}, \quad v_i^{a,b,c} = 1_{(|\varepsilon_i - x_{in}'b|\le\sigma c + n^{-1/2}ac)}.$$

1. *Decompose* $\mathcal{M}_n$. Write $\mathcal{M}_n = \mathcal{M}_{n,1} + \mathcal{M}_{n,2} + \mathcal{M}_{n,3}$, where

$$\mathcal{M}_{n,1} = n^{-1/2}\sum_{i=1}^{n} g_{in}\varepsilon_i^p 1_{(|\varepsilon_i|\le\sigma c)}, \quad \mathcal{M}_{n,2} = n^{-1/2}\sum_{i=1}^{n} g_{in}\mathsf{E}_{i-1}\varepsilon_i^p\{v_i^{a,b,c} - 1_{(|\varepsilon_i|\le\sigma c)}\},$$
$$\mathcal{M}_{n,3} = n^{-1/2}\sum_{i=1}^{n} g_{in}\varepsilon_i^p\{v_i^{a,b,c} - 1_{(|\varepsilon_i|\le\sigma c)}\} - n^{-1/2}\sum_{i=1}^{n} g_{in}\mathsf{E}_{i-1}\varepsilon_i^p\{v_i^{a,b,c} - 1_{(|\varepsilon_i|\le\sigma c)}\}.$$

Therefore, the first term in stochastic expansion is $\mathcal{M}_{n,1}$. We will linearize $\mathcal{M}_{n,2}$ to obtain the second term, and argue that $\mathcal{M}_{n,3}$ is small in probability.

2. *Linearize* $\mathcal{M}_{n,2}$. Note $\mathcal{M}_{n,2} = n^{1/2}\{\overline{\mathsf{G}}_n^{g,p}(a,b,c) - \overline{\mathsf{G}}_n^{g,p}(0,0,c)\}$, see (1.4.4). Apply Theorem 1.4.4 by Assumption 1.3.1(ia, iic) to get $\mathcal{M}_{n,2} = \overline{\mathcal{M}}_{n,2} + \mathrm{o_P}(n^{-2\eta})$, where

$$\overline{\mathcal{M}}_{n,2} = 2\sigma^{p-1}c^p\mathsf{f}(c)n^{-1/2}\sum_{i=1}^{n} g_{in}\{1_{(p \text{ is even})}n^{-1/2}ac + 1_{(p \text{ is odd})}x_{in}'b\}.$$

This reduces as desired by Assumption 1.3.1(iia). Note $0 < \eta \le 1/4$. Thus, we have $\mathcal{M}_{n,2} = \overline{\mathcal{M}}_{n,2} + \mathrm{o_P}(1)$ uniformly in $0 < c < \infty$ and $|a|, |b| \le n^{1/4-\eta}B$.

3. *Bounding* $\mathcal{M}_{n,3}$. Note $\mathcal{M}_{n,3} = \mathbb{G}_n^{g,p}(a,b,c) - \mathbb{G}_n^{g,p}(0,0,c)$, see (1.4.5). Due to Assumption 1.3.1(ia, iib, iic), Theorem 1.4.3 shows $\mathcal{M}_{n,3} = \mathrm{o_P}(1)$ uniformly in a, b, c. ■

1.8.2 Proofs of Theorem 1.3.1, 1.3.2 and 1.3.3

Proof of Theorem 1.3.1. The $m+1$ step estimators for β, σ^2 are defined in (1.2.4), (1.2.5). These are least squares estimators for the non-outlying observations and satisfy

$$N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = (\sum_{i=1}^{n} N'x_i x_i' N v_{i,c}^{(m)})^{-1}(\sum_{i=1}^{n} N'x_i\varepsilon_i v_{i,c}^{(m)}), \tag{1.8.2}$$

$$\begin{aligned} n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\} = \varsigma_c^{-2}(n^{-1}\sum_{i=1}^{n} v_{i,c}^{(m)})^{-1}n^{-1/2}\{\sum_{i=1}^{n}(\varepsilon_i^2 - \varsigma_c^2\sigma^2)v_{i,c}^{(m)} \\ - (\sum_{i=1}^{n}\varepsilon_i x_i' N v_{i,c}^{(m)})(\sum_{i=1}^{n} N'x_i x_i' N v_{i,c}^{(m)})^{-1}(\sum_{i=1}^{n} N'x_i\varepsilon_i v_{i,c}^{(m)})\}. \end{aligned} \tag{1.8.3}$$

We express the weight $v_{i,c}^{(m)}$ in (1.2.3) as

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \le \widehat{\sigma}_c^{(m)} c)} = 1_{(|\varepsilon_i - x_{in}'\widehat{b}_c^{(m)}| \le \sigma c + n^{-1/2}\widehat{a}_c^{(m)} c)} = v_i^{\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c},$$

where $\widehat{b}_c^{(m)} = N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$ and $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are estimation errors for β and σ in the m step of the algorithm.

Since $|\widehat{b}_c^{(m)}| + |\widehat{a}_c^{(m)}| = \mathrm{O_P}(1)$ and by Assumption 1.3.1(ia, ii), then Lemma 1.8.1 and Remark 1.8.1 with $\kappa = 0$, $\eta = 1/4$ show asymptotic expansions for product moments. Substitute these expansions into (1.8.2), (1.8.3) to first get

$$\widehat{b}_c^{(m+1)} = \frac{2c\mathsf{f}(c)}{\psi}\widehat{b}_c^{(m)} + (\psi\Sigma)^{-1}\sum_{i=1}^{n} N' x_i \varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)} + R_\beta(\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c),$$

where the remainder $R_\beta(a, b, c)$ vanishes uniformly in $c_0 \le c < \infty$ and $|a|, |b| \le B$. A key to this is that c is bounded away from zero and that Σ is positive definite by Assumption 1.3.1(iia) so that the denominator ψ, $\psi\Sigma$ is bounded away from zero.

Secondly, we get an expression for $\widehat{\sigma}_c^{(m+1)}$. By Taylor expansion, first note that

$$n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = \frac{1}{2\sigma} n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\} + n^{-1/2}\mathrm{O}[n\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\}^2].$$

Then apply arguments as above to get

$$\widehat{a}_c^{(m+1)} = \frac{c(c^2 - \varsigma_c^2)\mathsf{f}(c)}{\tau_2^c}\widehat{a}_c^{(m)} + \frac{1}{2\sigma\tau_2^c} n^{-1/2}\sum_{i=1}^{n}(\varepsilon_i^2 - \varsigma_c^2\sigma^2)1_{(|\varepsilon_i| \le \sigma c)} + R_\sigma(\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c),$$

where the remainder $R_\sigma(a, b, c)$ also vanishes uniformly. ■

To prove the tightness and fixed point result, let $|\cdot|$ refer to the usual Euclidean vector norm, while $\|M\| = \max\{\mathrm{eigen}(M'M)\}^{1/2}$ is the spectral norm for any matrix M. Note that the norms are compatible so that $|Mx| \le \|M\||x|$ for any vector x.

Proof of Theorem 1.3.2. Due to Assumption 1.3.1(ia, ii), Theorem 1.3.1 shows

$$\widehat{u}_c^{(m+1)} = \Gamma_c \widehat{u}_c^{(m)} + K_c + R_u(\widehat{u}_c^{(m)}, c), \tag{1.8.4}$$

where the remainder term satisfies $\sup_{c_0 \le c < \infty}\sup_{|u| \le B}|R_u(u, c)| = \mathrm{o_P}(1)$ and

$$\widehat{u}_c^{(m)} = \begin{pmatrix} \widehat{b}_c^{(m)} \\ \widehat{a}_c^{(m)} \end{pmatrix} = \begin{Bmatrix} N^{-1}(\widehat{\beta}_c^{(m)} - \beta) \\ n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) \end{Bmatrix}, \quad \Gamma_c = \begin{Bmatrix} \frac{2c\mathsf{f}(c)}{\psi} I_{\dim x} & 0 \\ 0 & \frac{c(c^2 - \varsigma_c^2)\mathsf{f}(c)}{\tau_2^c} \end{Bmatrix}, \tag{1.8.5}$$

$$K_c = \begin{Bmatrix} (\psi\Sigma)^{-1} & 0 \\ 0 & (2\sigma\tau_2^c)^{-1} \end{Bmatrix} \sum_{i=1}^{n} \begin{Bmatrix} N' x_i \varepsilon_i \\ n^{-1/2}(\varepsilon_i^2 - \varsigma_c^2\sigma^2) \end{Bmatrix} 1_{(|\varepsilon_i| \le \sigma c)}. \tag{1.8.6}$$

Apply the autoregressive equation (1.8.4) recursively to obtain the representation

$$\widehat{u}_c^{(m+1)} = \Gamma_c^{m+1}\widehat{u}_c^{(0)} + \sum_{l=0}^{m}\Gamma_c^l\{K_c + R_u(\widehat{u}_c^{(m-l)}, c)\}. \tag{1.8.7}$$

Use the triangle inequality and $|Mx| \leq \|M\||x|$ to get

$$|\widehat{u}_c^{(m+1)}| \leq \|\Gamma_c^{m+1}\||\widehat{u}_c^{(0)}| + \{|K_c| + \max_{0\leq l\leq m} |R_u(\widehat{u}_c^{(l)}, c)|\} \sum_{l=0}^{m} \|\Gamma_c^l\|.$$

Assumption 1.3.1(ia) shows $\sup_{c_0\leq c<\infty} \max\{|2c\mathsf{f}(c)/\psi|, |c(c^2-\varsigma_c^2)\mathsf{f}(c)/\tau_2^c|\} < 1$, see Johansen and Nielsen (2013, Theorem 3.5), so $\sup_{c_0\leq c<\infty} \|\Gamma_c\| < 1$. Thus by Gelfand's formula in Varga (2000, Theorem 3.4) that $\lim_{m\to\infty} \|M^m\|^{1/m} = \max|\text{eigen}(M)|$, for some ω such that $\sup_{c_0\leq c<\infty} \|\Gamma_c\| < \omega < 1$ there exists $m_0 > 0$ so for all $m > m_0$ then

$$\sup_{c_0\leq c<\infty} \|\Gamma_c^m\| < \omega^m < 1. \tag{1.8.8}$$

Also note $(I_{\dim x+1} - \Gamma_c)^{-1} = \sum_{l=0}^{\infty} \Gamma_c$. This in turn implies for some $1 < B_0 < \infty$

$$\sup_{0\leq m<\infty} \sup_{c_0\leq c<\infty} \|\Gamma_c^m\| < B_0, \quad \sup_{c_0\leq c<\infty} \|(I_{\dim x+1} - \Gamma_c)^{-1}\| \leq \sum_{l=0}^{\infty} \sup_{c_0\leq c<\infty} \|\Gamma_c^l\| < B_0. \tag{1.8.9}$$

Therefore, (1.8.9) shows for all $m \in [0, \infty)$

$$|\widehat{u}_c^{(m+1)}| < B_0\{|\widehat{u}_c^{(0)}| + |K_c| + \max_{0\leq l\leq m} |R_u(\widehat{u}_c^{(l)}, c)|\}. \tag{1.8.10}$$

For any $c \in [c_0, \infty)$, Assumption 1.3.1(iii) with $\eta = 1/4$ guarantees tightness of $\widehat{u}_c^{(0)}$, while the kernel K_c is tight by Johansen and Nielsen (2016a, Theorem 4.4) using Assumption 1.3.1(ia, iib, iic). Thus, for all $\epsilon, \delta > 0$ there exist $n_0, U_0 > 0$ so that for all $n > n_0$ the set

$$\mathcal{A}_n = \{B_0 \sup_{c_0\leq c<\infty} (|\widehat{u}_c^{(0)}| + |K_c|) \leq U_0/3, B_0 \sup_{c_0\leq c<\infty} \sup_{|u|\leq U_0} |R_u(u, c)| < \delta/2\} \tag{1.8.11}$$

has probability larger than $1 - \epsilon$.

Mathematical induction over m is used to show $\sup_{0\leq m<\infty} \sup_{c_0\leq c<\infty} |\widehat{u}_c^{(m)}| \leq U_0$ on the set $\mathcal{A}_n$. For $m = 0$ as induction starts, $\sup_{c_0\leq c<\infty} |\widehat{u}_c^{(0)}| \leq B_0^{-1}U_0/3 < U_0$ holds since $B_0 > 1$. The induction assumption is that $\sup_{0\leq l\leq m} \sup_{c_0\leq c<\infty} |\widehat{u}_c^{(l)}| \leq U_0$. This implies $B_0 \max_{0\leq l\leq m} |R_u(\widehat{u}_c^{(l)}, c)| < \delta/2$, and then the bound in (1.8.10) becomes $\sup_{c_0\leq c<\infty} |\widehat{u}_c^{(m+1)}| < 2U_0/3 + \delta/2 < U_0$ so that $\sup_{0\leq l\leq m+1} \sup_{c_0\leq c<\infty} |\widehat{u}_c^{(l)}| \leq U_0$. ■

Proof of Theorem 1.3.3. Based on Assumption 1.3.1(ia, ii, iii), Theorem 1.3.1 provides the 1-step equation (1.8.4). Theorem 1.3.2 shows $\sup_{0\leq m<\infty} \sup_{c_0\leq c<\infty} |\widehat{u}_c^{(m)}| = \mathsf{O_P}(1)$, so the remainder term in (1.8.4) is $\mathsf{o_P}(1)$. Thus, for $m, n \to \infty$ the fixed point should satisfy $\widehat{u}_c^* = \Gamma_c\widehat{u}_c^* + K_c$ so that

$$\widehat{u}_c^* = (I_{\dim x+1} - \Gamma_c)^{-1}K_c. \tag{1.8.12}$$

Substitute (1.8.5), (1.8.6) of $\widehat{u}_c^*$, Γ_c and K_c into (1.8.12) to obtain

$$\begin{Bmatrix} N^{-1}(\widehat{\beta}_c^* - \beta) \\ n^{1/2}(\widehat{\sigma}_c^* - \sigma) \end{Bmatrix} = \begin{bmatrix} \frac{1}{\psi - 2c\mathsf{f}(c)} \Sigma^{-1} \sum_{i=1}^n N' x_i \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} \\ \frac{1}{2\sigma\{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)\}} n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \varsigma_c^2 \sigma^2) 1_{(|\varepsilon_i| \leq \sigma c)} \end{bmatrix}.$$

Replace (1.8.7) and (1.8.12) into the deviation $\widehat{\Delta}_c^{(m+1)} = \widehat{u}_c^{(m+1)} - \widehat{u}_c^*$, and then apply $\sum_{l=0}^m \Gamma_c^l = (I_{\dim x+1} - \Gamma_c^{m+1})(I_{\dim x+1} - \Gamma_c)^{-1}$ to attain

$$\widehat{\Delta}_c^{(m+1)} = \Gamma_c^{m+1}\{\widehat{u}_c^{(0)} - (I_{\dim x+1} - \Gamma_c)^{-1} K_c\} + \sum_{l=0}^m \Gamma_c^l R_u(\widehat{u}_c^{(m-l)}, c).$$

To bound $\widehat{\Delta}_c^{(m+1)}$, use the triangle inequality and $|Mx| \leq \|M\||x|$ to get

$$|\widehat{\Delta}_c^{(m+1)}| \leq \|\Gamma_c^{m+1}\|\{|\widehat{u}_c^{(0)}| + \|(I_{\dim x+1} - \Gamma_c)^{-1}\||K_c|\} + \max_{0 \leq l \leq m} |R_u(\widehat{u}_c^{(l)}, c)| \sum_{l=0}^m \|\Gamma_c^l\|.$$

By Assumption 1.3.1(ia) and Gelfand's formula, (1.8.8) and the second inequality in (1.8.9) imply for $m > m_0$

$$|\widehat{\Delta}_c^{(m+1)}| < \omega^{m+1}(|\widehat{u}_c^{(0)}| + B_0|K_c|) + B_0 \max_{0 \leq l \leq m} |R_u(\widehat{u}_c^{(l)}, c)|.$$

On the set $\mathcal{A}_n$ as in (1.8.11), since $\sup_{0 \leq m < \infty} \sup_{c_0 \leq c < \infty} |\widehat{u}_c^{(m)}| \leq U_0$ by Theorem 1.3.2, then $\sup_{c_0 \leq c < \infty} |\widehat{\Delta}_c^{(m+1)}| < \omega^{m+1}(B_0^{-1}U_0/3 + U_0/3) + \delta/2 < \omega^{m+1}U_0 + \delta/2$. As $0 < \omega < 1$, ω^{m+1} declines exponentially so m_0 can be chosen sufficiently large that for all $m > m_0$ then $\omega^{m+1}U_0 < \delta/2$. Thus $\mathsf{P}(\sup_{c_0 \leq c < \infty} |\widehat{\Delta}_c^{(m+1)}| < \delta) > 1 - \epsilon$ for $m > m_0$, $n > n_0$. ■

1.8.3 Proof of Theorem 1.3.4

Proof of Theorem 1.3.4. Assumption 1.3.1(ia) implies $\mathsf{E}|\varepsilon_i/\sigma|^l < \infty$ for some $l > 4$. Apply (1.3.2) and the Chebyshev inequality to get $\lambda/n = \mathsf{P}(|\varepsilon_i| > \sigma c_n) \leq \mathsf{E}|\varepsilon_i/\sigma|^l c_n^{-l}$. Thus $c_n \leq (\mathsf{E}|\varepsilon_i/\sigma|^l)^{1/l} \lambda^{-1/l} n^{1/l}$ so that the divergence rate of c_n is $\mathrm{O}(n^{1/l}) = \mathrm{o}(n^{1/4})$.

1. *A bound on the sample space.* By Assumption 1.3.1(ia, ii, iii) with $\eta = 1/4$, Theorem 1.3.2 and 1.3.3 show that $\widehat{\beta}_{c_n}^{(m)}$, $(\widehat{\sigma}_{c_n}^{(m)})^2$ are tight. Assumption 1.3.1(iib) gives $\max_{1 \leq i \leq n} |x_{in}| = \mathrm{O}_{\mathsf{P}}(n^{\kappa - 1/2}) = \mathrm{o}_{\mathsf{P}}(n^{-1/4})$ for some $0 \leq \kappa < 1/4$. Thus, for all $\epsilon > 0$ there exists a large constant A_0 so that the set

$$\mathcal{B}_n = \{\sup_{0 \leq m < \infty} (|\widehat{b}_{c_n}^{(m)}| + |\widehat{a}_{c_n}^{(m)}|) + n^{1/4} \max_{1 \leq i \leq n} |x_{in}| \leq A_0\}$$

has the probability larger than $1 - \epsilon$ for all n. Note that $\widehat{b}_{c_n}^{(m)} = N^{-1}(\widehat{\beta}_{c_n}^{(m)} - \beta)$ and $\widehat{a}_{c_n}^{(m)} = n^{1/2}(\widehat{\sigma}_{c_n}^{(m)} - \sigma)$.

2. *Bound the indicator.* Define the random quantity,

$$s_{i,c_n}^{(m)} = \widehat{\sigma}_{c_n}^{(m)} c_n - y_i + x_i' \widehat{\beta}_{c_n}^{(m)} + \varepsilon_i = \sigma c_n + n^{-1/2} \widehat{a}_{c_n}^{(m)} c_n + x_{in}' \widehat{b}_{c_n}^{(m)}.$$

On the set $\mathcal{B}_n$ and as $c_n = \mathrm{o}(n^{1/4})$, we have for some $A_1 > 0$

$$s_{i,c_n}^{(m)} \leq \sigma c_n + n^{-1/2} A_0 c_n + n^{-1/4} A_0^2 \leq \sigma(c_n + n^{-1/4} A_1),$$
$$s_{i,c_n}^{(m)} \geq \sigma c_n - n^{-1/2} A_0 c_n - n^{-1/4} A_0^2 \geq \sigma(c_n - n^{-1/4} A_1).$$

Since the sets $(y_i - x_i' \widehat{\beta}_{c_n}^{(m)} > \widehat{\sigma}_{c_n}^{(m)} c_n)$ and $(\varepsilon_i > s_{i,c_n}^{(m)})$ are equal, we find

$$1_{(\varepsilon_i/\sigma > c_n + n^{-1/4} A_1)} \leq 1_{(y_i - x_i' \widehat{\beta}_{c_n}^{(m)} > \widehat{\sigma}_{c_n}^{(m)} c_n)} \leq 1_{(\varepsilon_i/\sigma > c_n - n^{-1/4} A_1)}.$$

A similar argument shows

$$1_{(\varepsilon_i/\sigma < -c_n - n^{-1/4} A_1)} \leq 1_{(y_i - x_i' \widehat{\beta}_{c_n}^{(m)} < -\widehat{\sigma}_{c_n}^{(m)} c_n)} \leq 1_{(\varepsilon_i/\sigma < -c_n + n^{-1/4} A_1)}.$$

Thus, we get the lower and upper bound for indicators uniformly in iteration m so

$$1_{(|\varepsilon_i/\sigma| > c_n + n^{-1/4} A_1)} \leq 1_{(|y_i - x_i' \widehat{\beta}_{c_n}^{(m)}| > \widehat{\sigma}_{c_n}^{(m)} c_n)} \leq 1_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)}. \tag{1.8.13}$$

3. *Expectation of indicator bounds.* The aim is to prove

$$n\mathsf{E}1_{(|\varepsilon_i/\sigma| > c_n + n^{-1/4} A_1)} \to \lambda, \qquad n\mathsf{E}1_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)} \to \lambda. \tag{1.8.14}$$

Since $n\mathsf{E}1_{(|\varepsilon_i/\sigma| > c_n)} = \lambda$ by (1.3.2), it suffices to show

$$\mathcal{E}_n = n\mathsf{E}\{1_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)} - 1_{(|\varepsilon_i/\sigma| > c_n + n^{-1/4} A_1)}\} \to 0.$$

Note $|\varepsilon_i/\sigma| \sim \mathsf{g}, \mathsf{G}$ and $\mathsf{g} = 2\mathsf{f}$, $\mathsf{G} = 2\mathsf{F} - 1$. By this and (1.3.2), $2\{1 - \mathsf{F}(c_n)\} = \lambda/n$. Write $\mathcal{E}_n$ as an integral, apply the mean value theorem and use the above identity to get

$$\mathcal{E}_n = n \int_{c_n - n^{-1/4} A_1}^{c_n + n^{-1/4} A_1} 2\mathsf{f}(u) du = 4nn^{-1/4} A_1 \mathsf{f}(\tilde{c}) = \frac{4\lambda n^{-1/4} A_1 \mathsf{f}(\tilde{c})}{2\{1 - \mathsf{F}(c_n)\}},$$

where $|\tilde{c} - c_n| \leq n^{-1/4} A_1$. Then, we find

$$\mathcal{E}_n = 2\lambda A_1 \frac{\mathsf{f}(\tilde{c})}{\mathsf{f}(c_n - n^{-1/4} A_1)} \frac{\mathsf{f}(c_n - n^{-1/4} A_1)}{\mathsf{f}(c_n)} \frac{\mathsf{f}(c_n)}{c_n\{1 - \mathsf{F}(c_n)\}} n^{-1/4} c_n.$$

Since $c_n - n^{-1/4} A_1 \leq \tilde{c}$ and f has the decreasing tail by Assumption 1.3.1(ia), the first ratio is bounded by 1. Since $c_n = \mathrm{o}(n^{1/4})$, Assumption 1.3.1($ib, ic$) shows the second and third ratio are bounded. Then use $n^{-1/4} c_n = \mathrm{o}(1)$ to get $\mathcal{E}_n = \mathrm{o}(1)$.

4. *Poisson approximation.* On the set $\mathcal{B}_n$, apply (1.8.13) to obtain

$$\sum_{i=1}^n 1_{(|\varepsilon_i/\sigma| > c_n + n^{-1/4} A_1)} \leq \sum_{i=1}^n 1_{(|y_i - x_i' \widehat{\beta}_{c_n}^{(m)}| > \widehat{\sigma}_{c_n}^{(m)} c_n)} \leq \sum_{i=1}^n 1_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)}.$$

Using (1.8.14), the Poisson limit theorem shows that the lower and upper bound have the Poisson limit with mean λ. Thus, by (1.3.3), $n\widehat{\gamma}_{c_n}^{(m)} \xrightarrow{\mathsf{D}} \text{Poisson}(\lambda)$ for all $0 \leq m < \infty$. ■

Chapter 2

Testing the Presence of Outliers in Regression Models

ABSTRACT[1]. Algorithms used to detect outliers in regression models have a positive probability of finding outliers even when the data generation process has no outliers. Deriving distributional results on the rate of falsely-discovered outliers, we propose two sets of tests for the overall presence of outliers. First, tests on whether the proportion of detected outliers differs from its expected value. Second, 'scaling' tests on whether the proportion of detected outliers decreases proportionally with the level of significance used for detection. We apply the tests to a model of economic growth and a difference-in-differences panel assessing the effectiveness of a carbon tax.

2.1 Introduction

The presence of outlying observations in a regression equation can be indicative of model misspecification. For example, estimating a linear model when the underlying data generating process (DGP) is non-linear, may result in a high number of outliers post-estimation, consequently, it is important to check for possible outlier contamination. However, algorithms used to detect outliers have a positive probability of finding outliers even when, in fact, the data generation process has no outliers. Using the concept of a gauge – the expected rate of falsely discovered outliers – we propose two sets of tests of the overall presence of outliers: first, a set of tests on whether the observed proportion and number of detected outliers deviate from their expected values in a well-specified

[1]This chapter is co-authored with Felix Pretis, Department of Economics, University of Victoria and Nuffield College, University of Oxford.

model. Second, a set of 'scaling' tests on whether the number of detected outliers decreases proportionally with the level of significance used to detect outliers. The first set of tests has power against the number of outliers present, while the second set has power against both outlier magnitude and number.

We consider outlier detection algorithms of the form of iterated 1-step Huber-skip M-estimators, in particular we focus on robustified least-squares (RLS) and impulse indicator saturation (IIS – see Hendry, Johansen, and Santos, 2008). In these algorithms, the rate of false-discovery (or gauge) can be controlled through a cut-off threshold used to define whether an observation is an outlier or not. IIS has seen increased use as a method to asses model mis-specification and detect outliers (see studies listed in Table 2.5.2 in section 2.5), however, due to a lack of distributional results on the gauge, there was no obvious cut-off rule or threshold of observed outliers relative to the expected number of outliers that would classify a model as misspecified (see the discussion by Doornik and Hendry, 2016). Here we derive the asymptotic distribution of the the gauge to construct a set of simple tests for the overall presence of outliers and model misspecification. We assess whether the observed proportion of observations classified as outliers significantly deviates from the proportion expected when there are no outliers. For example, if outlier detection takes place at a cut-off threshold such that we expect 1% of the sample to be classified as outliers by chance (even when there are none), and subsequently 2% of the sample are classified as outliers, is the difference between the nominal level (1%) and the observed proportion (2%) sufficiently large to reject the null hypothesis of no outliers?

This first set of tests focuses exclusively on the number of outliers, not taking outlier magnitude into account. We therefore propose a second set of tests, which we refer to as 'scaling' tests. These account for both the number and magnitude of outliers by assessing whether the detected proportion of outliers scales proportionally to the level of the cut-off in outlier detection. For example, if there are no outliers in the DGP, then detecting outliers at a significance level of 5% should result in a proportionally higher number of outliers detected relative to detection at 1%. By repeated testing at different levels of detection significance, this proportional scaling can be used to test the global hypothesis of no outliers being present.

We show that when there are no outliers present, the gauge of RLS and IIS follows an asymptotic normal distribution centred around the nominal level of significance used to detect outliers, while the number of outlying observations observed can be approximated

by the Poisson distribution parameterized by the expected outlier number. We further show that the scaling tests based on the sum of detected outliers follows an asymptotic normal distribution, with the supremum approximated by a Gaussian process for which the critical values can be simulated. The distributional results hold for regression models with covariates that can be random variables and deterministically or stochastically trending. Simulation results show that the standard normal test on the proportion of outliers exhibits a size close to the nominal level for large ($n \geq 100$) samples. In small samples, testing on a Poisson approximation of the number is preferred. The scaling tests exhibit a size close to the nominal level and have desirable power properties under a range of alternatives – both for increasing magnitude as well as number of outliers.

Our applications show the versatility of the proposed tests. In a cross-sectional model of economic growth (based on Mankiw, Romer, and Weil, 1992), we find little evidence of significant outlying countries. In a panel difference-in-differences model of the impacts of a carbon tax on CO_2 emissions we identify significant outliers in the un-taxed control group, the presence of which impacts the estimated effect of the tax. The tests of the presence of outliers can be straight-forwardly implemented using the freely-available R-package *gets* (Pretis, Reade, and Sucarrat, 2018) or the *Autometrics* algorithm in PcGive (Doornik and Hendry, 2013).

2.1.1 Related literature

The focus of the broader literature on outliers has remained predominantly on methods of detection, rather than the explicit testing of the overall presence of outliers. One-step estimators have been considered in Bickel (1975) and Ruppert and Carroll (1980). In particular the 1-step Huber-skip estimator was studied by Welsh and Ronchetti (2002). Asymptotic theory of 1-step M-estimators was established by Johansen and Nielsen (2009), while iteration of these estimators was investigated in Johansen and Nielsen (2013). Hoover and Perez (1999) originally introduced the idea of a gauge in a simulation study of general-to-specific variable selection algorithms, subsequently formally discussed by Hendry, Johansen, and Santos (2008) as the expected retention rate of irrelevant regressors in the context of model selection algorithms. Doornik (2009) presented a comprehensive simulation study on the gauge for the model selection algorithm Autometrics. Johansen and Nielsen (2016b) provide an asymptotic analysis for the gauge, where Jiao and Nielsen (2017) extended the pointwise asymptotics to the uniform convergence theory which allows the cut-off and the number of iterations both to

increase.

Comprehensive overviews of outlier detection approaches beyond iterated 1-step estimators are provided in the earlier surveys in Hodge and Austin (2004) and Chandola, Banerjee, and Kumar (2009). On the testing side, Grubbs (1969) provides one of the first formal tests for the presence of outliers by iteratively testing observations in a single series. In the context of regression modelling, Tietjen, Moore, and Beckman (1973) propose a test for a single outlier in simple linear regression based on standardised residuals. This approach was extended by Prescott (1975). A concern with post-estimation tests on standardized residuals is that the residuals themselves are determined during estimation which may be affected by the presence of outliers. More recently Srivastava and von Rosen (1998) provide likelihood ratio tests for single outliers, and Kaji (2018) considers the robustness of regression coefficients to removing outlying observations or winsorizing prior to estimation.

In contrast to the existing literature, here we construct overall tests for the presence of outliers based on the number or proportion of observations classified as outliers. Even though the focus in the indicator saturation and RLS literature has remained predominantly on time series, our proposed tests can be readily applied to cross-sectional or panel data. Further use extends to analysis of model output where the underlying model is unknown or analytically intractable such as large-scale simulated models in economics or climate research. The remainder of the chapter is structured as follows. Section 2.2 propose proportion and count tests by deriving the distributional results of the gauge for the two algorithms under consideration: RLS and IIS. Section 2.3 proposes the scaling tests and establishes their limiting distributions. The simulation performance of the tests is considered in section 2.4, while section 2.5 considers applications to a cross-sectional macroeconomic model of economic growth, a panel difference-in-differences model of a carbon tax, and re-visits a set of published studies using IIS.

2.2 Proportion and count tests

Assume the data $\{(y_i, x_i)\}_{i=1}^n$ satisfies the time series regression equation (1.1.1), and we keeping considering the outlier detection algorithms 1.2.1, 1.2.2 defined in Chapter 1. The gauge has been defined in (1.3.3), while this chapter continues to apply Assumption 1.3.1 to derive Gaussian approximations to the gauge.

A model of y_i may be misspecified in a number of ways (from incorrect functional

form, changing variance, to omitting variables, non-normality of the residuals, and mis-measured data) potentially resulting in a series of outlying observations in the estimated mis-specified model. The aim of the proposed tests here is to test the above DGP (1.1.1) for the presence of outliers identified by indicator variables. Outlier detection algorithms considered here use absolute residuals and then calculate robust least squares estimators from the non-outlying sample. This implicitly assumes symmetry, while non-symmetry leads to bias. All our asymptotic analysis is carried out under the null hypothesis that there are no outliers contained in the model matching the DGP (1.1.1).

Here we derive the asymptotic distribution of the gauge $\widehat{\gamma}_c^{(m)}$ under the null of no outliers to test whether the detected proportion of outliers significantly deviates from the expected proportion of outliers when there are no outliers present. The main result of asymptotic normality of the gauge is presented in theorems 2.2.3 and 2.2.4. In particular, theorem 2.2.3 shows the asymptotic normal distribution of the gauge for a robustified least squares estimator, and theorem 2.2.4 shows that the distribution of the gauge under IIS matches that of the robustified implementation.

With the normal approximation to the gauge and the Poisson approximation to the number of outliers, we propose a set of tests for the overall presence of outliers - Outlier proportion test: A standard normal test on the proportion of outliers; and outlier count test: A Poisson test on the number of outliers.

2.2.1 Gaussian approximation to the gauge

The distributional results are structured as follows: initially a stochastic expansion and tightness result is given for the gauge of iterated estimators defined in Algorithm 1.2.1 and 1.2.2, and a weak convergence theory follows. A Gaussian process theory arises when the proportion of detected outliers is controlled by choosing the cut-off value c as n increases, whereas a Poisson exceedence theory is established when the c is set to allow a fixed number of outliers regardless of the sample size n. The proof involves the empirical process theory developed by Jiao and Nielsen (2017). Thus we first present the asymptotic expansion of the following empirical product moments.

Lemma 2.2.1. *Suppose Assumption 1.3.1(ia, ii) holds. Let* $v_i^{a,b,c} = 1_{(|\varepsilon_i - x_{in}'b| \leq \sigma c + n^{-1/2}ac)}$. *Then we have an expansion*

$$n^{-1/2}\sum_{i=1}^{n} v_i^{a,b,c} = n^{-1/2}\sum_{i=1}^{n} 1_{(|\varepsilon_i| \leq \sigma c)} + 2\mathsf{f}(c)\frac{ac}{\sigma} + R(a,b,c).$$

Then for any $B > 0$ and as $n \to \infty$

$$\sup_{0<c<\infty} \sup_{|a|,|b| \leq n^{1/4-\eta}B} |R(a,b,c)| = \mathrm{o_P}(1).$$

The proof is given in the supplementary material.

Notice this result generalises Johansen and Nielsen (2009, 2016a) which did either allow variation in a, b but fixed c or variation in b, c but fixed a. Thus, built on their restricted expansion result, Johansen and Nielsen (2016b) only considered pointwise convergence theory in c for the gauge. One contribution of the present chapter is to extend pointwise asymptotics to convergence that is uniform in c. A stochastic expansion is given for the gauge of iterated 1-step Huber-skip M-estimator algorithms.

Theorem 2.2.1. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 1.2.1. Suppose Assumption 1.3.1(ia, ii) holds, and that $N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are $\mathrm{O_P}(1)$. Then uniformly in $m \in [0, \infty)$, $c \in \mathbb{R}_+$ and as $n \to \infty$*

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2} \sum_{i=1}^{n} \{1_{(|\varepsilon_i| > \sigma c)} - \gamma_c\} - 2c\mathsf{f}(c) n^{1/2}(\frac{\widehat{\sigma}_c^{(m)}}{\sigma} - 1) + \mathrm{o_P}(1).$$

The proof is given in the supplementary material.

The next result shows the gauge is tight uniformly in iteration $m \in [0, \infty)$ and in the cut-off value $c \in [c_0, \infty)$. Note $c_0 > 0$ is a finite number.

Theorem 2.2.2. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 1.2.1. Suppose Assumption 1.3.1(ia, ii, iii) holds with $\eta = 1/4$. Then as $n \to \infty$*

$$\sup_{0 \leq m < \infty} \sup_{c_0 \leq c < \infty} |n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)| = \mathrm{O_P}(1).$$

We control the proportion of discovered outliers by selecting the cut-off value c as n increases, and obtain a process of the gauge by varying c. A weak convergence theory follows from a finite dimensional convergence result derived by the expansion in Theorem 2.2.1 and tightness in Theorem 2.2.2, see Theorem 13.1 in Billingsley (1968). Here we analyse RLS and IIS by asymptotically approximating the process of their gauge as a Gaussian process (see the definition of Gaussian processes in Adler and Taylor, 2009, p. 27).

Theorem 2.2.3. ***Gaussian approximation to the gauge for Robustified Least Squares****: Consider the Robustified Least Squares, where the initial estimators $\widehat{\beta}_c^{(0)} =$*

$\widetilde{\beta}$, $(\widehat{\sigma}_c^{(0)})^2 = \widetilde{\sigma}^2$ *are the full sample least squares estimators. Suppose Assumption 1.3.1*(ia, ii) *holds. Let* $\mathbb{G}_n(c) = n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c)$ *for* $c \in [c_0, \infty)$*. Then as* $n \to \infty$

$$\mathbb{G}_n \rightsquigarrow \mathsf{GP}(0, \Sigma),$$

where $\mathsf{GP}(c)$ *for* $c \in [c_0, \infty)$ *is the Gaussian process with mean* $\mathsf{E}\{\mathsf{GP}(t)\} = 0$ *and variance-covariance structure* $\Sigma_{st} = \mathsf{Cov}\{\mathsf{GP}(s), \mathsf{GP}(t)\}$ *so*

$$\Sigma_{st} = \gamma_t(1 - \gamma_s) + s\mathsf{f}(s)(\tau_2^t + \gamma_t - 1) + t\mathsf{f}(t)(\tau_2^s + \gamma_s - 1) + st\mathsf{f}(s)\mathsf{f}(t)(\tau_4 - 1),$$

for any $c_0 \leq s \leq t < \infty$*. The proof is given in the supplementary material.*

Theorem 2.2.3 establishes the asymptotic normality of the gauge when detecting outliers using RLS. The following theorem 2.2.4 shows that the same results apply when detecting outliers using IIS.

Theorem 2.2.4. ***Gaussian approximation to the gauge for Impulse Indicator Saturation****: Consider the split-half Impulse Indicator Saturation in Algorithm 1.2.2 where* $n_1 = n_2 = n/2$*. Suppose Assumption 1.3.1*(ia, ii) *holds for each sub-sample set* $\mathcal{I}_1$*,* $\mathcal{I}_2$*. The sample gauge is*

$$\widehat{\gamma}_c^{(-1)} = n^{-1}\sum_{i=1}^{n}(1 - v_{i,c}^{(-1)}) = n^{-1}\{\sum_{i\in\mathcal{I}_1} 1_{(|y_i - x_i'\widehat{\beta}_2| > \widehat{\sigma}_2 c)} + \sum_{i\in\mathcal{I}_2} 1_{(|y_i - x_i'\widehat{\beta}_1| > \widehat{\sigma}_1 c)}\}.$$

Then as $n \to \infty$ *the process of the normalized gauge* $n^{1/2}(\widehat{\gamma}_c^{(-1)} - \gamma_c)$ *for* $c \in [c_0, \infty)$ *weakly converges to the same Gaussian process as for RLS reported in Theorem 2.2.3. The proof is given in the supplementary material.*

The above weak convergence result for the gauge does not depend on the stochastic properties of regressors, which can be stationary, or deterministically or stochastically trending, since only the variance estimator appears in the asymptotic expansion in Theorem 2.2.1. For the 1-step Huber-skip M-estimator either starting from the full sample or split-half sample least squares, the gauge has the same limiting Gaussian process. Denote $\widetilde{\gamma}_c$ as $\widehat{\gamma}_c^{(0)}$ in RLS or as $\widehat{\gamma}_c^{(-1)}$ in IIS. The weak convergence in Theorems 2.2.3, 2.2.4 implies the pointwise asymptotic result in Johansen and Nielsen (2016b) so that for any $c \in [c_0, \infty)$

$$n^{1/2}(\widetilde{\gamma}_c - \gamma_c) \xrightarrow{\mathsf{D}} \mathsf{N}\{0, \gamma_c(1 - \gamma_c) + 2c\mathsf{f}(c)(\tau_2^c + \gamma_c - 1) + c^2\mathsf{f}^2(c)(\tau_4 - 1)\}. \tag{2.2.1}$$

We have thus established the asymptotic distribution of the sample gauge $\widetilde{\gamma}_c$ under the null of no outliers. The sample gauge using RLS or IIS follows an asymptotic normal

distribution given by equation (2.2.1) centred around γ_c defined by the chosen cut-off c (e.g. $\gamma_c = 0.05$ for $c = 1.96$ if assume ε_i/σ follows the standard normal). A fixed point can also be derived for the gauge as in theorem 2.2.5.

Theorem 2.2.5. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 1.2.1. Suppose Assumption 1.3.1(ia, ii, iii) holds with $\eta = 1/4$. Then for all $\epsilon, \delta > 0$ a pair $m_0, n_0 > 0$ exists, so for all $m > m_0$ and $n > n_0$*

$$\mathsf{P}\{\sup_{c_0 \le c < \infty} |n^{1/2}(\widehat{\gamma}_c^{(m)} - \widehat{\gamma}_c^*)| > \delta\} < \epsilon,$$

where

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^* - \gamma_c) = n^{-1/2} \sum_{i=1}^{n} \{1_{(|\varepsilon_i| > \sigma c)} - \gamma_c\} \\ - \frac{c\mathsf{f}(c)}{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)} n^{-1/2} \sum_{i=1}^{n} (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|\varepsilon_i| \le \sigma c)} + \mathrm{o}_{\mathsf{P}}(1), \end{aligned}$$

uniformly in $c \in [c_0, \infty)$. Let $\mathbb{G}_n^*(c) = n^{1/2}(\widehat{\gamma}_c^* - \gamma_c)$ for $c \in [c_0, \infty)$. Then as $n \to \infty$

$$\mathbb{G}_n^* \rightsquigarrow \mathsf{GP}(0, \Sigma),$$

where $\mathsf{GP}(c)$ for $c \in [c_0, \infty)$ is the Gaussian process with mean $\mathsf{E}\{\mathsf{GP}(t)\} = 0$ and variance-covariance structure $\Sigma_{st} = \mathsf{Cov}\{\mathsf{GP}(s), \mathsf{GP}(t)\}$ so

$$\begin{aligned} \Sigma_{st} = \gamma_t(1 - \gamma_s) - \frac{t\mathsf{f}(t)}{\tau_2^t - t(t^2 - \varsigma_t^2)\mathsf{f}(t)} \{\varsigma_t^2(1 - \gamma_s) - \tau_2^s\} \\ + \frac{s\mathsf{f}(s)}{\tau_2^s - s(s^2 - \varsigma_s^2)\mathsf{f}(s)} \frac{t\mathsf{f}(t)}{\tau_2^t - t(t^2 - \varsigma_t^2)\mathsf{f}(t)} \{\tau_4^s - \frac{(\tau_2^s)^2}{1 - \gamma_s}\}, \end{aligned}$$

for any $c_0 \le s \le t < \infty$. The proof is given in the supplementary material.

While the gauge can be approximated by a normal distribution, the number of outliers (given by $n\widehat{\gamma}_c^{(m)}$) can be approximated by a Poisson distribution. Johansen and Nielsen (2016b) proved the Poisson approximation to the gauge for the finite step Huber-skip M-estimator, while the iterated result was established by Jiao and Nielsen (2017). A Poisson exceedence theory arises in the scenario where the cut-off value c is set to allow the fixed number λ of outliers regardless of the sample size n.

2.2.2 Outlier proportion test: A standard normal test

Using the distributional result in Theorems 2.2.3, 2.2.4 and equation (2.2.1), a simple test for whether the proportion of outliers is different from zero can be constructed. If

there are no outliers then the proposed test statistic z_γ in (2.2.2) follows an asymptotic standard normal distribution:

$$z_\gamma = \frac{\widetilde{\gamma}_c - \gamma_c}{\sqrt{\omega_{\gamma_c}}} \underset{H_0}{\overset{a}{\sim}} \mathsf{N}(0,1), \tag{2.2.2}$$

where the variance term ω_{γ_c} is given by[2]

$$\omega_{\gamma_c} = n^{-1}\{\gamma_c(1-\gamma_c) + 2c\mathsf{f}(c)(\tau_2^c + \gamma_c - 1) + c^2\mathsf{f}^2(c)(\tau_4 - 1)\}. \tag{2.2.3}$$

Note for a standard normal reference distribution, we have $\tau_2^c = 1 - \gamma_c - 2c\mathsf{f}(c)$, and $\tau_4 = 3$.

The proposed test procedure, referred to here as the 'proportion test', is then to estimate a model using RLS or IIS at a chosen selection significance level γ_c (e.g. 0.05 for $c = 1.96$ when ε_i/σ is standard normal), record the observed proportion of outliers $\widetilde{\gamma}_c$ by dividing the number of retained impulse indicators by the total number of indicators selected over, and subsequently construct and compare z_γ to the critical values of a standard normal distribution. It is worth emphasizing that there are two significance levels at play when testing on the proportion of outliers: First, selection in RLS or IIS at γ_c determines the proportion of outliers expected under the null hypothesis, second the choice of the significance level p for then testing on z_γ determines the null rejection frequency of the outlier proportion test.

For example, consider selecting in IIS at $\gamma_c = 0.01$ (with $c = 2.576$ for a standard normal reference) in a sample of $n = 100$ observations, resulting (by chance) in the retention of three impulse indicators corresponding to three outlying observations. Then the observed proportion of outliers is equal to $\widetilde{\gamma}_c = 3/100 = 0.03$ with approximate standard deviation of $\sqrt{\omega}_{\gamma_c=0.01} = 100^{-1/2}\,[0.01(1-0.01) - 0.0028]^{1/2} = 0.00844$. Comparing $z_{\gamma=0.01} = (0.03 - 0.01)/0.00844 = 2.37$ to the critical values of a standard normal distribution, the null hypothesis of no outliers is rejected for $p = 0.05$ ($z_{\gamma=0.01} > 1.96$) while for $p = 0.01$ ($z_{\gamma=0.01} < 2.576$) it is not.

2.2.3 Outlier count test: A Poisson test

An alternative is to consider choosing the critical value of selection to permit an acceptable number λ_γ of spuriously retained outliers rather than a proportion of the sample:

[2]Note that the variance expression (2.2.3) uses γ_c rather than $\widetilde{\gamma}_c$. While using $\widetilde{\gamma}_c$ may provide a better approximation for the variance of $\widetilde{\gamma}_c$ under the null, the power under the alternative would be drastically lower – consider the extreme case of selecting at $\gamma_c = 0.05$ and detecting $\widehat{\gamma} = 0.3$ of the sample as outliers, then $\widetilde{\gamma}_c(1-\widetilde{\gamma}_c) = 0.3 \times 0.7 = 0.21$ while $\gamma_c(1-\gamma_c) = 0.0475$.

in this case a Poisson approximation for the number of retained outliers under the null hypothesis of no outliers can be used (Theorem 1.3.4). Referred to here as the 'count test', the estimated number of outliers $n\widetilde{\gamma}_c$ can be evaluated against the critical values of a Poisson(λ_γ) distribution (where $\lambda_\gamma = n\gamma_c$):

$$k_\gamma = n\widetilde{\gamma}_c \underset{H_0}{\overset{a}{\sim}} \text{Poisson}(\lambda_\gamma). \tag{2.2.4}$$

Continuing the above example of having detected $n\widetilde{\gamma}_c = 3$ outliers in a sample of $n = 100$ observations, testing against a Poisson($\lambda_{\gamma=0.01}$) distribution (where $\lambda_{\gamma=0.01} = 0.01 \times 100 = 1$) yields a p-value of $p = 0.08$ for both a two-sided and one-sided test of the sample number of outliers exceeding the hypothesized number.

2.3 Scaling tests

Testing using the proportion or number of outliers by simple normal or Poisson distributions (as above in sections 2.2.2 and 2.2.3) ignores information about magnitudes of discovered outlying observations. A small number of extremely large outliers might be an indicator of model misspecification, however, the simple normal or Poisson tests lose the power in this scenario. For example, suppose the true proportion of large outliers in the DGP is 5%, then using the proportion outlier test at $\gamma_c = 0.05$ may lead to non-rejection of the null-hypothesis of no outliers because by chance the expected proportion of outliers coincides to the true proportion of outliers. A complete assessment ideally should not only take the proportion or number of detected outliers but also their magnitudes into account.

We therefore propose what we call 'scaling tests' to consider both the proportion and magnitudes of outliers by assessing whether the sample gauge scales proportionally to chosen levels of significance γ_c. Consider selection at increasingly conservative significance levels. Under the null H_0, it would be expected to observe a corresponding decrease in detecting proportion of outliers, whereas under the alternative H_1, especially in the case of a small number of large outliers, the sample gauge would maintain at a certain level and not have a large change when the level of significance is sufficiently small. We choose a set of K different levels of significance $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$ with corresponding cut-offs $\mathcal{C}_K = \{c_k\}_{k=1}^K$ through the one-to-one mapping $\gamma_c = 1 - \mathsf{G}(c)$ and $c = \mathsf{G}^{-1}(1-\gamma_c)$, where significance levels are chosen in an increasingly conservative way so $\gamma_{c_1} > \gamma_{c_2} > \cdots > \gamma_{c_K}$ and $c_1 < c_2 < \cdots < c_K$. Note c_1 is selected so $c_1 \geq c_0$ where c_0 is a small finite real value. Denote $\widetilde{\gamma}_c$ as $\widehat{\gamma}_c^{(0)}$ in RLS or $\widehat{\gamma}_c^{(-1)}$ in IIS. For each γ_{c_k} and

c_k, run RLS or IIS and record sample gauges $\widetilde{\gamma}_{c_k}$ for $k = 1, 2, \ldots, K$, and subsequently test whether the sample gauge deviates from the population gauge for the full range of significance levels.

We consider three types of scaling tests. The first scaling test sums all deviations to construct the test statistic as $\sum_{c \in \mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c)$. The second scaling test takes the supremum over all deviations so the test statistic is $\sup_{c \in \mathcal{C}_K} |n^{1/2}(\widetilde{\gamma}_c - \gamma_c)|$. The third type of test relies on testing the global hypothesis of no outliers based on repeated testing using the proportion or count tests, applying corrections for multiple-hypothesis testing. Intuitively, the sum test considers deviations from the expected proportion of outliers, the supremum test considers the highest deviation from the expected proportion, and the global test conducts multiple hypothesis tests corrected for repeated testing.

The next two theorems provide asymptotic distributions for the above proposed test statistics under the null H_0. Notice that the limiting distributions of two test statistics do not depend on the stochastic properties of regressors.

2.3.1 Scaling sum test

The proposed scaling sum test is to construct the test statistic S_{sum} given in equation (2.3.1) by summing over the entire set of deviations between the observed and expected proportion of detect outliers:

$$S_{sum} = \sum_{c \in \mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c), \tag{2.3.1}$$

and then compare it against the critical values of a normal distribution with variance given in theorem 2.3.1.

Theorem 2.3.1. *Consider Robustified Least Squares or Impulse Indicator Saturation. Suppose Assumption 1.3.1(ia, ii) holds. Let $\widetilde{\gamma}_c$ be $\widehat{\gamma}_c^{(0)}$ in RLS or $\widehat{\gamma}_c^{(-1)}$ in IIS. Choose K different significance levels $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$ with corresponding cut-offs $\mathcal{C}_K = \{c_k\}_{k=1}^K$ such that $\gamma_{c_1} > \gamma_{c_2} > \cdots > \gamma_{c_K}$ and $c_1 < c_2 < \cdots < c_K$. Then under H_0 and as $n \to \infty$*

$$S_{sum} = \sum_{c \in \mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \mathsf{Var}_{\mathcal{C}_K}), \tag{2.3.2}$$

where

$$\begin{aligned}\mathsf{Var}_{\mathcal{C}_K} = &\sum_{k=1}^{K} \gamma_{c_k}(1 - \gamma_{c_k}) + 2 \sum_{1 \le k < l \le K} (\gamma_{c_l} - \gamma_{c_k}\gamma_{c_l}) + \{\sum_{k=1}^{K} c_k \mathsf{f}(c_k)\}^2(\tau_4 - 1) \\ &+ 2\{\sum_{k=1}^{K} c_k \mathsf{f}(c_k)\}\{\sum_{k=1}^{K} (\tau_2^{c_k} + \gamma_{c_k} - 1)\}.\end{aligned}$$

The proof is given in the supplementary material.

Once applying RLS or IIS with different cut-offs $\mathcal{C}_K = \{c_k\}_{k=1}^K$ (or significance levels $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$) to obtain a set of sample gauge $\widetilde{\mathcal{S}}_K = \{\widetilde{\gamma}_{c_k}\}_{k=1}^K$, we can propose three type of tests by exploring either L_1, L_2 or L_∞ norm to summarize the distance between two vectors $\widetilde{\mathcal{S}}_K$ (observed gauge) and $\mathcal{S}_K$ (expected gauge). The limiting distribution of these three tests can be obtained by applying the continuous mapping theorem to the weak limit of the process $\widetilde{\mathcal{S}}_K$ given by Theorem 2.2.3, 2.2.4. In practice, three tests will have distince power performances in the finite sample under different circumstances. This chapter proceeds by applying the L_∞ norm (taking supremum over all deviations) in the next subsection to propose the scaling supremum test.

The scaling sum test sums over all deviations by $S_{sum} = \sum_{c\in\mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c)$ instead of using the L_1 norm with the test statistic $\sum_{c\in\mathcal{C}_K} |n^{1/2}(\widetilde{\gamma}_c - \gamma_c)|$. One advantage is that its limit is obtained analytically as the normal distribution whereas the weak limit of L_1 version test has to be simulated from the Gaussian process, although the S_{sum} test might lose power in some cases relative to the L_1 test. Shown by the simulation study in section 2.4, the scaling sum test performs at least as good as the scaling supremum test in the finite sample.

2.3.2 Scaling supremum test

The proposed scaling supremum test is to construct the test statistic S_{sup} considering the maximum over the entire set of deviations between the observed and expected proportion of detected outliers as given in equation (2.3.3):

$$S_{sup} = \sup_{c\in\mathcal{C}_K} |n^{1/2}(\widetilde{\gamma}_c - \gamma_c)|. \tag{2.3.3}$$

The asymptotic distribution of the scaling sup test is derived in theorem 2.3.2. In practice, critical values of the multivariate Gaussian process can be obtained via simulation.

Theorem 2.3.2. *Consider Robustified Least Squares or Impulse Indicator Saturation. Suppose Assumption 1.3.1(ia, ii) holds. Let $\widetilde{\gamma}_c$ be $\widehat{\gamma}_c^{(0)}$ in RLS or $\widehat{\gamma}_c^{(-1)}$ in IIS. Choose K different significance levels $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$ with corresponding cut-offs $\mathcal{C}_K = \{c_k\}_{k=1}^K$ such that $\gamma_{c_1} > \gamma_{c_2} > \cdots > \gamma_{c_K}$ and $c_1 < c_2 < \cdots < c_K$. Then under H_0 and as $n \to \infty$*

$$S_{sup} = \sup_{c\in\mathcal{C}_K} |n^{1/2}(\widetilde{\gamma}_c - \gamma_c)| \xrightarrow{\mathsf{D}} \sup_{c\in\mathcal{C}_K} |\mathsf{GP}(c; 0, \Sigma)|, \tag{2.3.4}$$

where $\mathsf{GP}(c)$ *for* $c \in [c_0, \infty)$ *is the Gaussian process with mean* $\mathsf{E}\{\mathsf{GP}(t)\} = 0$ *and variance-covariance structure* $\Sigma_{st} = \mathsf{Cov}\{\mathsf{GP}(s), \mathsf{GP}(t)\}$ *so*

$$\Sigma_{st} = \gamma_t(1-\gamma_s) + s\mathsf{f}(s)(\tau_2^t + \gamma_t - 1) + t\mathsf{f}(t)(\tau_2^s + \gamma_s - 1) + st\mathsf{f}(s)\mathsf{f}(t)(\tau_4 - 1),$$

for any $c_0 \leq s \leq t < \infty$.

Proof of Theorem 2.3.2. Apply the continuous mapping theorem to the weak convergence results in Theorem 2.2.3, 2.2.4 by Assumption 1.3.1(ia, ii), see Billingsley (1969, Theorem 5.1) and van der Vaart and Wellner (1996, Theorem 1.3.6), then the limiting distribution of the test statistic follows. ■

To perform scaling sum and supremum tests, we have to assume the explicit form of the distribution of $\varepsilon_i/\sigma \sim \mathsf{f}$, and it is statistically difficult to estimate the density f in practice. This is because, without knowing the bias correction factor ς_c^2 for estimating variance as f is unknown, the standard estimate of σ^2 will be biased downward, which will subsequently produce upward biased standardised residuals for errors ε_i although residuals can be estimated consistently based on beta estimates. Then the estimated density of f using the biased standardised residuals will be inconsistent by whatever methods. In this chapter, assumption on the known density f is not avoidable. Unlike scaling sum and sup tests, the bootstrap does not require assuming the known form of the density f. However, if the specification assumed on f is correct, S_{sum} and S_{sup} tests will have the higher power under the alternative where outliers in fact present, since the bootstrap has a certain probability not to sample outlying observations under the alternative especially when we are concerned with a small number of outliers which have large magnitude, and then S_{sum} and S_{sup} would perform as if under the null such that the distribution of their test statistics is distorted.

2.3.3 Scaling global test

The third proposed scaling test for the presence of outliers is based on repeated testing using the proportion or count outlier tests from sections 2.2 and then using the Simes (1986) (based on the Bonferroni procedure, see also Seeger, 1968) correction for repeated testing. Similar to the sum and supremum scaling tests, run RLS or IIS at levels of significance $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$. For each γ_{c_k} conduct either a proportion (section 2.2.2) or count (section 2.2.3) test of the individual null hypothesis of no outliers. Let $\widetilde{p}_k$ denote the kth observed p-value of the K hypothesis tests using either the proportion or count

tests. Order the observed p-values of each test in ascending order as $\widetilde{p}_{(1)} \leq \widetilde{p}_{(2)} \leq \cdots \leq \widetilde{p}_{(K)}$. The rejection rule of the global hypothesis is then: reject the global hypothesis of no outliers if there exists a k in $\{1, 2, \ldots, K\}$ for which:

$$\widetilde{p}_{(k)} \leq \frac{k}{K} p_{glob}, \tag{2.3.5}$$

where p_{glob} is the chosen level of significance when testing the global hypothesis. In other words, reject the overall null hypothesis of no outliers H_0 if there exists an observed p-value from one of the tests that is lower than the threshold p-value adjusted for multiple testing.

For repeated independent tests, Simes (1986) proved the real size of this approach can be controlled exactly at the nominal level p_{glob} for any finite sample. Under dependency, Sarkar and Chang (1997) showed the false rejection rate is upper bounded by the test significance p_{glob} under a general class of distributions, where a regularity condition requires the positive regression dependence structure for joint distribution of all single test statistics (see the concept of positive regression dependence for bivariate distribution in Lehmann (1966) and for multivariate distribution in Sarkar (1969)).

This regularity condition does not automatically hold for the proposed scaling tests, meaning that we can expect size distortions in small samples. To see this, consider choosing the set of selection significance $\{\gamma_{c_k}\}_{k=1}^{K}$ and cut-offs $\{c_k\}_{k=1}^{K}$ in an increasing conservative way and consider the Simes procedure using the individual normal proportion test with test statistics and limiting distribution given in (2.2.2). Under assumption $\varepsilon_i/\sigma \sim \mathsf{N}(0,1)$ where $\tau_2^c = \psi_c - 2c\mathsf{f}(c)$ and $\tau_4 = 3$, Theorem 2.2.3 and 2.2.4 show that the test statistic vector $\{n^{1/2}(\widetilde{\gamma}_{c_1} - \gamma_{c_1}), n^{1/2}(\widetilde{\gamma}_{c_2} - \gamma_{c_2}), \ldots, n^{1/2}(\widetilde{\gamma}_{c_K} - \gamma_{c_K})\}$ has the joint Gaussian distribution with mean 0 and covariance $\Sigma_{c_j c_k} = \gamma_{c_k}(1 - \gamma_{c_j}) - 2c_j c_k \mathsf{f}(c_j)\mathsf{f}(c_k)$ for $c_1 \leq c_j < c_k \leq c_K$. In the multivariate normal setting, positive regression dependence requires each element in the convariance matrix to be positive, i.e. $\Sigma_{c_j c_k} > 0$ for any $c_1 \leq c_j < c_k \leq c_K$, see Example 1 in Sarkar and Chang (1997) and Case 1 in Benjamini and Yekutieli (2001). However $\Sigma_{c_j c_k}$ is not necessarily positive everywhere, so the regularity condition does not hold automatically in our context. For instance when the sample size $n = 100$, and we choose the set of expected number of outliers as $(10, 9, \ldots, 1)$ so the selection significance $\{\gamma_{c_k}\}_{k=1}^{K}$ is $(0.10, 0.09, \ldots, 0.01)$ and the cut-offs $\{c_k\}_{k=1}^{K}$ as $(1.64, 1.70, \ldots, 2.58)$ for $K = 10$. Then $\Sigma_{c_1 c_{10}} = -0.004 < 0$ as $\gamma_1 = 0.10$, $\gamma_{10} = 0.01$, $c_1 = 1.64$, $c_{10} = 2.58$, $\mathsf{f}(c_1) = 0.103$, and $\mathsf{f}(c_{10}) = 0.014$. Thus, the real size might not be bounded by the nominal level in small samples. This is consistent with our simulation results that the size of the global scaling test is not well controlled in small

samples. Nevertheless, the improved Bonferroni procedure is still a useful computational approach in practice, as the size is close to the nominal test level and power tends to one when the sample size becomes large (see simulation section 2.4).

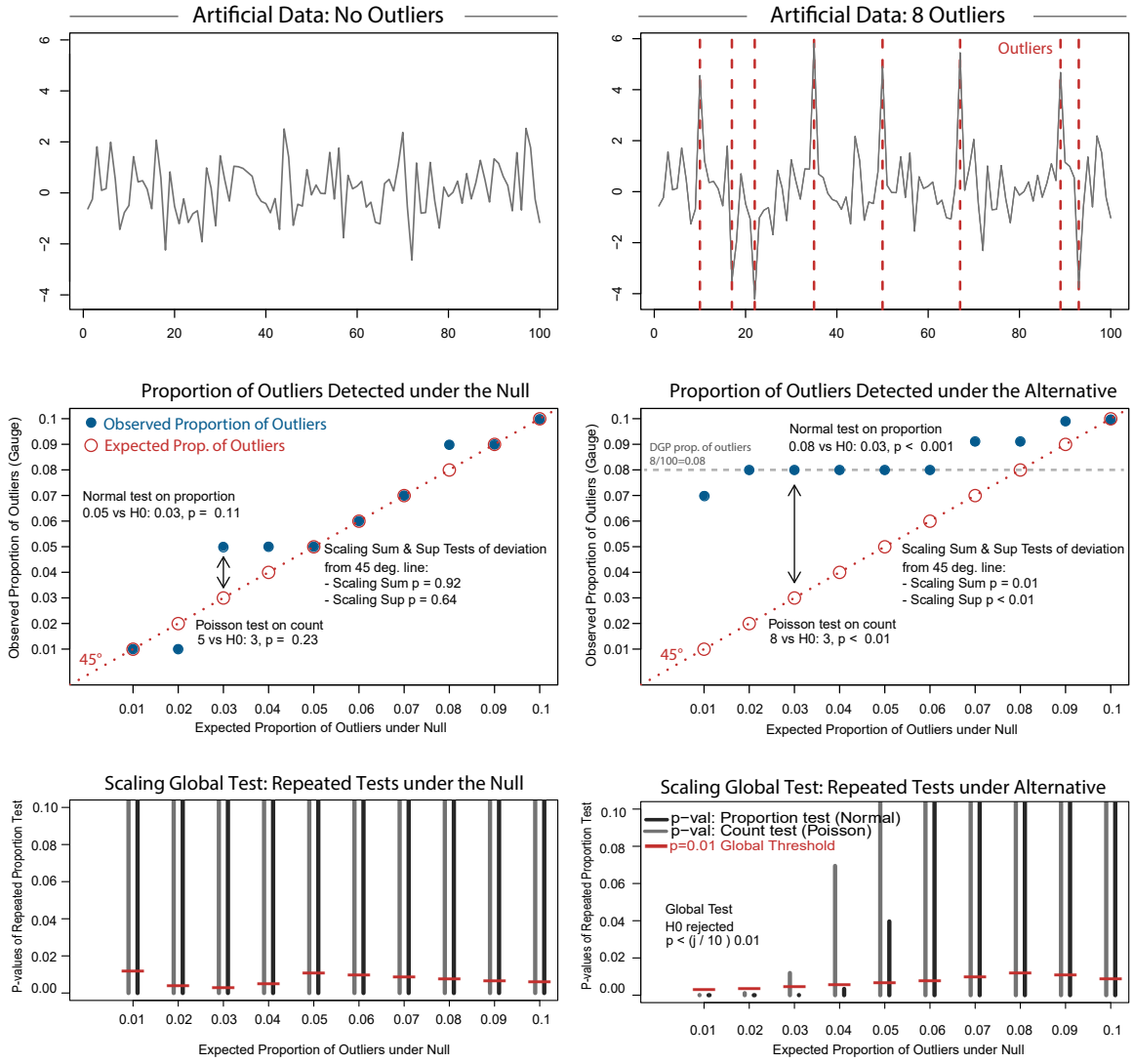


Figure 2.3.1: Illustrating the Proportion, Count, and Scaling (Sum, Sup., and Global) tests when there are no outliers (left) and in the presence of 8 outliers (right) in artificial data. Top panels show artificial data, middle panels show the detected proportion of outliers using IIS plotted against the expected proportion of outliers for the different levels of significance of selection. The bottom panels show the observed p-values of repeated tests together with the corresponding thresholds (red bars) for the scaling global test.

2.3.4 Implementation and use of the tests

Here we provide an illustration of the proposed proportion, count, scaling sum, supremum, and global tests under the null of no outliers as well as under an alternative. Subsequently we investigate their performances in simulations in section 2.4 to provide guidance on which test to apply in which situations in practise.

Figure 2.3.1 shows an example of the proposed tests using artificial data. We first model the series shown in the top panels using algorithm 1.2.2 in the R-package *gets* using the `isat` function at a given level of significance of detection, for example $\gamma_c = 0.03$. Under the null of no outliers we therefore expect a rate of falsely-discovered outliers of 0.03. The proportion and count tests assess the difference between expected (red empty circles) and observed (blue solid points) proportions of outliers for a single level of selection significance in RLS or IIS – the results shown for a test when 3 outliers are expected as marked by the arrow in Figure 2.3.1. To implement the scaling tests we repeatedly run IIS at varying significance levels of detection, here these range from $\gamma_c = 0.01$ to 0.1. The sum and supremum scaling tests assess the entire deviation of the observed (blue solid) from the expected (red hollow, along 45 the degree line) proportion of outliers for varying levels of selection significance γ_c. The global scaling test (bottom panel) compares the p-values of multiple individual hypothesis tests using either the count (light gray) or proportion (dark gray) tests repeatedly against the threshold given by the correction for repeated testing (red lines). The scaling tests have power where the proportion or count tests do not, as can be seen when for example testing for a deviation from 8 outliers under the alternative (right) where observed $\widetilde{\gamma}_c = 0.09$ is close to the expected value under the null $\gamma_c = 0.08$ despite the presence of outliers.

The proportion and count outlier tests require a single run of RLS or IIS, while the scaling and global tests require multiple (K) runs of RLS or IIS which can be computationally intensive in large samples. IIS can be applied using the algorithm *Autometrics* (Doornik, 2009) in PcGive (Doornik and Hendry, 2013) or using the `isat` function in the R-package *gets* (Pretis, Reade, and Sucarrat, 2018). Code to implement the gauge test and compute the approximate variance is available from the authors compatible with the R-package *gets*. The most direct use of the test is to assess the presence of outlying observations in regression models occurring for any number of reasons: from measurement errors, non-normality, to model misspecification. Alternatively one can consider y_i as a series of forecast errors (difference between predicted and actual observations) and thus the tests lend themselves to assess forecast misspecification: if forecasts (e.g.

from economic dynamic stochastic general equilibrium models, or general circulation models in climate) track observations well, then only a small number of forecast errors will be classified as outliers, while if the forecasts fail, a large number of outliers will be detected (see Ericsson, 2017 and Pretis, Mann, and Kaufmann, 2015).

2.4 Simulation study

Here we consider the performance of the proportion test (2.2.2), the count test (2.2.4), as well as the scaling sum (2.3.2), supremum (2.3.3), and global, (2.3.5) tests in a set of Monte Carlo simulations under the null hypothesis of no outliers and under different alternatives. All simulations are coded using the indicator saturation implementation `isat` in the R-package *gets* (Pretis, Reade, and Sucarrat, 2018) with $10,000$ replications for the proportion and count tests, and $5,000$ replications for the scaling tests. Note that relative to the simulation study of Croux and Wilms (2016) and previous simulation assessments of IIS (e.g. Marczak and Proietti, 2016) and step-indicator saturation (Castle et al., 2015), here we assess the properties of testing the presence of outliers on the proportion and number of outliers itself, rather than assessing what fraction of outliers are recovered in simulations (often referred to as the potency in previous simulations). Thus, the proposed tests here are more classical tests in the sense of them being evaluated by studying their size and power.

2.4.1 Size under the null

To study the tests under the null hypothesis, the simulated IIS model nests the DGP and there are no outlying observations. For the proportion and count tests different sample lengths n, specifications for x_i, and levels of significance of selection γ_c are explored. For the scaling tests we additionally consider different levels K of selection significance where γ_{c_k} are chosen such that the expected number of outliers equals distinct integers $\lambda_{\gamma_k} = 1, 2, ..., K$ so $n\gamma_{c_k} = \lambda_{\gamma_k}$. The global scaling test is assessed using both the proportion and count tests repeatedly. The DGPs considered are static white noise (equivalent to a cross-section or static panel, $x_i = 1$ for $\forall i$), as well as stationary first-order autoregression (AR) $x_i = y_{i-1}$ with coefficient $\beta = 0.5$ and a stochastically-trending process with $\beta = 1$.

Proportion and count tests under the null Simulation results of the proportion and count tests under the null are shown in Figure 2.4.1 Consistent with theory - the estimated proportion of outliers in IIS approaches the theory gauge γ_c as the sample size increases and the theory standard deviation (2.2.3) matches the simulation standard deviation closely for large samples. While the standard normal test on the proportion is oversized for small samples due to both a higher than expected gauge and variance, the size is close to the nominal level for larger samples (approx. $n \geq 100$) for both static and dynamic DGPs. The small-sample effect is particularly pronounced for low levels of significance of selection γ_c in IIS. This is intuitive, as in small samples the gauge can only take a small range of possible values. The Poisson approximation yields a size close to the nominal level of the test for small sample sizes, however, appears under-sized for large samples. The suggestion is thus to rely on the count test in small samples, and the proportion test in large samples to yield a stable size close to the nominal level.

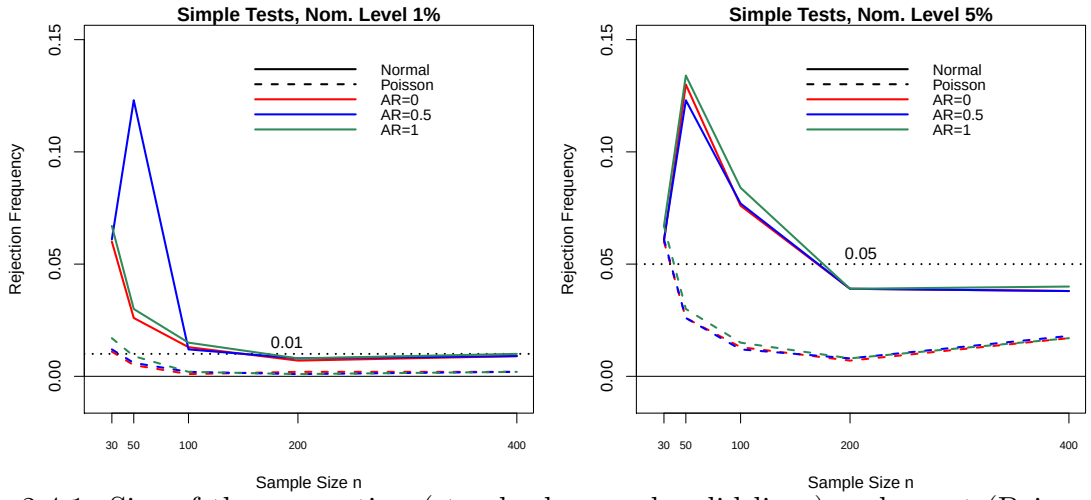


Figure 2.4.1: Size of the proportion (standard normal, solid lines) and count (Poisson, dashed lines) outlier tests under the null hypothesis of well-specified models without outliers for selection significance $\gamma_c = 0.01$ for static and dynamic models AR $\beta = (0, 0.5, 1)$ with error variance 1 for sample size $(30, 50, 100, 200, 400)$ at nominal test level of 0.01 (left) and 0.05 (right) with 10,000 replications.

Scaling tests under the null Simulation results of the scaling sum, supremum, and global tests under the null are shown in Figures 2.4.2 and 2.4.3 The size of the scaling tests closely matches the nominal level for larger sample sizes for both testing at $K = 3$, $K = 5$, and $K = 10$, for static and dynamic models. As expected, the scaling tests over-reject in small samples. The global test exhibits a size close to the nominal level when

based on repeated testing using the count test in small samples, and the proportion test in large samples.

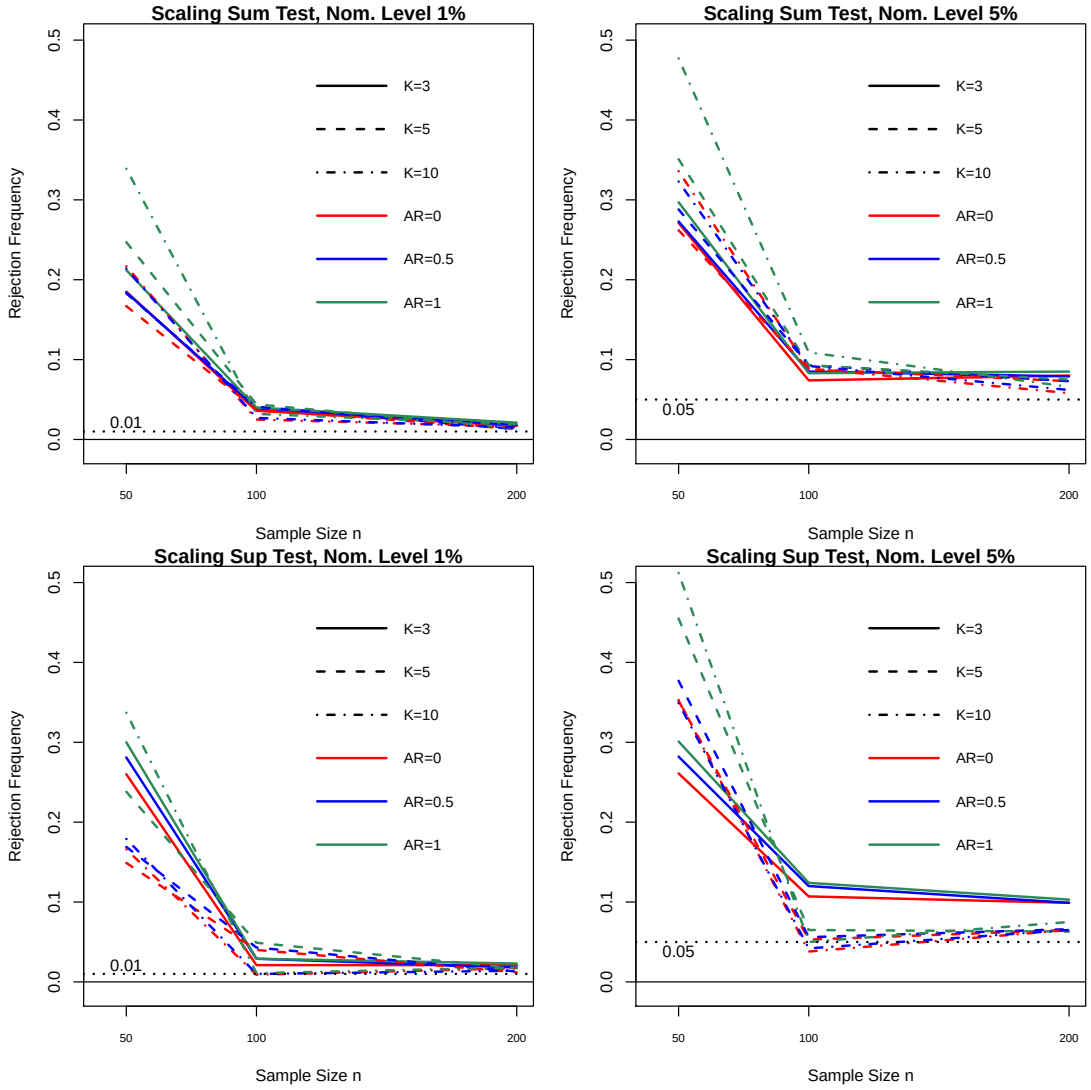


Figure 2.4.2: Size of the scaling sum and sup tests under the null hypothesis of no outliers for scale levels $K = (3, 5, 10)$ for static and dynamic models AR $\beta = (0, 0.5, 1)$ with error variance 1 for sample size $n = (50, 100, 200)$ for nominal test levels $p = (0.01, 0.05)$ with 5000 replications.

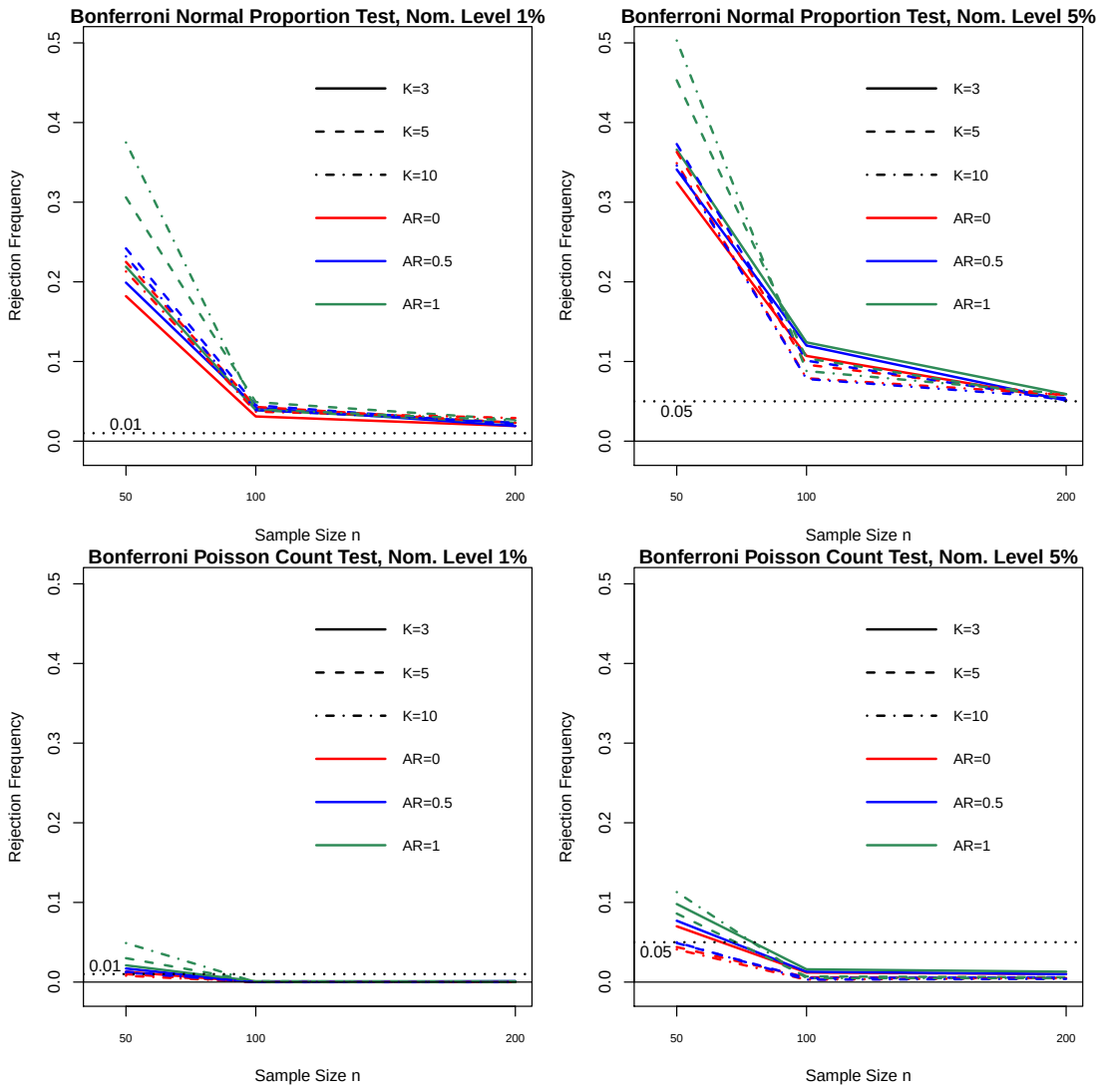


Figure 2.4.3: Size of Bonferroni correction for normal proportion and poisson count tests under the null hypothesis of no outliers for scale levels $K = (3, 5, 10)$ for static and dynamic models AR $\beta = (0, 0.5, 1)$ with error variance 1 for sample size $n = (50, 100, 200)$ for nominal test levels $p = (0.01, 0.05)$ with 5000 replications.

2.4.2 Power under alternatives

To assess the power of rejecting the null hypothesis of no outliers when there are actually outliers in the DGP,[3] the simulation DGPs are specified with varying proportions and

[3]This is notably different to studying the proportion of outliers correctly identified – referred to as potency in previous simulation studies.

magnitudes of (randomly placed with the same signs or alternating signs) additive[4] outliers. The percentage of outlying observations is increased from 5% to 15% of the sample, with outlier magnitudes of $\delta/\sigma = 2, 3, 4, 6$ where $\sigma = 1$.

Proportion and count tests under alternatives Simulation results of the proportion and count tests under the alternative are shown in Figure 2.4.4. The power of the proportion and count tests increases with the sample size n. As expected, the standard normal proportion test exhibits higher power than the count test (though note that the power results are not size-corrected). The tests also exhibit increasing power as the proportion of outliers increases, however, once the proportion of outliers becomes very large (here simulated as 15% of the sample), the null-rejection frequency decreases as the high number of outliers can be mis-interpreted as a higher variance, thus leading to low retention of outliers. The tests perform very differently for the unit root process as compared with static and stationary models. When $\beta = 1$ in AR(1) process, the tests have the highest power for the 5% proportion of outliers while have the lowest power when the proportion of outliers becomes 15%. Since the unit root process has the perfect memory, the relative small proportion of oultiers (as 10% in static and stationary models) now becomes sufficiently high number of outliers, which can be mis-understood as a high variance so that only few outliers would be detected by the algorithm leading to the lower power.

Scaling and global tests under alternatives Simulation results of the scaling tests under the alternative are shown in Figure 2.4.5 for the sum and supremum scaling tests, and in Figure 2.4.6 for the scaling global test. The scaling tests exhibits good power even for a small number of outliers (5% of the sample), and their power increases in the magnitude of outliers. The global test using repeated proportion testing performs better than the global test using repeated count testing, consistent with the latter being under-sized in larger samples under the null.

Selection of scaling step K affects the power performance of scaling sum and sup tests when the different number of outliers exists in the model, whereas global tests is more robust to selection of levels K. In the situation of 5 outliers (5% of 100 observations)

[4]IIS can detect both innovational and additive outliers in time series – the first by a single indicator, the latter by multiple subsequent indicators with opposing signs. In a setting with observations independent over i (e.g. cross-section), additive and innovational outliers are identical and can be matched by a single indicator. For a discussion on additive relative to innovational outliers see Nielsen (2004).

contained in the data, scaling sum and sup tests with level $K = 3$ has the highest power while $K = 10$ exhibits much lower even for large outlier magnitude. This is because if the scaling level K (say 10) is chosen larger than the number of outlying observations (say 5), then the scaling steps larger than the proportion of data contamination (between 10 and 6) does not provide useful information for the test statistic, since in the case of scaling level above 5 deviances between the expected and observed gauge are rather small, however the limiting distribution expects the high deviance due to the large number of K. In the situation where outliers become sufficiently large (15% of the whole sample), the power recovers for the selection of scaling level $K = 10$. In addition, when consider the unit root process that AR $\beta = 1$ and proportion of outliers as 15%, the power increases in scaling level K. This is because the algorithm can mis-interpret the high proportion of outlying observations as a higher variance when the tight selection significance is chosen, thus leading to low retention rate of outliers.

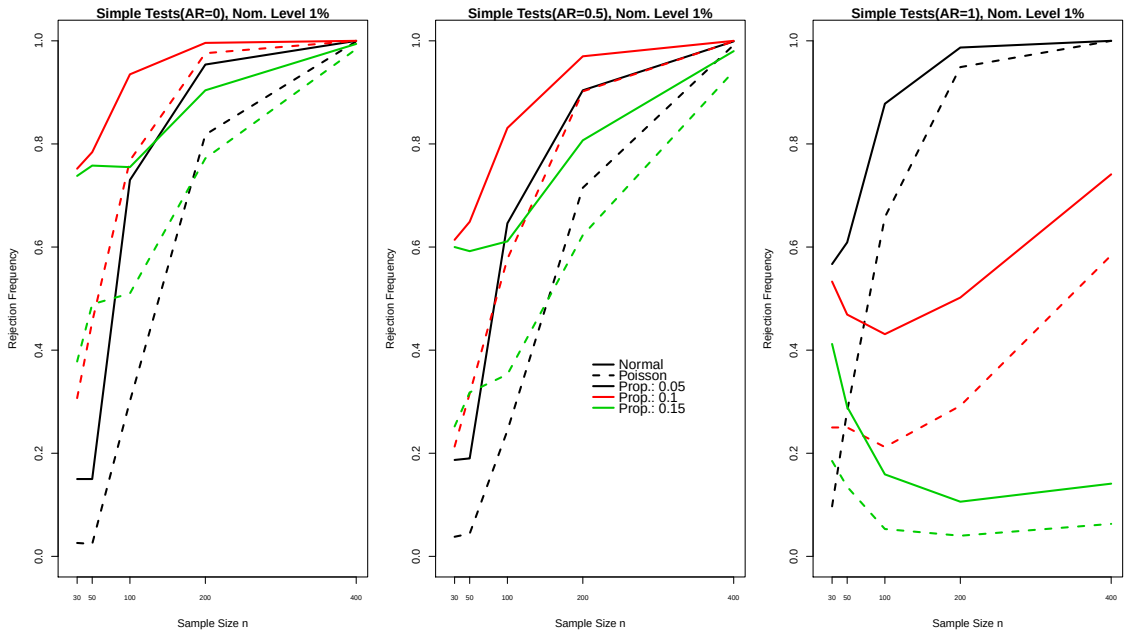


Figure 2.4.4: Power of the proportion (standard normal, solid lines) and count (Poisson, dashed lines) outlier tests for the selection significance $\gamma_c = 0.01$ for static and dynamic models AR $\beta = (0, 0.5, 1)$ with error variance 1 for sample size $(30, 50, 100, 200, 400)$ for proportion of outliers 0.05 (black), 0.10 (red), 0.15 (green) with the alternating sign and with outlier magnitude $\delta/\sigma = 4$ at nominal test level 0.01 with $10,000$ replications.

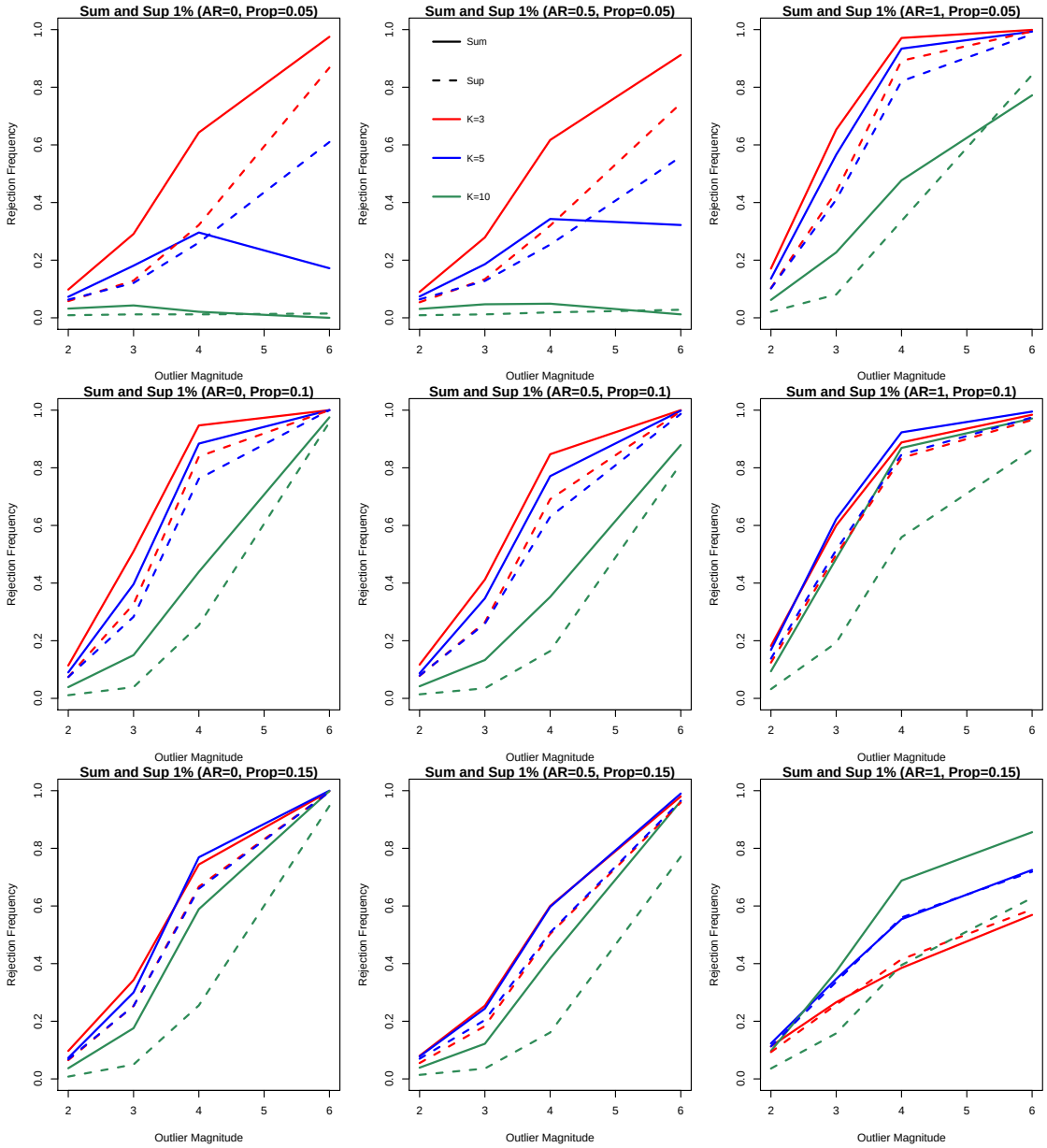


Figure 2.4.5: Power of scale sum and sup tests under the alternative hypothesis for scale levels $K = (3, 5, 10)$ for static and dynamic models AR $\beta = (0, 0.5, 1)$ with error variance 1 for sample size $n = 100$ for proportion of outliers $(0.05, 0.10, 0.15)$ and magnitude of outliers $(2, 3, 4, 6)$ for nominal test levels $p = 0.01$ with 5000 replications.

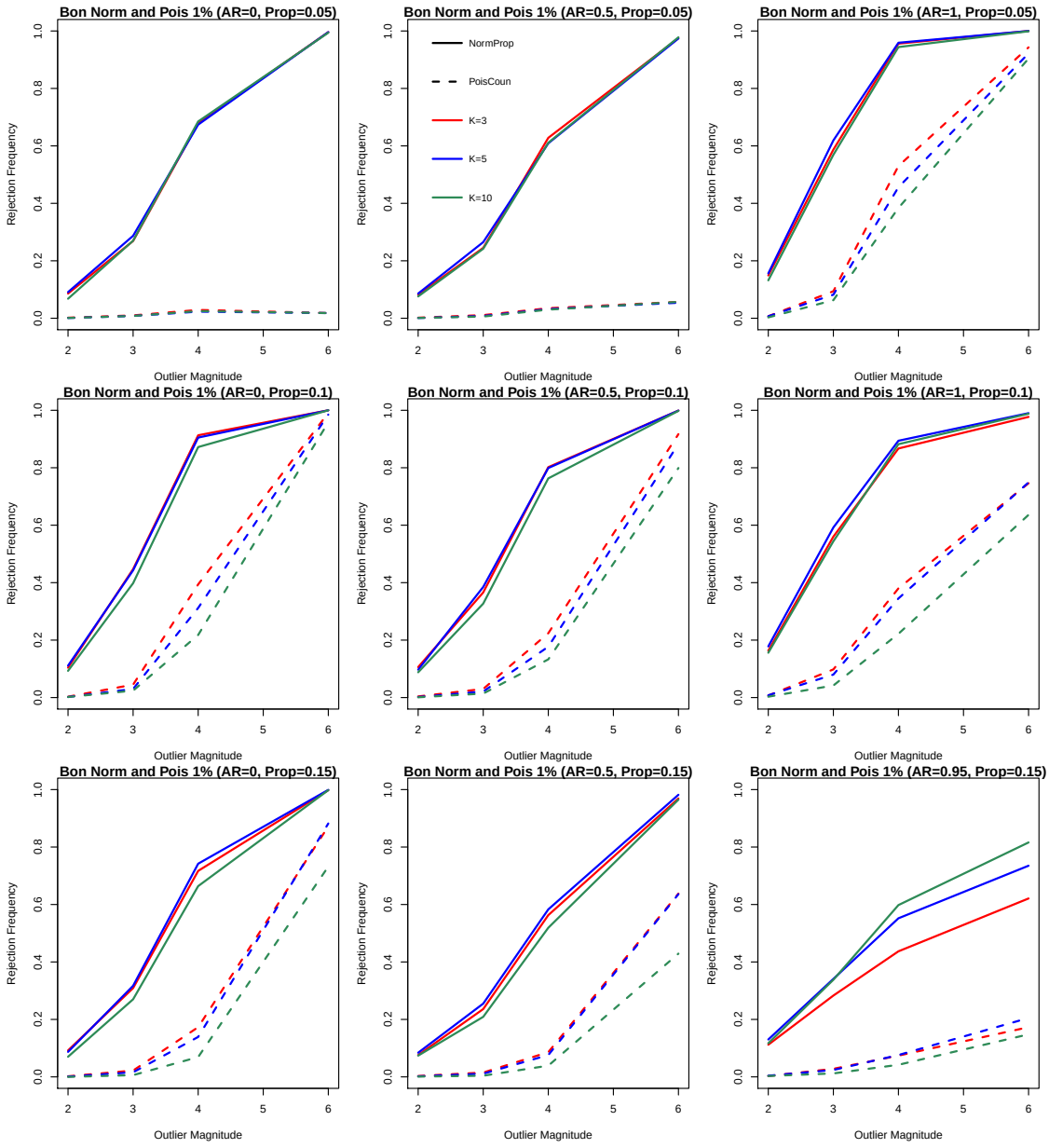


Figure 2.4.6: Power of Bonferroni correction for normal proportion and poisson count tests under the alternative hypothesis for scale levels $K = (3, 5, 10)$ for static and dynamic models AR $\beta = (0, 0.5, 1)$ with error variance 1 for sample size $n = 100$ for proportion of outliers $(0.05, 0.10, 0.15)$ and magnitude of outliers $(2, 3, 4, 6)$ for nominal test levels $p = 0.01$ with 5000 replications.

2.4.3 Recommendations based on simulation results

In summary, simulations show that the standard normal test of the proportion of outliers performs well in large ($n \geq 100$) samples with size close to the nominal level and desirable power properties. The count test against a Poisson distribution yields a size close to the nominal level in small samples, however is under-sized (and subsequently exhibits low power) in large samples.

The scaling tests are over-sized in small samples ($n \approx 50$), but exhibit size close to the nominal level for larger samples ($n \geq 100$), and show good power against both the magnitude and number of outliers. The scaling sum test shows higher power in practise than the scaling test based on the supremum. There is little difference between the scaling global and sum tests in terms of power. The global test exhibits a size close to nominal levels when using the count test in small samples, and the proportion test in large samples. When taking outlier magnitude into account and handling larger samples, the simulations therefore suggest to rely on the sum or global scaling tests (based on repeated testing using the proportion test).

In practice, we suggest to carry out a simple gauge test (normal or poisson) first based on a loose selection significance level. If the null is not rejected, then subsequently the more computationally-intensive scaling tests can be applied to further check the overall presence of outliers.

2.5 Applications

2.5.1 Economic growth in a cross-sectional model

In a cross-sectional analysis we re-visit the study by Mankiw, Romer, and Weil (1992) (henceforth MRW) by estimating their augmented cross-country Solow growth model using IIS to assess model specification through the detection of outliers when modelling GDP per capita as a function of the investment share, growth of effective labour, and human capital. We first test their augmented Solow model for the presence of outliers by assessing the proportion and number of detected outliers, second we apply the scaling and global tests to also take outlier magnitude into account.

Testing on the number of outliers (*proportion* and *count* Tests) in MRW Applying IIS at a significance level of selection of $\gamma_c = 0.05$ results in seven outlying

countries[5] denoted as I_{Country} in equation (2.5.1) identified in the sample of countries excluding oil producers ($n = 98$):[6]

$$\begin{aligned}\log(\text{GDPpc})_i = &\underset{(0.91)}{-3.82} + \underset{(0.12)}{0.60}\text{Inv}_i - \underset{(0.36)}{1.79}ngd_i + \underset{(0.06)}{0.67}\text{School}_i - \underset{(0.44)}{1.26}I_{\text{Ghana}} - \underset{(0.43)}{0.91}I_{\text{Togo}} \\ &- \underset{(0.44)}{1.42}I_{\text{Zaire}} - \underset{(0.44)}{0.95}I_{\text{Zambia}} - \underset{(0.44)}{1.01}I_{\text{Burma}} + \underset{(0.44)}{1.06}I_{\text{HK}} - \underset{(0.43)}{1.05}I_{\text{India}}\end{aligned} \tag{2.5.1}$$

This corresponds to an estimated outlier proportion of $\widetilde{\gamma}_c = 7/98 = 0.071$. Testing whether the observed proportion (0.071) significantly differs from the expected proportion ($\gamma_c = 0.05$), the corresponding test statistic is given by $z_\gamma = (0.071 - 0.05)/\sqrt{\omega_\gamma} = 0.019/0.0144 = 1.46$. A two-tailed comparison against the standard normal distribution yields a p-value of 0.24, suggesting no rejecting of the null-hypothesis of the proportion of outlying observations differing from the expected proportion. Testing whether the number of observed outliers differs from the expected number using the Poisson count test equally does not lead to a rejection of the null hypothesis: a two-sided Poisson test of $n\widetilde{\gamma}_c = 7$ against the expected number $n\gamma_c = 4.9$ yields a p-value of $p = 0.36$.

The coefficients on investment (Inv.), effective labour growth (ngd), and human capital (School) are similar to the original estimates of MRW (with t-values for tests of difference of -0.25, -0.17, and -0.17 respectively) further suggesting that applying IIS on a well specified model has little effect on other included covariates.[7] The detected outliers in the non-oil producing sample overlap in parts with the outyling countries in the robust estimation of the MRW model in Temple (1998), however, the tests presented here suggest that the number of outliers detected do not differ significantly from their expected value under the null hypothesis of no outliers.[8]

[5]These are: Ghana, Togo, Zaire, Zambia, Burma, Hong Kong, and India. Selection at $\gamma = 0.01$ results in one outlier, Zaire, being detected.

[6]The variable Inv_i refers to investment relative to GDP_i, ngd is the growth in effective units in labour plus depreciation, and School_i refers to the percentage of working age population in secondary schooling – see MRW 1992 for more details.

[7]When both an efficiency correction on the standard errors and a consistency correction for the estimate of σ are applied – see Johansen and Nielsen (2009, 2016b). In current further research we are investigating formal tests of the distorting effects of outliers comparing robust to non-robust estimates – see the heuristic apoproach briefly discussed in section 2.5.2.

[8]Temple (1998) considers robust estimation of multiple sub-samples beyond the example considered here, and finds additional outliers in these samples.

Testing both number and magnitude of outliers through *scaling Sum*, *supremum*, and *global* tests in MRW The above tests based on the estimates of $\widetilde{\gamma}_c$ and $n\widetilde{\gamma}_c$ use a single level of significance of selection γ_c in IIS and have power only against the number of outliers. Here we apply the set of scaling tests (sum, supremum, global) to take outlier magnitude into account by running IIS repeatedly using multiple levels of selection significance γ_{c_k} and recording $\widetilde{\gamma}_{c_k}$ at each run. Table 2.5.1 and Figure 2.5.1 show the number of outliers detected together with the corresponding $K = 10$ different values of selection significance γ_{c_k}, chosen such that the expected number of outliers are distinct integers ($n\gamma_{c_k} = 1, 2, ..., 10$). The observed proportion of outliers $\widetilde{\gamma}_{c_k}$ is compared to the expected number γ_{c_k} using the sum, supremum, and global tests (see Figure 2.5.1).

To apply the scaling sum test, we compute $S_{sum} = 1.91$ using equation (2.3.1) and compare it against the critical values of the normal distribution with the variance given by $\mathsf{Var}_{\mathsf{C_K}}$ in equation (2.3.2). This yields a p-value of $p = 0.046$, suggesting no rejection of the null hypothesis of no outliers when testing at the 1% level. For the scaling supremum test we take the supremum using (2.3.3) to construct $S_{sup} = 0.40$ which is assessed against the critical values obtained by simulating the Gaussian process in equation (2.3.4), which yields a p-value of 0.06. Finally, applying the global scaling test, for each γ_{c_k} we conduct a proportion outlier test, the recorded p-values of these tests are shown in Table 2.5.1 and Figure 2.5.1 (bottom panel). These are ordered in ascending order and compared to the corresponding threshold given by (2.3.5) (red bars in Figure 2.5.1 and bottom row in Table 2.5.1). All observed p-values exceed their corresponding threshold value, thus the global null hypothesis of no outliers is not rejected.

All scaling tests (sum, supremum, and global) tests show that the number of detected outliers scales proportionally to the level of significance of selection consistent with the null hypothesis of no outliers and the model not being misspecified. Overall, tests on both the number and magnitude of detected outliers suggest that the model is well specified and outlying countries are found consistent with expected values observed by chance under the null.

2.5.2 The effects of a carbon tax on transport emissions in a difference-in-differences panel model

Carbon taxes have emerged as a policy tool to tackle anthropogenic CO_2 emissions, however, they are rarely evaluated for their effectiveness in reducing emissions. Pretis

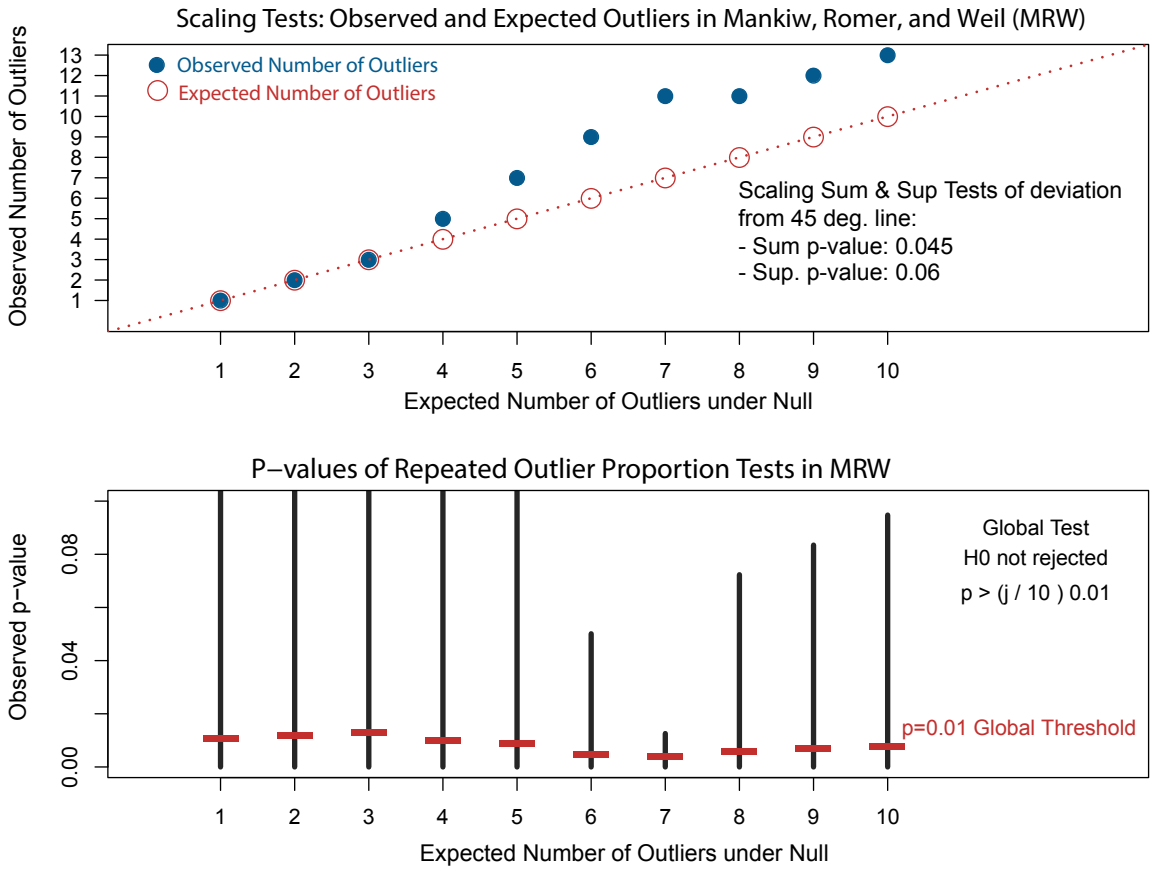


Figure 2.5.1: Scaling tests on the Mankiw-Romer-Weil model of economic growth. Top panel shows observed (blue solid) and expected (red empty) number of outliers in the MRW augmented Solow Model estimated using IIS at $K = 10$ selection significance levels (see Table 2.5.1 for full results). Under the null of no outliers the observed outliers are expected to fall on the 45 degree line (red). Using the IIS scaling sum and sup tests the null hypothesis of no outliers is not rejected (p=0.045 for sum, and p=0.06 for the supremum tests). Bottom panel shows the p-values from repeated proportion tests (gray) together with the global test thresholds at 1% (red bars). All observed p-values exceed their respective thresholds and the global null hypothesis of no outliers is not rejected.

(2019) studies the emission impacts of a carbon tax introduced in British Columbia (BC, Canada) in a difference-in-differences panel, but does not formally test for the presence of outliers in the model. Here we consider the impacts of the introduction of the BC carbon tax on transport-related (predominantly vehicle) CO_2 emissions.

We estimate a fixed effects panel model of 7 Canadian provinces over 26 years in

Lev. of Sign.: γ_{c_k}	0.010	0.020	0.031	0.041	0.051	0.061	0.071	0.082	0.092	0.102
Detected Prop.: $\widetilde{\gamma}_{c_k}$	0.010	0.020	0.031	0.051	0.071	0.091	0.112	0.112	0.122	0.133
Expected Num.: $n\gamma_{c_k}$	1	2	3	4	5	6	7	8	9	10
Detected Num: $n\widetilde{\gamma}_{c_k}$	1	2	3	5	7	9	11	11	12	13
Scal. Sum: 1.91 ($p \approx 0.046$)										
Scal. Sup.: 0.40 ($p \approx 0.06$)										
Prop. Test p-value $\widetilde{p}^g$:	1.000	1.000	1.000	0.462	0.168	0.050	0.013	0.072	0.083	0.095
Global threshold (1%)	0.008	0.009	0.010	0.007	0.006	0.002	0.001	0.003	0.004	0.005
$\widetilde{p}^g \leq$ threshold:	x	x	x	x	x	x	x	x	x	x

Table 2.5.1: Detected ($\widetilde{\gamma}_{c_k}$) and expected (γ_{c_k}) proportion of outliers in the MRW augmented Solow growth model for $K = 10$ levels of significance γ_{c_k} used to construct the scaling tests. Lower rows show p-values of repeated proportion test and the corresponding threshold of the global test. All proportion test values exceed their global thresholds and the null hypothesis of no outliers is not rejected.

a difference-in-differences specification including province and year fixed effects. The variable of interest (Tax) is the interaction of a dummy variable for the taxed region and the time period following the introduction of the carbon tax in 2008.[9] Data on emissions, population, and GDP is available from the National Inventory Report from Environment Canada. We estimate the model using IIS (Algorithm 1.2.2) at a significance level of $\gamma_c = 0.01$. Model estimates[10] are shown in equation (2.5.2):

$$\begin{aligned}\log(\mathrm{CO_2Emiss})_{i,t} = &\underset{(0.04)}{-0.08}\mathrm{Tax}_{i,t} - \underset{(0.15)}{0.15}\log(\mathrm{Pop})_{i,t} + \underset{(0.15)}{0.84}\log(\mathrm{GDP})_{i,t} + \underset{(0.073)}{0.19}\, I_{\mathrm{Sk},2001} \\ &+ \underset{(0.073)}{0.26}\, I_{\mathrm{Sk},2002} + \underset{(0.073)}{0.24}\, I_{\mathrm{Sk},2003} + \underset{(0.073)}{0.19}\, I_{\mathrm{Sk},2005} + \underset{(0.073)}{0.34}\, I_{\mathrm{Sk},2006} + \underset{(0.073)}{0.35}\, I_{\mathrm{Sk},2007} \\ &+ \underset{(0.073)}{0.32}\, I_{\mathrm{Sk},2008} + \underset{(0.073)}{0.30}\, I_{\mathrm{Sk},2009} \end{aligned} \tag{2.5.2}$$

Using IIS (Algorithm 1.2.2) at $\gamma_c = 0.01$ we identify 8 outliers relative to the model conditional on population, GDP and fixed effects. All outliers correspond to a single province (Saskatchewan) in the un-taxed control group over a period of 2001–2009. We formally test the significance of these outliers using the proportion and count tests, as well as the proposed scaling tests. Test results are shown in Figure 2.5.2 together with the identified outlying observations (Panel B) and estimated impacts of the carbon tax (Panel C) with and without controlling for the identified outlying observations.

Both the proportion and count tests reject the null hypothesis of the observed

[9] We omit Ontario, Alberta, and Quebec, from the control group as these provinces introduced their own carbon pricing schemes.

[10] Individual and year fixed effects are included in the model but not reported here.

number of outliers equalling the expected number. The observed outlier proportion is 8/182=0.044 which is significantly different from the expected proportion $\gamma_c = 0.01$ (proportion test-statistic of 5.427, $p < 0.0001$). Similarly the count of 8 observed outliers significantly differs from the expected rate of 1.82 using the Poisson test.

These individual tests are confirmed when using the scaling tests from section 2.3. The sum and supremum test statistics are 5.764 ($p < 0.0001$) and 0.667 ($p = 0.0001$) respectively. The global test equally rejects the null hypothesis of no outliers.

The identified outlying observations correspond to rapid growth in emissions in a single province in the control group. These results suggest that for one province in the control group, transport emissions are not well approximated by GDP, Population and individual fixed effects in the regression equation. Not controlling for these significant outliers likely over-states the emission reductions as a result of the introduction of the carbon tax. The estimated tax impact is -16% when not controlling for these outlying observations, compared to -8% when controlling for them using the robust estimates.

We use a heuristic approach to compare ordinary estimates without considering outlier contamination to robust estimates from IIS by constructing the following test statistic: $|\widehat{\beta}_{\text{tax,ordinary}} - \widehat{\beta}_{\text{tax,robust}}|/\text{se}(\widehat{\beta}_{\text{tax,robust}})$ which we compare against the standard normal distribution to assess whether the difference between robust and non-robust estimates are statistically different. Here $\widehat{\beta}_{\text{tax,ordinary}} = -0.16$, $\widehat{\beta}_{\text{tax,robust}} = -0.08$ while $\text{se}(\widehat{\beta}_{\text{tax,robust}}) = 0.04$ so the test statistics is given as $0.08/0.04 = 2$ yielding a p-value of 0.046. The null hypothesis of no coefficient distortion is then rejected at a 5% significance level. This suggests that the impact of the introduction of a carbon tax on emissions is overestimated by not controlling for these outliers.[11].

2.5.3 Re-visiting studies using impulse indicator saturation

As the proportion and count tests only require the detected number of outliers, sample size, and the nominal significance level of selection γ_c in IIS or RLS, both tests can readily be applied to existing studies already using indicator saturation without requir-

[11] Although this heuristic approach is informative and frequently used by empirical researchers, its approximation by the normal distribution is not correct asymptotically. Current ongoing research focuses on establishing formal tests with correct asymptotics under the null and desirable power properties under alternatives for outlier distortion in coefficients when using Algorithm 1.2.1 such as robustified least squares or impulse indicator saturation. For coefficient distortion in winsorizing see also the work by Kaji (2018).

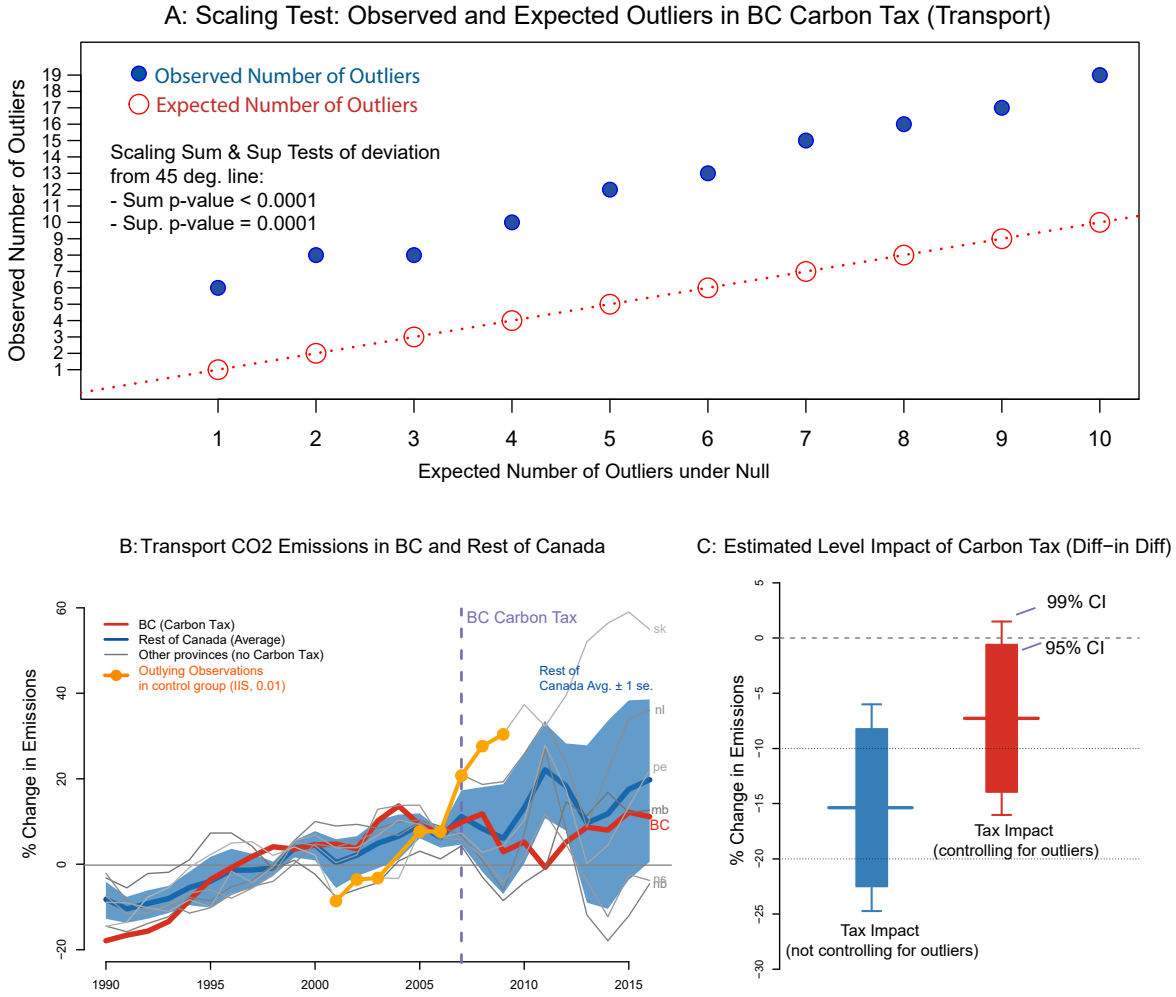


Figure 2.5.2: Scaling tests on the Difference-in-Differences panel model of the emissions impact of the BC carbon tax. Top panel shows observed (blue solid) and expected (red empty) number of outliers in the panel model estimated using IIS (Algorithm 1.2.2) at $K = 10$ selection significance levels. Under the null of no outliers the observed outliers are expected to fall on the 45 degree line (red). Using the scaling sum and sup tests the null hypothesis of no outliers is rejected ($p \leq 0.001$). Panel B shows the percentage change in CO2 emissions relative to the 1990-2008 average with the treated region (BC) shown in red, and identified outliers in the control group highlighted in orange. Panel C shows the estimated impact of the carbon tax on transport emissions when controlling for outliers using IIS (Algorithm 1.2.2) (right, red) and not controlling for outliers (left, blue).

ing re-estimation.[12] For example, Ericsson (2017) (building on Martinez, 2015) uses IIS to assess systematic forecast bias in forecasts of US government debt. In a sample of $n = 29$ forecast error observations applying IIS at $\gamma_c = 0.01$, Ericsson (2017) detects 4 ($\widetilde{\gamma}_c = 4/29 = 0.14$), 5 ($\widetilde{\gamma}_c = 5/29 = 0.17$), and 7 ($\widetilde{\gamma}_c = 7/29 = 0.24$) outlying observations in the government forecasts of three different agencies respectively. We can use these detected outlying observations directly to assess whether the number of discovered outliers exceeds the expected rate under the null. Applying the proportion test results in test statistics of 8.16 ($p \approx 0$), 10.36 ($p \approx 0$), and 14.76 ($p \approx 0$) respectively, strongly rejecting the null of no outliers, or no forecast mis-specification in the present example. Simulations in Section 2.4 suggest, however, that the standard normal test over-rejects in small samples. Thus, to address concerns about the small number of forecast observations ($n = 29$) the count test is also considered. Under the null hypothesis the expected rate is equal to $n\gamma_c = 29 \times 0.01 = 0.29$. A two-sided test against a $\mathsf{Poisson}(n\gamma_c = 0.29)$ distribution given the observed rates of $n\widetilde{\gamma}_c = (4, 5, 7)$ yields the p-values of $p = 0.0002, p \approx 0$, and $p \approx 0$ respectively. Thus, both the standard normal proportion as well as the Poisson count test reject that the observed proportion (number) of outliers equals the expected proportion (number). These results provide strong evidence of forecast mis-specification in the series of debt-forecasts considered by Ericsson (2017).

Expanding on the example of re-visiting Ericsson's (2017) results, we apply the outlier count and proportion tests to a set of 12 published studies which used IIS in a range of applications: from modelling UK wages to forecasts of food prices and the economic impacts of climate change. Table 2.5.2 provides an overview of the studies together with the results of applying the proportion and count tests. Overall, interpreted model misspecification by the authors closely coincides with the null rejection of the outlier test.

2.6 Conclusion

The numbers of outliers detected in regression models can be used as a test of the overall presence of outliers and model misspecification. We propose two sets of testing procedures for the null hypothesis of no outliers based on robustified least-squares and

[12] The scaling tests require estimation of the models at multiple levels of significance, therefore requires re-estimation and is thus less straightforward to apply to published studies.

Authors	**IIS Application**	$\widetilde{\gamma}_c$	γ_c	***Prop.***	***Count***
Castle and Hendry (2009)	UK Wage Equation	6/141=0.043	0.001	p≈0	p≈0
Hendry (1999)	US Food Expenditure (results from dynamic model)	7/72=0.097	0.01	p≈0	p≈0
Hendry (2011)	UK Consumer Expenditure	3/65=0.046	0.025	p=0.15	p=0.22
Castle and Hendry (2012)	US Returns to Education	301/5173=0.06	0.001	p≈0	p≈0
Hendry and Pretis (2013)	Atmospheric CO_2 Concentrations	1/246=0.004	0.001	p=0.12	p=0.21
Dreger and Wolters (2014)	Euro Area Money Demand	6/112=0.054	0.05	p=0.79	p=0.83
Anundsen (2015)	US Rent Prices	4/140=0.028	0.01	p=0.01	p=0.054
Nymoen and Sparrman (2015)	Unemployment Rate in Panel of OECD Countries	157/994=0.158	0.025	p≈0	p≈0
Stillwagon (2016)	US/GBP Exchange Rate	2/373=0.005	0.001	p=0.006	p=0.05
Ahumada and Cornejo (2016)	Food Price Forecasts (reported for soybean prices, 1-step fixed sample)	4/28 = 0.143	0.01	p≈0	p≈0
Martinez (2019)	Cross-sectional Model of Hurricane Damages.	7/98 = 0.07	0.01	p≈0	p≈0
Pretis et al. (2018)	Panel Estimation of Climate Impacts on Economic Growth	81/7007 = 0.012	0.001	p≈0	p≈0

Table 2.5.2: The normal and Poisson outlier tests applied to 12 published studies using impulse indicator saturation (IIS): $\widetilde{\gamma}$ shows the estimated proportion of outliers as the number $n\widetilde{\gamma}_c$ of impulse indicators retained over the number of observations, γ_c the nominal significance level of selection, and "Prop." and "Count" show the p-values of applying the *proportion* and *count* tests to the authors' results.

indicator saturation: first, a set of tests on the proportion and count of detected outliers, and second, a set of scaling tests. Tests on the proportion and count have power against the number of outliers, while scaling tests have power against both the number and magnitude of outliers. Both sets of tests are valid for stationary or stochastically trending regressors, and can also directly be applied to cross-sectional or panel models. It is worth highlighting that rejection of the null hypothesis of the observed outlier proportion (or number) equalling the expected proportion (or number) is indicative of model misspecification, but non-rejection does not rule out that individual outliers (even in a small proportion) denote actual outlying observations, as the outlier magnitude is not considered when testing on the proportion alone. The outlier magnitude can be taken into account through the proposed scaling tests by assessing the number of outliers at varying levels of the nominal level of significance used to detect outliers. These tests perform well with size close to nominal levels and high power against both magnitude and number of outliers.

The testing procedures are readily available through *Autometrics* in PcGive and

the R-package *gets*, and can be applied in a wide range of settings: from evaluating model misspecification, to the accuracy of forecasts in simulated models, to assessing the presence of outliers in time series, panel, or cross sectional models.

2.7 Appendix: Proofs

Here we provide the proofs of the lemmas and theorems in the main text.

Proof of Lemma 2.2.1. The term of interest is $\mathcal{M}_n = n^{-1/2}\sum_{i=1}^n v_i^{a,b,c}$.

1. *Decompose* $\mathcal{M}_n$. Write $\mathcal{M}_n = \mathcal{M}_{n,1} + \mathcal{M}_{n,2} + \mathcal{M}_{n,3}$, where

$$\mathcal{M}_{n,1} = n^{-1/2}\sum_{i=1}^n 1_{(|\varepsilon_i|\leq\sigma c)}, \quad \mathcal{M}_{n,2} = n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}\{v_i^{a,b,c} - 1_{(|\varepsilon_i|\leq\sigma c)}\},$$
$$\mathcal{M}_{n,3} = n^{-1/2}\sum_{i=1}^n \{v_i^{a,b,c} - 1_{(|\varepsilon_i|\leq\sigma c)}\} - n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}\{v_i^{a,b,c} - 1_{(|\varepsilon_i|\leq\sigma c)}\}.$$

Therefore, the first term in stochastic expansion is $\mathcal{M}_{n,1}$. We will linearize $\mathcal{M}_{n,2}$ to obtain the second term, and argue that $\mathcal{M}_{n,3}$ is small in probability.

2. *Linearize* $\mathcal{M}_{n,2}$. By Assumption 1.3.1(ia, iic) Theorem 10 in Jiao and Nielsen (2017) shows $\mathcal{M}_{n,2} = \overline{\mathcal{M}}_{n,2} + \mathrm{O}_\mathsf{P}(n^{-2\eta})$, where $\overline{\mathcal{M}}_{n,2} = 2c\mathsf{f}(c)a/\sigma$. Note $0 < \eta \leq 1/4$. Thus, we have $\mathcal{M}_{n,2} = \overline{\mathcal{M}}_{n,2} + \mathrm{o}_\mathsf{P}(1)$ uniformly in $0 < c < \infty$ and $|a|, |b| \leq n^{1/4-\eta}B$.

3. *Bounding* $\mathcal{M}_{n,3}$. Due to Assumption 1.3.1(ia, iib, iic), Theorem 9 in Jiao and Nielsen (2017) shows $\mathcal{M}_{n,3} = \mathrm{o}_\mathsf{P}(1)$ uniformly in a, b, c. ■

Remark 2.7.1. *To prove Lemma 2.2.1, Jiao and Nielsen (2017) and Johansen and Nielsen (2016a) requires the higher moment condition $q = 1 + 2^{r+1}$ for which $r \geq 2$ so $2^{r-1} \geq 1 + (1/4 + \kappa - \eta)(1 + \dim x)$ in Assumption 1.3.1. Berenguer-Rico and Nielsen (2018, Theorem 2.1) apply the truncation argument on regressors x_i with the purpose of reducing the moment condition. In fact the proof here adapts Berenguer-Rico and Nielsen (2018) argument so only needs the lower moment condition $q = 8$ which does not depend on the dimension of x_i.*

Proof of Theorem 2.2.1. The definition (1.1.5) of the sample gauge gives

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^n \{(1 - v_{i,c}^{(m)}) - \gamma_c\}.$$

Then express the weight $v_{i,c}^{(m)}$ in (1.2.3) as

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \leq \widehat{\sigma}_c^{(m)} c)} = 1_{(|\varepsilon_i - x_{in}'\widehat{b}_c^{(m)}| \leq \sigma c + n^{-1/2}\widehat{a}_c^{(m)} c)} = v_i^{\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c},$$

where $\widehat{b}_c^{(m)} = N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$ and $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are the m step estimation errors for β and σ. Since $|\widehat{b}_c^{(m)}| + |\widehat{a}_c^{(m)}| = \mathrm{O_P}(1)$ and by Assumption 1.3.1(ia, ii), apply the asymptotic expansion in Lemma 2.2.1 to obtain for any $0 \leq m < \infty$

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} - 2c\mathsf{f}(c)n^{1/2}(\frac{\widehat{\sigma}_c^{(m)}}{\sigma} - 1) + R(\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c),$$

where the remainder term $R(a, b, c)$ vanishes in probability uniformly in $c \in \mathbb{R}_+$ and $|a|, |b| \leq B$. ■

Proof of Theorem 2.2.2. By Assumption 1.3.1(ia, ii, iii) with $\eta = 1/4$, Theorem 2 in Jiao and Nielsen (2017) shows

$$\sup_{0\leq m<\infty}\sup_{c_0\leq c<\infty}|N^{-1}(\widehat{\beta}_c^{(m)} - \beta)| + |n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)| = \mathrm{O_P}(1).$$

Thus apply Theorem 2.2.1 to get

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} - 2c\mathsf{f}(c)n^{1/2}(\frac{\widehat{\sigma}_c^{(m)}}{\sigma} - 1) + \mathrm{o_P}(1),$$

where, uniformly in $m \in [0, \infty)$ and $c \in [c_0, \infty)$, the first term is $\mathrm{O_P}(1)$ due to the empirical process CLT, see Billingsley (1968, Theorem 16.4) or van der Vaart and Wellner (Theorem 2.12.1), while the second term is $\mathrm{O_P}(1)$ since $\sup_{c_0\leq c<\infty} c\mathsf{f}(c) < \infty$ by Assumption 1.3.1(ia) and tightness of $\widehat{\sigma}_c^{(m)}$. Therefore, the above term of interest is bounded in probability uniformly in m, c. ■

Proof of Theorem 2.2.3. Since $\widehat{\beta}_c^{(0)} = \widetilde{\beta}$, $(\widehat{\sigma}_c^{(0)})^2 = \widetilde{\sigma}^2$ are full sample least squares and by moment conditions for regressors in Assumption 1.3.1(ii), the initial estimators satisfy $N^{-1}(\widehat{\beta}_c^{(0)} - \beta) = \mathrm{O_P}(1)$, $n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) = \mathrm{O_P}(1)$ and

$$n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) = n^{-1/2}(2\sigma)^{-1}\sum_{i=1}^{n}(\varepsilon_i^2 - \sigma^2) + \mathrm{O_P}(n^{-1/2}).$$

Further by Assumption 1.3.1(ia, ii), the asymptotic expansion in Theorem 2.2.1 shows

$$n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} - c\mathsf{f}(c)n^{-1/2}\sum_{i=1}^{n}(\frac{\varepsilon_i^2}{\sigma^2} - 1) + \mathrm{o_P}(1),$$

uniformly in $c \in [c_0, \infty)$. The finite dimensional convergence follows by CLT so for any $c \in [c_0, \infty)$ then

$$\mathbb{G}_n(c) = n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) \xrightarrow{\mathsf{D}} \mathsf{N}\{0, \gamma_c(1-\gamma_c) + 2c\mathsf{f}(c)(\tau_2^c + \gamma_c - 1) + c^2\mathsf{f}^2(c)(\tau_4 - 1)\}.$$

Moreover, with the tightness result of $\mathbb{G}_n$ in Theorem 2.2.2 by Assumption 1.3.1(ia, ii), weak convergence arises so

$$\mathbb{G}_n \rightsquigarrow \mathsf{GP}(0, \Sigma).$$

The rest is to calculate the covariance for the above process between the time s and t where $c_0 \leq s \leq t < \infty$. Replace c by s and t respectively in the above stochastic expansion of $n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c)$, and then compute asymptotic covariance as

$$\Sigma_{st} = \gamma_t(1-\gamma_s) + s\mathsf{f}(s)(\tau_2^t + \gamma_t - 1) + t\mathsf{f}(t)(\tau_2^s + \gamma_s - 1) + st\mathsf{f}(s)\mathsf{f}(t)(\tau_4 - 1). \quad \blacksquare$$

Proof of Theorem 2.2.4. Since Assumption 1.3.1(ii) holds for $\mathcal{I}_1$, $\mathcal{I}_2$, the initial estimators satisfy $N_j^{-1}(\widehat{\beta}_j - \beta) = \mathrm{O}_\mathsf{P}(1)$, $n_j^{1/2}(\widehat{\sigma}_j - \sigma) = \mathrm{O}_\mathsf{P}(1)$ and

$$n_j^{1/2}(\widehat{\sigma}_j - \sigma) = n_j^{-1/2}(2\sigma)^{-1}\sum_{i\in\mathcal{I}_j}(\varepsilon_i^2 - \sigma^2) + \mathrm{O}_\mathsf{P}(n_j^{-1/2}),$$

for $j = 1, 2$. Insert this into the asymptotic expansion in Theorem 2.2.1 by Assumption 1.3.1(ia, ii) and use $n_1 = n_2 = n/2$ to get

$$\begin{aligned} n_j^{-1/2}\sum_{i\in\mathcal{I}_j}\{1_{(|y_i - x_i'\widehat{\beta}_{3-j}| > \widehat{\sigma}_{3-j}c)} - \gamma_c\} &= n_j^{-1/2}\sum_{i\in\mathcal{I}_j}\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} \\ &\quad - c\mathsf{f}(c)n_j^{-1/2}\sum_{i\in\mathcal{I}_{3-j}}(\frac{\varepsilon_i^2}{\sigma^2} - 1) + \mathrm{o}_\mathsf{P}(1), \end{aligned}$$

for $j = 1, 2$ and uniformly in $c \in [c_0, \infty)$. Combine counts of outliers for two subsamples $\mathcal{I}_1$, $\mathcal{I}_2$ to obtain

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^{(-1)} - \gamma_c) &= n^{-1/2}\sum_{j=1}^{2}\sum_{i\in\mathcal{I}_j}\{1_{(|y_i - x_i'\widehat{\beta}_{3-j}| > \widehat{\sigma}_{3-j}c)} - \gamma_c\} \\ &= n^{-1/2}\sum_{j=1}^{2}[\sum_{i\in\mathcal{I}_j}\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} - c\mathsf{f}(c)\sum_{i\in\mathcal{I}_{3-j}}(\frac{\varepsilon_i^2}{\sigma^2} - 1)] + \mathrm{o}_\mathsf{P}(1), \end{aligned}$$

uniformly in c. This reduces to the expansion for the Robustified Least Squares. Thus argue along the lines of Theorem 2.2.3 to attain the same limiting Gaussian process. $\blacksquare$

Proof of Theorem 2.2.5. By Assumption 1.3.1(ia, ii, iii) with $\eta = 1/4$, Theorem 3 in Jiao and Nielsen (2017) shows, when m, n are large, that the iterated 1-step Huber-skip M-estimator converges in probability to the fixed point $\widehat{\beta}_c^*$, $\widehat{\sigma}_c^*$ where $N^{-1}(\widehat{\beta}_c^* - \beta) = \mathrm{O}_\mathsf{P}(1)$, $n^{1/2}(\widehat{\sigma}_c^* - \sigma) = \mathrm{O}_\mathsf{P}(1)$ and

$$n^{1/2}(\widehat{\sigma}_c^* - \sigma) = \frac{1}{2\sigma\{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)\}}n^{-1/2}\sum_{i=1}^{n}(\varepsilon_i^2 - \varsigma_c^2\sigma^2)1_{(|\varepsilon_i|\leq\sigma c)} + \mathrm{o}_\mathsf{P}(1),$$

uniformly in c. Thus the gauge also has the fixed point $\widehat{\gamma}_c^* = n^{-1}\sum_{i=1}^n 1_{(|y_i - x_i'\widehat{\beta}_c^*| > \widehat{\sigma}_c^* c)}$, and then substitute $n^{1/2}(\widehat{\sigma}_c^* - \sigma)$ into the stochastic expansion in Theorem 2.2.1 to obtain

$$n^{1/2}(\widehat{\gamma}_c^* - \gamma_c) = n^{-1/2}\sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c)} - \gamma_c\} \\ - \frac{c\mathsf{f}(c)}{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)} n^{-1/2}\sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|\varepsilon_i| \leq \sigma c)} + \mathsf{o_P}(1),$$

uniformly in $c \in [c_0, \infty)$. CLT shows finite dimensional convergence so for any c

$$\mathbb{G}_n^*(c) = n^{1/2}(\widehat{\gamma}_c^* - \gamma_c) \xrightarrow{\mathsf{D}} \mathsf{N}[0, \gamma_c(1-\gamma_c) + \{\frac{c\mathsf{f}(c)}{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)}\}^2 \{\tau_4^c - \frac{(\tau_2^c)^2}{1-\gamma_c}\}].$$

Moreover, with tightness of $\mathbb{G}_n^*$ by Theorem 2.2.2, weak convergence is established so

$$\mathbb{G}_n^* \rightsquigarrow \mathsf{GP}(0, \Sigma).$$

For $c_0 \leq s \leq t < \infty$, replace c by s and t respectively in the expansion of $n^{1/2}(\widehat{\gamma}_c^* - \gamma_c)$, and then compute their asymptotic covariance as

$$\Sigma_{st} = \gamma_t(1-\gamma_s) - \frac{t\mathsf{f}(t)}{\tau_2^t - t(t^2 - \varsigma_t^2)\mathsf{f}(t)}\{\varsigma_t^2(1-\gamma_s) - \tau_2^s\} \\ + \frac{s\mathsf{f}(s)}{\tau_2^s - s(s^2 - \varsigma_s^2)\mathsf{f}(s)} \frac{t\mathsf{f}(t)}{\tau_2^t - t(t^2 - \varsigma_t^2)\mathsf{f}(t)}\{\tau_4^s - \frac{(\tau_2^s)^2}{1-\gamma_s}\}. \qquad \blacksquare$$

Proof of Theorem 2.3.1. Notice first $\sum_{c \in \mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c) = \sum_{k=1}^K n^{1/2}(\widetilde{\gamma}_{c_k} - \gamma_{c_k})$. Due to Assumption 1.3.1(ia, ii), proofs of Theorem 2.2.3, 2.2.4 show the asymptotic expansion

$$n^{1/2}(\widetilde{\gamma}_c - \gamma_c) = n^{-1/2}\sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c)} - \gamma_c\} - c\mathsf{f}(c) n^{-1/2}\sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - 1) + \mathsf{o_P}(1),$$

uniformly in $c \in [c_0, \infty)$. Note c_1 is chosen in a way so that $c_1 \geq c_0$. Then replace c by $c_k, k = 1, 2, \ldots, K$ respectively in the above expansion and insert these into the test statistic to obtain

$$\sum_{k=1}^K n^{1/2}(\widetilde{\gamma}_{c_k} - \gamma_{c_k}) = \sum_{k=1}^K n^{-1/2}\sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c_k)} - \gamma_{c_k}\} \\ - \{\sum_{k=1}^K c_k \mathsf{f}(c_k)\}\{n^{-1/2}\sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - 1)\} + \mathsf{o_P}(1).$$

Therefore CLT shows

$$\sum_{c \in \mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \mathsf{Var}_{\mathcal{C}_K}),$$

where

$$\mathsf{Var}_{\mathcal{C}_K} = \sum_{k=1}^{K} \gamma_{c_k}(1-\gamma_{c_k}) + 2 \sum_{1\le k<l\le K} (\gamma_{c_l} - \gamma_{c_k}\gamma_{c_l}) + \{\sum_{k=1}^{K} c_k \mathsf{f}(c_k)\}^2(\tau_4 - 1)$$
$$+ 2\{\sum_{k=1}^{K} c_k \mathsf{f}(c_k)\}\{\sum_{k=1}^{K} (\tau_2^{c_k} + \gamma_{c_k} - 1)\}. \qquad \blacksquare$$

Chapter 3

A Simple Robust Procedure in Instrumental Variables Regression

ABSTRACT[1]. A frequent concern in applied economics is that key empirical findings may be driven by a tiny set of outliers. To perform outlier robustness checks in practical applications of instrumental variables regressions, the common practice is first to run ordinary two stage least squares and remove observations classified as outliers with residuals beyond a chosen cut-off value. Subsequently 2SLS is re-calculated based on non-outlying observations, and this procedure can also be iterated until robust results are obtained. In this chapter, we analyze this simple robust algorithm asymptotically, then provide consistent estimation and valid inferential procedures on structural parameters for practical implementation given the cut-off value. Moreover, this chapter provides asymptotic theory for setting the cut-off, which is chosen to control the gauge (proportion of outliers wrongly discovered). Finally, we construct a formal test of outlier robustness based on the Hausman type test statistics comparing between the ordinary and robust estimators. Asymptotics are derived under the null hypothesis that there is no contamination in the cross-sectional i.i.d. data. The established weak convergence result involves empirical processes and the fixed point arguments. Thus, this chapter also proves the uniform and weak law for a new class of weighted and marked empirical processes of residuals in IVs regressions, allowing for estimation errors of structural parameters. An empirical application to Acemoglu et al. (2017) shows the utility of the proposed theory.

[1]This chapter is wholly my own work and becomes my job market paper

3.1 Introduction

In applied economics research, all valid empirical findings should share one essential feature: that the results of analyses are not affected by a small portion of atypical observations to a nonnegligible degree; see Young (2017). Thus, empirical researchers are frequently concerned that their crucial findings are driven by a tiny set of outliers. For example, on estimating risk aversion coefficients, Toda and Walsh (2015) found that trimming a small sub-sample of observations drastically changes their estimates, making otherwise insignificant statistical inference significant.

For the regression or instrumental variables regression work in the empirical economics and finance literature, a recurrent critique is to ask whether the paper examines the robustness of its conclusion, especially with respect to outlying observations. For instance, considering De Long and Summers' (1991) paper investigating the effect of equipment investment on economic growth, Auerbach, Hassett, and Oliner (1994) questioned whether the positive capital effect is due to outliers. De Long and Summers (1994) agreed that they faced sample-selection issues and their findings are particularly sensitive to one outlier. Similarly in estimating the institutional impact on economic growth, Acemoglu, Johnson, and Robinson (2001) attracted extensive comments[2] on whether the validity of instruments by mortality rates is ruined by some outlying observations. Guthrie, Sokolowsky, and Wan's (2012) critique of Chhaochharia and Grinstein (2009) was that the removal of a few observations invalidates their empirical results, while the similar comment on Reinhart and Rogoff (2010) can be found in Herndon, Ash, and Pollin (2014).

In parallel to the empirical work, some recent theoretical papers in econometrics are also concerned about outlier contamination, particularly its effect on the performance of statistical inference. Taking Berenguer-Rico, Johansen and Nielsen (2019a, b), Berenguer-Rico and Nielsen (2018), and Berenguer-Rico and Wilms (2018) as examples, they showed that outlying observations substantially influence the outcome of specification checks which test if residuals are i.i.d. normal.

In response to the long-standing issue that potential outliers may drive key empirical findings, the statistics literature[3] then designed some robust methods in regression which are less sensitive to outlying observations. In the context of instrumental vari-

[2] see for example Albouy (2012) and Acemoglu, Johnson, and Robinson (2012).

[3] see the thorough treatment in Rousseeuw and Leroy (1987) and Maronna, Martin, and Yohai (2006).

ables regression, Young (2018) carried out a comprehensive empirical study of 1400 IVs regressions in 32 papers published in journals of the American Economic Association, and found that the two stage least squares is more sensitive to outliers than the ordinary least squares. Thus, two types of robust methods (quantile and rank regression) were proposed and subsequently extended to the IVs setting (see Honoré and Hu, 1994).

The quantile idea was first successfully introduced to regression by Koenker and Bassett (1978) to study the quantile effect (the whole distribution rather than just the mean effect) of treatment on the outcome of interest, and afterwards Chernozhukov and Hansen (2005) further generalized this method to simultaneous equations involving instruments. Meanwhile, within the local average treatment effect framework of Imbens and Angrist (1994), Imbens and Newey (2009) provided another quantile approach through IVs based on Abadie, Angrist, and Imbens (2002)[4]. It is well known that the quantile approach is not robust to bad leverage points where the joint distribution of dependent and independent variables shift, see Rousseeuw and Leroy (1987). Compared to the quantile method, the rank regression by Adichie (1967) has been used much less often to perform IVs in economics (see Honoré and Powell, 1994).

Even though there are two existing IVs robust methods in econometrics, the common practice in empirical research for outlier analysis is to first run ordinary two stage least squares and to find outliers by examining if residuals are higher than the selected cut-off. 2SLS is then re-computed with the non-outlying observations, and this procedure can be iterated until the robust result is obtained. In practical applications of IVs regressions, many economists apply this simple procedure to carry out robustness checks by comparing the updated estimates from original ones relative to standard errors, see for instance De Long and Summers (1991, 1994), Auerbach, Hassett, and Oliner (1994), Acemoglu, Johnson, and Robinson (2001, 2012), Albouy (2012), Fabrizio, Rose, and Wolfram (2007), Toda and Walsh (2015), and Acemoglu et al. (2017).

Analyzing this common practice is significantly important for outlier analysis in empirical IVs research, since asymptotic theory will give a guidance on better implementing the algorithm to identify outliers, to obtain consistent and robust estimates, to make valid inference, and to perform a formal test of outlier robustness examining whether the robust estimate is distinct from the ordinary one. However, this heuristic approach lacks formal justification, and it is technically demanding to build up the probability tools (empirical processes) for establishing its statistical properties.

[4]see discussion in Wüthrich (2018) for a connection between estimands of these two quantile models in IVs.

The main contribution of this chapter is to analyse this simple robust IVs algorithm asymptotically, then provide procedures for consistent estimation and valid inference on structural parameters given the cut-off value. Moreover, this chapter contributes to providing asymptotic theory for setting the cut-off using the concept of gauge (the proportion of outliers falsely detected), like the tuning parameters in non-parametric econometrics. Lastly, we apply the established weak convergence results to construct a formal test of outlier robustness based on the Hausman type test statistics checking whether the robust estimate is equal to the ordinary one. Asymptotics in this chapter are derived under the null hypothesis that there is no contamination in the cross-sectional i.i.d. data. To carry out asymptotic analysis, we establish a theory for a new class of marked and weighted empirical processes of residuals, which allows for variation in location, scale and quantile in the structural equation.

The heuristic procedure described above can be formalized. Suppose we have independent and identically distributed data $\{(y_i, x_i, z_i)\}_{i=1}^n$ where the structural equation holds for $\{(y_i, x_i)\}_{i=1}^n$ such that

$$y_i = x_i'\beta + u_i, \quad i = 1, 2, \ldots, n,$$

while $\{(x_i, z_i)\}_{i=1}^n$ satisfy the projection equation

$$x_i' = z_i'\Pi + r_i', \quad i = 1, 2, \ldots, n.$$

Outliers are pairs of observations that do not conform with the structural equation so given an initial estimator $(\widetilde{\beta}, \widetilde{\sigma}^2)$ the robust algorithm decides which observations are outlying through $1_{(|y_i - x_i'\widetilde{\beta}| > \widetilde{\sigma}c)}$, where the choice of the cut-off c is related to the known reference density $u_i/\sigma \sim \mathsf{f}_u$. For those observations that are not outliers, re-run two stage least squares regression to get the updated robust estimator $(\widehat{\beta}, \widehat{\sigma})$. Notice that the usual estimator of σ^2 need to be corrected, so we will introduce ς^2 in section §2 as the bias correction factor that allows consistent estimation for σ^2. The one-step robust procedure can be iterated and the iteration may also be initiated by a robust estimator. More simply, two special algorithms heavily applied by empirical researchers are either starting with the full sample estimate, see for example Acemoglu et al. (2017); or the split half sample estimate, see for example Hendry (1999).

Analysis of robust methods in the iterated one-step frame is by no means new. In the classical regression, one-step (Huber, L, and M) estimators have been considered by Bickel (1975), Ruppert and Carroll (1980), and Welsh and Ronchetti (2002). The idea of iterating one-step estimators can be found in Dollinger and Staudte (1991),

who applied an influence function argument to demonstrate convergence of iteratively reweighted least squares with smooth weights. Notwithstanding we are interested in binary weights, the spirit of their argument still provides a guidance for our IVs robust procedure. Other related work includes for example Cavaliere and Georgiev (2013), who explored the first order autoregression with infinite variance. For iterated one-step Huber-skip M and L-estimators, asymptotic theory has been recently established by Hendry, Johansen, and Santos (2008), Berenguer-Rico and Nielsen (2018), Johansen and Nielsen (2009, 2013, 2016a, 2016b, 2018), and Jiao and Nielsen (2017)[5]. The simple robust procedure studied in this chapter mimics the iterated one-step Huber-skip M-estimator but in the IVs setting. Thus, this chapter extends asymptotics of iterated one-step Huber-skip M-estimators to the IVs regression where the exogeneity restriction on regressors breaks down, i.e. $\mathsf{E}x_i u_i \neq 0_{d_x}$.

The leading purpose of this work is to build up an asymptotic theory for the simple robust procedure commonly applied by empirical researchers in the IVs regression. With the derived weak convergence result of the algorithm, we propose a consistent updated estimator of σ^2 and provide a valid inference on β. After removing outliers, the standard method will underestimate σ^2 using the sub-sample of non-outlying observations due to the sample truncation, though the updated β estimator is consistent. Moreover, without the correct limiting distribution of the outlier removal estimator of β, the heuristic procedure uses the usual 2SLS asymptotics to perform inference with the selecting non-outlying observations. Thus, the standard errors are erroneously calculated and the standard inference method is invalid. To fix the problem, we introduce the bias correction factor ς^2 to adjust the updated σ^2 estimator and derive the weak limit of the updated β estimator, which together allow us to perform valid inference on structural parameters β.

Another contribution of this chapter is to construct a formal test of checking whether the empirical results are driven by outliers. We suggest using the Durbin-Hausman-Wu test statistics to measure the distance between the outlier removal and the full sample estimators, and then show that it converges to a chi-square distribution by analytically deriving the joint limiting distribution of two estimators (the ordinary and robust). It thus enables us to test whether the parameter of interest changes its value significantly before and after removing outliers, allowing formal statistical analysis of

[5]Maronna and Yohai (1997) and Hansen, Heaton, and Yaron (1996) also studied iterated (three stage least squares, GMM) estimators in the IVs regression but with the distinct concern on asymptotic efficiency when non-i.i.d. structural errors occur.

outlier robustness checks.

Kaji (2018) applies the nonparametric bootstrap by randomly sampling observations i with replacement from the data $\{(y_i, x_i, z_i)\}_{i=1}^n$ to draw the joint distribution of the outlier removal and the full sample estimators so as to obtain a distribution of the test statistics of outlier robustness. To prove its validity, he represents the outlier removal estimator as L-statistics (integrals of transformations of empirical quantile functions with respect to random sample selection measures), and develops a new empirical process theory and a functional delta method for its quantile process tailored for these statistics, with an important innovation to the choice of L_1 norm (instead of the standard L_∞ norm in the literature). As compared with our test, the bootstrapping procedure does not require a distributional assumption on $u_i/\sigma \sim \mathsf{f}_u$, however, it has the lower power under alternatives where there are significant outliers in the data. This could be because the drawn distribution is distorted by bootstrapping iterations that sample and include outlying observations from the original data, and so is distinct from what it should be under the null of no outliers. Another outlier robustness test frequently used is the so-called heuristic approach, which only considers the marginal distribution of the baseline full sample estimator when construct its test statistics. Thus, by ignoring the natural comovement between the ordinary and robust estimator, it is invalid though informative in the preliminary outlier robustness analysis.

We re-explore the outlier robustness analysis in Acemoglu et al. (2017) as our empirical application to show the utility of the established theory. Their paper estimates the effect of democracy on log GDP per capita, and examines the sensitivity of their estimators to the removal of outliers on the structural residuals through $1_{(|y_i - x_i'\widetilde{\beta}| > \widetilde{\sigma}c)}$, where they choose the full sample 2SLS for $(\widetilde{\beta}, \widetilde{\sigma})$ and $c = 1.96$. We perform valid inference on β based on the robust method after removing outlying observations and conducts a formal Hausman type test of outlier robustness to statistically check whether trimming off outliers would dramatically change the estimates. For most coefficients, the outlier robustness tests are rejected at 5% significance level, indicating the outlier removal estimator is distinct from the full sample one at least in magnitutde. We find that the rejections are not discovered by either the heuristic or the Kaji (2018) nonparametric bootstrapping testing procedures. The adjusted inference based on the robust estimator cannot reject the hypothesis that the coefficient on democracy equals zero, thus the result of outlier analysis does not robustly support that democracy has a strictly positive effect on GDP per capita, though there is little evidence for the view that democracy is

a constraint for economic development.

The argument involves a theory for a new class of weighted and marked empirical processes of residuals as well as the fixed point theory. Since we will apply the martingale argument to analyze the empirical processes, so construct the filtration as follows

$$\mathcal{F}_{i-1} = \sigma(z_1, \ldots, z_i, r_1, \ldots, r_{i-1}, u_1, \ldots, u_{i-1}).$$

The empirical process is defined from the generalized empirical distribution function

$$\widehat{\mathsf{F}}_{u,n}^{w,p}(a, b, c) = \frac{1}{n}\sum_{i=1}^{n} w_{in} u_i^p 1_{(u_i \leq \sigma c + n^{-1/2} ac + z_{in}' \Pi b + n^{-1/2} r_i' b)},$$

where the weights w_{in} are combinations of the normalized $\mathcal{F}_{i-1}$ measurable instruments z_{in} and u_i^p are the $\mathcal{F}_i$ adapted marks, while a, b represent the normalized estimation errors for σ, β. We derive asymptotic expansions that are uniform in a, b, c and allow for a near $n^{1/4}$ inefficiency in the estimation uncertainties a, b.

Following the classical idea, Koul and Ossiander (1994) used the entropy argument to first build up the theory for the weighted empirical process in the autoregressive model[6]. Recently, empirical processes including both weights and marks have been analyzed by a series of papers, see Engler and Nielsen (2009), Johansen and Nielsen (2009, 2016a), Jiao and Nielsen (2017), Jiao (2018), Berenguer-Rico and Nielsen (2018), and Berenguer-Rico, Johansen, and Nielsen (2019a, b). However, all these papers restrict attention to the classical regression model where regressors x_i are assumed to be orthogonal to errors u_i, i.e. $\mathsf{E}x_i u_i = 0_{d_x}$, so the filtration at $i-1$ can be simply constructed as $\mathcal{H}_{i-1} = \sigma(x_1, \ldots, x_i, u_1, \ldots, u_{i-1})$. But in the IVs setting further dependence structure is involved between x_i and u_i or r_i and u_i making the problem complicated, and thus their argument does not go through and we have to reconstruct the martingale based on filtration $\mathcal{F}_{i-1}$ instead[7].

[6]also see Koul (2002) and Koul and Ling (2006).

[7]In the classical setting where $\mathsf{E}x_i u_i = 0_{d_x}$ holds such that the simple filtration $\mathcal{H}_{i-1}$ applies, Nielsen and Qian (2018) faced the related problem when investigating the empirical process of temporal differenced residuals

$$\widehat{\mathsf{H}}_n(a, b, c) = \frac{1}{n}\sum_{i=1}^{n} 1_{(\nabla u_i \leq \sqrt{2}\sigma c + \sqrt{2} n^{-1/2} ac + \nabla x_{in}' b)},$$

where $\nabla u_i = u_i - u_{i+1}$, $\nabla x_{in} = x_{in} - x_{i+1,n}$, and x_{in} are normalized regressors due to its stochastic property, while a, b have the same meaning as in our setup. At time $i-1$ only x_{in} is $\mathcal{H}_{i-1}$ adapted while ∇u_i is independent of $\mathcal{H}_{i-1}$. Then dependence arises between ∇u_i and ∇x_{in} since u_i in ∇u_i is correlated with $x_{i+1,n}$ in ∇x_{in} at $\mathcal{H}_{i-1}$ although $\mathsf{E}x_i u_i = 0_{d_x}$. But their issue is different from this chapter since we relax independence between regressors x_i and errors u_i in the process.

One of our contributions is to show uniform convergence and tightness for the empirical process constructed by $n^{1/2}\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,c)$, which are essential for building up the weak convergence theory. Further we provide the linearization for the compensator of $n^{1/2}\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,c)$. While $n^{1/2}$-convergence results are demonstrated for the process $n^{1/2}\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,c)$, we finally establish n-convergence for a new type of empirical processes with marks r_i instead of u_i.

The paper proceeds as follows: §3.2 proposes a class of robust algorithms in the instrumental variables regression. §3.3 presents the main results and §3.4 re-explores outlier robustness analysis in Acemoglu et al. (2017), whilst §3.5 establishes the uniform and weak law for a new class of the weighted and marked empirical processes with proofs in Appendix §3.7, 3.8. Lastly, proofs of the main theorems in §3.3 are shown in Appendix §3.9.

3.2 Model and a robust procedure

The instrumental variables (IVs) regression with some notations is described first. It is widely known that two stage least squares estimator is sensitive to outliers. We review a robust procedure related to the iterated one-step Huber-skip M-estimator but in an IVs setting. At last, define the concept of gauge to measure the rate of false detection of outlier detection algorithms. Then the cut-off value of such algorithms, like tuning parameters in nonparametrics, can be chosen indirectly from the gauge.

3.2.1 Model

Suppose in the cross-sectional settings we have independently and identically distributed data $\{(y_i, x_i, z_i)\}_{i=1}^n$ across individuals, where y_i is univariate and x_i, z_i are multivariate with dimension d_x, d_z. Assume the data $\{(y_i, x_i)\}_{i=1}^n$ satisfies the structural equation

$$y_i = x_i'\beta + u_i, \quad i = 1, 2, \ldots, n. \tag{3.2.1}$$

Structural errors $\{u_i\}_{i=1}^n$ are univariate i.i.d. random variables with scale σ so that u_i/σ has the known density $\mathsf{f}_u(y)$ and distribution function $\mathsf{F}_u(y) = \mathsf{P}(u_i/\sigma \leq y)$ for $y \in \mathbb{R}$ with mean 0 and variance 1. In the structural model some elements of x_i are endogenous, i.e. $\mathsf{E}x_i u_i \neq 0_{d_x}$, instruments z_i are required for estimating the parameter $\beta \in \mathbb{R}^{d_x}$ consistently. The first stage regression holds for $\{(x_i, z_i)\}_{i=1}^n$

$$x_i' = z_i'\Pi + r_i', \quad i = 1, 2, \ldots, n. \tag{3.2.2}$$

Instruments z_i are assumed to have the zero mean so $\mathsf{E}z_i = 0_{d_z}$ and to be orthogonal to errors r_i in the first stage equation so $\mathsf{E}z_i r_i' = 0_{d_z \times d_x}$. Innovations $\{r_i\}_{i=1}^n$ are d_x-variate i.i.d. random vector with symmetric and positive definite dispersion matrix $\Sigma \in \mathbb{R}^{d_x \times d_x}$ so $\Sigma^{-1/2} r_i$ follows the density $\mathsf{f}_r(x)$ and distribution function $\mathsf{F}_r(x)$ for $x \in \mathbb{R}^{d_x}$ with mean 0_{d_x} and identity variance-covariance matrix I_{d_x}. The parameter Π lies in $\mathbb{R}^{d_z \times d_x}$ and we suppose $d_x \leq d_z$ meaning the number of regressors is less than or equal to the number of instruments. Furthermore, to identify structural parameters β, instruments z_i need to be valid, i.e. $\mathsf{E}z_i u_i = 0_{d_z}$, and informative, i.e. rank condition for identification ($\operatorname{rank} \Pi = d_x$ and $\operatorname{rank} \mathsf{E}z_i z_i' = d_z$).

To apply the martingale argument, we need to construct the filtration

$$\mathcal{F}_{i-1} = \sigma(z_1, \ldots, z_i, r_1, \ldots, r_{i-1}, u_1, \ldots, u_{i-1}) \tag{3.2.3}$$

such that u_i, r_i are $\mathcal{F}_i$ measurable and independent of $\mathcal{F}_{i-1}$ while z_i is adapted to $\mathcal{F}_{i-1}$. Notice in the IVs setting u_i is correlated with r_i. Assume the scaled errors $(u_i/\sigma, \Sigma^{-1/2} r_i)$ have the joint density $\mathsf{f}_{u,r}(y, x)$ and distribution function $\mathsf{F}_{u,r}(y, x)$ for $y \in \mathbb{R}$, $x \in \mathbb{R}^{d_x}$. Note the joint distribution $\mathsf{f}_{u,r}, \mathsf{F}_{u,r}$ does also depend on the covariance $\Omega = \mathsf{Cov}(u_i/\sigma, \Sigma^{-1/2} r_i) = \mathsf{E}(\Sigma^{-1/2} r_i u_i/\sigma)$, however for simplicity $\Omega \in \mathbb{R}^{d_x}$ is suppressed in the notation of joint density. Further suppose the joint density can be decomposed into the conditional and marginal ones so $\mathsf{f}_{u,r}(y, x) = \mathsf{f}_{u|r}(y|x)\mathsf{f}_r(x) = \mathsf{f}_{r|u}(x|y)\mathsf{f}_u(y)$. Then denote the conditional expected value

$$\xi_y = \mathsf{E}(\Sigma^{-1/2} r_i | u_i/\sigma = y) = \int_{\mathbb{R}^{d_x}} x \mathsf{f}_{r|u}(x|y)(dx). \tag{3.2.4}$$

Notice $\xi_y \in \mathbb{R}^{d_x}$ is related to the covariance Ω between u_i/σ and $\Sigma^{-1/2} r_i$. In practice $\mathsf{f}_{u,r}$, $\mathsf{F}_{u,r}$ will often be assumed to be $(1 + d_x)$-variate normal, so the scaled error vector $(u_i/\sigma, \Sigma^{-1/2} r_i)$ follows

$$\begin{pmatrix} u_i/\sigma \\ \Sigma^{-1/2} r_i \end{pmatrix} \overset{\mathsf{D}}{=} \mathsf{N}\left\{ \begin{pmatrix} 0 \\ 0_{d_x} \end{pmatrix}, \begin{pmatrix} 1 & \Omega' \\ \Omega & I_{d_x} \end{pmatrix} \right\}. \tag{3.2.5}$$

As the marginal and conditional distribution of multivariate normal are still normally distributed, then $\Sigma^{-1/2} r_i | u_i/\sigma \sim \mathsf{f}_{r|u}$ follows normal $\mathsf{N}(\Omega u_i/\sigma, I_{d_x} - \Omega\Omega')$. Thus the conditional expectation in (3.2.4) has the form $\xi_y = \Omega y$. Denote

$$\zeta_y^+ = \xi_y + \xi_{-y}, \qquad \zeta_y^- = \xi_y - \xi_{-y}. \tag{3.2.6}$$

Under normality in (3.2.5), we have $\xi_{-y} = -\Omega y = -\xi_y$ so $\zeta_y^+ = 0_{d_x}$ and $\zeta_y^- = 2\xi_y = 2\Omega y$.

The robust procedure we analyzed detects outliers through checking if absolute standardised residuals in the structural equation (3.2.1) are beyond the chosen cut-off value c and then calculating the robust two stage least squares estimator from the non-outlying sample. This implicitly assumes symmetry of u_i/σ, while non-symmetry leads to specific forms of bias. We assume symmetry for f_u when analyzing the robust algorithm in §3.3, but not for the general empirical process results in §3.5. Let the absolute error $|u_i|/\sigma$ in (3.2.1) have a density $\mathsf{g}_u(y)$ and distribution function $\mathsf{G}_u(y) = \mathsf{P}(|u_i|/\sigma \leq y)$ for $y > 0$. With a symmetry assumption, $\mathsf{G}_u(y) = 2\mathsf{F}_u(y) - 1$ and $\mathsf{g}_u(y) = 2\mathsf{f}_u(y)$. Define $\psi = \mathsf{G}_u(c)$ so the probability of exceeding the cut-off c is $\gamma = 1 - \psi$. Suppose the k-th moment of the density f_u exists, then introduce

$$\tau_k = \int_{-\infty}^{\infty} y^k \mathsf{f}_u(y)dy, \qquad \tau_k^c = \int_{-c}^{c} y^k \mathsf{f}_u(y)dy. \tag{3.2.7}$$

Thus $\tau_0^c = \psi$, $\tau_2 = 1$ while $\tau_k = \tau_k^c = 0$ for odd k when assuming symmetry for f_u. Define the conditional variance of u_i/σ given $(|u_i|/\sigma \leq c)$ as

$$\varsigma_c^2 = \frac{\tau_2^c}{\psi} = \frac{\int_{-c}^{c} y^2 \mathsf{f}_u(y)dy}{\mathsf{P}(|u_i| \leq \sigma c)}. \tag{3.2.8}$$

This will be used as a bias correction factor for the variance estimate computed from the selected non-outlying sample. With normality assumption (3.2.5), we have $u_i/\sigma \sim \mathsf{f}_u$ follows standard normal $\mathsf{N}(0,1)$ then $\tau_2^c = \psi - 2c\mathsf{f}_u(c)$, $\tau_4 = 3$, $\tau_4^c = 3\psi - 2c(c^2+3)\mathsf{f}_u(c)$.

For matrices M, choose the spectral norm $|M| = \max\{\text{eigen}(M'M)\}^{1/2}$ so that for vectors x then $|x|$ is the Euclidean norm. The spectral norm is compatible with respect to the Euclidean norm so $|Mx| \leq |M||x|$. Notice (dM) represents the exterior product of all elements in the matrix M to describe the multivariate integral, see the chapter 2 of Muirhead (1982). For instance consider a vector $x \in \mathbb{R}^{d_x}$, then (dx) is the exterior product of some measures on $\mathbb{R}^{d_x}$.

3.2.2 Robust statistical algorithms to detect outliers

We first define an iterated version of robust two stage least squares estimators. Two specific examples with different initial estimators are then described.

To carry out outlier analysis, Acemoglu et al. (2017) first compute the full sample 2SLS and choose it as the initial estimator for the parameter β, then calculate all the residuals in the structural equation and detect outliers if the absolute value of standardised residuals in (3.2.1) are beyond the selected cut-off value 1.96. Running two stage least squares based on the non-outlying observations gives them the updated robust

estimator for β. Finally, they compare two β estimates for outlier robustness checks. In principle, a different initial estimators can be chosen and this robust procedure can also be iterated. We define a formal version of the algorithm as follows.

Algorithm 3.2.1. *(**Iterated 2SLS**). Choose a cut-off $c > 0$.*

1. Select initial estimators $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ and let $m = 0$.

2. Define indicator variables for selecting non-outlying observations

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \leq \widehat{\sigma}_c^{(m)} c)}. \tag{3.2.9}$$

3. Calculate least squares estimators for Π

$$\widehat{\Pi}_c^{(m+1)} = (\sum_{i=1}^{n} z_i z_i' v_{i,c}^{(m)})^{-1} (\sum_{i=1}^{n} z_i x_i' v_{i,c}^{(m)}). \tag{3.2.10}$$

4. Compute two stage least squares estimators for β and σ^2

$$\widehat{\beta}_c^{(m+1)} = (\widehat{\Pi}_c^{(m+1)\prime} \sum_{i=1}^{n} z_i z_i' v_{i,c}^{(m)} \widehat{\Pi}_c^{(m+1)})^{-1} (\widehat{\Pi}_c^{(m+1)\prime} \sum_{i=1}^{n} z_i y_i v_{i,c}^{(m)}), \tag{3.2.11}$$

$$(\widehat{\sigma}_c^{(m+1)})^2 = \varsigma_c^{-2} (\sum_{i=1}^{n} v_{i,c}^{(m)})^{-1} \{\sum_{i=1}^{n} (y_i - x_i'\widehat{\beta}_c^{(m+1)})^2 v_{i,c}^{(m)}\}. \tag{3.2.12}$$

5. Let $m = m + 1$ and repeat 2, 3, and 4.

Algorithm 3.2.1 is a similar procedure as iterated one-step Huber-skip M-estimators[8] but in the instrumental variables context instead of the classical regression setting. Algorithm 3.2.1 chooses the same sub-sample of non-outlying observations by the indicators $v_{i,c}^{(m)}$ for estimating Π in the first stage regression (3.2.2) and β, σ^2 in the structural equation (3.2.1). Notice the fact, that the sub-sample selected for (3.2.2) is the same as for (3.2.1), is essential for demonstrating consistency of estimators for the structural parameters β, σ^2, see the discussion in Remark 3.9.2 (Appendix 3.9.2). Furthermore, to estimate σ^2 consistently the bias correction factor ς_c^2 is required to adjust in (3.2.12), otherwise without ς_c^2 corrected in the variance estimator it will converge in probability to $\varsigma_c^2\sigma^2$ instead, see Remark 3.9.3 (Appendix 3.9.2).

In §3.3 we analyze this algorithm asymptotically and show how to choose the cut-off c indirectly using the concept of gauge. The algorithm could start with a robust estimator,

[8] one-step updated estimator mimics the Huber (1964) skip estimator, which has criterion function $\rho(y) = \min(y^2, c^2)/2$ as opposed to the Huber estimator with criterion function $\rho(y) = y^2/2$ for $|y| \leq c$ and $\rho(y) = c|y| - c^2/2$ otherwise, see Hampel et al. (1986, p. 104) and Jurečková, Sen, and Picek (2013, p. 175).

while as the procedure applied in Acemoglu et al. (2017) the so called Robustified 2SLS is initiated using the full sample 2SLS. The following algorithm formally defines Robustified 2SLS.

Algorithm 3.2.2. *(**Robustified 2SLS**). Choose a cut-off $c > 0$.*

1.1. Calculate least squares estimator for Π based upon the whole sample

$$\widehat{\Pi}_c^{(0)} = (\sum_{i=1}^{n} z_i z_i')^{-1} (\sum_{i=1}^{n} z_i x_i'). \tag{3.2.13}$$

1.2. Compute two stage least squares estimator for β and σ^2 for the whole sample

$$\widehat{\beta}_c^{(0)} = (\widehat{\Pi}_c^{(0)\prime} \sum_{i=1}^{n} z_i z_i' \widehat{\Pi}_c^{(0)})^{-1} (\widehat{\Pi}_c^{(0)\prime} \sum_{i=1}^{n} z_i y_i), \qquad (\widehat{\sigma}_c^{(0)})^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' \widehat{\beta}_c^{(0)})^2. \tag{3.2.14}$$

1.3. Choose $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ through (3.2.14), and let $m = 0$.

2. Follow the step 2, 3, 4, and 5 in Algorithm 3.2.1.

Algorithm 3.2.2 is not very robust with respect to high leverage points when we have i.i.d. cross-sectional data. Thus, to deal with the high leverage points, another typical example of Algorithm 3.2.1 starting with the split sample 2SLS was first proposed but in the classical regression by Hendry (1999) in his empirical work[9]. This saturated idea is to divide the full sample into two sub-samples and use 2SLS calculated from each sub-sample to detect outliers in the other sub-sample. The following algorithm formally defines Saturated 2SLS.

Algorithm 3.2.3. *(**Saturated 2SLS**). Choose a cut-off $c > 0$.*

1.1. Split full sample into two sets $\mathcal{I}_j$, $j = 1, 2$ of n_j observations where $\sum_{j=1}^{2} n_j = n$.

1.2. Calculate least squares estimators for Π based upon each sub-sample $\mathcal{I}_j$ for $j = 1, 2$

$$\widehat{\Pi}_j = (\sum_{i \in \mathcal{I}_j} z_i z_i')^{-1} (\sum_{i \in \mathcal{I}_j} z_i x_i'). \tag{3.2.15}$$

1.3. Compute two stage least squares estimators for β and σ^2 for sub-sample $\mathcal{I}_j$, $j = 1, 2$

$$\widehat{\beta}_j = (\widehat{\Pi}_j' \sum_{i \in \mathcal{I}_j} z_i z_i' \widehat{\Pi}_j)^{-1} (\widehat{\Pi}_j' \sum_{i \in \mathcal{I}_j} z_i y_i), \qquad \widehat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} (y_i - x_i' \widehat{\beta}_j)^2. \tag{3.2.16}$$

1.4. Define the initial indicator variables for selecting non-outlying observations

$$v_{i,c}^{(0)} = 1_{(i \in \mathcal{I}_1)} 1_{(|y_i - x_i' \widehat{\beta}_2| \leq \widehat{\sigma}_2 c)} + 1_{(i \in \mathcal{I}_2)} 1_{(|y_i - x_i' \widehat{\beta}_1| \leq \widehat{\sigma}_1 c)}. \tag{3.2.17}$$

[9] see also Hendry, Johansen, and Santos (2008) and Hendry and Doornik (2014).

1.5. Compute $\widehat{\Pi}_c^{(1)}$, $\widehat{\beta}_c^{(1)}$, $(\widehat{\sigma}_c^{(1)})^2$ *using* (3.2.10), (3.2.11), (3.2.12) *with* $m = 0$, *and let* $m = 1$.

2. Follow the step 2, 3, 4, and 5 in Algorithm 3.2.1.

Algorithm 3.2.3 is possibly more robust than the Robustified Two Stage Least Squares when we have prior knowledge that outliers are located in a particular subset of the whole sample. Since the location of contaminated observations is unknown in most practical situations, the choice of the initial sets $\mathcal{I}_1$ and $\mathcal{I}_2$ should be iterated.

3.3 The main results

We start by listing assumptions, then asymptotics for Iterated 2SLS and weak convergence for Robustified 2SLS and Saturated 2SLS. It is followed by comparing efficiency between robust and non-robust 2SLS under the null of no outliers. Finally, I provide the valid inference procedure on β when implementing these trimming methods and propose a new Hausman type test to check the presence of influential outliers.

3.3.1 Assumptions

We list the sufficient assumptions for asymptotic theory of Algorithm 3.2.1, 3.2.2, 3.2.3. To simplify analysis, this section focuses on somewhat stronger conditions than they need to be. In section §3.5 on the one-sided empirical process, we will introduce some weaker assumptions. For example, we impose the symmetricity assumption on f_u in this section but not in section §3.5, though the idea of reasoning is the same and the results in §3.3 can be extended to the asymmetric case using the argument in Johansen and Nielsen (2009).

Since $\{(y_i, x_i, z_i)\}_{i=1}^n$ are i.i.d., instruments z_i can be normalized by the rate $n^{-1/2}$ so denote $z_{in} = n^{-1/2} z_i$, then we have $M_{zz,n} = \sum_{i=1}^n z_{in} z_{in}' = n^{-1} \sum_{i=1}^n z_i z_i' \xrightarrow{\mathsf{P}} \mathsf{E} z_i z_i' = M_{zz}$ by Law of Large Numbers assuming $\mathsf{E}|z_i|^2 < \infty$. Denote the infeasible ideal fitted value $\tilde{x}_i = \Pi' z_i$ while the feasible counterpart is attained by estimating Π consistently. The normalized $\tilde{x}_i$ is defined as $\tilde{x}_{in} = n^{-1/2} \tilde{x}_i = \Pi' z_{in}$ such that

$$M_{\tilde{x}\tilde{x},n} = \sum_{i=1}^n \tilde{x}_{in} \tilde{x}_{in}' = \Pi' M_{zz,n} \Pi \xrightarrow{\mathsf{P}} \Pi' M_{zz} \Pi = \mathsf{E} \tilde{x}_i \tilde{x}_i' = M_{\tilde{x}\tilde{x}}.$$

If $d_x \leq d_z$ and rank condition for identification hold so that $\operatorname{rank} \Pi = d_x$, $M_{zz} > 0$, then we have $M_{\tilde{x}\tilde{x}} > 0$ as well.

To carry out asymptotic analysis, the scaled error vector $(u_i/\sigma, \Sigma^{-1/2}r_i)$, instruments z_i, and the initial estimator $(\widetilde{\beta}, \widetilde{\sigma}^2)$ must satisfy the following conditions.

Assumption 3.3.1. *Let $\mathcal{F}_{i-1} = \sigma(z_1, \ldots, z_i, r_1, \ldots, r_{i-1}, u_1, \ldots, u_{i-1})$ be an increasing sequence of σ-fields so u_{i-1}, r_{i-1}, z_i are $\mathcal{F}_{i-1}$ measurable while r_i, u_i are independent of $\mathcal{F}_{i-1}$. Suppose $(u_i/\sigma, \Sigma^{-1/2}r_i)$ have continuously differentiable joint, conditional, and marginal densities $\mathsf{f}_{u,r}(y,x) = \mathsf{f}_{u|r}(y|x)\mathsf{f}_r(x) = \mathsf{f}_{r|u}(x|y)\mathsf{f}_u(y)$ which are positive on $y \in \mathbb{R}$, $x \in \mathbb{R}^{d_x}$. Assume $d_x \leq d_z$ and $\operatorname{rank}\Pi = d_x$. For $0 \leq \kappa < \eta \leq 1/4$, choose an integer $s \geq 2$ such that $2^{s-1} > 1 + (1/4 - \eta)(1 + d_x)$. Let $e = 1 + 2^{s+1}$. Suppose*

(i) the marginal density $\mathsf{f}_u(y)$ is symmetric and satisfies for $y \in \mathbb{R}$

(a) tail monotonicity: $y^e\mathsf{f}_u(y), |y^{e+1}\dot{\mathsf{f}}_u(y)|$ are decreasing for large y;

(ii) the marginal density $\mathsf{f}_r(x)$ satisfies for $x \in \mathbb{R}^{d_x}$

(a) moments: $\int_{x\in\mathbb{R}^{d_x}} |x|^4\mathsf{f}_r(x)(dx) < \infty$;

(iii) the joint and conditional densities $\mathsf{f}_{u,r}(y,x), \mathsf{f}_{u|r}(y|x)$ satisfy for $y \in \mathbb{R}, x \in \mathbb{R}^{d_x}$

(a) boundedness: $\sup_{y\in\mathbb{R},x\in\mathbb{R}^{d_x}} |(1+y)y^{e-2}\mathsf{f}_{u|r}(y|x) + y^{e-1}\dot{\mathsf{f}}_{u|r,y}(y|x)| < \infty$;

(iv) the instruments z_i satisfy

(a) $M_{zz,n} = \sum_{i=1}^n z_{in}z_{in}' \xrightarrow{\mathsf{P}} M_{zz} > 0$;

(b) $\max_{1\leq i\leq n} |n^{1/2-\kappa}z_{in}| = \mathrm{O}_{\mathsf{P}}(1)$;

(c) $n^{-1}\mathsf{E}\sum_{i=1}^n |n^{1/2}z_{in}|^e = \mathsf{E}|n^{1/2}z_{in}|^e = \mathrm{O}(1)$;

(v) the initial estimators $(\widetilde{\beta}, \widetilde{\sigma}^2)$ satisfy

(a) $n^{1/2}(\widetilde{\beta} - \beta) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$;

(b) $n^{1/2}(\widetilde{\sigma}^2 - \sigma^2) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$.

The conditions $(ia, iia, iiia)$ are satisfied in a range of situations. In particular (ia) is satisfied by the normal and t distribution; see Johansen and Nielsen (2016a, Example 3.1), while $(iia, iiia)$ holds for the normal distribution. Condition (iv) is standard for stationary instruments, see Johansen and Nielsen (2016a, Example 3.2). Condition (v) allows the standardised estimation errors to diverge at a rate of $n^{1/4-\eta}$ rather than being bounded in probability. In particular, $\eta = 1/4$ can be chosen for estimators with standard convergence rates.

Notice κ is not present in the moment condition $2^{s-1} > 1 + (1/4 - \eta)(1 + d_x)$, thus there is a trade-off between the convergence rate η of initial estimators and the required number s of moments for the density $\mathsf{f}_{u,r}$. In standard situations the normalized estimator is bounded so $\eta = 1/4$, then the moment condition to s reduces to $s = 2$. Otherwise when the initial estimator diverges at the rate η, then the number of moments

s grows linearly with the dimension of the regressors d_x[10].

3.3.2 Asymptotics of Algorithm 3.2.1

This subsection shows asymptotic theory for Algorithms 3.2.1. Before establishing asymptotic properties of the iterated estimators for β, σ^2, we first demonstrate that the updated estimator for Π is consistent uniformly in $c \in [c_+, \infty)$ for a finite number $c_+ > 0$, given tightness of previous estimators for β, σ^2.

Theorem 3.3.1. *Consider Algorithm 3.2.1. Suppose Assumption 3.3.1$(ia, iia, iiia, iv)$ holds, and that $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are $\mathrm{O}_{\mathsf{P}}(1)$ for any $m \in [0, \infty)$. Then as $n \to \infty$*

$$\sup_{c_+ \leq c < \infty} |\widehat{\Pi}_c^{(m+1)} - \Pi| = \mathrm{o}_{\mathsf{P}}(1).$$

The proof of the above theorem uses empirical process theory studied in §3.5. Then with uniform consistency of location estimator for Π in the first stage regression (3.2.2) we build a one-step stochastic expansion of the updated estimators for structural parameters β, σ^2 in (3.2.1) in terms of its original estimators, kernels, and small remainder terms.

Theorem 3.3.2. *Consider Algorithm 3.2.1. Suppose Assumption 3.3.1$(ia, iia, iiia, iv)$ holds, and that $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are $\mathrm{O}_{\mathsf{P}}(1)$ for any $m \in [0, \infty)$. Then as $n \to \infty$ and uniformly in $c \in [c_+, \infty)$*

$$\begin{aligned}
n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) &= \frac{2c\mathsf{f}_u(c)}{\psi} n^{1/2}(\widehat{\beta}_c^{(m)} - \beta) + (M_{\tilde{x}\tilde{x},n}\psi)^{-1} \sum_{i=1}^n \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} + \mathrm{o}_{\mathsf{P}}(1), \\
n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) &= \frac{c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)}{\tau_2^c} n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) + \frac{\sigma}{2\tau_2^c} n^{-1/2} \sum_{i=1}^n (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i| \leq \sigma c)} \\
&\quad + \frac{(c^2 - \varsigma_c^2)\mathsf{f}_u(c)\zeta_c^{-\prime}\Sigma^{1/2}}{2\tau_2^c} n^{1/2}(\widehat{\beta}_c^{(m)} - \beta) + \mathrm{o}_{\mathsf{P}}(1).
\end{aligned}$$

Remark 3.3.1. *The above theorem explores expansion of Algorithm 3.2.1 (Iterated 2SLS) in the IVs setting where $\mathsf{E}x_i u_i \neq 0_{d_x}$, which is the generalized result of iterated 1-step Huber-skip M-estimators in the classical regression; see Theorem 1 in Jiao and Nielsen (2017). To be specific, the expansion of β estimator only depends on its own estimation error in the previous step; see Proof of Theorem 3.3.2 (Appendix 3.9.2) why $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ does not occur, and moreover if $\tilde{x}_{in}$ is replaced by x_{in} then it is the same*

[10] also see discussion in Berenguer-Rico, Johansen, and Nielsen (2019a, Remark 3.1, 3.2).

as that in the classical regression. While the expansion of σ^2 estimator does also depend on the estimation error $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$, this term disappears in the classical regression where $\mathsf{E}x_i u_i = 0_{d_x}$ and $\mathsf{E}r_i u_i = 0_{d_x}$ such that $\xi_c = 0_{d_x}$, $\zeta_c^+ = \zeta_c^- = 0_{d_x}$. Thus, the second expansion degenerates to that in Jiao and Nielsen (2017).

Remark 3.3.2. *Under normality (3.2.5) for $(u_i/\sigma, \Sigma^{-1/2} r_i)$ it follows $\zeta_c^- = 2\Omega c$ then*

$$\frac{(c^2 - \varsigma_c^2)\mathsf{f}_u(c)\zeta_c^{-\prime}\Sigma^{1/2}}{2\tau_2^c} = \frac{c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)}{\tau_2^c}\Omega'\Sigma^{1/2},$$

so for the expansion of variance estimator the fraction above appearing in the coefficient of $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$ equals that of $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$. Moreover, observe two expansions and find their autoregressive coefficients $2c\mathsf{f}_u(c)/\psi$, $c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)/\tau_2^c$. By Johansen and Nielsen (2013, Theorem 3.5), also see Jiao and Nielsen (2017, Theorem 2), Assumption 3.3.1(ia) implies they are strictly bounded by one so

$$\sup_{c_+ \leq c < \infty} \max\{|\frac{2c\mathsf{f}_u(c)}{\psi}|, |\frac{c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)}{\tau_2^c}|\} < 1. \tag{3.3.1}$$

This in fact shows the spectral radius of the autoregressive coefficient matrix is smaller than one in the iterative system in Theorem 3.3.2, which is significant to establish the tightness and fixed point result, see the unit cycle boundedness condition (3.9.10) in the one-step stochastic expansion system (3.9.6), (3.9.7), (3.9.8) in Proof of Theorem 3.3.3 (Appendix 3.9.2).

Assumption 3.3.1(v) with $\eta = 1/4$ corresponds to a standard convergence rate for the initial estimator. Theorem 3.3.2 provides an iterative equation between the updated and original estimators, while its autoregressive coefficient has a spectral radius strictly bounded by the unit cycle, see (3.3.1) in Remark 3.3.2. Thus a geometric argument and mathematical induction are then used to show $\widehat{\beta}_c^{(m)}$, $\widehat{\sigma}_c^{(m)}$ are tight in iteration $m \in [0, \infty)$ and in the cut-off value $c \in [c_+, \infty)$.

Theorem 3.3.3. *Consider Algorithm 3.2.1. Suppose Assumption 3.3.1 holds with $\eta = 1/4$. Then as $n \to \infty$*

$$\sup_{0 \leq m < \infty} \sup_{c_+ \leq c < \infty} |n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)| + |n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)| = \mathrm{O}_{\mathsf{P}}(1).$$

The above tightness result is required for building the weak convergence theory of iterated estimators for β, σ^2. Further combined with Theorem 3.3.1 a corollary follows, to demonstrate uniform consistency of $\widehat{\Pi}_c^{(m)}$ in m and c.

Corollary 3.3.1. *Consider Algorithm 3.2.1. Suppose Assumption 3.3.1 holds with* $\eta = 1/4$. *Then as* $n \to \infty$

$$\sup_{0\le m<\infty}\sup_{c_+\le c<\infty}|\widehat{\Pi}_c^{(m)} - \Pi| = \mathrm{o_P}(1).$$

Since tightness results for iterated estimators of β, σ^2 have been established by Theorem 3.3.3, we can apply the one-step expansion in Theorem 3.3.2 recursively. Then for any $m \in [0,\infty)$ the stochastic expansions of $m+1$ step estimators are explored in terms of the initial estimators, kernels, and small remainder terms.

Theorem 3.3.4. *Consider Algorithm 3.2.1. Suppose Assumption 3.3.1 holds with* $\eta = 1/4$. *Then as* $n \to \infty$ *and uniformly in* $c \in [c_+,\infty)$ *we have for any* $m \in [0,\infty)$

$$\begin{aligned}
n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) &= \varrho_{\beta\beta,c}^{(m+1)} n^{1/2}(\widehat{\beta}_c^{(0)} - \beta) + \varrho_{\beta\tilde{x}u,c}^{(m+1)} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^n \tilde{x}_{in} u_i 1_{(|u_i|\le\sigma c)} + \mathrm{o_P}(1),\\
n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) &= \varrho_{\sigma\sigma,c}^{(m+1)} n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) + \varrho_{\sigma uu,c}^{(m+1)} n^{-1/2} \sum_{i=1}^n (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i|\le\sigma c)}\\
&\quad + \varrho_{\sigma\beta,c}^{(m+1)\prime} n^{1/2}(\widehat{\beta}_c^{(0)} - \beta) + \varrho_{\sigma\tilde{x}u,c}^{(m+1)\prime} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^n \tilde{x}_{in} u_i 1_{(|u_i|\le\sigma c)} + \mathrm{o_P}(1),
\end{aligned}$$

where coefficients have expressions

$$\begin{aligned}
\varrho_{\beta\beta,c}^{(m+1)} &= \{\frac{2c\mathsf{f}_u(c)}{\psi}\}^{m+1}, \qquad \varrho_{\beta\tilde{x}u,c}^{(m+1)} = \frac{\psi^{m+1} - \{2c\mathsf{f}_u(c)\}^{m+1}}{\psi^{m+1}\{\psi - 2c\mathsf{f}_u(c)\}},\\
\varrho_{\sigma\sigma,c}^{(m+1)} &= \{\frac{c(c^2-\varsigma_c^2)\mathsf{f}_u(c)}{\tau_2^c}\}^{m+1}, \qquad \varrho_{\sigma uu,c}^{(m+1)} = \sigma\frac{(\tau_2^c)^{m+1} - \{c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}^{m+1}}{2(\tau_2^c)^{m+1}\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}},\\
\varrho_{\sigma\beta,c}^{(m+1)} &= \sum_{l=0}^m (\frac{2}{\psi})^{m-l}(\frac{c^2-\varsigma_c^2}{\tau_2^c})^{l+1}\frac{\{c\mathsf{f}_u(c)\}^{m+1}\Sigma^{1/2}\zeta_c^-}{2c},\\
\varrho_{\sigma\tilde{x}u,c}^{(m+1)} &= [\frac{(\tau_2^c)^{m+1} - \{c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}^{m+1}}{(\tau_2^c)^m\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}} - \sum_{l=0}^m (\frac{2}{\psi})^{m-l}(\frac{c^2-\varsigma_c^2}{\tau_2^c})^l\{c\mathsf{f}_u(c)\}^m]\\
&\quad \times \frac{(c^2-\varsigma_c^2)\mathsf{f}_u(c)\Sigma^{1/2}\zeta_c^-}{2\tau_2^c\{\psi - 2c\mathsf{f}_u(c)\}}.
\end{aligned}$$

For $m \in [0,\infty)$ the above expansion of $m+1$ step estimators generalizes the one-step expansion in Theorem 3.3.2. If $m = 0$ in Theorem 3.3.4, then we have

$$\begin{aligned}
&\varrho_{\beta\beta,c}^{(1)} = \frac{2c\mathsf{f}_u(c)}{\psi}, \qquad \varrho_{\beta\tilde{x}u,c}^{(1)} = \frac{1}{\psi}, \qquad (3.3.2)\\
&\varrho_{\sigma\sigma,c}^{(1)} = \frac{c(c^2-\varsigma_c^2)\mathsf{f}_u(c)}{\tau_2^c}, \quad \varrho_{\sigma uu,c}^{(1)} = \frac{\sigma}{2\tau_2^c}, \quad \varrho_{\sigma\beta,c}^{(1)} = \frac{(c^2-\varsigma_c^2)\mathsf{f}_u(c)\Sigma^{1/2}\zeta_c^-}{2\tau_2^c}, \quad \varrho_{\sigma\tilde{x}u,c}^{(1)} = 0_{d_x},
\end{aligned}$$

so the expansion reduces to the one-step case shown in Theorem 3.3.2.

Initially the tight estimator is assumed to be available. This is iterated through the one-step equation presented in Theorem 3.3.2 with the spectral radius of its autoregressive coefficient bounded by one, see (3.3.1) in Remark 3.3.2, so when the iteration step becomes sufficiently large the iterated estimator converges in probability to the fixed point. To explore convergence in iteration, let $m \to \infty$ in Theorem 3.3.4 to get

$$\begin{aligned} \varrho^{(\infty)}_{\beta\beta,c} &= 0, & \varrho^{(\infty)}_{\beta\tilde{x}u,c} &= \frac{1}{\psi - 2c\mathsf{f}_u(c)}, \\ \varrho^{(\infty)}_{\sigma\sigma,c} &= 0, & \varrho^{(\infty)}_{\sigma uu,c} &= \frac{\sigma}{2\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}}, \\ \varrho^{(\infty)}_{\sigma\beta,c} &= 0_{d_x}, & \varrho^{(\infty)}_{\sigma\tilde{x}u,c} &= \frac{(c^2-\varsigma_c^2)\mathsf{f}_u(c)\Sigma^{1/2}\zeta_c^-}{2\{\psi - 2c\mathsf{f}_u(c)\}\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}}, \end{aligned} \tag{3.3.3}$$

then the fixed point $n^{1/2}(\widehat{\beta}_c^{(*)} - \beta) = n^{1/2}(\widehat{\beta}_c^{(\infty)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(*)} - \sigma) = n^{1/2}(\widehat{\sigma}_c^{(\infty)} - \sigma)$ follows.

Theorem 3.3.5. *Consider Algorithm 3.2.1. Suppose Assumption 3.3.1 holds with $\eta = 1/4$. Then for all $\epsilon, \delta > 0$ a pair $n_0 > 0, m_0 > 0$ exists, so for $n > n_0$ and $m > m_0$*

$$\mathsf{P}\{\sup_{c_+ \le c < \infty} |n^{1/2}(\widehat{\beta}_c^{(m)} - \widehat{\beta}_c^{(*)})| + |n^{1/2}(\widehat{\sigma}_c^{(m)} - \widehat{\sigma}_c^{(*)})| > \delta\} < \epsilon,$$

where

$$\begin{aligned} n^{1/2}(\widehat{\beta}_c^{(*)} - \beta) &= \frac{1}{\psi - 2c\mathsf{f}_u(c)} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^n \tilde{x}_{in} u_i 1_{(|u_i| \le \sigma c)}, \\ n^{1/2}(\widehat{\sigma}_c^{(*)} - \sigma) &= \frac{\sigma}{2\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}} n^{-1/2} \sum_{i=1}^n (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i| \le \sigma c)} \\ &\quad + \frac{(c^2-\varsigma_c^2)\mathsf{f}_u(c)\zeta_c^{-\prime}\Sigma^{1/2}}{2\{\psi - 2c\mathsf{f}_u(c)\}\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^n \tilde{x}_{in} u_i 1_{(|u_i| \le \sigma c)}. \end{aligned}$$

Based on Theorem 3.3.3, if the initial estimator is bounded in a large compact set with large probability, then any iterated estimator takes values in the same compact set. The proof of Theorem 3.3.5 is to further argue that the deviation between the m-fold iterated estimator and the fixed point is the sum of two terms vanishing exponentially and in probability respectively when m and n tend to infinity.

Algorithm 3.2.1 mimics the Huber (1964) skip estimator in the IVs context. By investigating Theorem 3.3.5 we reason that through infinite iterations the fixed point of the algorithm approximates the Huber-skip IVs estimator in the sense that they have the same asymptotic expansion in terms of kernels and thus follow the same limiting distribution (see Theorem 3.3.8).

Similar to Remark 3.3.1, Theorems 3.3.4 and 3.3.5 extend the $(m+1)$-fold expansion and the fixed point result of the iterated 1-step Huber-skip M-estimator in classical regression to the IVs setting where $\mathsf{E}x_iu_i \neq 0_{d_x}$ is allowed[11]. In practice we assume $(u_i/\sigma, \Sigma^{-1/2}r_i)$ follows normality (3.2.5) then $\xi_c = \Omega c$, $\zeta_c^+ = 0_{d_x}$, $\zeta_c^- = 2\Omega c$, so coefficients appearing in expansions in Theorems 3.3.4 and 3.3.5 can be further reduced as in Remark 3.3.2.

3.3.3 Weak convergence of Algorithm 3.2.2 and 3.2.3

Algorithms 3.2.2 and 3.2.3 are special versions of Algorithm 3.2.1 with different starting points. Their initial estimators are either the full sample or split sample two stage least squares which do not depend on the cut-off, and so satisfy the tightness property. Therefore, theorems and corollaries in the previous subsection apply for these two algorithms as well. Moreover, since Algorithms 3.2.2 and 3.2.3 start with two stage least squares estimators whose statistical properties are well known, the asymptotic distribution can then be established for these two robust procedures.

When researchers carry out empirical work using instrumental variables regression, their interest mainly lies in making inference on β whereas they are only concerned about the consistency of estimators for σ^2, Π. Theorem 3.3.3 and Corollary 3.3.1 have already shown that iterated estimators of σ^2, Π are consistent, so in order to perform inference on structural parameters in (3.2.1) we next build distributional theory for the iterated estimator of β in Algorithms 3.2.2 and 3.2.3.

Choosing the distinct cut-off value c in the interval $[c_+, \infty)$, we then get a process $\mathbb{G}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta)$ for any $m \in [0, \infty)$. A weak convergence theory for $\mathbb{G}_n^{(m+1)}$ follows from a finite dimensional convergence derived by the expansion in Theorem 3.3.4 and tightness in Theorem 3.3.3 (see Theorem 13.1 in Billingsley, 1968). The iterated estimator of β as a process is asymptotically approximated by the Gaussian process[12].

We first analyze Algorithm 3.2.2 by providing asymptotics for the full sample two stage least squares estimator $\widetilde{\beta}$ defined through its initial estimate $\widehat{\beta}_c^{(0)}$ in (3.2.14).

Lemma 3.3.1. *Consider the full sample two stage least squares*

$$\widetilde{\beta} = (\widetilde{\Pi}' \sum_{i=1}^n z_i z_i' \widetilde{\Pi})^{-1} (\widetilde{\Pi}' \sum_{i=1}^n z_i y_i), \quad \textit{where} \quad \widetilde{\Pi} = (\sum_{i=1}^n z_i z_i')^{-1} (\sum_{i=1}^n z_i x_i').$$

[11] replace $\tilde{x}_{in}$ by x_{in} and set $\mathsf{E}x_iu_i = 0_{d_x}$ so $\mathsf{E}r_iu_i = 0_{d_x}$, we then have $\xi_c = 0_{d_x}$, $\zeta_c^+ = \zeta_c^- = 0_{d_x}$ thus all results degenerate to those shown in Jiao and Nielsen (2017).

[12] see the definition of Gaussian processes in Adler and Taylor (2009, p. 27).

Suppose Assumption 3.3.1(ia, iia, iva, ivc) holds. Then as $n \to \infty$

$$n^{1/2}(\widetilde{\beta} - \beta) = M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + \mathsf{o}_{\mathsf{P}}(1).$$

Furthermore, we have

$$n^{1/2}(\widetilde{\beta} - \beta) \xrightarrow{\mathsf{D}} \mathsf{N}(0_{d_x}, \sigma^2 M_{\tilde{x}\tilde{x}}^{-1}).$$

The next step is to establish the weak convergence theory for the process $\mathbb{G}_n^{(m+1)}$ of the iterated β estimator where $m \in [0, \infty)$.

Theorem 3.3.6. *Consider Algorithm 3.2.2. Suppose Assumption 3.3.1(ia, iia, iiia, iv) holds. Denote the process* $\mathbb{G}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta)$ *for* $c \in [c_+, \infty)$ *and* $m \in [0, \infty)$. *Then as* $n \to \infty$ *we have*

$$\mathbb{G}_n^{(m+1)}(c) = \varrho_{\beta\beta,c}^{(m+1)} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + \varrho_{\beta\tilde{x}u,c}^{(m+1)} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \le \sigma c)} + \mathsf{o}_{\mathsf{P}}(1),$$

where $\varrho_{\beta\beta,c}^{(m+1)}$, $\varrho_{\beta\tilde{x}u,c}^{(m+1)}$ *are defined in Theorem 3.3.4. Furthermore,* $\mathbb{G}_n^{(m+1)}$ *weakly converges to a zero mean Gaussian process* $\mathbb{G}^{(m+1)}$ *with variance given as*

$$\mathsf{Var}\{\mathbb{G}^{(m+1)}(c)\} = \{(\varrho_{\beta\beta,c}^{(m+1)})^2 + 2\tau_2^c \varrho_{\beta\beta,c}^{(m+1)} \varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c (\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\} \sigma^2 M_{\tilde{x}\tilde{x}}^{-1}.$$

The one-step updated estimator from the full sample 2SLS is of particular interest in Algorithm 3.2.2. To explore its asymptotics let $m = 0$ in Theorem 3.3.6 such that $\varrho_{\beta\beta,c}^{(1)} = 2c\mathsf{f}_u(c)/\psi$ and $\varrho_{\beta\tilde{x}u,c}^{(1)} = \psi^{-1}$ as in (3.3.2), then the corollary below follows.

Corollary 3.3.2. *Consider Algorithm 3.2.2. Suppose Assumption 3.3.1(ia, iia, iiia, iv) holds. Denote the process* $\mathbb{G}_n^{(1)}(c) = n^{1/2}(\widehat{\beta}_c^{(1)} - \beta)$ *for* $c \in [c_+, \infty)$. *Then as* $n \to \infty$ *we have*

$$\mathbb{G}_n^{(1)}(c) = \frac{2c\mathsf{f}_u(c)}{\psi} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + (M_{\tilde{x}\tilde{x},n}\psi)^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \le \sigma c)} + \mathsf{o}_{\mathsf{P}}(1).$$

Furthermore, $\mathbb{G}_n^{(1)}$ *weakly converges to a zero mean Gaussian process* $\mathbb{G}^{(1)}$ *with variance given as*

$$\mathsf{Var}\{\mathbb{G}^{(1)}(c)\} = \frac{4c^2\mathsf{f}_u^2(c) + 4\tau_2^c c\mathsf{f}_u(c) + \tau_2^c}{\psi^2} \sigma^2 M_{\tilde{x}\tilde{x}}^{-1}.$$

Then we analyze Algorithm 3.2.3 and find asymptotically it is equivalent to Algorithm 3.2.2 although they start with distinct initial estimates when implementing Algorithm 3.2.1.

Theorem 3.3.7. *Consider Algorithm 3.2.3 where $n_1 = n_2 = n/2$. Suppose Assumption 3.3.1$(ia, iia, iiia, iv)$ holds for each sub-sample set $\mathcal{I}_1$, $\mathcal{I}_2$. Then as $n \to \infty$ and for any $m \in [0, \infty)$, the process of the estimator $n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta)$ for $c \in [c_+, \infty)$ has the same asymptotic expansion as for Algorithm 3.2.2 and thus weakly converges to the identical Gaussian process reported in Theorem 3.3.6.*

The proof involves checking whether the expansion of the first step updated estimator for Algorithm 3.2.3 is the same as Algorithm 3.2.2 using arguments in Theorems 3.3.1 and 3.3.2. Since the two algorithms are iterated from the split half or full sample 2SLS respectively, they have the same asymptotics if their $\widehat{\beta}_c^{(1)}$ perform identically when $n \to \infty$.

Finally, when $n \to \infty$ and $m \to \infty$, the process $\mathbb{G}_n^{(m)}$ uniformly converges to the limiting process $\mathbb{G}^{(\infty)} = \mathbb{G}^{(*)}$ in probability. Set $\varrho_{\beta\beta,c}^{(\infty)} = 0$, $\varrho_{\beta\tilde{x}u,c}^{(\infty)} = 1/\{\psi - 2c\mathsf{f}_u(c)\}$ as in (3.3.3), then Theorems 3.3.6 and 3.3.7 immediately show the weak convergence of the fixed point process $\mathbb{G}^{(*)}$ for Algorithms 3.2.1, 3.2.2, and 3.2.3.

Theorem 3.3.8. *Consider Algorithm 3.2.1 with tight initial estimators, such as Algorithm 3.2.2, 3.2.3. Suppose Assumption 3.3.1$(ia, iia, iiia, iv)$ holds. When $m \to \infty$ let the fixed point process $\mathbb{G}_n^{(*)}(c) = n^{1/2}(\widehat{\beta}_c^{(*)} - \beta)$ for $c \in [c_+, \infty)$ as in Theorem 3.3.5. Then as $n \to \infty$ the process $\mathbb{G}_n^{(*)}$ has the asymptotic expansion shown in Theorem 3.3.5 and thus weakly converges to a zero mean Gaussian process $\mathbb{G}^{(*)}$ with variance given as*

$$\mathsf{Var}\{\mathbb{G}^{(*)}(c)\} = \frac{\tau_2^c}{\{\psi - 2c\mathsf{f}_u(c)\}^2}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}.$$

Consider Algorithms 3.2.2 and 3.2.3. Weak convergence in Theorems 3.3.6 and 3.3.7 implies pointwise convergence, so for any $c \in [c_+, \infty)$, $m \in [0, \infty)$, and as $n \to \infty$ it holds

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) \xrightarrow{\mathsf{D}} \mathsf{N}[0_{d_x}, \{(\varrho_{\beta\beta,c}^{(m+1)})^2 + 2\tau_2^c \varrho_{\beta\beta,c}^{(m+1)} \varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c (\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}], \quad (3.3.4)$$

where $\varrho_{\beta\beta,c}^{(m+1)}$, $\varrho_{\beta\tilde{x}u,c}^{(m+1)}$ are defined in Theorem 3.3.4.

Johansen and Nielsen (2009, 2013) built up the asymptotic distributions for the one-step Huber-skip M-estimator and its infinite iteration either starting from full sample or split sample least squares in the classical setting where $\mathsf{E}x_i u_i = 0_{d_x}$[13]. The distributional results for Algorithms 3.2.2 and 3.2.3 are suprisingly the same as the iterated one-step

[13]also see Johansen and Nielsen (2016b) for the asymptotic distribution of the $(m+1)$-step estimator where $m \in [0, \infty)$.

Huber-skip M-estimators if $\tilde{x}_{in}$ is replaced by x_{in} even though dependence is allowed between regressors and errors in the IVs regression. This is due to the fact that the stochastic expansion for the updated β estimator only depends on its own previous step, not on the σ estimator (see Theorem 3.3.2 and Remark 3.3.1). Furthermore, this result is because ζ_c^+ defined in (3.2.6, §3.2.1) does not appear in the β expansion since the data is demeaned such that $\mathsf{E}z_i = 0_{d_z}$[14].

3.3.4 Efficiency comparison

Outlier analysis is trade-off between efficiency and robustness: Algorithms 3.2.2 and 3.2.3 will lose efficiency under the null of no outliers relative to the non-robust ordinary 2SLS since it is with positive probability to wrongly detect outliers, whereas they are indeed more robust under the alternative when there is data contamination. This subsection mainly investigates efficiency loss of two algorithms in return for gaining robustness.

Distributions of $\widehat{\beta}_c^{(m+1)}$ are given by (3.3.4) while Lemma 3.3.1 shows the limit of the ordinary two stage least squares estimator $\widetilde{\beta}$. Thus, we can now compare the efficiency of the robust estimator $\widehat{\beta}_c^{(m+1)}$ with respect to the non-robust estimator $\widetilde{\beta}$ under the null of no outliers. Relative efficiency is defined by

$$\begin{aligned}\vartheta_c^{(m+1)} &= \text{efficiency}(\widehat{\beta}_c^{(m+1)}, \widetilde{\beta}) = \{\mathsf{avar}(\widehat{\beta}_c^{(m+1)})\}^{-1}\{\mathsf{avar}(\widetilde{\beta})\} \\ &= \{(\varrho_{\beta\beta,c}^{(m+1)})^2 + 2\tau_2^c\varrho_{\beta\beta,c}^{(m+1)}\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}^{-1} I_{d_x}. \qquad (3.3.5)\end{aligned}$$

In the rest of the paper we treat $\vartheta_c^{(m+1)}$ as a scalar factor though it is a scaled identity matrix. We are interested in efficiency comparison for two special cases: when $m = 0$ and $m \to \infty$. Substitute (3.3.2) and (3.3.3) into (3.3.5) to obtain

$$\vartheta_c^{(1)} = \text{efficiency}(\widehat{\beta}_c^{(1)}, \widetilde{\beta}) = \frac{\psi^2}{4c^2\mathsf{f}_u^2(c) + 4\tau_2^c c\mathsf{f}_u(c) + \tau_2^c} I_{d_x}, \qquad (3.3.6)$$

$$\vartheta_c^{(*)} = \text{efficiency}(\widehat{\beta}_c^{(*)}, \widetilde{\beta}) = \frac{\{\psi - 2c\mathsf{f}_u(c)\}^2}{\tau_2^c} I_{d_x}. \qquad (3.3.7)$$

Using the standard normal density for f_u, we then plot (3.3.6), (3.3.7) in terms of critical values $c \in [c_+, \infty)$ in Figure 3.3.1 and plot (3.3.5) in terms of iteration steps $m \in [0, \infty)$ for the cut-offs $c_1 = 1.645(10\%)$, $c_2 = 1.96(5\%)$, $c_3 = 2.576(1\%)$ in Figure 3.3.2.

[14] see the details in Proof of Theorem 3.3.2, Appendix 3.9.2.

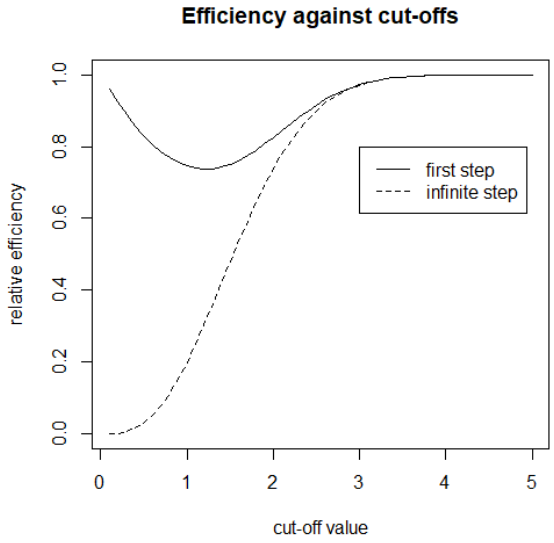


Figure 3.3.1: The efficiency of $\widehat{\beta}_c^{(1)}$ (solid line) and $\widehat{\beta}_c^{(*)}$ (dashed line) with respect to the two stage least squares $\widetilde{\beta}$ for f_u equal to the standard normal density and $c \in [0.1, 5]$.

Asymptotics in this chapter are derived under the null hypothesis of no outliers where Algorithms 3.2.2 and 3.2.3 have a probability to falsely trim the non-outlying observations given the cut-off c. Thus robust estimates have a higher asymptotic variance relative to the ordinary 2SLS, and so the relative efficiency plots lie in the interval $(0, 1)$ as in Figure 3.3.1 and 3.3.2.

When critical values c become larger, the chance of wrongly detecting outliers tends to be smaller. Subsequently, robust estimates will have the smaller variance, and their efficiency become larger and closer to one, and vice versa if c decreases. First observe the efficiency plot of the infinite step estimator $\widehat{\beta}_c^{(*)}$ in Figure 3.3.1, and find it is consistent with what we discussed. Then check the solid line of the first step estimator $\widehat{\beta}_c^{(1)}$, and find that its efficiency tends to one when $c \to 0$ as well as when $c \to \infty$. It is intuitive that for large c the efficiency increases as c increases. However we also find for small c the efficiency increases as c decreases. This is due to the fact that the updated estimator $\widehat{\beta}_c^{(1)}$ is asymptotically equivalent to 2SLS when $c \to 0$ as stated in Remark 3.3.3.

Remark 3.3.3. *Let $m = 0$ and $n \to \infty$, Theorem 3.3.4 gives the equation*

$$n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = \frac{2c\mathsf{f}_u(c)}{\psi} n^{1/2}(\widetilde{\beta} - \beta) + (M_{\tilde{x}\tilde{x},n}\psi)^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \le \sigma c)} + \mathsf{o}_{\mathsf{P}}(1),$$

uniformly in $c \in [c_+, \infty)$. Note that $\widetilde{\beta}$ is 2SLS so $\widehat{\beta}_c^{(1)}$ is its updated estimator. Suppose f_u follows the standard normal then $\psi = 2\mathsf{F}_u(c) - 1$, $\tau_2^c = \psi - 2c\mathsf{f}_u(c)$, and $\dot{\mathsf{f}}_u(c) = -c\mathsf{f}_u(c)$.

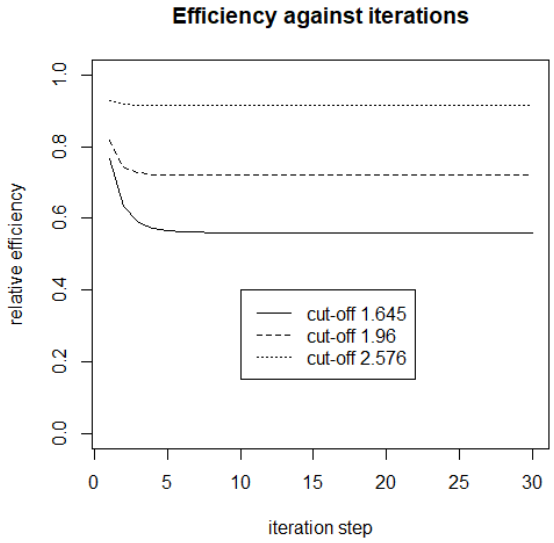


Figure 3.3.2: The efficiency of $\widehat{\beta}_c^{(m+1)}$ with $m \in [0, 29)$ relative to the two stage least squares $\widetilde{\beta}$ for f_u equal to the standard normal density and $c_1 = 1.645$ (solid line), $c_2 = 1.96$ (dashed line), $c_3 = 2.576$ (dotted line).

Let $c_+ > 0$ be sufficiently small so that we can investigate the situation where $c \to 0$. Lemma 3.9.6 (Appendix 3.9.3) shows the weak limit of the kernel term such that

$$(M_{\tilde{x}\tilde{x},n}\psi)^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \le \sigma c)} \xrightarrow{\mathsf{D}} \mathsf{N}(0_{d_x}, \frac{\tau_2^c}{\psi^2} \sigma^2 M_{\tilde{x}\tilde{x}}^{-1}).$$

By L'Hôpital's rule, we have $\tau_2^c/\psi^2 \to 0$ as $c \to 0$ so the kernel term vanishes. Again when $c \to 0$, L'Hôpital's rule gives $2c\mathsf{f}_u(c)/\psi \to 1$, thus $n^{1/2}(\widehat{\beta}_c^{(1)} - \beta)$ is asymptotically equivalent to $n^{1/2}(\widetilde{\beta} - \beta)$ indicating an efficiency of unity.

In fact, check the efficiency formula given by (3.3.6). When $c \to 0$, use $\tau_2^c/\psi^2 \to 0$ and $2c\mathsf{f}_u(c)/\psi \to 1$ to get $\psi^2/\{4c^2\mathsf{f}_u^2(c) + 4\tau_2^c c\mathsf{f}_u(c) + \tau_2^c\} \to 1$, we can then find efficiency$(\widehat{\beta}_c^{(1)}, \widetilde{\beta}) \to 1$.

Figure 3.3.1 also indicatess the efficiency of the first iteration estimate dominates the infinite step. This fact is consistent with Figure 3.3.2 showing in general that the efficiency decreases as iteration increases regardless of cut-off values c. Figure 3.3.2 reveals the efficiency drops dramatically within the first few steps and converges to the fixed point upon through large iterations (also demonstrated by Theorem 3.3.8). Intuition is that as iteration increases algorithms will throw away more observations falsely classified as outliers so loss more efficiency until attaining the fixed point. It is

not surprising that the critical value $c_3 = 2.576(1\%)$ has the highest relative efficiency and the $c_1 = 1.645(10\%)$ has the lowest while $c_2 = 1.96(5\%)$ is in between.

Figure 3.3.1 and 3.3.2 imply Algorithms 3.2.2 and 3.2.3 do not lose too much efficiency compared to the ordinary 2SLS if the cut-off value c and iteration step m are appropriately chosen under the null of no outliers, whereas these two procedures will provide the more robust result under the alternative when there is in fact presence of outliers.

3.3.5 Valid inference on structural parameters β

Assume that the density f_u is known, say standard normal $\mathsf{N}(0,1)$, and choose the cut-off value c and iteration step m, say $c = 1.96(5\%)$ and $m = 0$ or ∞, when carrying out Algorithms 3.2.2 and 3.2.3. The standard method of conducting inference based on ordinary 2SLS asymptotics[15] are not reliable for three reasons. Firstly, variance σ^2 cannot be estimated consistently without bias correction factor ς_c^2 introduced in (3.2.8, §3.2.1)[16]. Secondly, the standard method fails to take into account of the relative efficiency factor $\vartheta_c^{(m+1)}$ appearing in the asymptotic variance of $\widehat{\beta}_c^{(m+1)}$, see (3.3.4) and (3.3.5). Thirdly, to obtain the standard error of $\widehat{\beta}_c^{(m+1)}$ the usual inferential procedure incorrectly divides the estimated asymptotic variance by the sub-sample size $\sum_{i=1}^n \upsilon_{i,c}^{(m)}$ of all non-outlying observations retained, while the full sample size n should have been used.

On implementing Algorithms 3.2.2 and 3.2.3 to perform valid inference on the structural parameters β given the cut-off value c and iteration step m, we have to assume the explicit form of f_u in order to obtain the bias correction factor ς_c^2 and relative efficiency factor $\vartheta_c^{(m+1)}$. In practice, it is statistically difficult to estimate the density f_u. This is because, without knowing ς_c^2 as f_u is unknown, the standard estimate of σ^2 will be biased downward, which will subsequently produce upward biased standardised residuals for structural errors u_i although residuals can be estimated consistently based on beta estimates. Then the estimated density of f_u using the biased standardised residuals will be inconsistent by whatever methods. Thus, assumption on the known density f_u is not avoidable to conduct valid inference on β in this chapter. Proving validity of bootsrap Kaji (2018) applied nonparametric bootstrap to obtain standard errors of beta estimates when conducting inference, which however does not require the assumed known form of

[15] see ordinary 2SLS asymptotics in Lemma 3.3.1.

[16] see the details in Remark 3.9.3 why ς_c^2 is required to estimate σ^2 consistently.

the density f_u.

Weak limits of beta estimates in the step $m+1$ are provided by (3.3.4) such that

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) \xrightarrow{\mathsf{D}} \mathsf{N}\{0_{d_x}, (\vartheta_c^{(m+1)})^{-1}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}\},$$

where $\vartheta_c^{(m+1)}$ is defined by (3.3.5) and the cut-off c is given. By assuming on the form of f_u, $\vartheta_c^{(m+1)}$ is known as well as ς_c^2 so σ^2 can be estimated consistently by $(\widehat{\sigma}_c^{(m+1)})^2$, see Theorem 3.3.3. Corollary 3.3.1 implies we can consistently estimate $M_{\tilde{x}\tilde{x}} = \Pi'\mathsf{E}z_i z_i'\Pi$ by

$$\widehat{M}_{\tilde{x}\tilde{x},c}^{(m+1)} = (\sum_{i=1}^{n} v_{i,c}^{(m)})^{-1}(\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^{n} z_i z_i' v_{i,c}^{(m)}\widehat{\Pi}_c^{(m+1)}). \tag{3.3.8}$$

Notice that σ^2 and $M_{\tilde{x}\tilde{x}}$ above are estimated using the robust estimator with the subsample after trimming all outliers detected, since the estimator of standard errors should be robust itself, otherwise the inference on β might lead to incorrect results under the alternative hypothesis where data is contaminated. Then, the beta estimates are asymptotically approximated by

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) \overset{a}{\sim} \mathsf{N}\{0_{d_x}, (\vartheta_c^{(m+1)})^{-1}(\widehat{\sigma}_c^{(m+1)})^2(\widehat{M}_{\tilde{x}\tilde{x},c}^{(m+1)})^{-1}\}.$$

For conducting valid inference, the above estimated asymptotic variance can then be further divided by n to obtain standard errors so that

$$\widehat{\beta}_c^{(m+1)} \overset{a}{\sim} \mathsf{N}\{\beta, n^{-1}(\vartheta_c^{(m+1)})^{-1}(\widehat{\sigma}_c^{(m+1)})^2(\widehat{M}_{\tilde{x}\tilde{x},c}^{(m+1)})^{-1}\}.$$

The standard inferential procedure fails to take into consideration of $(\vartheta_c^{(m+1)})^{-1}$, so mistakenly applies usual 2SLS asymptotics to get the asymptotic variance of $\widehat{\beta}_c^{(m+1)}$ by $\text{avar}_{\text{2SLS}} = \sigma^2 M_{\tilde{x}\tilde{x}}^{-1}$. Thus, to address this problem the relative efficiency factor $\vartheta_c^{(m+1)}$ should be adjusted to $\text{avar}_{\text{2SLS}}$, then the normal weak limit is rearranged as

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) \xrightarrow{\mathsf{D}} \mathsf{N}\{0_{d_x}, (\vartheta_c^{(m+1)})^{-1}\text{avar}_{\text{2SLS}}\}.$$

Furthermore, without considering the consistency factor ς_c^{-2} by the standard method σ^2 is inconsistently estimated by $\varsigma_c^2(\widehat{\sigma}_c^{(m+1)})^2$. Then asymptotic variance of $\widehat{\beta}_c^{(m+1)}$ is falsely estimated by $\widehat{\text{avar}}_{\text{2SLS},c}^{(m+1)} = \varsigma_c^2(\widehat{\sigma}_c^{(m+1)})^2(\widehat{M}_{\tilde{x}\tilde{x},c}^{(m+1)})^{-1}$, which should be rectified by $\iota_c^{(m+1)} = (\vartheta_c^{(m+1)}\varsigma_c^2)^{-1}$ so as to fix the inferential problem. Therefore, the correct asymptotic variance is the term $\widehat{\text{avar}}_{\text{2SLS},c}^{(m+1)}$ estimated by usual asymptotics but adjusted by $\iota_c^{(m+1)}$, then error-free normal approximation can now be expressed as

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) \overset{a}{\sim} \mathsf{N}(0_{d_x}, \iota_c^{(m+1)}\widehat{\text{avar}}_{\text{2SLS},c}^{(m+1)}).$$

In addition, the standard method incorrectly estimates the variance of $\widehat{\beta}_c^{(m+1)}$ by $\widehat{\text{var}}_{2\text{SLS},c}^{(m+1)}$ through dividing $\widehat{\text{avar}}_{2\text{SLS},c}^{(m+1)}$ by the sub-sample size $\sum_{i=1}^n v_{i,c}^{(m)}$ of non-outlying observations rather than the full sample size n. Thus, the erroneous standard errors are computed as the square root of $\widehat{\text{var}}_{2\text{SLS},c}^{(m+1)} = \varsigma_c^2(\widehat{\sigma}_c^{(m+1)})^2(\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_i z_i' v_{i,c}^{(m)}\widehat{\Pi}_c^{(m+1)})^{-1}$. Therefore, correct standard errors should be based on $\widehat{\text{var}}_{2\text{SLS},c}^{(m+1)}$ by usual 2SLS asymptotics, however rectified by $n^{-1}\sum_{i=1}^n v_{i,c}^{(m)}$ and $\iota_c^{(m+1)}$. In summary, the valid inference should rely on the following rearranged normal approximation

$$\widehat{\beta}_c^{(m+1)} \overset{a}{\sim} \mathsf{N}\{\beta, (n^{-1}\sum_{i=1}^n v_{i,c}^{(m)})\iota_c^{(m+1)}\widehat{\text{var}}_{2\text{SLS},c}^{(m+1)}\}.$$

Consider the standard normal $\mathsf{N}(0,1)$ for f_u. Choose the cut-off value c as 1.645(10%), 1.96(5%), 2.576(1%) and set the iteration step m as 0, ∞. Notice that $n^{-1}\sum_{i=1}^n v_{i,c}^{(m)}$ is the proportion of non-outlying observations retained, which has the probability limit ψ under the null of no outliers due to Lemma 3.9.2 and LLN. List in Table 3.3.1 the probability ψ within the cut-off, the bias correction factor ς_c^2 for estimating σ^2, the relative efficiency factor $\vartheta_c^{(m+1)}$ for asymptotic variance of $\widehat{\beta}_c^{(m+1)}$, the adjusting factor $\iota_c^{(m+1)}$ for estimating asymptotic variance, and the adjusting factor $\psi\iota_c^{(m+1)}$ for estimating variance of $\widehat{\beta}_c^{(m+1)}$ under the null of no outliers for large samples.

c	ψ	ς_c^2	$\vartheta_c^{(1)}$	$\vartheta_c^{(*)}$	$\iota_c^{(1)}$	$\iota_c^{(*)}$	$\psi\iota_c^{(1)}$	$\psi\iota_c^{(*)}$
1.645	0.900	0.623	0.767	0.561	2.093	2.862	1.884	2.576
1.960	0.950	0.759	0.818	0.721	1.612	1.828	1.531	1.737
2.576	0.990	0.925	0.927	0.916	1.167	1.181	1.155	1.169

Table 3.3.1: Set f_u equal to the standard normal density and the cut-offs c as 1.645, 1.96, 2.576. The probability ψ of the absolute value of u_i/σ within the cut-off c, the bias correction factor ς_c^2 for estimating σ^2, the relative efficiency factor $\vartheta_c^{(1)}$, $\vartheta_c^{(*)}$ for asymptotic variance of $\widehat{\beta}_c^{(1)}$, $\widehat{\beta}_c^{(*)}$, the adjusting factor $\iota_c^{(1)}$, $\iota_c^{(*)}$ for estimating asymptotic variance, and the adjusting factor $\psi\iota_c^{(1)}$, $\psi\iota_c^{(*)}$ for estimating variance of $\widehat{\beta}_c^{(1)}$, $\widehat{\beta}_c^{(*)}$ under the null of no outliers for large samples.

It is fundamental to note that ψ, ς_c^2, $\vartheta_c^{(1)}$, $\vartheta_c^{(*)}$ are between 0 and 1, so $\iota_c^{(1)} = (\vartheta_c^{(1)}\varsigma_c^2)^{-1}$, $\iota_c^{(*)} = (\vartheta_c^{(*)}\varsigma_c^2)^{-1}$ and $\psi\iota_c^{(1)}$, $\psi\iota_c^{(*)}$ are greater than 1, since the trimming method underestimates the variance σ^2 and has the lower efficiency (larger asymptotic variance) than ordinary 2SLS such that the adjusted (estimated) asymptotic variance and standard errors of beta estimates are larger than those provided by usual asymptotics.

Consistent with Figures 3.3.1 and 3.3.2, as the cut-off c increases then ψ, ς_c^2, $\vartheta_c^{(1)}$, $\vartheta_c^{(*)}$ increase and approach to 1 so $\iota_c^{(1)}$, $\iota_c^{(*)}$ decrease and approach to 1 along with the fact that $\psi\iota_c^{(1)}$, $\psi\iota_c^{(*)}$ approach to $\iota_c^{(1)}$, $\iota_c^{(*)}$. Moreover, $\vartheta_c^{(1)}$ in the first step dominates the infinite step iteration $\vartheta_c^{(*)}$ so that magnitudes $\iota_c^{(1)}$, $\psi\iota_c^{(1)}$ in the first step is smaller and closer to 1 than that of the infinite step $\iota_c^{(*)}$, $\psi\iota_c^{(*)}$.

Nuisance parameters ξ_c, ζ_c^+, ζ_c^-[17] measure the dependence between x_i and u_i (between r_i and u_i) in the IVs setup. They are significantly important to evaluate the degree of endogeneity in the structural equation, but difficult to estimate in practice. However, performing valid inference on β does not require estimating these nuisance parameters, since they all vanish asymptotically and thus do not contribute to standard errors. Therefore, Algorithms 3.2.2 and 3.2.3 can be easily implemented not only to identify outliers and to obtain a robust estimator, but also to carry out valid inference.

3.3.6 A new Hausman type test of outlier robustness

A frequent concern in empirical economics is whether a tiny set of outliers may have invalidated empirical results. The common practice is to carry out robustness checks by redoing the analyses with the sample after trimming all outliers detected by Algorithm 3.2.1 and comparing results from the original ones with the full sample. Instead of heruistically checking the difference between robustified and ordinary 2SLS without knowing whether it is statistically significant, this chapter formalizes the outlier robustness check as a new type of Durbin (1954)-Hausman (1978)-Wu (1973) test.

The test is based on trade-off between robustness and efficiency and enables to judge whether the least squares estimation is appropriate or the robust method should be preferred. As discussed in §3.3.4, the robust estimator produced by Algorithm 3.2.1 is consistent both under the null and the alternative (although less efficient under the null), whereas ordinary 2SLS is efficient (and consistent) under the null, but inconsistent otherwise. The test statistics is looking on statistically significant difference between robustified and ordinary 2SLS. If the model is correctly specified, the Hausman type test statistics should be rather small under the null of no outliers, since two consistent methods should produce estimates that are very close. On the contrary when outliers have large influence on least squares estimation, the robust trimming method should be very different from the ordinary estimate, so 2SLS should be rejected and the robustified 2SLS should be preferred even if less efficient. The proposed test thus will evaluate

[17] see ξ_c in (3.2.4, §3.2.1) and ζ_c^+, ζ_c^- in (3.2.6, §3.2.1).

whether the gain in robustness is more valuable than the corresponding loss in efficiency.

In practice, empirical researchers run the full sample 2SLS $\widetilde{\beta}$ and compare to the robustified 2SLS $\widehat{\beta}_c^{(m+1)}$ on the sub-sample with all outliers removed by Algorithm 3.2.2 or 3.2.3. The new type of Hausman test can detect whether two estimates are significantly distinct by looking on the L2 norm of difference between $\widetilde{\beta}$ and $\widehat{\beta}_c^{(m+1)}$. Thus, it is essential first to derive the stochastic expansion and weak limit of a sequence of processes $\mathbb{H}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ for $c \in [c_+, \infty)$ and $m \in [0, \infty)$.

Theorem 3.3.9. *Consider Algorithm 3.2.2 or 3.2.3. Suppose Assumption 3.3.1*$(ia, iia, iiia, iv)$ *holds. For* $c \in [c_+, \infty)$ *and* $m \in [0, \infty)$ *denote the process* $\mathbb{H}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$. *Then as* $n \to \infty$ *we have*

$$\mathbb{H}_n^{(m+1)}(c) = (\varrho_{\beta\beta,c}^{(m+1)} - 1) M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + \varrho_{\beta\tilde{x}u,c}^{(m+1)} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} + \mathsf{o}_{\mathsf{P}}(1),$$

where $\varrho_{\beta\beta,c}^{(m+1)}$, $\varrho_{\beta\tilde{x}u,c}^{(m+1)}$ *are defined in Theorem 3.3.4. Furthermore,* $\mathbb{H}_n^{(m+1)}$ *weakly converges to a zero mean Gaussian process* $\mathbb{H}^{(m+1)}$ *with variance given as*

$$\mathsf{Var}\{\mathbb{H}^{(m+1)}(c)\} = \{(\varrho_{\beta\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta\beta,c}^{(m+1)} - 1)\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}.$$

The proof of the above theorem follows from stochastic expansions of $\widehat{\beta}_c^{(m+1)}$ in Theorem 3.3.6, 3.3.7 and $\widetilde{\beta}$ in Lemma 3.3.1. The next corollary establishes a new type of Hausman test for outlier robustness checks, which is based on pointwise convergence of $\mathbb{H}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ implied by its weak limit in Theorem 3.3.9.

Corollary 3.3.3. *Consider Algorithm 3.2.2 or 3.2.3. Suppose Assumption 3.3.1*$(ia, iia, iiia, iv)$ *holds. For* $c \in [c_+, \infty)$ *and* $m \in [0, \infty)$ *then as* $n \to \infty$ *we have*

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}\{0_{d_x}, \mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})\},$$

where $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = \{(\varrho_{\beta\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta\beta,c}^{(m+1)} - 1)\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}$. *Then the new type of Hausman test statistics has the weak limit*

$$H_{n,c}^{(m+1)} = n(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})'\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \xrightarrow{\mathsf{D}} \chi_{d_x}^2.$$

Hausman (1978) argued that, when two estimators are correlated (one is always consistent but inefficient under the null, the other efficient but not consistent under the alternative), the asymptotic variance of their difference is given by the difference of their respective asymptotic variances. However, his argument requires additional regularity conditions that might not be satisfied in our context. The next remark is to show that the argument does work so $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = \mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta})$ when $\mathsf{f}_u \sim \mathsf{N}(0, 1)$.

Remark 3.3.4. *Rearrange the expression of* $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ *in Corollary 3.3.3 to achieve*

$$\{(\varrho_{\beta\beta,c}^{(m+1)})^2 - 2\varrho_{\beta\beta,c}^{(m+1)} + 1 + 2\tau_2^c\varrho_{\beta\beta,c}^{(m+1)}\varrho_{\beta\tilde{x}u,c}^{(m+1)} - 2\tau_2^c\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}.$$

Assume $\mathsf{f}_u \sim \mathsf{N}(0,1)$*, then* $\tau_2^c = \psi - 2c\mathsf{f}_u(c)$*. Recall from Theorem 3.3.4 that*

$$\varrho_{\beta\beta,c}^{(m+1)} = \{\frac{2c\mathsf{f}_u(c)}{\psi}\}^{m+1}, \qquad \varrho_{\beta\tilde{x}u,c}^{(m+1)} = \frac{\psi^{m+1} - \{2c\mathsf{f}_u(c)\}^{m+1}}{\psi^{m+1}\{\psi - 2c\mathsf{f}_u(c)\}}.$$

Apply these terms to attain $-2\varrho_{\beta\beta,c}^{(m+1)} + 1 - 2\tau_2^c\varrho_{\beta\tilde{x}u,c}^{(m+1)} = -1$*. Further notice that* $\mathsf{avar}(\widehat{\beta}_c^{(m+1)}) = \{(\varrho_{\beta\beta,c}^{(m+1)})^2 + 2\tau_2^c\varrho_{\beta\beta,c}^{(m+1)}\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}$ *and* $\mathsf{avar}(\widetilde{\beta}) = \sigma^2 M_{\tilde{x}\tilde{x}}^{-1}$*, see (3.3.4) and Lemma 3.3.1. Thus, we finally shows*

$$\begin{aligned}
\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) &= \{(\varrho_{\beta\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta\beta,c}^{(m+1)} - 1)\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1} \\
&= \{(\varrho_{\beta\beta,c}^{(m+1)})^2 + 2\tau_2^c\varrho_{\beta\beta,c}^{(m+1)}\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1} - \sigma^2 M_{\tilde{x}\tilde{x}}^{-1} \\
&= \mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta}).
\end{aligned}$$

Remark 3.3.4 demonstrates that, in practical applications, it is equivalent either to estimate $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ or $\mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta})$ so as to obtain test statistics of outlier robustness developed in Corollary 3.3.3. If $u_i/\sigma \sim \mathsf{f}_u$ follows $\mathsf{N}(0,1)$ according to which the cut-off value c is chosen, terms τ_2^c, $\varrho_{\beta\beta,c}^{(m+1)}$, $\varrho_{\beta\tilde{x}u,c}^{(m+1)}$ are known for any iteration step m, so the only item left to estimate in the asymptotic variance is $\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}$.

As discussed in §3.3.5, the estimator of asymptotic variance should be robust itself, so $\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}$ should be estimated by $(\widehat{\sigma}_c^{(m+1)})^2$, $\widehat{M}_{\tilde{x}\tilde{x},c}^{(m+1)}$, see (3.2.12), (3.3.8), based on the sub-sample with all outliers removed instead of using the full sample. Otherwise the test might lose power and lead to incorrect results under the alternative, where estimates using the full sample are inconsistent though they are consistent (efficient) under the null of no outliers. Notice that to obtain test statistics many empirical researchers wrongly estimate $\mathsf{avar}(\widetilde{\beta})$ by using the full sample estimates of σ^2, $M_{\tilde{x}\tilde{x}}$, which might produce $\widehat{H}_{n,c}^{(m+1)} \leq 0$ as $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)}) - \widehat{\mathsf{avar}}(\widetilde{\beta}) \leq 0$ and make the test meaningless under the alternative, see discussion in detail in our empirical application to Acemoglu et al. (2017) in §3.4. Thus, in practice it is important to avoid this false procedure by applying the robust estimator of asymptotic variance of difference between ordinary and robustified 2SLS.

Once achieving the estimator $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ of asymptotic variance by

$$\{(\varrho_{\beta\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta\beta,c}^{(m+1)} - 1)\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}(\widehat{\sigma}_c^{(m+1)})^2(\widehat{M}_{\tilde{x}\tilde{x},c}^{(m+1)})^{-1}, \quad (3.3.9)$$

which is consistent under the null and robust under the alternative, then we have

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \overset{a}{\sim} \mathsf{N}\{0_{d_x}, \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})\},$$

and

$$\widehat{H}_{n,c}^{(m+1)} = n(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})'\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \overset{a}{\sim} \chi^2_{d_x}.$$

Thus, a new type of Hausman test for outlier robustness has now been established, and it can either be performed as the one-sided test with the chi-squared limit using estimates of the whole parameter vector or as the two-sided test with the normal limit using the corresponding estimates of individual scalar parameters.

Note that the asymptotic variance in (3.3.9) is invertible, since the first scalar term in the expression $(\varrho_{\beta\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta\beta,c}^{(m+1)} - 1)\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2 \neq 0$ and its inverse is

$$\{(\varrho_{\beta\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta\beta,c}^{(m+1)} - 1)\varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c(\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\}^{-1}(\widehat{\sigma}_c^{(m+1)})^{-2}\widehat{M}_{\tilde{x}\tilde{x},c}^{(m+1)}.$$

We therefore avoid the rank deficiency problem described and addressed through generalized inverses in Hausman and Taylor (1981) and Holly (1982).

On implementing Algorithm 3.2.2 or 3.2.3, we are particularly interested in the trimming estimator just updated from the original 2SLS and in the fixed point estimator iterated upon through infinite steps. The next corollary provides two special cases of the test for outlier robustness frequently used in empirics, checking whether $\widetilde{\beta}$ is distinct from $\widehat{\beta}_c^{(1)}$ when $m = 0$ or from $\widehat{\beta}_c^{(*)}$ when $m = \infty$.

Corollary 3.3.4. *Consider Algorithm 3.2.2 or 3.2.3. Suppose Assumption 3.3.1*$(ia, iia, iiia, iv)$ *holds. For* $c \in [c_+, \infty)$ *and a large* n *then we have for* $m = 0$

$$n^{1/2}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) \overset{a}{\sim} \mathsf{N}\{0_{d_x}, \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})\},$$

and

$$\widehat{H}_{n,c}^{(1)} = n(\widehat{\beta}_c^{(1)} - \widetilde{\beta})'\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) \overset{a}{\sim} \chi^2_{d_x},$$

where

$$\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) = \frac{\{2c\mathsf{f}_u(c) - \psi\}^2 + 2\tau_2^c\{2c\mathsf{f}_u(c) - \psi\} + \tau_2^c}{\psi^2}(\widehat{\sigma}_c^{(1)})^2(\widehat{M}_{\tilde{x}\tilde{x},c}^{(1)})^{-1}.$$

In addition, we have for $m = \infty$

$$n^{1/2}(\widehat{\beta}_c^{(*)} - \widetilde{\beta}) \overset{a}{\sim} \mathsf{N}\{0_{d_x}, \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(*)} - \widetilde{\beta})\},$$

and

$$\widehat{H}_{n,c}^{(*)} = n(\widehat{\beta}_c^{(*)} - \widetilde{\beta})'\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(*)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(*)} - \widetilde{\beta}) \overset{a}{\sim} \chi^2_{d_x},$$

where

$$\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(*)} - \widetilde{\beta})^{18} = \frac{\{2c\mathsf{f}_u(c) - \psi\}^2 + 2\tau_2^c\{2c\mathsf{f}_u(c) - \psi\} + \tau_2^c}{\{2c\mathsf{f}_u(c) - \psi\}^2}(\widehat{\sigma}_c^{(*)})^2(\widehat{M}_{\tilde{x}\tilde{x},c}^{(*)})^{-1}.$$

Without knowing the asymptotic theory derived in this section, some empirical researchers simply use the similar form of the Hausman test statistics but incorrectly replace the standard error of the difference of two estimates $\widetilde{\beta}$ and $\widehat{\beta}_c^{(m+1)}$ by the standard error of the marginal distribution of the baseline estimate $\widetilde{\beta}$. They subsequently compare this test statistics with the critical value either drawn from the standard normal for the two-sided test or from the chi-square with d_x degree of freedom for the one-sided test. The above test procedure is commonly referred to as the heuristic method. Since the heuristic test statistics is constructed in a mistaken way, its size does not converge to the nominal level under the null and it has low statistical power under the alternative. Thus, even the heuristic method is informative for preliminary analysis, it is not consistent in theory and not reliable in practice.

Kaji (2018) applies the non-parametric bootstrap by randomly sampling observations i from the data $\{(y_i, x_i, z_i)\}_{i=1}^n$ with replacement to draw the distribution of the L_1 norm of the difference between $\widehat{\beta}_c^{(m+1)}$ and $\widetilde{\beta}$. To prove the validity of the bootstrap, he first characterises the weak convergence of a class of integrable empirical processes and then build the functional delta method for the corresponding quantile processes. Unlike the asymptotic theory in this chapter, the bootstrap in Kaji (2018) does not require assuming the known form of the density f_u. However, if the specification assumed on f_u is correct, our Hausman type test would have the higher power than the bootstrapping method under the alternative where outliers in fact present. This could be because under the alternative the bootstrap has a certain probability to sample outlying observations that have large influence on β estimation. For these bootstrapping iterations having sampled large outliers from the original data, the test statistics on difference between the robust estimate $\widehat{\beta}_c^{(m+1)}$ and the baseline estimate $\widetilde{\beta}$ would then become rather large, so the distribution of the test statistics drawn from the bootstrap sample would be highly distorted and significantly distinct from what it should be under the null of no outliers.

[18] If $\mathsf{f}_u \sim \mathsf{N}(0,1)$, then $\tau_2^c = \psi - 2c\mathsf{f}_u(c)$ so $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})$ and $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(*)} - \widetilde{\beta})$ can be further simplified.

3.4 Empirical application to Acemoglu et al. (2017)

This chapter now applies the robust procedure proposed in §3.2.2 and its theory to re-carry out outlier analysis in Acemoglu et al. (2017). Their paper tries to address the long-standing debate of whether democracy does cause economic growth through an empirical study. Acemoglu et al. (2017) find evidence that democracy has a significant and robust positive effect on GDP per capita and further suggest that democracy increases GDP by encouraging investment, increasing schooling, inducing economic reforms and openness, improving the provision of public goods, and reducing social unrest.

To investigate the causal effect of democracy on GDP per capita, Acemoglu et al. (2017) collect an annual panel that comprises 175 countries around the world from the year 1960 to 2010. The panel is unbalanced in the sense that not all observations of GDP and democracy are available for the entire sample. In order to reduce measurement error, they create a consolidated and dichotomous measure of democracy from several datasets and document robustness of their results to other measures of GDP and democracy.

Acemoglu et al. (2017) conduct three specifications that control for the rich dynamics of GDP and endogenous selection into democracy, which otherwise confound the effect of democracy on economic growth. Three specifications consist a dynamic panel model for log GDP per capita, a semi-parametric treatment effects model which does not rely on a parametric assumption for GDP dynamics, the same dynamic panel model but instrumenting democracy by regional waves of democratizations and reversals. We focus on the first and third specification (a dynamic panel model) in Acemoglu et al. (2017) where they carry out outlier analysis for OLS and 2SLS fixed effect estimates.

Their baseline preferred specification constructs a dynamic panel that models log GDP per capita as outcome variable, democracy as treatment variable, and four lags of log GDP per capita as other covariates to control GDP dynamics. They start with running the fixed effect regression implemented as OLS estimator by adding dummy variable regressors for each country and year. Although the fixed effect estimator suffers the Nickell (1981) bias of order one over time horizon resulting from the failure of strict exogenity in dynamic panel model, this bias should be small in their setup where the time dimension of panel is fairly large (each country is on average observed 38.8 times). This motivates their use of the OLS fixed effect estimator as a natural starting point even in dynamic panel models. Acemoglu et al. (2017) then apply Arellano and Bond (1991) GMM estimator that deal with Nickell bias, and further use an even more refined Hahn, Hausman, and Kuersteiner (2002) GMM estimator that address "too many instruments"

problem faced by Arellano and Bond estimator. Consistent with their expectation, both type of GMM estimates are similar to the OLS fixed effec estimate such that the coefficient on democracy variable is estimated to be positive and highly significant.

Their third specification is identical to the dynamic panel model above, but they treat democracy as endogenous and exploit IVs strategy by using regional waves of democratization and transitions to nondemocracy as source of exogenous variation in democracy. The instrument regional waves of democratization is constructed as the jackknifed average of democracy for countries in the same region with the same initial regime cell, which leaves out the own-country observation. Although there is no consensus on the factors that create such waves, the existing evidence suggests that they are not explained by regional economic trends. Thus, Acemoglu et al. (2017) argue that regional waves are valid as instruments such that their exclusion restriction assumption is plausible. They further show from the first stage regression that regional waves are informative such that rank conditions hold for identification. By applying 2SLS fixed effect and Hahn, Hausman, and Kuersteiner (2002) estimator, they reassure that the causal effect of democracy on log GDP per capita is significant and positive.

Three empirical strategies tackle different possibilities and find very similar results, which bolsters their confidence to demonstrate the positive causal effect of democracy. To defend against the critique that their empirical findings are driven by outliers, Acemoglu et al. (2017) explores the sensitivity of their results to outliers for their baseline OLS and 2SLS fixed effect regression.

For each observation in the model (3.2.1) and (3.2.2), we now have the dependent variable, regressors, and instruments

$$
\begin{aligned}
y_{i,t} &:= \log \mathrm{GDP}_{i,t}, \\
x_{i,t} &:= (\mathrm{Democracy}_{i,t}, \log \mathrm{GDP}_{i,t-1}, \ldots, \log \mathrm{GDP}_{i,t-4}, \\
&\qquad 1_{i=1}, \ldots, 1_{i=175}, 1_{t=1960}, \ldots, 1_{t=2010})', \\
z_{i,t} &:= (\mathrm{Wave}_{i,t-1}, \ldots, \mathrm{Wave}_{i,t-4}, \log \mathrm{GDP}_{i,t-1}, \ldots, \log \mathrm{GDP}_{i,t-4}, \\
&\qquad 1_{i=1}, \ldots, 1_{i=175}, 1_{t=1960}, \ldots, 1_{t=2010})'.
\end{aligned}
$$

In regressors, only $\mathrm{Democracy}_{i,t}$ is treated as endogenous, while the rest of $x_{i,t}$ are assumed to be orthogonal to structural errors. Then the structural and first stage

regression parameters are

$$\beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_4 \\ \theta_1 \\ \vdots \\ \theta_{175} \\ \phi_{1960} \\ \vdots \\ \phi_{2010} \end{pmatrix}, \qquad \Pi := \begin{pmatrix} \Pi_1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Pi_4 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \pi_1 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \pi_4 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vartheta_1 & 0 & \cdots & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vartheta_{175} & 0 & \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ \varphi_{1960} & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \varphi_{2010} & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

In the structural parameter vector β, we are particulary interested in β_0 on the effect of democracy as well as $\beta_1, \ldots, \beta_4$ indicating the dynamics of the GDP process[19].

Implicitly assuming normal distribution for structural errors, i.e. $\mathsf{f}_u \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$, Acemoglu et al. (2017) choose the cut-off value $c = 1.96$ for 5% significance level such that outliers are identified among observations whose absolute values of standardized residuals of structural errors are beyond 1.96. They then perform the same OLS and 2SLS fixed effect regression but exclude these outlying observations[20]. Investigating the difference between the ordinary (full sample) and robust (sub-sample without outliers) estimates, they argue robustness of their results to outliers.

The above robust method is in fact Algorithm 3.2.2 (Robustified 2SLS), which select the cut-off $c = 1.96$ (5%) by assuming $\mathsf{f}_u \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$ and iterates only once by choosing $m = 0$. However, Acemoglu et al. (2017) do not take into consideration the case that observations are classified as outliers and removed just by chance, so consistency does

[19] Aside from structural parameters $\beta_0, \beta_1, \ldots, \beta_4$, Acemoglu et al. (2017) report three more parameters: (1) $\beta_5 := \beta_0/(1 - \beta_1 - \beta_2 - \beta_3 - \beta_4)$ (long-run equilibrium effect of democracy); (2) $\beta_6 := e_{25}$, where $e_j = \beta_0 + \beta_1 e_{j-1} + \beta_2 e_{j-2} + \beta_3 e_{j-3} + \beta_4 e_{j-4}$ and $e_0 = e_{-1} = e_{-2} = e_{-3} = 0$ (effect of transition to democracy after 25 years); (3) $\beta_7 := \beta_1 + \beta_2 + \beta_3 + \beta_4$ (persistence of the GDP process).

[20] This is the first robust method that Acemoglu et al. (2017) apply, and its results are reported in column (2), Table A8 for OLS fixed effect regression; and in column (2), Table A13 for 2SLS fixed effect regression in their paper. They also use other robust methods, such as Cook's distance, Li's (1985) procedure, and Huber M-estimator; and take into account the influence of outliers in both the first and second stage; see their reported results in other columns in Table A8, A13.

not hold for the updated estimator of the structural error variance if without adjusting the bias correction factor (3.2.8), although estimation of structural parameters is still consistent. In addition, by not adjusting standard errors calculated from the usual asymptotoics that does not have the correct asymptotic variance of robust estimates, inference on regression coefficients is invalid. At last, without a formal statistical test we do not know whether the robust estimate is significantly distinct from the ordinary one, thus even when we end up with the robustness conclusion, our argument is not rooted in a correct statistical reasoning.

To solve all these problems, this chapter applies Algorithm 3.2.2 to re-explore outlier robustness analysis in Acemoglu et al. (2017). As in their paper, we first compute the full sample 2SLS $\widetilde{\beta}$, then choose the cut-off $c = 1.96$ to run a single iteration $m = 0$ of the algorithm to obtain $\widehat{\beta}^{(1)}_{1.96}$. Besides selecting $m = 0$, we continue to run the algorithm upon through infinite iterations $m = \infty$ until reaching the fixed point $\widehat{\beta}^{(*)}_{1.96}$. In adjusting the standard errors of $\widehat{\beta}^{(1)}_{1.96}$ and $\widehat{\beta}^{(*)}_{1.96}$ by the factors listed in Table 3.3.1, this chapter performs valid inference on structural parameters β and conducts the new Hausman type tests to check whether $\widetilde{\beta}$ is different from $\widehat{\beta}^{(1)}_{1.96}$ or $\widehat{\beta}^{(*)}_{1.96}$. Our re-examination of Acemoglu et al. (2017) finds that the effect of democracy estimated with the full sample $\widetilde{\beta}_0$ is distinct in magnitude from the robust estimates $\widehat{\beta}^{(1)}_{0,1.96}$ or $\widehat{\beta}^{(*)}_{0,1.96}$ produced by Algorithm 3.2.2, although we find little evidence for the view that democracy is a constraint on economic growth ($\beta_0 < 0$).

Tables 3.4.1 and 3.4.2 list results of the OLS and 2SLS fixed effect regressions together with formal statistical tests for the hypothesis that outliers have no effect on parameters. The first three columns show the OLS (2SLS) fixed effect and robust estimates in Table 3.4.1 (Table 3.4.2). Select the cut-off value $c = 1.96$ and set the iteration step $m = -1$, $m = 0$, $m = \infty$, then run Algorithm 3.2.2 to produce three estimates $\widetilde{\beta}$, $\widehat{\beta}^{(1)}_{1.96}$, $\widehat{\beta}^{(*)}_{1.96}$. In Column 1, $\widetilde{\beta}$ represents the full sample OLS (2SLS) estimate in Table 3.4.1 (Table 3.4.2). In Column 2, we remove observations whose absolute values of standardized residuals are beyond the cut-off 1.96 (in the second stage) and re-compute OLS (2SLS) to obtain the robust estimate $\widehat{\beta}^{(1)}_{1.96}$ in Table 3.4.1 (Table 3.4.2). In Column 3, this procedure is iterated until the fixed point $\widehat{\beta}^{(*)}_{1.96}$ is attained. Values reported below the estimates in parentheses () are standard errors[21]. Note that in Columns 2 and 3 in Tables 3.4.1 and 3.4.2 we report the adjusted standard errors[22] by applying the theory in §3.3.5.

[21] Acemoglu et al. (2017) also report standard errors robust against heteroskedasticity and serial correlation at the country level in their Table A8, A13.

[22] Acemoglu et al. (2017) do not take the adjustments to the standard errors of $\widehat{\beta}^{(1)}_{1.96}$ reported in

In Table 3.4.1 (Table 3.4.2), ordinary standard errors are multiplied by $\sqrt{6044/6336}$ ($\sqrt{6019/6309}$) and $\sqrt{\iota^{(1)}_{1.96}}$ in Column 2; and by $\sqrt{5213/6336}$ ($\sqrt{5202/6309}$) and $\sqrt{\iota^{(*)}_{1.96}}$ in Column 3. Table 3.3.1 in §3.3.5 shows $\iota^{(1)}_{1.96} = 1.612$ and $\iota^{(*)}_{1.96} = 1.828$, so in order to produce the reported standard errors in Table 3.4.1 (Table 3.4.2) we apply the adjusting factor 1.240 (1.240) in Column 2 and 1.226 (1.228) in Column 3.

In Tables 3.4.1 and 3.4.2, Columns $4-7$ present formal statistical tests for outlier robustness analysis. Columns $4-5$ give test results for the null hypothesis that the two parameters estimated by $\widetilde{\beta}$ and $\widehat{\beta}^{(1)}_{1.96}$ are identical. Columns $6-7$ present test results for the same hypothesis but comparing $\widetilde{\beta}$ with $\widehat{\beta}^{(*)}_{1.96}$. Columns 5 and 7 show the new Hausman type test statistics $\widehat{H}^{(1)}_{1.96}$ and $\widehat{H}^{(*)}_{1.96}$ that use the standard errors of the difference of two estimators $\widehat{\mathsf{avar}}(\widehat{\beta}^{(1)}_{1.96} - \widetilde{\beta})$ and $\widehat{\mathsf{avar}}(\widehat{\beta}^{(*)}_{1.96} - \widetilde{\beta})$[23], see Corollary 3.3.4 in §3.3.6 for construction of the Hausman test statistics. We first apply the two-sided type Hausman test with the normal limit for each individual scalar parameter β_0, β_1, ..., and β_4, then perform the one-sided test with the chi-squared limit for the whole parameter vector $(\beta_0, \beta_1, \ldots, \beta_4)'$. Columns 4 and 6 show the heuristic test statistics $\widehat{\mathrm{Heu}}^{(1)}_{1.96}$ and $\widehat{\mathrm{Heu}}^{(*)}_{1.96}$ for the same hypothesis, which however use the standard error of the marginal distribution of the baseline full sample estimate $\widetilde{\beta}$. P-values are reported in brackets [] below test statistics.

We see from Tables 3.4.1 and 3.4.2 that β_0 on democracy is estimated by 2SLS to have the larger value than OLS fixed effect regressions while the effect of GDP dynamics remains roughly unchanged, which is consistent with our expectation that the IVs strategy affects mostly the coefficient of the endogenous variable democracy. The standard errors of the estimates $\widetilde{\beta}$, $\widehat{\beta}^{(1)}_{1.96}$, $\widehat{\beta}^{(*)}_{1.96}$ decrease with iteration, suggesting that it is under the alternative hypothesis where the data is contaminated by outlying observations driving up the standard error of the full sample estimate $\widetilde{\beta}$. Otherwise under the null of no outliers standard errors of $\widehat{\beta}^{(1)}_{1.96}$, $\widehat{\beta}^{(*)}_{1.96}$ should become larger than $\widetilde{\beta}$, since by gaining robustness the estimates $\widehat{\beta}^{(1)}_{1.96}$, $\widehat{\beta}^{(*)}_{1.96}$ lose efficiency relative to $\widetilde{\beta}$, leading

Column (2) in their Table A8, A13.

[23]Consider $c = 1.96$ and $m = 0, \infty$ and note $\mathsf{avar}(\widehat{\beta}^{(m+1)}_c - \widetilde{\beta}) = \mathsf{avar}(\widehat{\beta}^{(m+1)}_c) - \mathsf{avar}(\widetilde{\beta})$. To construct the standard errors of the difference of two estimators, some empirical researchers estimate $\mathsf{avar}(\widetilde{\beta})$ using the full sample and estimate $\mathsf{avar}(\widehat{\beta}^{(m+1)}_c)$ using the sub-sample with all outliers removed. This produces the negative estimate of $\mathsf{avar}(\widehat{\beta}^{(m+1)}_c - \widetilde{\beta})$ in this application as $\widehat{\mathsf{avar}}(\widehat{\beta}^{(m+1)}_c) - \widehat{\mathsf{avar}}(\widetilde{\beta}) \leq 0$ clearly shown in Tables 3.4.1 and 3.4.2, making the Hausman test meaningless. As suggested in §3.3.6, the estimator of $\mathsf{avar}(\widehat{\beta}^{(m+1)}_c - \widetilde{\beta})$ should be robust itself to ensure the power performance under the alternative. In other words, we need to robustly estimate $\mathsf{avar}(\widetilde{\beta})$ using the sub-sample of all non-outlying observations.

		OLS Estimate			Test: $\widehat{\beta}^{(1)}_{1.96}$, $\widetilde{\beta}$		Test: $\widehat{\beta}^{(*)}_{1.96}$, $\widetilde{\beta}$	
	β	$\widetilde{\beta}$	$\widehat{\beta}^{(1)}_{1.96}$	$\widehat{\beta}^{(*)}_{1.96}$	$\widehat{\text{Heu}}^{(1)}_{1.96}$	$\widehat{H}^{(1)}_{1.96}$	$\widehat{\text{Heu}}^{(*)}_{1.96}$	$\widehat{H}^{(*)}_{1.96}$
Democracy	β_0	0.787 (0.228)	0.556 (0.186)	0.142 (0.129)	−1.011 [0.312]	−2.908 [0.004]	−2.821 [0.005]	−9.421 [0.000]
GDP 1st lag	β_1	1.238 (0.013)	1.226 (0.011)	1.264 (0.009)	−0.992 [0.321]	−2.537 [0.011]	2.088 [0.037]	5.501 [0.000]
GDP 2nd lag	β_2	−0.207 (0.020)	−0.198 (0.018)	−0.235 (0.013)	0.435 [0.664]	1.142 [0.254]	−1.456 [0.145]	−4.081 [0.000]
GDP 3rd lag	β_3	−0.026 (0.019)	−0.027 (0.016)	−0.030 (0.012)	−0.038 [0.970]	−0.104 [0.917]	−0.198 [0.843]	−0.579 [0.563]
GDP 4th lag	β_4	−0.043 (0.012)	−0.030 (0.010)	−0.019 (0.008)	1.088 [0.277]	2.975 [0.003]	1.985 [0.047]	5.741 [0.000]
Whole vector					18.347 [0.003]	140.514 [0.000]	62.262 [0.000]	645.087 [0.000]
Sample size		6336	6044	5213				

Table 3.4.1: This table presents the OLS fixed effect estimates of the simple regression exploring the effects of democracy and four lags of log GDP per capita on log GDP per capita. Then, formal statistical tests are shown to achieve outlier robustness analysis. Columns $1-3$ represent OLS estimates with standard errors reported below in parentheses (). The cut-off is chosen as $c = 1.96$, then Algorithm 3.2.2 is iterated at the step $m = -1$, $m = 0$, $m = \infty$ to produce three estimates $\widetilde{\beta}$, $\widehat{\beta}^{(1)}_{1.96}$, $\widehat{\beta}^{(*)}_{1.96}$. In Column 1 we show the baseline full sample OLS. In Column 2 we remove observations with a standardized residual estimated above 1.96 or below -1.96. In Column 3 we iterate the above procedure until reaching the fixed point. In Columns $2-3$ standard errors of $\widehat{\beta}^{(1)}_{1.96}$ and $\widehat{\beta}^{(*)}_{1.96}$ are adjusted by the theory in §3.3.5. Columns $4-5$ represent outlier robustness tests comparing $\widetilde{\beta}$ with $\widehat{\beta}^{(1)}_{1.96}$ and Columns $6-7$ comparing $\widetilde{\beta}$ with $\widehat{\beta}^{(*)}_{1.96}$. Columns 4 and 6 show heuristic test statistics $\widehat{\text{Heu}}^{(1)}_{1.96}$ and $\widehat{\text{Heu}}^{(*)}_{1.96}$ that use the standard errors of $\widetilde{\beta}$. Columns 5 and 7 show the new Hausman type test statistics $\widehat{H}^{(1)}_{1.96}$ and $\widehat{H}^{(*)}_{1.96}$ that use the standard errors of $\widehat{\beta}^{(1)}_{1.96} - \widetilde{\beta}$ and $\widehat{\beta}^{(*)}_{1.96} - \widetilde{\beta}$ developed in §3.3.6. P-values are reported in brackets [] below test statistics.

to the larger asymptotic variance. The number of observations used in the OLS and 2SLS estimates are 6336 and 6309. Even when the model contains no outliers, Algorithm 3.2.2 throws away around 5% of the whole observations with a large probability for the cut-off value 1.96 to produce $\widehat{\beta}^{(1)}_{1.96}$ and $\widehat{\beta}^{(*)}_{1.96}$. Thus, under the null of no outliers the expected number of observations retained by Algorithm 3.2.2 should be around 6019 and 5994 that are sub-sample sizes of 95% of the full OLS and 2SLS sample. We notice that

		2SLS Estimate			Test: $\widehat{\beta}^{(1)}_{1.96}$, $\widetilde{\beta}$		Test: $\widehat{\beta}^{(*)}_{1.96}$, $\widetilde{\beta}$	
	β	$\widetilde{\beta}$	$\widehat{\beta}^{(1)}_{1.96}$	$\widehat{\beta}^{(*)}_{1.96}$	$\widehat{\text{Heu}}^{(1)}_{1.96}$	$\widehat{H}^{(1)}_{1.96}$	$\widehat{\text{Heu}}^{(*)}_{1.96}$	$\widehat{H}^{(*)}_{1.96}$
Democracy	β_0	1.149 (0.607)	0.941 (0.487)	0.605 (0.332)	−0.343 [0.732]	−1.000 [0.317]	−0.896 [0.370]	−3.099 [0.002]
GDP 1st lag	β_1	1.238 (0.013)	1.224 (0.012)	1.258 (0.009)	−1.103 [0.270]	−2.820 [0.005]	1.531 [0.126]	4.059 [0.000]
GDP 2nd lag	β_2	−0.205 (0.020)	−0.194 (0.018)	−0.225 (0.013)	0.564 [0.573]	1.483 [0.138]	−0.961 [0.336]	−2.702 [0.007]
GDP 3rd lag	β_3	−0.029 (0.019)	−0.032 (0.017)	−0.036 (0.013)	−0.142 [0.887]	−0.389 [0.697]	−0.349 [0.727]	−1.022 [0.307]
GDP 4th lag	β_4	−0.040 (0.012)	−0.027 (0.010)	−0.016 (0.008)	1.158 [0.247]	3.157 [0.002]	2.001 [0.045]	5.797 [0.000]
Whole vector					18.215 [0.003]	138.445 [0.000]	56.027 [0.000]	571.720 [0.000]
Sample size		6309	6019	5202				

Table 3.4.2: This table presents the 2SLS fixed effect estimates of the IVs regression exploring the causal effects of democracy and four lags of log GDP per capita on log GDP per capita. Then, formal statistical tests are shown to achieve outlier robustness analysis. Columns $1-3$ represent 2SLS estimates with standard errors reported below in parentheses (). The cut-off is chosen as $c = 1.96$, then Algorithm 3.2.2 is iterated at the step $m = -1$, $m = 0$, $m = \infty$ to produce three estimates $\widetilde{\beta}$, $\widehat{\beta}^{(1)}_{1.96}$, $\widehat{\beta}^{(*)}_{1.96}$. In Column 1 we show the baseline full sample 2SLS. In Column 2 we remove observations with a standardized residual estimated above 1.96 or below -1.96 in the second stage. In Column 3 we iterate the above procedure until reaching the fixed point. In Columns $2-3$ standard errors of $\widehat{\beta}^{(1)}_{1.96}$ and $\widehat{\beta}^{(*)}_{1.96}$ are adjusted by the theory in §3.3.5. Columns $4-5$ represent outlier robustness tests comparing $\widetilde{\beta}$ with $\widehat{\beta}^{(1)}_{1.96}$ and Columns $6-7$ comparing $\widetilde{\beta}$ with $\widehat{\beta}^{(*)}_{1.96}$. Columns 4 and 6 show heuristic test statistics $\widehat{\text{Heu}}^{(1)}_{1.96}$ and $\widehat{\text{Heu}}^{(*)}_{1.96}$ that use the standard errors of $\widetilde{\beta}$. Columns 5 and 7 show the new Hausman type test statistics $\widehat{H}^{(1)}_{1.96}$ and $\widehat{H}^{(*)}_{1.96}$ that use the standard errors of $\widehat{\beta}^{(1)}_{1.96} - \widetilde{\beta}$ and $\widehat{\beta}^{(*)}_{1.96} - \widetilde{\beta}$ developed in §3.3.6. P-values are reported in brackets [] below test statistics.

the observed number of observations retained by $\widehat{\beta}^{(*)}_{1.96}$ are much less than the expected number for both of the OLS and 2SLS fixed effect regressions, although the sub-sample size of $\widehat{\beta}^{(1)}_{1.96}$ for OLS and 2SLS are not far away from the expected number. This fact again indicates that the model does contain outliers.

By only computing point estimates $\widetilde{\beta}$, $\widehat{\beta}^{(1)}_{1.96}$, $\widehat{\beta}^{(*)}_{1.96}$ for the OLS or 2SLS fixed effect regressions without applying formal statistical tests of outlier robustness, we cannot

draw any conclusion on whether or not the empirical findings are driven by outliers, since the information contained in the standard errors also needs to be considered in order to evaluate the distance between ordinary and robust estimates. For either the simple regression in Table 3.4.1 or IVs specification in Table 3.4.2, both of the heuristic and Hausman tests reject the null hypothesis that the parameter vector β estimated by $\widetilde{\beta}$ is identical to the robust estimate $\widehat{\beta}^{(1)}_{1.96}$ or $\widehat{\beta}^{(*)}_{1.96}$. It is obvious to note from Columns $4-7$ in Tables 3.4.1 and 3.4.2 that the Hausman tests yield strictly the lower p-values than the heuristic tests, which is concordant with our theory in §3.3.6 that the Hausman test has the better power performance than the heuristics. In the simple regressions, the Hausman tests reject the null of outlier robustness at the 5% significance level for most coefficients except β_2, β_3 when compare $\widetilde{\beta}$ with $\widehat{\beta}^{(1)}_{1.96}$; except β_3 when compare $\widetilde{\beta}$ with $\widehat{\beta}^{(*)}_{1.96}$. In the IVs regressions, the Hausman tests mostly give the same result but the identity of β_0 is not rejected by comparing $\widetilde{\beta}$ with $\widehat{\beta}^{(1)}_{1.96}$.

Kaji (2018) also re-examines outlier robustness of the results found in Acemoglu et al. (2017). Without the weak limit of the Hausman type test statistics derived in our paper, he carries out the non-parametric bootstrap to draw the distribution of the L_1 norm of the difference between the baseline full sample estimate and the robust estimate. The bootstrap is implemented by randomly sampling countries i with replacement. Each draw of country i adds its time dimensional number of observations to the bootstrap sample. The bootstrap consists of 10000 iterations. By comparing between the baseline OLS estimate $\widetilde{\beta}$ and outlier-removed estimate $\widehat{\beta}^{(1)}_{1.96}$, the bootstrap does not reject the null that two estimates are identical, indicating robustness to outliers of the empirical findings in Acemoglu et al. (2017). The reason why the null of outlier robustness is not rejected by Kaji (2018) could be the low power issue faced by the bootstrap under alternatives; see discussion in the last paragraph in §3.3.6. Note that Kaji (2018) only applies the tests for checking the difference for each coefficient $\beta_0, \beta_1, \ldots, \beta_4$, but not for the whole vector β.

Table 3.4.3 focuses on β_0 the effect of democracy on log GDP per capita by summarizing regression results from Tables 3.4.1 and 3.4.2. We present estimates of β_0 and compare standard errors with and without the adjustments together with their corresponding inference including 95% confidence intervals and p-values of the t-test for the null hypothesis that $\beta_0 = 0$. The adjusted standard errors and inference results are reported in parentheses () after the ordinary ones.

We see from Table 3.4.3 that the modified standard errors are uniformly larger than

		Est.	Std. Err.(A)	95% Confidence Interval(A)	P-value(A)
OLS	$\widetilde{\beta}_0$	0.787	0.228	$[0.339, 1.234]$	0.001
	$\widehat{\beta}^{(1)}_{0,1.96}$	0.556	0.150(0.186)	$[0.262, 0.850]([0.191, 0.920])$	0.000(0.003)
	$\widehat{\beta}^{(*)}_{0,1.96}$	0.142	0.106(0.129)	$[-0.064, 0.349]([-0.111, 0.396])$	0.177(0.271)
2SLS	$\widetilde{\beta}_0$	1.149	0.607	$[-0.041, 2.340]$	0.058
	$\widehat{\beta}^{(1)}_{0,1.96}$	0.941	0.393(0.487)	$[0.170, 1.712]([-0.014, 1.897])$	0.017(0.054)
	$\widehat{\beta}^{(*)}_{0,1.96}$	0.605	0.271(0.332)	$[0.074, 1.136]([-0.046, 1.257])$	0.025(0.069)

Table 3.4.3: This table concentrates on β_0 the effect of democracy on log GDP per capita. We present estimates of β_0, standard errors, 95% confidence intervals, and p-values of the t-test for the null hypothesis $\beta_0 = 0$. To compare standard errors and inference with and without the adjustments, we first report the ordinary standard errors and inference, followed by in parentheses () the ones adjusted by the theory in §3.3.5. Rows $1-3$ represent OLS fixed effect regressions and Rows $4-6$ 2SLS fixed effect regressions. The cut-off is chosen as $c = 1.96$, then Algorithm 3.2.2 is iterated at the step $m = -1$, $m = 0$, $m = \infty$ to produce three estimates $\widetilde{\beta}_0$, $\widehat{\beta}^{(1)}_{0,1.96}$, $\widehat{\beta}^{(*)}_{0,1.96}$ of β_0 in simple or IVs regressions. In Row 1 we show the baseline full sample ordinary least squares, and instrument democracy in Row 4. In Row 2 we remove observations with a standardized residual estimated above 1.96 or below -1.96, and the same outlier selection is based on residuals in the second stage in Row 5. In Rows 3 and 6 we iterate the above procedure until reaching the fixed point.

those without taking the adjustments, so the adjusted confidence bands are uniformly wider and p-values are uniformly larger; see the theory in §3.3.5. Recall that the baseline full sample and robust estimates of β_0 are statistically different, indicated by the Hausman tests. In the OLS regressions, inference based on $\widehat{\beta}^{(*)}_{0,1.96}$ shows that β_0 is not significantly distinct from 0. In the 2SLS regressions, we cannot reject $\beta_0 = 0$ at the 5% significance level by performing the adjusted inference either based on the baseline estimate $\widetilde{\beta}$ or the robust estimates $\widehat{\beta}^{(1)}_{0,1.96}$ and $\widehat{\beta}^{(*)}_{0,1.96}$. Thus, we cannot find the evidence robustly supporting the result $\beta_0 > 0$ that the effect of democracy is strictly positive on economic growth.

3.5 Weighted and marked empirical process

Consider the weighted and marked empirical distribution function

$$\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,c) = \frac{1}{n}\sum_{i=1}^{n} w_{in} u_i^p 1_{(u_i \leq \sigma c + n^{-1/2} ac + z'_{in}\Pi b + n^{-1/2} r'_i b)}, \tag{3.5.1}$$

with $\mathcal{F}_{i-1}$ adapted weights w_{in} and $\mathcal{F}_i$ measurable marks u_i^p. The filtration $\mathcal{F}_{i-1}$ is generated by $(z_1, \ldots, z_i, r_1, \ldots, r_{i-1}, u_1, \ldots, u_{i-1})$ so $z_{in} \in \mathcal{F}_{i-1}$ while $u_i, r_i \in \mathcal{F}_i$ are independent of $\mathcal{F}_{i-1}$. The observable data $\{(y_i, x_i, z_i)\}_{i=1}^n$ has i.i.d. structure in this chapter, so $a \in \mathbb{R}$, $b \in \mathbb{R}^{d_x}$ represent normalized estimation errors $\widetilde{a} = n^{1/2}(\widetilde{\sigma} - \sigma)$, $\widetilde{b} = n^{1/2}(\widetilde{\beta} - \beta)$, and $c \in \mathbb{R}$ is the quantile, while σ, Π, Σ are true parameters for the variance of the structural error, for the location coefficient in the first stage regression, and for the variance of the first stage error respectively. Let normalized instruments $z_{in} = n^{-1/2} z_i$ such that $M_{zz,n} = \sum_{i=1}^n z_{in} z'_{in} = n^{-1}\sum_{i=1}^n z_i z'_i$ converges in probability to $\mathsf{E} z_i z'_i = M_{zz}$ by Law of Large Numbers when $\mathsf{E}|z_i|^2 < \infty$. Our interest focuses on weights w_{in} given as either of 1, $n^{1/2} z_{in} = z_i$, $n z_{in} z'_{in} = z_i z'_i$ and p as either of 0, 1, 2. To form the empirical process, introduce the compensator

$$\overline{\mathsf{F}}^{w,p}_{u,n}(a,b,c) = \frac{1}{n}\sum_{i=1}^{n} w_{in} \mathsf{E}_{i-1} u_i^p 1_{(u_i \leq \sigma c + n^{-1/2} ac + z'_{in}\Pi b + n^{-1/2} r'_i b)}, \tag{3.5.2}$$

where $\mathsf{E}_{i-1}(\cdot) = \mathsf{E}(\cdot|\mathcal{F}_{i-1})$. Note that $\overline{\mathsf{F}}^{1,0}_{u,n}(0,0,c) = \mathsf{F}_u(c) = \mathsf{P}(u_i \leq \sigma c)$.

We embed these processes into the space $D[0,1]$ that are processes continuous from the right and with limits of left, where the space is endowed with the Skorokhod metric. We do this as follows. The indicator $1_{(u_i \leq \sigma c)}$ and the distribution function $\mathsf{F}_u(c)$ can be defined as 0 or 1 when c takes the values $-\infty$ and ∞ respectively. We can then define quantiles $c_\psi = \mathsf{F}_u^{-1}(\psi)$ for $0 \leq \psi \leq 1$. Correspondingly we can continously extend the definition of the weighted and marked empirical distribution function and its compensator by chosing $\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,-\infty) = \overline{\mathsf{F}}^{w,p}_{u,n}(a,b,-\infty) = 0$ while we then have $\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,\infty) = n^{-1}\sum_{i=1}^n w_{in} u_i^p$ and $\overline{\mathsf{F}}^{w,p}_{u,n}(a,b,\infty) = n^{-1}\sum_{i=1}^n w_{in} \mathsf{E}_{i-1} u_i^p$. We now define the empirical process for $0 \leq \psi \leq 1$

$$\mathbb{F}^{w,p}_{u,n}(a,b,c_\psi) = n^{1/2}\{\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,c_\psi) - \overline{\mathsf{F}}^{w,p}_{u,n}(a,b,c_\psi)\}. \tag{3.5.3}$$

In the following we first present assumptions, and then results follow for the one-sided process and finally extend to the absolute case.

3.5.1 Assumptions

In this section, the density f_u is not necessarily symmetric. The conditions listed below are weaker than Assumption 3.3.1.

Assumption 3.5.1. *Let $\mathcal{F}_{i-1} = \sigma(z_1, \ldots, z_i, r_1, \ldots, r_{i-1}, u_1, \ldots, u_{i-1})$ be an increasing sequence of σ-fields so u_{i-1}, r_{i-1}, z_i, w_{in} are $\mathcal{F}_{i-1}$ measurable while r_i, u_i are independent of $\mathcal{F}_{i-1}$. Suppose $(u_i/\sigma, \Sigma^{-1/2} r_i)$ have continuously differentiable joint, conditional, and marginal densities $\mathsf{f}_{u,r}(y,x) = \mathsf{f}_{u|r}(y|x)\mathsf{f}_r(x) = \mathsf{f}_{r|u}(x|y)\mathsf{f}_u(y)$ which are positive on $y \in \mathbb{R}$, $x \in \mathbb{R}^{d_x}$. Let p, η, κ be given so $p \in \mathbb{N}_0$, $0 \leq \kappa < \eta \leq 1/4$. Choose $0 < \nu \leq 1$, $s \in \mathbb{N}_0$ such that*

$$2^{s-1} > 1 + (1/4 - \eta)(1 + d_x). \tag{3.5.4}$$

Suppose

(i) the marginal density $\mathsf{f}_u(y)$ satisfies for $y \in \mathbb{R}$

(a) moments: $\int_{-\infty}^{\infty} |y|^{2^s p/\nu} \mathsf{f}_u(y) dy < \infty$;

(b) boundedness: $\sup_{y \in \mathbb{R}} |y^{2^s p+1} \mathsf{f}_u(y) + y^{2^s p+2} \dot{\mathsf{f}}_u(y)| < \infty$;

(c) smoothness: a $C_{\mathrm{H}} \in \mathbb{N}$ exists so that for all $\epsilon > 0$

$$\frac{\sup_{y \geq \epsilon}(1 + y^{2^s p})\mathsf{f}_u(y)}{\inf_{0 \leq y \leq \epsilon}(1 + y^{2^s p})\mathsf{f}_u(y)} \leq C_{\mathrm{H}}, \qquad \frac{\sup_{y \leq -\epsilon}(1 + y^{2^s p})\mathsf{f}_u(y)}{\inf_{-\epsilon \leq y \leq 0}(1 + y^{2^s p})\mathsf{f}_u(y)} \leq C_{\mathrm{H}};$$

(ii) the marginal density $\mathsf{f}_r(x)$ satisfies for $x \in \mathbb{R}^{d_x}$

(a) moments: $\int_{x \in \mathbb{R}^{d_x}} |x|^4 \mathsf{f}_r(x)(dx) < \infty$;

(iii) the joint and conditional densities $\mathsf{f}_{u,r}(y,x), \mathsf{f}_{u|r}(y|x)$ satisfy for $y \in \mathbb{R}, x \in \mathbb{R}^{d_x}$

(a) boundedness: $\sup_{y \in \mathbb{R}, x \in \mathbb{R}^{d_x}} |(1+y) y^{2^s p-1} \mathsf{f}_{u|r}(y|x) + y^{2^s p} \dot{\mathsf{f}}_{u|r,y}(y|x)| < \infty$;

(iv) the instruments z_i satisfy

(a) $\max_{1 \leq i \leq n} |n^{1/2-\kappa} z_{in}| = \mathrm{O}_{\mathsf{P}}(1)$;

(v) the weights w_{in} satisfy

(a) $n^{-1} \mathsf{E} \sum_{i=1}^n |w_{in}|^{2^s} (1 + |n^{1/2} z_{in}|) = \mathrm{O}(1)$;

(b) $n^{-1} \sum_{i=1}^n |w_{in}|(1 + |n^{1/2} z_{in}|^2) = \mathrm{O}_{\mathsf{P}}(1)$.

Remark 3.5.1. *Assumption 3.3.1$(ia, iia, iiia, ivb, ivc)$ implies Assumption 3.5.1 with $s \geq 2$ satisfying (3.5.4) when w_{in} is either of 1, $n^{1/2} z_{in} = z_i$, $n z_{in} z_{in}' = z_i z_i'$ and p is either of 0, 1, 2. Details are given in Lemma 3.8.1 in the appendix.*

3.5.2 Asymptotic results for empirical process

We present three asymptotic results. The first theorem shows that the estimation error for the scale and regression parameters is negligible uniformly in the quantile.

Theorem 3.5.1. *Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1 holds with $\nu = 1$ and $s \geq 2$ satisfying (3.5.4). Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B} |\mathbb{F}_{u,n}^{w,p}(a,b,c_\psi) - \mathbb{F}_{u,n}^{w,p}(0,0,c_\psi)| = \mathrm{o}_{\mathsf{P}}(1).$$

The proof involves a chaining argument. For this, we apply an iterated martingale inequality, see Lemma 3.7.4, to explore the tail behaviour of the maximum of a family of martingales. We can, however, split this argument into two parts. First, we keep b fixed and consider variation in a, c_ψ. Then we keep a fixed and consider variation in b, c_ψ. Combining these two arguments will finally finish the proof.

The second result provides a linearization of the compensator.

Theorem 3.5.2. *Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1$(ia, ib, iia, iiia, vb)$ holds with $\nu = 1$, $s = 0$. Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B} |n^{1/2}\{\overline{\mathsf{F}}_{u,n}^{w,p}(a,b,c_\psi) - \overline{\mathsf{F}}_{u,n}^{w,p}(0,0,c_\psi)\} - \mathcal{B}_{\mathsf{F}_u,n}(a,b,c_\psi)| = \mathrm{O}_{\mathsf{P}}(n^{-2\eta}),$$

where the bias term, with $\xi_{c_\psi} = \mathsf{E}(\Sigma^{-1/2}r_i|u_i/\sigma = c_\psi)$, is defined as

$$\mathcal{B}_{\mathsf{F}_u,n}(a,b,c_\psi) = \sigma^{p-1}c_\psi^p \mathsf{f}_u(c_\psi)n^{-1/2}\sum_{i=1}^n w_{in}(n^{-1/2}ac_\psi + n^{-1/2}\xi_{c_\psi}'\Sigma^{1/2}b + z_{in}'\Pi b).$$

Since x_i and u_i are correlated so for u_i and r_i, this dependent structure requires a more intricate analysis for the compensator. Similar as the proof for uniform convergence in Theorem 3.5.1, we proceed by first considering variation in a, c_ψ while $b = 0$. We then approximate the compensator by combining with the argument for variation in b, c_ψ while $a = 0$.

Finally, we argue the empirical process $\mathbb{F}_{u,n}^{w,p}(0,0,c_\psi)$ is tight when viewed as a sequence in n of processes on $D[0,1]$. By Billingsley (1968, Theorem 13.2), two conditions need to be checked. First, it holds by construction that $\mathbb{F}_{u,n}^{w,p}(0,0,c_0) = 0$ where $c_0 = \mathsf{F}_u^{-1}(0) = -\infty$. Second, the next theorem shows that the modulus of continuity is small. The proof uses a dyadic argument and then apply an iterated martingale exponential inequality, see Lemma 3.7.5, to explore the tail probability of the maximum of a family of martingales.

Theorem 3.5.3. *Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1(ia, va) holds with $0 < \nu < 1$, $s = 2$. Then for all $\epsilon > 0$*

$$\lim_{\phi\downarrow 0}\limsup_{n\to\infty}\mathsf{P}\{\sup_{0\leq\psi\leq\psi^\dagger\leq 1:\psi^\dagger-\psi\leq\phi} |\mathbb{F}_{u,n}^{w,p}(0,0,c_{\psi^\dagger}) - \mathbb{F}_{u,n}^{w,p}(0,0,c_\psi)| > \epsilon\} \to 0.$$

The empirical distribution functions $\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,c)$ and $\widehat{\Pi}^{(m+1)}_c$ involve in the updated estimators $\widehat{\beta}^{(m+1)}_c$, $(\widehat{\sigma}^{(m+1)}_c)^2$, see (3.2.11), (3.2.12). While $n^{1/2}$-convergence results have been established for the empirical process $n^{1/2}\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,c)$ by Theorem 3.5.1, 3.5.2, 3.5.3, consistency for the estimator of Π is then required to build up the asymptotic distribution theory for the iterated estimators of β, σ^2. A product moment appearing in $\widehat{\Pi}^{(m+1)}_c$ but not included in $\widehat{\mathsf{F}}^{w,p}_{u,n}(a,b,c)$ is $n^{-1}\sum_{i=1}^n z_i r'_i v^{(m)}_{i,c}$, see (3.2.9), (3.2.10). Thus, proving consistency of the Π estimator, we need the following n-convergence theorem for a new class of empirical processes with the weight $n^{1/2}z_{in} = z_i$ and mark r'_i instead of mark u_i in the previous case.

Theorem 3.5.4. *Let $c_\psi = \mathsf{F}^{-1}_u(\psi)$. Suppose Assumption 3.5.1 holds with $\nu = 1$ and $s \geq 2$ satisfying (3.5.4). Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B}|n^{-1/2}\sum_{i=1}^n z_{in}r'_i\{1_{(u_i\leq\sigma c_\psi+n^{-1/2}ac_\psi+z'_{in}\Pi b+n^{-1/2}r'_i b)}-1_{(u_i\leq\sigma c_\psi)}\}| = o_{\mathsf{P}}(1).$$

The proof first considers uniform convergence for the empirical process without weights $n^{1/2}z_{in} = z_i$ and marks r_i. Since we establish n-convergence result instead of with the normalization $n^{1/2}$, the Hölder inequality is then used to prove Theorem 3.5.4 by adding the weights and marks into the process with only indicators.

3.5.3 A result for the two-sided empirical process

Estimators in Algorithm 3.2.1, 3.2.2, 3.2.3 involves indicators depending on the absolute value of residuals of structural errors. We therefore present some results for a class of two-sided weighted and marked empirical processes.

Define the weighted and marked absolute empirical distribution function

$$\widehat{\mathsf{G}}^{w,p}_{u,n}(a,b,c) = \frac{1}{n}\sum_{i=1}^n w_{in}u_i^p 1_{(|u_i-z'_{in}\Pi b-n^{-1/2}r'_i b|\leq\sigma c+n^{-1/2}ac)}. \tag{3.5.5}$$

We suppose a such that $\sigma + n^{-1/2}a > 0$, in which case it suffices to consider $c \geq 0$. This restriction on a is satisfied when choosing a as $\widetilde{a} = n^{1/2}(\widetilde{\sigma}-\sigma)$ so that $\sigma+n^{-1/2}\widetilde{a} = \widetilde{\sigma} > 0$. Introduce the compensator of $\widehat{\mathsf{G}}^{w,p}_{u,n}(a,b,c)$

$$\overline{\mathsf{G}}^{w,p}_{u,n}(a,b,c) = \frac{1}{n}\sum_{i=1}^n w_{in}\mathsf{E}_{i-1}u_i^p 1_{(|u_i-z'_{in}\Pi b-n^{-1/2}r'_i b|\leq\sigma c+n^{-1/2}ac)}. \tag{3.5.6}$$

Note that $\overline{\mathsf{G}}^{1,0}_{u,n}(0,0,c) = \mathsf{G}_u(c) = \mathsf{P}(|u_i| \leq \sigma c)$. Similar as defining the one-sided empirical process $\mathbb{F}^{w,p}_{u,n}(a,b,c_\psi)$ on the space $D[0,1]$ equipped with the Skorokhod metric, let

$c_\psi = \mathsf{G}_u^{-1}(\psi)$ for $0 \leq \psi \leq 1$ so $c_\psi \geq 0$ and the absolute empirical process is

$$\mathbb{G}_{u,n}^{w,p}(a,b,c_\psi) = n^{1/2}\{\widehat{\mathsf{G}}_{u,n}^{w,p}(a,b,c_\psi) - \overline{\mathsf{G}}_{u,n}^{w,p}(a,b,c_\psi)\}. \tag{3.5.7}$$

We can now derive asymptotic theory for the absolute empirical process from Theorems 3.5.1, 3.5.2, 3.5.3, 3.5.4. These results are presented under more restrictive Assumption 3.3.1, where the distribution f_u of structural error is symmetric, see Remark 3.5.1. In this section, we only consider w_{in} chosen as 1, $n^{1/2}z_{in} = z_i$, $nz_{in}z_{in}' = z_i z_i'$ and p as 0, 1, 2.

The first theorem states the absolute processes with estimation errors converge to the one without estimation errors uniformly in the quantile.

Theorem 3.5.5. *Let $c_\psi = \mathsf{G}_u^{-1}(\psi)$. Suppose Assumption 3.3.1$(ia, iia, iiia, ivb, ivc)$ holds. Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B} |\mathbb{G}_{u,n}^{w,p}(a,b,c_\psi) - \mathbb{G}_{u,n}^{w,p}(0,0,c_\psi)| = \mathsf{o_P}(1).$$

Then we provide the first-order approximation to the absolute compensator.

Theorem 3.5.6. *Let $c_\psi = \mathsf{G}_u^{-1}(\psi)$. Suppose Assumption 3.3.1$(ia, iia, iiia, ivc)$ holds. Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B} |n^{1/2}\{\overline{\mathsf{G}}_{u,n}^{w,p}(a,b,c_\psi) - \overline{\mathsf{G}}_{u,n}^{w,p}(0,0,c_\psi)\} - \mathcal{B}_{\mathsf{G}_u,n}(a,b,c_\psi)| = \mathsf{O_P}(n^{-2\eta}),$$

where the bias term, with $\xi_{c_\psi} = \mathsf{E}(\Sigma^{-1/2}r_i|u_i/\sigma = c_\psi)$, is defined as

$$\begin{aligned}\mathcal{B}_{\mathsf{G}_u,n}(a,b,c_\psi) = \sigma^{p-1}c_\psi^p \mathsf{f}_u(c_\psi)n^{-1/2}\sum_{i=1}^n w_{in}[\{1+(-1)^p\}n^{-1/2}ac_\psi \\ + n^{-1/2}\{\xi_{c_\psi} - (-1)^p\xi_{-c_\psi}\}'\Sigma^{1/2}b + \{1-(-1)^p\}z_{in}'\Pi b].\end{aligned}$$

The next theorem extends the tightness result to the case of empirical processes with absolute indicators.

Theorem 3.5.7. *Let $c_\psi = \mathsf{G}_u^{-1}(\psi)$. Suppose Assumption 3.3.1(ia, ivc) holds. Then for all $\epsilon > 0$*

$$\lim_{\phi\downarrow 0}\limsup_{n\to\infty}\mathsf{P}\{\sup_{0\leq\psi\leq\psi^\dagger\leq 1:\psi^\dagger-\psi\leq\phi} |\mathbb{G}_{u,n}^{w,p}(0,0,c_{\psi^\dagger}) - \mathbb{G}_{u,n}^{w,p}(0,0,c_\psi)| > \epsilon\} \to 0.$$

While the above theorems deal with $n^{1/2}$-convergence results, the final result considers n-convergence for a new type of absolute processes used to prove consistency of the estimator for Π in the first stage regression (3.2.2).

Theorem 3.5.8. *Let* $c_\psi = \mathsf{G}_u^{-1}(\psi)$. *Suppose Assumption 3.3.1*$(ia, iia, iiia, ivb, ivc)$ *holds. Then for any* $B > 0$ *and as* $n \to \infty$

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B}|n^{-1/2}\sum_{i=1}^{n} z_{in}r_i'\{1_{(|u_i-z_{in}'\Pi b-n^{-1/2}r_i'b|\leq\sigma c_\psi+n^{-1/2}ac_\psi)}-1_{(|u_i|\leq\sigma c_\psi)}\}| = o_{\mathsf{P}}(1).$$

The proofs of empirical processes are given in Appendix 3.8, but first we discuss a metric on $\mathbb{R}$ and provide some useful inequalities applied repeatedly in the chaining argument in Appendix 3.7. Finally, the proofs of the main results follow in Appendix 3.9.

3.6 Conclusion

Since most empirical analyses are affected by atypical observations, when many applied economists apply IVs regressions they first run ordinary 2SLS and use the resulting residuals to find non-outlying observations. They then re-run 2SLS on this subset, subsequently iterating this procedure until they obtain robust results.

In this chapter we analyze this simple robust algorithm asymptotically, and provide the consistent estimation and valid inferential procedures given the cut-off value. We concentrate on cross-sectional i.i.d. data in this chapter, though our analysis can be extended to time series, whether stationary, deterministic trend, or unit root. The results are all derived under the null hypothesis that there are no outliers in the model. Since in this case there is still a positive probability to find outliers using such algorithms, we then propose the concept of gauge - the expected retention rate of falsely discovered outliers - to evaluate their performance.

Future work should focus on building asymptotic theory for the gauge, to give empirical researchers an indirect way of choosing the cut-off. Furthermore, some specification tests could be devised to check whether outliers truly exist with respect to the reference model (3.2.1, 3.2.2), as in Jiao and Pretis (2018).

3.7 Appendix A: A metric on $\mathbb{R}$ and some inequalities

The asymptotic theory uses a chaining argument. This involves a partitioning of the quantile axis using a metric, which is presented first. Then follows some preliminary inequalities including an iterated exponential martingale inequality.

For $x, y \in \mathbb{R}$ define the function

$$J_{i,p}(x,y) = (\frac{u_i}{\sigma})^p \{1_{(u_i/\sigma \le y)} - 1_{(u_i/\sigma \le x)}\}. \tag{3.7.1}$$

Our interest focus on $J_{i,p}(x,y)$ of order 2^s with $s \in \mathbb{N}$. For $y, z \in \mathbb{R}$ note $z^{2^s p}$ is non-negative since $2^s p$ is even for $p \in \mathbb{N}_0$ and $s \in \mathbb{N}$, so introduce a positive and increasing function

$$\mathsf{H}_s(y) = \int_{-\infty}^{y} (1 + z^{2^s p}) \mathsf{f}_u(z) dz. \tag{3.7.2}$$

The derivative of this function is $\dot{\mathsf{H}}_s(y) = (1 + y^{2^s p}) \mathsf{f}_u(y)$. Then, denote the constant

$$H_s = \mathsf{H}_s(\infty) = \int_{-\infty}^{\infty} (1 + z^{2^s p}) \mathsf{f}_u(z) dz, \tag{3.7.3}$$

which is assumed to be finite. Selection of the specific $s \in \mathbb{N}$ will be more clear in proofs of the empirical process results. The intuition of $\mathsf{H}_s(y)$ is obtained through setting $p = 0$ so that $\mathsf{H}_s(y) = 2\mathsf{F}_u(y)$, $\dot{\mathsf{H}}_s(y) = 2\mathsf{f}_u(y)$ and $H_s = 2$. Therefore, $\mathsf{H}_s(y)$ is the generalization of the distribution $\mathsf{F}_u(y) \sim u_i/\sigma$. Define the metric d on $x, y \in \mathbb{R}$ as

$$\mathsf{d}(x,y) = |\mathsf{H}_s(x) - \mathsf{H}_s(y)|. \tag{3.7.4}$$

Notice the metric d is totally bounded on $\mathbb{R}$ since $\mathsf{d}(-\infty, \infty) = H_s < \infty$, and it is a special case of the pseudo metric proposed by Koul and Ossiander (1994). For $x \le y$ and $1 \le q \le s$,

$$0 \le \mathsf{E}_{i-1} J_{i,p}(x,y)^{2^q} = \mathsf{E} J_{i,p}(x,y)^{2^q} < \mathsf{H}_s(y) - \mathsf{H}_s(x) = \mathsf{d}(x,y), \tag{3.7.5}$$

as $|z|^q < 1 + |z|^s$ for $0 \le q \le s$.

In the context of chaining, partition the range of $\mathsf{H}_s(c)$ into K intervals of equal size H_s/K. In other words, use the metric d to partition $\mathbb{R}$ into K intervals by endpoints

$$-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = \infty, \tag{3.7.6}$$

with $c_{-k} = c_0$ for $k \in \mathbb{N}$ so that $\mathsf{d}(c_{k-1}, c_k) = H_s/K$ for $1 \le k \le K$.

We first prove an inequality for the difference of two indicators, which will be used to control variation within the chaining interval.

Lemma 3.7.1. *For any e, e_0 on the u-axis and $d \ge |e - e_0|$, we have an inequality*

$$|1_{(u \le e)} - 1_{(u \le e_0)}| \le 1_{(e_0 - d \le u \le e_0 + d)}.$$

Proof of Lemma 3.7.1. From $|e - e_0| \leq d$ we find $e_0 - d \leq e \leq e_0 + d$. Suppose $e \geq e_0$, then using the monotonicity of the indicator function in e and e_0, we have

$$|1_{(u \leq e)} - 1_{(u \leq e_0)}| = 1_{(e_0 \leq u \leq e)} \leq 1_{(e_0 - d \leq u \leq e_0 + d)}.$$

The similar argument applies for $e < e_0$, so we also have

$$|1_{(u \leq e)} - 1_{(u \leq e_0)}| = 1_{(e \leq u \leq e_0)} \leq 1_{(e_0 - d \leq u \leq e_0 + d)}. \qquad \blacksquare$$

The next lemma is the bias correction for differences in expectations involving two dependent random variables, which will be used to linearize the compensator.

Lemma 3.7.2. *Let (Y, X) be a random vector with dimension $1 + d_x$ with the joint density $\mathsf{m}(y, x)$ for $y \in \mathbb{R}, x \in \mathbb{R}^{d_x}$. Suppose the joint density decomposes into the conditional and marginal densities as $\mathsf{m}(y, x) = \mathsf{m}_{Y|X}(y|x)\mathsf{m}_X(x) = \mathsf{m}_{X|Y}(x|y)\mathsf{m}_Y(y)$ and its derivative for y exists so $\dot{\mathsf{m}}_y(y, x) = \dot{\mathsf{m}}_{Y|X,y}(y|x)\mathsf{m}_X(x)$. Let $b_1 \in \mathbb{R}$, $b_2 \in \mathbb{R}^{d_x}$, and $p \in \mathbb{N}_0$. Assume $\mathsf{E}|X|^2 < \infty$, then*

$$\begin{aligned}
&\sup_{c \in \mathbb{R}} |\mathsf{E}Y^p\{1_{(Y \leq c + b_1 + X'b_2)} - 1_{(Y \leq c)}\} - c^p \mathsf{m}_Y(c)\{b_1 + \mathsf{E}(X'b_2 | Y = c)\}| \\
\leq &(\frac{1}{2}b_1^2 + |b_1||b_2|\mathsf{E}|X| + \frac{1}{2}|b_2|^2\mathsf{E}|X|^2) \sup_{y \in \mathbb{R}, x \in \mathbb{R}^{d_x}} |py^{p-1}\mathsf{m}_{Y|X}(y|x) + y^p \dot{\mathsf{m}}_{Y|X,y}(y|x)|.
\end{aligned}$$

Proof of Lemma 3.7.2. Let $\mathcal{E} = \mathsf{E}Y^p\{1_{(Y \leq c + b_1 + X'b_2)} - 1_{(Y \leq c)}\}$, and write the expectation as an integral

$$\mathcal{E} = \iint_{y \in \mathbb{R}, x \in \mathbb{R}^{d_x}} y^p 1_{(c \leq y \leq c + b_1 + x'b_2)} \mathsf{m}(y, x) dy(dx) = \int_{x \in \mathbb{R}^{d_x}} \int_c^{c + b_1 + x'b_2} y^p \mathsf{m}(y, x) dy(dx).$$

Apply Taylor expansion at point c to the inner integral to obtain

$$\int_c^{c + b_1 + x'b_2} y^p \mathsf{m}(y, x) dy = (b_1 + x'b_2)c^p \mathsf{m}(c, x) + \frac{1}{2}(b_1 + x'b_2)^2\{p\tilde{c}^{p-1}\mathsf{m}(\tilde{c}, x) + \tilde{c}^p \dot{\mathsf{m}}_y(\tilde{c}, x)\},$$

where $|\tilde{c} - c| \leq |b_1 + x'b_2|$. Then $\mathcal{E}$ is approximated by the first order term $\mathcal{E}_1$ with the remainder term $\mathcal{E}_2$ left so $\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2$ where

$$\begin{aligned}
\mathcal{E}_1 &= \int_{\mathbb{R}^{d_x}} (b_1 + x'b_2)c^p \mathsf{m}(c, x)(dx), \\
\mathcal{E}_2 &= \frac{1}{2} \int_{\mathbb{R}^{d_x}} (b_1 + x'b_2)^2\{p\tilde{c}^{p-1}\mathsf{m}(\tilde{c}, x) + \tilde{c}^p \dot{\mathsf{m}}_y(\tilde{c}, x)\}(dx).
\end{aligned}$$

Analyze the first order term $\mathcal{E}_1$ to get

$$\mathcal{E}_1 = c^p \mathsf{m}_Y(c)b_1 + c^p \mathsf{m}_Y(c) \int_{\mathbb{R}^{d_x}} x'b_2 \mathsf{m}_{X|Y}(x|c)(dx) = c^p \mathsf{m}_Y(c)\{b_1 + \mathsf{E}(X'b_2 | Y = c)\}.$$

For the remainder term $\mathcal{E}_2 = \mathcal{E} - \mathcal{E}_1$, apply the triangle inequality so

$$|\mathcal{E}_2| \leq \frac{1}{2}\int_{\mathbb{R}^{d_x}} \{b_1^2 + 2|b_1 x' b_2| + (x'b_2)^2\}|p\tilde{c}^{p-1}\mathsf{m}_{Y|X}(\tilde{c}|x) + \tilde{c}^p \dot{\mathsf{m}}_{Y|X,y}(\tilde{c}|x)|\mathsf{m}_X(x)(dx).$$

As $|b_1 x' b_2| \leq |b_1||b_2||x|$ and $(x'b_2)^2 \leq |x|^2|b_2|^2$, so

$$|\mathcal{E}_2| \leq (\frac{1}{2}b_1^2 + |b_1||b_2|\mathsf{E}|X| + \frac{1}{2}|b_2|^2\mathsf{E}|X|^2) \sup_{y\in\mathbb{R},x\in\mathbb{R}^{d_x}} |py^{p-1}\mathsf{m}_{Y|X}(y|x) + y^p\dot{\mathsf{m}}_{Y|X,y}(y|x)|. \ \blacksquare$$

The next inequality provides a tightness type result for the metric d, which will be used to prove the tightness property of empirical process $\mathbb{F}_{u,n}^{w,p}(0,0,c_\psi)$.

Lemma 3.7.3. *(Johansen and Nielsen, 2016a, Lemma B.3). Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1(ia) holds with $0 < \nu < 1$, $s \in \mathbb{N}_0$. Then there exist C_ν, $0 < \phi_0 < 1$ so that for $0 \leq \phi \leq \phi_0$ it follows $\sup_{0\leq\psi\leq 1-\phi} \mathsf{d}(c_\psi, c_{\psi+\phi}) \leq C_\nu \phi^{1-\nu}$.*

The chaining argument involves the tail behaviour of the maximum of a family of martingales which can be controlled using the iterated exponential martingale inequality from Johansen and Nielsen (2016a). It builds on an exponential martingale inequality derived by Bercu and Touati (2008, Theorem 2.1) which relaxes the need of boundedness for martingale difference series but required in Freedman (1975) inequality. The following are two special cases of iterated exponential martingale inequalities, where the number of elements in the martingale family is increasing and where it is fixed. The first inequality will be used to demonstrate the uniform convergence result for empirical processes, while the second is for handling the tightness property.

Lemma 3.7.4. *(Johansen and Nielsen, 2016a, Theorem 5.2) For l so $1 \leq l \leq L$, let $z_{l,i}$ be $\mathcal{F}_i$ adapted satisfying $\mathsf{E}z_{l,i}^{2^s} < \infty$ for some $s \in \mathbb{N}$. Let $D_q = \max_{1\leq l\leq L}\sum_{i=1}^n \mathsf{E}_{i-1}z_{l,i}^{2^q}$ for $1 \leq q \leq s$. Suppose, for some $\varsigma \geq 0$, $\lambda > 0$, that $L = \mathrm{O}(n^\lambda)$ and $\mathsf{E}D_q = \mathrm{O}(n^\varsigma)$ for $1 \leq q \leq s$. If $\upsilon > 0$ is chosen such that*

(i) $\varsigma < 2\upsilon$;

(ii) $\varsigma + \lambda < \upsilon 2^s$;

then, we have for all $\kappa > 0$

$$\lim_{n\to\infty} \mathsf{P}\{\max_{1\leq l\leq L} |\sum_{i=1}^n (z_{l,i} - \mathsf{E}_{i-1}z_{l,i})| > \kappa n^\upsilon\} = 0.$$

Lemma 3.7.5. *(Johansen and Nielsen, 2016a, Theorem 5.3) For l so $1 \leq l \leq L$, let $z_{l,i}$ be $\mathcal{F}_i$ adapted satisfying $\mathsf{E}z_{l,i}^4 < \infty$. Let $D_q = \max_{1\leq l\leq L}\sum_{i=1}^n \mathsf{E}_{i-1}z_{l,i}^{2q}$ and suppose $\mathsf{E}D_q \leq Dn$ for $q = 1, 2$ and some $D > 0$. Then for all $\theta, \kappa > 0$*

$$\mathsf{P}\{\max_{1\leq l\leq L} |\sum_{i=1}^n (z_{l,i} - \mathsf{E}_{i-1}z_{l,i})| > \kappa n^{1/2}\} \leq \frac{(L+1)\theta^3 D}{\kappa n} + \frac{\theta D}{\kappa} + 4L\exp(-\frac{\kappa\theta}{14}).$$

3.8 Appendix B: Proofs of the empirical process results

We first prove uniform $n^{1/2}$-convergence and tightness for the one-sided empirical processes, then linearize the compensator term and show n-convergence for a specific type of the process. Finally, extend all results to the processes considering absolute residuals.

3.8.1 Uniform $n^{1/2}$-convergence of empirical process

We first consider Theorem 3.8.1 which allows variation in a, b but c fixed. Then use the argument in Johansen and Nielsen (2016b, Theorem 4.1) to extend the result to Theorem 3.8.2 where we have variation in b, c but $a = 0$, also see Berenguer-Rico, Johansen, and Nielsen (2019a, Theorem 3.1), while use Jiao and Nielsen (2017, Theorem 5) to prove Theorem 3.8.3 showing convergence uniformly in a, c but $b = 0$. Finally, by combining Theorem 3.8.2, 3.8.3 the uniform convergence in a, b, c follows.

Theorem 3.8.1. *Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1$(ia, iia, iiia, v)$ holds with $\nu = 1$ and $s \geq 1$ such that $2^{s-1} > 1/2 + (1/4 - \eta)(2 + d_x)$. Then for any $0 \leq \psi \leq 1$, $B > 0$ and as $n \to \infty$*

$$\sup_{|a|,|b| \leq n^{1/4-\eta}B} |\mathbb{F}_{u,n}^{w,p}(a, b, c_\psi) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_\psi)| = \mathsf{o_P}(1).$$

Proof of Theorem 3.8.1. Without loss of generality let $\sigma = 1$, $\Sigma^{1/2} = I_{d_x}$. Denote $R_n(a, b, c_\psi) = \mathbb{F}_{u,n}^{w,p}(a, b, c_\psi) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_\psi)$ so we want to show $\mathcal{R}_{\psi,n} = \mathsf{o_P}(1)$ for any $0 \leq \psi \leq 1$ as $n \to \infty$ where $\mathcal{R}_{\psi,n} = \sup_{|a|,|b| \leq n^{1/4-\eta}B} |R_n(a, b, c_\psi)|$. Throughout denote $C > 0$ as usual a constant which may have different values in different expressions.

1. *Construct a, b-balls.* For $\delta, n > 0$, cover the set $|a|, |b| \leq n^{1/4-\eta}B$ with balls of radius δ and centers a_j, b_m. The numbers of a, b-balls are $J = n^{1/4-\eta}B/\delta = \mathrm{O}(n^{1/4-\eta}/\delta)$ and $M = (n^{1/4-\eta}B/\delta)^{d_x} = \mathrm{O}(n^{(1/4-\eta)d_x}/\delta)$ respectively. Thus, for any a, b there exist a_j, b_m so that $|a - a_j| \leq \delta$ and $|b - b_m| \leq \delta$.

2. *Apply chaining.* Write $R_n(a, b, c_\psi) = R_n(a_j, b_m, c_\psi) + \{R_n(a, b, c_\psi) - R_n(a_j, b_m, c_\psi)\}$ where $R_n(a_j, b_m, c_\psi)$ is a discrete point term while $R_n(a, b, c_\psi) - R_n(a_j, b_m, c_\psi)$ is a local oscillation term. The triangle inequality gives $\mathcal{R}_{\psi,n} \leq \mathcal{R}_{\psi,n,1} + \mathcal{R}_{\psi,n,2}$ where

$$\mathcal{R}_{\psi,n,1} = \max_{1 \leq j \leq J, 1 \leq m \leq M} |R_n(a_j, b_m, c_\psi)|,$$

$$\mathcal{R}_{\psi,n,2} = \max_{1 \leq j \leq J, 1 \leq m \leq M} \sup_{|a - a_j| \leq \delta, |b - b_m| \leq \delta} |R_n(a, b, c_\psi) - R_n(a_j, b_m, c_\psi)|.$$

It then suffices to show that $\mathcal{R}_{\psi,n,1}, \mathcal{R}_{\psi,n,2}$ are $\mathsf{o_P}(1)$.

3. *The discrete point term $\mathcal{R}_{\psi,n,1}$ is* $\mathsf{o_P}(1)$. Recall the notation $J_{i,p}$ in (3.7.1), then write $R_n(a_j, b_m, c_\psi) = n^{-1/2}\sum_{i=1}^n (z_{l,i} - \mathsf{E}_{i-1} z_{l,i})$ where

$$z_{l,i} = w_{in} J_{i,p}(c_\psi, c_\psi + n^{-1/2} a_j c_\psi + z_{in}' \Pi b_m + n^{-1/2} r_i' b_m).$$

Use Lemma 3.7.4 with $\upsilon = 1/2$, index $l = (j, m)$ and $L = JM$ to prove $\mathcal{R}_{\psi,n,1} = \mathsf{o_P}(1)$, and we need to verify its conditions.

The moment $\mathsf{E} z_{l,i}^{2^s} < \infty$ for $s \in \mathbb{N}$. Bounding the difference of indicator functions by unity and using independence of u_i with $\mathcal{F}_{i-1}$ gives $\mathsf{E} z_{l,i}^{2^s} \leq \mathsf{E}|w_{in}|^{2^s} u_i^{2^s p} = \mathsf{E}|w_{in}|^{2^s} \mathsf{E} u_i^{2^s p}$. This is finite by Assumption 3.5.1(ia, va).

The parameter λ. The set of indices l has the size $L = JM$. Since $J = \mathrm{O}(n^{1/4-\eta}/\delta)$ and $M = \mathrm{O}(n^{(1/4-\eta)d_x}/\delta)$ while δ is now treated as a constant then $L = \mathrm{O}(n^\lambda)$ where $\lambda = (1/4 - \eta)(1 + d_x)$.

The parameter ς. Consider $1 \leq q \leq s$. Notice we have

$$|1_{(u_i \leq c_\psi + n^{-1/2} a_j c_\psi + z_{in}' \Pi b_m + n^{-1/2} r_i' b_m)} - 1_{(u_i \leq c_\psi)}| \leq 1_{(\underline{c}_{i\psi jm,r_i} \leq u_i \leq \overline{c}_{i\psi jm,r_i})},$$

where we denote

$$\begin{aligned}
\underline{c}_{i\psi jm,r_i} &= c_\psi - n^{-1/2}|a_j||c_\psi| - |z_{in}||\Pi||b_m| - n^{-1/2}|r_i||b_m|, \\
\overline{c}_{i\psi jm,r_i} &= c_\psi + n^{-1/2}|a_j||c_\psi| + |z_{in}||\Pi||b_m| + n^{-1/2}|r_i||b_m|.
\end{aligned}$$

The above follows from Lemma 3.7.1 by $e = c_\psi + n^{-1/2} a_j c_\psi + z_{in}' \Pi b_m + n^{-1/2} r_i' b_m$, $e_0 = c_\psi$, $d = n^{-1/2}|a_j||c_\psi| + |z_{in}||\Pi||b_m| + n^{-1/2}|r_i||b_m| \geq |n^{-1/2} a_j c_\psi + z_{in}' \Pi b_m + n^{-1/2} r_i' b_m| = |e - e_0|$. Let $\mathcal{E}_i = \mathsf{E}_{i-1} J_{i,p}^{2^q}(c_\psi, c_\psi + n^{-1/2} a_j c_\psi + z_{in}' \Pi b_m + n^{-1/2} r_i' b_m)$, then by applying the above inequality concerning the difference of indicator functions we have

$$\mathcal{E}_i \leq \mathsf{E}_{i-1} J_{i,p}^{2^q}(\underline{c}_{i\psi jm,r_i}, \overline{c}_{i\psi jm,r_i}) = \iint_{y \in \mathbb{R}, x \in \mathbb{R}^{d_x}} y^{2^q p} 1_{(\underline{c}_{i\psi jm,x} \leq y \leq \overline{c}_{i\psi jm,x})} \mathsf{f}_{u,r}(y, x) dy (dx).$$

Decompose the joint density into the conditional and marginal ones, then apply the mean value theorem to the inner integral to get

$$\begin{aligned}
\mathcal{E}_i &\leq \int_{x \in \mathbb{R}^{d_x}} \{ \int_{\underline{c}_{i\psi jm,x}}^{\overline{c}_{i\psi jm,x}} (1 + y^{2^s p}) \mathsf{f}_{u|r}(y|x) dy \} \mathsf{f}_r(x) (dx) \\
&= \int_{x \in \mathbb{R}^{d_x}} (\overline{c}_{i\psi jm,x} - \underline{c}_{i\psi jm,x})(1 + \tilde{c}^{2^s p}) \mathsf{f}_{u|r}(\tilde{c}|x) \mathsf{f}_r(x) (dx),
\end{aligned}$$

for an intermediate point $\tilde{c}$ such that $\underline{c}_{i\psi jm,x} \leq \tilde{c} \leq \overline{c}_{i\psi jm,x}$. Insert $\underline{c}_{i\psi jm,x}, \overline{c}_{i\psi jm,x}$ to get

$$\mathcal{E}_i \leq 2 \sup_{y \in \mathbb{R}, x \in \mathbb{R}^{d_x}} (1 + y^{2^s p}) \mathsf{f}_{u|r}(y|x) \int_{x \in \mathbb{R}^{d_x}} n^{-1/2} (|a_j||c_\psi| + |n^{1/2} z_{in}||\Pi||b_m| + |x||b_m|) \mathsf{f}_r(x) (dx).$$

Note $|a_j|, |b_m| \leq n^{1/4-\eta}B$ for each $1 \leq j \leq J$, $1 \leq m \leq M$ while c_ψ, Π are fixed. Furthermore, since Assumption 3.5.1$(iia, iiia)$ that $\int_{x\in\mathbb{R}^{d_x}} |x|\mathsf{f}_r(x)(dx) < \infty$ and that $\sup_{y\in\mathbb{R},x\in\mathbb{R}^{d_x}}(1+y^{2^s p})\mathsf{f}_{u|r}(y|x) < \infty$, then there exists $C > 0$ so uniformly in j, m

$$\mathcal{E}_i \leq Cn^{-1/4-\eta}\int_{x\in\mathbb{R}^{d_x}}(1+|n^{1/2}z_{in}|+|x|)\mathsf{f}_r(x)(dx) \leq Cn^{-1/4-\eta}(1+|n^{1/2}z_{in}|).$$

Insert the bound of $\mathcal{E}_i$ to get

$$D_q = \max_{1\leq l\leq L}\sum_{i=1}^n \mathsf{E}_{i-1}z_{l,i}^{2q} = \max_{1\leq j\leq J, 1\leq m\leq M}\sum_{i=1}^n |w_{in}|^{2^q}\mathcal{E}_i \leq Cn^{-1/4-\eta}\sum_{i=1}^n |w_{in}|^{2^q}(1+|n^{1/2}z_{in}|).$$

Since $|w_{in}|^{2q} \leq (1+|w_{in}|^{2^s})$ and $\mathsf{E}\sum_{i=1}^n(1+|w_{in}|^{2^s})(1+|n^{1/2}z_{in}|) = \mathrm{O}(n)$ by Assumption 3.5.1(va), then we have $\mathsf{E}D_q = \mathrm{O}(n^\varsigma)$ where $\varsigma = 3/4-\eta$.

Condition (i) is that $\varsigma < 2\upsilon$. This holds since $\eta > 0$ so $\varsigma = 3/4-\eta < 1 = 2\upsilon$.

Condition (ii) is that $\varsigma + \lambda < \upsilon 2^s$. We have

$$\varsigma + \lambda = 3/4-\eta+(1/4-\eta)(1+d_x) = 1/2+(1/4-\eta)(2+d_x).$$

We choose s so that $\varsigma+\lambda < \upsilon 2^s = 2^{s-1}$.

4. *Decompose the oscillation term* $\mathcal{R}_{\psi,n,2}$. As the terms $\mathbb{F}_{u,n}^{w,p}(0,0,c_\psi)$ in $R_n(a,b,c_\psi)$ and $R_n(a_j,b_m,c_\psi)$ cancel out, then we have

$$R_n(a,b,c_\psi) - R_n(a_j,b_m,c_\psi) = \mathbb{F}_{u,n}^{w,p}(a,b,c_\psi) - \mathbb{F}_{u,n}^{w,p}(a_j,b_m,c_\psi).$$

Let $s_{i,p}(a,a_j,b,b_m,c_\psi)$ denote the term

$$w_{in}u_i^p\{1_{(u_i\leq c_\psi+n^{-1/2}ac_\psi+z_{in}'\Pi b+n^{-1/2}r_i'b)} - 1_{(u_i\leq c_\psi+n^{-1/2}a_jc_\psi+z_{in}'\Pi b_m+n^{-1/2}r_i'b_m)}\},$$

it then follows

$$R_n(a,b,c_\psi) - R_n(a_j,b_m,c_\psi) = n^{-1/2}\sum_{i=1}^n\{s_{i,p}(a,a_j,b,b_m,c_\psi) - \mathsf{E}_{i-1}s_{i,p}(a,a_j,b,b_m,c_\psi)\}.$$

Choose $e = c_\psi+n^{-1/2}ac_\psi+z_{in}'\Pi b+n^{-1/2}r_i'b$ and $e_0 = c_\psi+n^{-1/2}a_jc_\psi+z_{in}'\Pi b_m+n^{-1/2}r_i'b_m$, then find $|e-e_0| = |n^{-1/2}(a-a_j)c_\psi + z_{in}'\Pi(b-b_m)+n^{-1/2}r_i'(b-b_m)|$. For any j,m it holds $|a-a_j| \leq \delta$, $|b-b_m| \leq \delta$, so $|e-e_0| \leq n^{-1/2}\delta(|c_\psi|+|n^{1/2}z_{in}||\Pi|+|r_i|) = d$. Thus, apply Lemma 3.7.1 to bound the difference of two indicators so that

$$|1_{(u_i\leq c_\psi+n^{-1/2}ac_\psi+z_{in}'\Pi b+n^{-1/2}r_i'b)} - 1_{(u_i\leq c_\psi+n^{-1/2}a_jc_\psi+z_{in}'\Pi b_m+n^{-1/2}r_i'b_m)}| \leq 1_{(\underline{c}_{i\psi jm,r_i}\leq u_i\leq \bar{\bar{c}}_{i\psi jm,r_i})},$$

where we deonote

$$\underline{c}_{i\psi jm,r_i} = c_\psi + n^{-1/2}a_jc_\psi + z_{in}'\Pi b_m + n^{-1/2}r_i'b_m - n^{-1/2}\delta(|c_\psi|+|n^{1/2}z_{in}||\Pi|+|r_i|),$$
$$\bar{\bar{c}}_{i\psi jm,r_i} = c_\psi + n^{-1/2}a_jc_\psi + z_{in}'\Pi b_m + n^{-1/2}r_i'b_m + n^{-1/2}\delta(|c_\psi|+|n^{1/2}z_{in}||\Pi|+|r_i|).$$

Apply the above inequality to get $|s_{i,p}(a,a_j,b,b_m,c_\psi)| \leq |w_{in}||J_{i,p}(\underline{c}_{\equiv i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})|$, then it follows by the triangle and Jensen's inequality that

$$\begin{aligned}&|R_n(a,b,c_\psi) - R_n(a_j,b_m,c_\psi)| \\ \leq& n^{-1/2}\sum_{i=1}^n |w_{in}|\{|J_{i,p}(\underline{c}_{\equiv i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})| + \mathsf{E}_{i-1}|J_{i,p}(\underline{c}_{\equiv i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})|\}.\end{aligned}$$

Decompose the oscillation term $\mathcal{R}_{\psi,n,2}$ into a martingale and two compensators such that $\mathcal{R}_{\psi,n,2} \leq \widetilde{\mathcal{R}}_{\psi,n,2} + 2\overline{\mathcal{R}}_{\psi,n,2}$, where

$$\begin{aligned}\widetilde{\mathcal{R}}_{\psi,n,2} &= \max_{1\leq j\leq J,1\leq m\leq M} n^{-1/2}\sum_{i=1}^n |w_{in}|\{|J_{i,p}(\underline{c}_{\equiv i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})| - \mathsf{E}_{i-1}|J_{i,p}(\underline{c}_{\equiv i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})|\} \\ \overline{\mathcal{R}}_{\psi,n,2} &= \max_{1\leq j\leq J,1\leq m\leq M} n^{-1/2}\sum_{i=1}^n |w_{in}|\mathsf{E}_{i-1}|J_{i,p}(\underline{c}_{\equiv i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})|.\end{aligned}$$

Thus, it suffices to prove that $\widetilde{\mathcal{R}}_{\psi,n,2}$ and $\overline{\mathcal{R}}_{\psi,n,2}$ are $\mathsf{o_P}(1)$.

5. *Bounding the compensator* $\overline{\mathcal{R}}_{\psi,n,2}$. Let $\overline{\mathcal{E}}_i = \mathsf{E}_{i-1}|J_{i,p}(\underline{c}_{\equiv i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})|$, then it is clear to see $\overline{\mathcal{E}}_i \leq \mathcal{H}_i = \mathsf{E}_{i-1}(1+u_i^{2^sp})1_{(\underline{c}_{\equiv i\psi jm,r_i}\leq u_i\leq\overline{\overline{c}}_{i\psi jm,r_i})}$ for $s \geq 1$, and thus our aim now is to bound $\mathcal{H}_i$. Write $\mathcal{H}_i$ as an integral and decompose the joint density so that

$$\begin{aligned}\mathcal{H}_i &= \iint_{y\in\mathbb{R},x\in\mathbb{R}^{d_x}} (1+y^{2^sp})1_{(\underline{c}_{\equiv i\psi jm,x}\leq y\leq\overline{\overline{c}}_{i\psi jm,x})}\mathsf{f}_{u,r}(y,x)dy(dx) \\ &= \int_{x\in\mathbb{R}^{d_x}}\{\int_{\underline{c}_{\equiv i\psi jm,x}}^{\overline{\overline{c}}_{i\psi jm,x}} (1+y^{2^sp})\mathsf{f}_{u|r}(y|x)dy\}\mathsf{f}_r(x)(dx).\end{aligned}$$

Apply the mean value theorem to the inner integral to get

$$\mathcal{H}_i = \int_{x\in\mathbb{R}^{d_x}} (\overline{\overline{c}}_{i\psi jm,x} - \underline{c}_{\equiv i\psi jm,x})(1+\tilde{\tilde{c}}^{2^sp})\mathsf{f}_{u|r}(\tilde{\tilde{c}}|x)\mathsf{f}_r(x)(dx),$$

where $\tilde{\tilde{c}}$ is an intermediate point so that $\underline{c}_{\equiv i\psi jm,x} \leq \tilde{\tilde{c}} \leq \overline{\overline{c}}_{i\psi jm,x}$. Insert $\underline{c}_{\equiv i\psi jm,x}$, $\overline{\overline{c}}_{i\psi jm,x}$ into $\mathcal{H}_i$ and bound the function $(1+y^{2^sp})\mathsf{f}_{u|r}(y|x)$, then it follows

$$\mathcal{H}_i \leq 2n^{-1/2}\delta \sup_{y\in\mathbb{R},x\in\mathbb{R}^{d_x}} (1+y^{2^sp})\mathsf{f}_{u|r}(y|x)\int_{x\in\mathbb{R}^{d_x}} (|c_\psi| + |n^{1/2}z_{in}||\Pi| + |x|)\mathsf{f}_r(x)(dx).$$

Due to Assumption 3.5.1$(iia, iiia)$ and that c_ψ, Π are fixed, it holds uniformly in $1 \leq j \leq J$, $1 \leq m \leq M$ that $\mathcal{H}_i \leq Cn^{-1/2}\delta(1+|n^{1/2}z_{in}|)$. Therefore, recall $\overline{\mathcal{E}}_i \leq \mathcal{H}_i$ then we find

$$\overline{\mathcal{R}}_{\psi,n,2} = \max_{1\leq j\leq J,1\leq m\leq M} n^{-1/2}\sum_{i=1}^n |w_{in}|\overline{\mathcal{E}}_i \leq Cn^{-1/2}\delta n^{-1/2}\sum_{i=1}^n |w_{in}|(1+|n^{1/2}z_{in}|).$$

Assumption 3.5.1(vb) that $n^{-1}\sum_{i=1}^{n}|w_{in}|(1+|n^{1/2}z_{in}|)=\mathrm{O}_{\mathsf{P}}(1)$ shows $\overline{\mathcal{R}}_{\psi,n,2}=\delta\mathrm{O}_{\mathsf{P}}(1)$, where the $\mathrm{O}_{\mathsf{P}}(1)$-term does not depend on δ. Choose δ sufficiently small then it follows $\overline{\mathcal{R}}_{\psi,n,2}=\mathrm{o}_{\mathsf{P}}(1)$.

6. *The martingale $\widetilde{\mathcal{R}}_{\psi,n,2}$ is* $\mathrm{o}_{\mathsf{P}}(1)$. Note that once δ is selected in item 5, it is then treated as constant so does not play any particular role in other items. To show $\widetilde{\mathcal{R}}_{\psi,n,2}=\mathrm{o}_{\mathsf{P}}(1)$, use Lemma 3.7.4 with $\upsilon=1/2$ and $z_{l,i}=|w_{in}||J_{i,p}(\underline{c}_{\vec{=}i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})|$ where the index $l=(j,m)$ has the size $L=JM$. Two conditions need to shown.

The moment $\mathsf{E}z_{l,i}^{2^s}<\infty$ for $s\in\mathbb{N}$. This can be shown by the same argument used in the item 3.

The parameter λ. The size $L=\mathrm{O}(n^{\lambda})$ where $\lambda=(1/4-\eta)(1+d_x)$, since $L=JM$ and $J=\mathrm{O}(n^{1/4-\eta})$, $M=\mathrm{O}(n^{(1/4-\eta)d_x})$.

The parameter ς. Consider $1\leq q\leq s$ and let $\widetilde{\mathcal{E}}_i=\mathsf{E}_{i-1}J_{i,p}^{2^q}(\underline{c}_{\vec{=}i\psi jm,r_i},\overline{\overline{c}}_{i\psi jm,r_i})$. Item 5 shows the bound of $\mathcal{H}_i$, so we find $\widetilde{\mathcal{E}}_i\leq\mathcal{H}_i\leq Cn^{-1/2}(1+|n^{1/2}z_{in}|)$ uniformly in j,m. Insert the bound of $\widetilde{\mathcal{E}}_i$ to get

$$D_q=\max_{1\leq l\leq L}\sum_{i=1}^{n}\mathsf{E}_{i-1}z_{l,i}^{2^q}=\max_{1\leq j\leq J,1\leq m\leq M}\sum_{i=1}^{n}|w_{in}|^{2^q}\widetilde{\mathcal{E}}_i\leq Cn^{-1/2}\sum_{i=1}^{n}|w_{in}|^{2^q}(1+|n^{1/2}z_{in}|).$$

Due to $|w_{in}|^{2^q}\leq(1+|w_{in}|^{2^s})$ and Assumption 3.5.1(va), we have $\mathsf{E}D_q=\mathrm{O}(n^{\varsigma})$ where $\varsigma=1/2$.

Condition (i) holds, since $\varsigma=1/2<1=2\upsilon$.

Condition (ii) holds, since

$$\varsigma+\lambda=1/2+(1/4-\eta)(1+d_x)\leq 1/2+(1/4-\eta)(2+d_x)<2^{s-1}=\upsilon 2^s.$$

Notice that the first inequality above is due to $\eta\leq 1/4$ and the number of moments s is selected in the item 3 such that $2^{s-1}\geq 1/2+(1/4-\eta)(2+d_x)$. ■

Theorem 3.8.2. *Let $c_\psi=\mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1($i,iia,iiia,iva,va$) holds with $\nu=1$ and $s\geq 2$ satisfying (3.5.4). Then for any $B>0$ and as $n\to\infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|b|\leq n^{1/4-\eta}B}|\mathbb{F}_{u,n}^{w,p}(0,b,c_\psi)-\mathbb{F}_{u,n}^{w,p}(0,0,c_\psi)|=\mathrm{o}_{\mathsf{P}}(1).$$

Proof of Theorem 3.8.2. Set $a=0$ and then apply the argument by Johansen and Nielsen (2016a, Theorem 4.1) and Berenguer-Rico, Johansen, and Nielsen (2019a, Theorem 3.1) to extend pointwise convergence shown in Theorem 3.8.1 to convergence uniformly in ψ. It further requires chaining for the c_ψ-axis, so for $\delta,n>0$ use the metric d to partition the quantile axis as laid out in (3.7.6) with $K=\mathrm{int}(H_s n^{1/2}/\delta)$ where

$s \geq 2$ such that (3.5.4) $2^{s-1} > 1 + (1/4 - \eta)(1 + d_x)$ holds. Notice $s \geq 2$ satisfying $2^{s-1} > 1 + (1/4 - \eta)(1 + d_x)$ implies $s \geq 1$ and $2^{s-1} > 1/2 + (1/4 - \eta)(2 + d_x)$, so we need the higher moment condition on s for proving the uniform convergence. ■

Theorem 3.8.3. *Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1(ia, ib, v) holds with $\nu = 1$, $s = 2$. Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0 \leq \psi \leq 1} \sup_{|a| \leq n^{1/4-\eta} B} |\mathbb{F}_{u,n}^{w,p}(a, 0, c_\psi) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_\psi)| = \mathsf{o}_{\mathsf{P}}(1).$$

Proof of Theorem 3.8.3. For $\delta, n > 0$ use the metric d to partition the c_ψ-axis as laid out in (3.7.6) with $K = \text{int}(H_s n^{1/2}/\delta)$ where $s = 2$, then argue as the proof of Theorem 5 in Jiao and Nielsen (2017) to show convergence uniformly in ψ, a. ■

Proof of Theorem 3.5.1. The term of interest is $\mathcal{W} = \mathbb{F}_{u,n}^{w,p}(a, b, c_\psi) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_\psi)$. Denote $c_{\psi^\dagger} = c_\psi(1 + n^{-1/2}a/\sigma)$. Notice that $\mathbb{F}_{u,n}^{w,p}(a, b, c_\psi) = \mathbb{F}_{u,n}^{w,p}(0, b, c_{\psi^\dagger})$ so it follows that $\mathcal{W} = \mathbb{F}_{u,n}^{w,p}(0, b, c_{\psi^\dagger}) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_\psi)$. Add and subtract $\mathbb{F}_{u,n}^{w,p}(a, 0, c_\psi) = \mathbb{F}_{u,n}^{w,p}(0, 0, c_{\psi^\dagger})$ and apply the triangle inequality to get

$$|\mathcal{W}| \leq |\mathbb{F}_{u,n}^{w,p}(0, b, c_{\psi^\dagger}) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_{\psi^\dagger})| + |\mathbb{F}_{u,n}^{w,p}(a, 0, c_\psi) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_\psi)|.$$

Thus, the problem reduces to showing

$$\sup_{0 \leq \psi^\dagger \leq 1} \sup_{|b| \leq n^{1/4-\eta} B} |\mathbb{F}_{u,n}^{w,p}(0, b, c_{\psi^\dagger}) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_{\psi^\dagger})| = \mathsf{o}_{\mathsf{P}}(1), \tag{3.8.1}$$

$$\sup_{0 \leq \psi \leq 1} \sup_{|a| \leq n^{1/4-\eta} B} |\mathbb{F}_{u,n}^{w,p}(a, 0, c_\psi) - \mathbb{F}_{u,n}^{w,p}(0, 0, c_\psi)| = \mathsf{o}_{\mathsf{P}}(1). \tag{3.8.2}$$

Then (3.8.1) was considered by Theorem 3.8.2 using Assumption 3.5.1$(i, iia, iiia, iva, va)$ with $\nu = 1$ and $s \geq 2$ such that (3.5.4) holds. Further, (3.8.2) follows by Theorem 3.8.3, which requires Assumption 3.5.1(ia, ib, v) with $\nu = 1$, $s = 2$. ■

3.8.2 Linearization of compensator

Theorem 3.5.2 is proved by further considering two separate cases where variation in a, c_ψ but $b = 0$ and where in b, c_ψ but $a = 0$. The next theorem shows the bias correction term in the first situation where b is set to 0 and only a, c_ψ vary. The proof has the similar spirit as linearization of compensator in Jiao and Nielsen (2017, Theorem 8).

Theorem 3.8.4. *Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1(ia, ib, vb) holds with $\nu = 1$, $s = 0$. Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0 \leq \psi \leq 1} \sup_{|a| \leq n^{1/4-\eta} B} |n^{1/2}\{\overline{\mathsf{F}}_{u,n}^{w,p}(a, 0, c_\psi) - \overline{\mathsf{F}}_{u,n}^{w,p}(0, 0, c_\psi)\} - \mathcal{B}_{\mathsf{F}_u,n}(a, 0, c_\psi)| = \mathsf{O}_{\mathsf{P}}(n^{-2\eta}),$$

where the bias term is defined as

$$\mathcal{B}_{\mathsf{F}_u,n}(a,0,c_\psi) = \sigma^{p-1}c_\psi^p \mathsf{f}_u(c_\psi) n^{-1/2} \sum_{i=1}^n w_{in} n^{-1/2} a c_\psi.$$

Proof of Theorem 3.8.4. See the proof of Theorem 8 in Jiao and Nielsen (2017) which is the generalized version of the above theorem, so follow their argument but set $b = 0$ in their context. ■

We state next theorem to handle variation in b, c_ψ while a is set to 0.

Theorem 3.8.5. *Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1$(ia, iia, iiia, vb)$ holds with $\nu = 1$, $s = 0$. Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\le\psi\le 1} \sup_{|b|\le n^{1/4-\eta}B} |n^{1/2}\{\overline{\mathsf{F}}_{u,n}^{w,p}(0,b,c_\psi) - \overline{\mathsf{F}}_{u,n}^{w,p}(0,0,c_\psi)\} - \mathcal{B}_{\mathsf{F}_u,n}(0,b,c_\psi)| = \mathrm{O}_{\mathsf{P}}(n^{-2\eta}),$$

where the bias term, with $\xi_{c_\psi} = \mathsf{E}(\Sigma^{-1/2} r_i | u_i/\sigma = c_\psi)$, is defined as

$$\mathcal{B}_{\mathsf{F}_u,n}(0,b,c_\psi) = \sigma^{p-1}c_\psi^p \mathsf{f}_u(c_\psi) n^{-1/2} \sum_{i=1}^n w_{in}(n^{-1/2}\xi'_{c_\psi}\Sigma^{1/2}b + z'_{in}\Pi b).$$

Proof of Theorem 3.8.5. The object of interest is

$$D_n(0,b,c_\psi) = n^{1/2}\{\overline{\mathsf{F}}_{u,n}^{w,p}(0,b,c_\psi) - \overline{\mathsf{F}}_{u,n}^{w,p}(0,0,c_\psi)\} - \mathcal{B}_{\mathsf{F}_u,n}(0,b,c_\psi),$$

where $\overline{\mathsf{F}}_{u,n}^{w,p}(0,b,c_\psi)$ is well-defined due to Assumption 3.5.1(ia) as the indicator function in expectation is bounded by 1. Let $g_i^{0,b,c_\psi} = 1_{(u_i\le\sigma c_\psi + z'_{in}\Pi b + n^{-1/2}r'_i b)} - 1_{(u_i\le\sigma c_\psi)}$. Then denote $h_i(0,b,c_\psi) = (n^{-1/2}\xi'_{c_\psi}\Sigma^{1/2}b + z'_{in}\Pi b)/\sigma$ and $s(c_\psi) = c_\psi^p \mathsf{f}_u(c_\psi)$. We define that $S_i(0,b,c_\psi) = \mathsf{E}_{i-1}u_i^p g_i^{0,b,c_\psi} - \sigma^p s(c_\psi)h_i(0,b,c_\psi)$ such that $D_n(0,b,c_\psi)$ is expressed as $n^{-1/2}\sum_{i=1}^n w_{in}S_i(0,b,c_\psi)$. To apply Lemma 3.7.2, let $Y = u_i/\sigma$ and $X = \Sigma^{-1/2}r_i$ so their joint density is $\mathsf{f}_{u,r}(y,x)$ for $y \in \mathbb{R}$ and $x \in \mathbb{R}^{d_x}$ while Y, X are independent of $\mathcal{F}_{i-1}$ and $z_i \in \mathcal{F}_{i-1}$. Also let $b_1 = z'_{in}\Pi b/\sigma \in \mathbb{R}$, $b_2 = n^{-1/2}\Sigma^{1/2}b \in \mathbb{R}^{d_x}$, and $c = c_\psi$. By Assumption 3.5.1(iii) that the joint density can be decomposed into conditional and marginal ones, the lemma gives

$$\begin{aligned}|S_i(0,b,c_\psi)| \le \sigma^p\{&\frac{1}{2}(z'_{in}\Pi b/\sigma)^2 + |z'_{in}\Pi b/\sigma||n^{-1/2}\Sigma^{1/2}b|\mathsf{E}|\Sigma^{-1/2}r_i| \\ &+ \frac{1}{2}|n^{-1/2}\Sigma^{1/2}b|^2\mathsf{E}|\Sigma^{-1/2}r_i|^2\} \sup_{y\in\mathbb{R},x\in\mathbb{R}^{d_x}} |py^{p-1}\mathsf{f}_{u|r}(y|x) + y^p\dot{\mathsf{f}}_{u|r,y}(y|x)|,\end{aligned}$$

uniformly in ψ. Note $|z'_{in}\Pi b| \le |z_{in}||\Pi||b|$ and $|\Sigma^{1/2}b| \le |\Sigma^{1/2}||b|$ due to the consistency property of the spectral norm. As $|b| \le n^{1/4-\eta}B$ and Assumption 3.5.1$(iia, iiia)$ with

$s = 0$ that $\mathsf{E}|\Sigma^{-1/2}r_i|, \mathsf{E}|\Sigma^{-1/2}r_i|^2 < \infty$ and $\sup_{y\in\mathbb{R},x\in\mathbb{R}^{d_x}} |py^{p-1}\mathsf{f}_{u|r}(y|x) + y^p\dot{\mathsf{f}}_{u|r,y}(y|x)| < \infty$, we have $|S_i(0,b,c_\psi)| = \mathrm{O}(n^{-1/2-2\eta})(1+|n^{1/2}z_{in}|^2)$. Then the triangle inequality gives

$$|D_n(0,b,c_\psi)| \leq n^{-1/2}\sum_{i=1}^{n}|w_{in}||S_i(0,b,c_\psi)| = \mathrm{O}(n^{-2\eta})n^{-1}\sum_{i=1}^{n}|w_{in}|(1+|n^{1/2}z_{in}|^2).$$

By Assumption 3.5.1(vb), the term $|D_n(0,b,c_\psi)|$ has order $\mathrm{O}_\mathsf{P}(n^{-2\eta})$ uniformly in ψ, b. ■

Then we combine terms $\mathcal{B}_{\mathsf{F}_u,n}(a,0,c_\psi)$ and $\mathcal{B}_{\mathsf{F}_u,n}(0,b,c_\psi)$ to correct the bias in the case where variation in a, b, c_ψ is jointly considered.

Proof of Theorem 3.5.2. The interest is

$$D_n(a,b,c_\psi) = n^{1/2}\{\overline{\mathsf{F}}^{w,p}_{u,n}(a,b,c_\psi) - \overline{\mathsf{F}}^{w,p}_{u,n}(0,0,c_\psi)\} - \mathcal{B}_{\mathsf{F}_u,n}(a,b,c_\psi),$$

where $\overline{\mathsf{F}}^{w,p}_{u,n}(a,b,c_\psi)$ is well-defined by Assumption 3.5.1(ia). Let $c_{\psi^\dagger} = c_\psi(1+n^{-1/2}a/\sigma)$. Note $\overline{\mathsf{F}}^{w,p}_{u,n}(a,b,c_\psi) = \overline{\mathsf{F}}^{w,p}_{u,n}(0,b,c_{\psi^\dagger})$ and $\mathcal{B}_{\mathsf{F}_u,n}(a,b,c_\psi) = \mathcal{B}_{\mathsf{F}_u,n}(0,b,c_\psi) + \mathcal{B}_{\mathsf{F}_u,n}(a,0,c_\psi)$. Add and subtract $n^{1/2}\overline{\mathsf{F}}^{w,p}_{u,n}(a,0,c_\psi) = n^{1/2}\overline{\mathsf{F}}^{w,p}_{u,n}(0,0,c_{\psi^\dagger})$ and $\mathcal{B}_{\mathsf{F}_u,n}(0,b,c_{\psi^\dagger})$ and apply the trianlge inequality to obtain

$$|D_n(a,b,c_\psi)| \leq |D_n(0,b,c_{\psi^\dagger})| + |D_n(a,0,c_\psi)| + |\mathcal{B}_{\mathsf{F}_u,n}(0,b,c_{\psi^\dagger}) - \mathcal{B}_{\mathsf{F}_u,n}(0,b,c_\psi)|.$$

Since c_ψ varies in $\mathbb{R}$ and $|a| \leq n^{1/4-\eta}B$, then $c_{\psi^\dagger}$ also moves on the whole real line so variation in ψ, a transfers to variation in $\psi^\dagger$. Furthermore, $c_{\psi^\dagger}$ converges to c_ψ as $n \to \infty$. Theorem 3.8.5 shows $|D_n(0,b,c_{\psi^\dagger})| = \mathrm{O}_\mathsf{P}(n^{-2\eta})$ uniformly in $\psi^\dagger, b$ by Assumption 3.5.1($ia, iia, iiia, vb$). While due to Assumption 3.5.1(ia, ib, vb) Theorem 3.8.4 demonstrates $|D_n(a,0,c_\psi)| = \mathrm{O}_\mathsf{P}(n^{-2\eta})$ uniformly in ψ, a. Notice

$$\xi_y = \mathsf{E}(\Sigma^{-1/2}r_i|\frac{u_i}{\sigma} = y) = \int_{\mathbb{R}^{d_x}} x\mathsf{f}_{r|u}(x|y)(dx),$$

then ξ_y is continuous uniformly in y due to the property of integral and Assumption 3.5.1(iii) that the conditional density $\mathsf{f}_{r|u}(x|y)$ is continuous uniformly in $x \in \mathbb{R}^{d_x}, y \in \mathbb{R}$. Also note y^p, $\mathsf{f}_u(y)$ are uniformly continuous in $y \in \mathbb{R}$ by Assumption 3.5.1(i). Since $c_{\psi^\dagger}$ only involves in $\mathcal{B}_{\mathsf{F}_u,n}(0,b,c_{\psi^\dagger})$ through $c^p_{\psi^\dagger}$, $\mathsf{f}_u(c_{\psi^\dagger})$, $\xi_{c_{\psi^\dagger}}$, then $\mathcal{B}_{\mathsf{F}_u,n}(0,b,c_{\psi^\dagger})$ converges to $\mathcal{B}_{\mathsf{F}_u,n}(0,b,c_\psi)$ uniformly in all its arguments as $c_{\psi^\dagger} \to c_\psi$, and subsequently the third term in the inequality of $|D_n(a,b,c_\psi)|$ vanishes. ■

3.8.3 Tightness of empirical process

The proof has the similar sprit as the asymptotic equi-continuity argument in Johansen and Nielsen (2016a, Theorem 4.4).

Proof of Theorem 3.5.3. Let $R(c_\psi, c_{\psi^\dagger}) = \mathbb{F}^{w,p}_{u,n}(0,0,c_{\psi^\dagger}) - \mathbb{F}^{w,p}_{u,n}(0,0,c_\psi)$. The aim is to bound the probability $\mathcal{P} = \mathsf{P}\{\mathcal{R} > \epsilon\}$ where $\mathcal{R} = \sup_{0\le\psi\le\psi^\dagger\le 1:\psi^\dagger-\psi\le\phi} |R(c_\psi, c_{\psi^\dagger})|$. Throughout, denote $C > 0$ as usual a constant not depending on ϵ, n, ϕ, which may have different values in different expressions. Assume $\sigma = 1$ without loss of generality and use the metric d built on the distance function $\mathsf{H}_s(x)$.

1. *Coefficients ϵ, ϕ, s.* Let $s = 2$. Note $0 < \epsilon < 1$ and $0 < \phi < 1$. Take ϵ, n as given and choose ϕ such that $\phi^{(1-\nu)/4} \le \epsilon^2$ for some $0 < \nu < 1$. A dyadic argument will be used, so given ϵ, n, ϕ we will select numbers $\overline{m}$, $\underline{m}$ and derive a bound to the probability $\mathcal{P}$ not depending on $\overline{m}$, $\underline{m}$.

2. *Fine grid.* Choose $\overline{m}$ so $2^{-\overline{m}} H_s \le n^{-1/2}\epsilon\phi^{(1-\nu)/4} \le 2^{1-\overline{m}} H_s$ where $H_s < \infty$ due to Assumption 3.5.1(ia).

3. *Coarse grid.* Choose $\underline{m}$ so $2^{-\underline{m}-1} H_s \le C\phi^{1-\nu} \le 2^{-\underline{m}} H_s$. For large n, $\overline{m} > \underline{m}$.

4. *Partition the support.* For each of $m = \underline{m}, \ldots, \overline{m}$, use the metric d to partition $\mathbb{R}$ as laid out in (3.7.6) with $K_m = 2^m$, since $H_s < \infty$ by Assumption 3.5.1(ia).

5. *Assign c_ψ and $c_{\psi^\dagger}$ to the partitioned support.* For each of $m = \underline{m}, \ldots, \overline{m}$, there exist $k_m \le k^\dagger_m$ and grid points $c_{k_m-1,m}$, $c_{k_m,m}$ and $c_{k^\dagger_m-1,m}$, $c_{k^\dagger_m,m}$ so $c_{k_m-1,m} < c_\psi \le c_{k_m,m}$ and $c_{k^\dagger_m-1,m} < c_{\psi^\dagger} \le c_{k^\dagger_m,m}$. Let $\underline{c}_m = c_{k_m-1,m}$, $\overline{c}_m = c_{k_m,m}$ and $\underline{c}^\dagger_m = c_{k^\dagger_m-1,m}$, $\overline{c}^\dagger_m = c_{k^\dagger_m,m}$ so $\underline{c}_m < c_\psi \le \overline{c}_m$ and $\underline{c}^\dagger_m < c_{\psi^\dagger} \le \overline{c}^\dagger_m$. Then $\overline{c}_{m-1} = c_{k_{m-1},m-1}$ equals either $\overline{c}_m = c_{k_m,m}$ or $c_{k_m+1,m}$ so $\mathsf{d}(\overline{c}_m, \overline{c}_{m-1})$ is either 0 or $2^{-m} H_s$. Since the bound $C\phi^{1-\nu} \le 2^{-\underline{m}} H_s$ in item 3 and Lemma 3.7.3 using Assumption 3.5.1(ia) with some $0 < \nu < 1$, there exist $C_\nu, \phi_0 > 0$ so for $0 \le \phi \le \phi_0$

$$\sup_{0\le\psi\le\psi^\dagger\le 1:\psi^\dagger-\psi\le\phi} \mathsf{d}(c_\psi, c_{\psi^\dagger}) \le C_\nu \phi^{1-\nu} \le 2^{-\underline{m}} H_s.$$

Thus there is at most one $\underline{m}$-grid point in the interval $(c_\psi, c_{\psi^\dagger})$.

6. *Apply chaining.* Relate c_ψ to the nearest right fine grid point $\overline{c}_{\overline{m}}$ and $c_{\psi^\dagger}$ to the nearest left fine grid point $\underline{c}^\dagger_{\overline{m}}$. Then split the interval $(c_\psi, c_{\psi^\dagger})$ into the three intervals $(c_\psi, \overline{c}_{\overline{m}})$, $(\overline{c}_{\overline{m}}, \underline{c}^\dagger_{\overline{m}})$, $(\underline{c}^\dagger_{\overline{m}}, c_{\psi^\dagger})$. If c_ψ, $c_{\psi^\dagger}$ are in the neighbouring $\overline{m}$-interval then $\overline{c}_{\overline{m}} = \underline{c}^\dagger_{\overline{m}}$, and if they are in the same $\overline{m}$-interval then $\overline{c}_{\overline{m}} > \underline{c}^\dagger_{\overline{m}}$ so $\underline{c}_{\overline{m}} = \underline{c}^\dagger_{\overline{m}}$ and $\overline{c}_{\overline{m}} = \overline{c}^\dagger_{\overline{m}}$. Thus

$$R(c_\psi, c_{\psi^\dagger}) = R(c_\psi, \overline{c}_{\overline{m}}) + R(\underline{c}^\dagger_{\overline{m}}, c_{\psi^\dagger}) - 1_{(\overline{c}_{\overline{m}} > \underline{c}^\dagger_{\overline{m}})} R(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}}) + 1_{(\overline{c}_{\overline{m}} < \underline{c}^\dagger_{\overline{m}})} R(\overline{c}_{\overline{m}}, \underline{c}^\dagger_{\overline{m}}).$$

An iterative argument can be made for the fourth term. Since $\overline{c}_{\overline{m}} < \underline{c}^{\dagger}_{\overline{m}}$, the coarser $(\overline{m}-1)$-grid points satisfy $\overline{c}_{\overline{m}} \le \overline{c}_{\overline{m}-1} \le \underline{c}^{\dagger}_{\overline{m}-1} \le \underline{c}^{\dagger}_{\overline{m}}$ so that

$$R(\overline{c}_{\overline{m}}, \underline{c}^{\dagger}_{\overline{m}}) = R(\overline{c}_{\overline{m}}, \overline{c}_{\overline{m}-1}) + R(\overline{c}_{\overline{m}-1}, \underline{c}^{\dagger}_{\overline{m}-1}) + R(\underline{c}^{\dagger}_{\overline{m}-1}, \underline{c}^{\dagger}_{\overline{m}}).$$

If $\overline{c}_{\overline{m}-1} = \underline{c}^{\dagger}_{\overline{m}-1}$, then $R(\overline{c}_{\overline{m}-1}, \underline{c}^{\dagger}_{\overline{m}-1}) = 0$ and iteration stops. In this case for $m < \overline{m}-1$ the m-grid points cross over as $\overline{c}_m \ge \overline{c}_{\overline{m}-1} = \underline{c}^{\dagger}_{\overline{m}-1} \ge \underline{c}^{\dagger}_m$. If $\overline{c}_{\overline{m}-1} < \underline{c}^{\dagger}_{\overline{m}-1}$, the argument can be made again for $R(\overline{c}_{\overline{m}-1}, \underline{c}^{\dagger}_{\overline{m}-1})$. In the m-th step, iteration continues if $\overline{c}_m < \underline{c}^{\dagger}_m$, and noting that if there are no other m-grid points between $\overline{c}_m$, $\underline{c}^{\dagger}_m$, the contribution from the $(m-1)$-th step is zero. Item 5 shows there is at most one $\underline{m}$-grid point in the interval $(c_\psi, c_{\psi^\dagger})$, so the $\underline{m}$-th step either gives a zero contribution or the grid points have crossed over at an earlier stage. Therefore, the fourth term satisfies

$$1_{(\overline{c}_{\overline{m}} < \underline{c}^{\dagger}_{\overline{m}})} R(\overline{c}_{\overline{m}}, \underline{c}^{\dagger}_{\overline{m}}) = \sum_{m=\underline{m}+1}^{\overline{m}} 1_{(\overline{c}_m < \underline{c}^{\dagger}_m)} \{R(\overline{c}_m, \overline{c}_{m-1}) + R(\underline{c}^{\dagger}_{m-1}, \underline{c}^{\dagger}_m)\}.$$

Then apply the triangle inequality to get

$$\begin{aligned} |R(c_\psi, c_{\psi^\dagger})| \le{}& |R(c_\psi, \overline{c}_{\overline{m}})| + |R(\underline{c}^{\dagger}_{\overline{m}}, c_{\psi^\dagger})| + |1_{(\overline{c}_{\overline{m}} > \underline{c}^{\dagger}_{\overline{m}})} R(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})| \\ &+ |\sum_{m=\underline{m}+1}^{\overline{m}} 1_{(\overline{c}_m < \underline{c}^{\dagger}_m)} R(\overline{c}_m, \overline{c}_{m-1})| + |\sum_{m=\underline{m}+1}^{\overline{m}} 1_{(\overline{c}_m < \underline{c}^{\dagger}_m)} R(\underline{c}^{\dagger}_{m-1}, \underline{c}^{\dagger}_m)|. \end{aligned}$$

Thus, it is argued that $\mathcal{R} \le \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3 + \mathcal{R}_4 + \mathcal{R}_5$ where

$$\begin{aligned} \mathcal{R}_1 &= \sup_{0 \le \psi \le \psi^\dagger \le 1 : \psi^\dagger - \psi \le \phi} |R(c_\psi, \overline{c}_{\overline{m}})|, \\ \mathcal{R}_2 &= \sup_{0 \le \psi \le \psi^\dagger \le 1 : \psi^\dagger - \psi \le \phi} |R(\underline{c}^{\dagger}_{\overline{m}}, c_{\psi^\dagger})|, \\ \mathcal{R}_3 &= \sup_{0 \le \psi \le \psi^\dagger \le 1 : \psi^\dagger - \psi \le \phi} |1_{(\overline{c}_{\overline{m}} > \underline{c}^{\dagger}_{\overline{m}})} R(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})|, \\ \mathcal{R}_4 &= \sup_{0 \le \psi \le \psi^\dagger \le 1 : \psi^\dagger - \psi \le \phi} |\sum_{m=\underline{m}+1}^{\overline{m}} 1_{(\overline{c}_m < \underline{c}^{\dagger}_m)} R(\overline{c}_m, \overline{c}_{m-1})|, \\ \mathcal{R}_5 &= \sup_{0 \le \psi \le \psi^\dagger \le 1 : \psi^\dagger - \psi \le \phi} |\sum_{m=\underline{m}+1}^{\overline{m}} 1_{(\overline{c}_m < \underline{c}^{\dagger}_m)} R(\underline{c}^{\dagger}_{m-1}, \underline{c}^{\dagger}_m)|. \end{aligned}$$

By Boole's inequality, it follows that $\mathcal{P} \le \mathcal{P}_1 + \mathcal{P}_2 + \mathcal{P}_3 + \mathcal{P}_4 + \mathcal{P}_5$ where $\mathcal{P}_1 = \mathsf{P}(\mathcal{R}_1 > \epsilon/5)$, $\mathcal{P}_2 = \mathsf{P}(\mathcal{R}_2 > \epsilon/5)$, $\mathcal{P}_3 = \mathsf{P}(\mathcal{R}_3 > \epsilon/5)$, $\mathcal{P}_4 = \mathsf{P}(\mathcal{R}_4 > \epsilon/5)$, $\mathcal{P}_5 = \mathsf{P}(\mathcal{R}_5 > \epsilon/5)$. It then suffices to find bounds to $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$, $\mathcal{P}_4$, $\mathcal{P}_5$.

7. *Decompose the term* $\mathcal{R}_1$. Since $\underline{c}_{\overline{m}} < c_\psi \le \overline{c}_{\overline{m}}$ and $\underline{c}_{\overline{m}} = c_{k_{\overline{m}}-1,\overline{m}}$, $\overline{c}_{\overline{m}} = c_{k_{\overline{m}},\overline{m}}$, we have $|J_{i,p}(c_\psi, \overline{c}_{\overline{m}})| \le |J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})| = |J_{i,p}(c_{k_{\overline{m}}-1,\overline{m}}, c_{k_{\overline{m}},\overline{m}})|$. Then by the triangle

inequality, it follows

$$\mathcal{R}_1 \le \max_{1\le k_{\overline{m}}\le K_{\overline{m}}} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\{|J_{i,p}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}})| + \mathsf{E}_{i-1}|J_{i,p}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}})|\}.$$

Thus a martingale decomposition gives $\mathcal{R}_1 \le \widetilde{\mathcal{R}}_1 + 2\overline{\mathcal{R}}_1$, where

$$\widetilde{\mathcal{R}}_1 = \max_{1\le k_{\overline{m}}\le K_{\overline{m}}} |n^{-1/2}\sum_{i=1}^{n}|w_{in}|\{|J_{i,p}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}})| - \mathsf{E}_{i-1}|J_{i,p}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}})|\}|,$$

$$\overline{\mathcal{R}}_1 = \max_{1\le k_{\overline{m}}\le K_{\overline{m}}} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\mathsf{E}_{i-1}|J_{i,p}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}})|.$$

By Boole's inequality, it shows that $\mathcal{P}_1 \le \widetilde{\mathcal{P}}_1 + 2\overline{\mathcal{P}}_1$ where we have $\widetilde{\mathcal{P}}_1 = \mathsf{P}(\widetilde{\mathcal{R}}_1 > \epsilon/15)$, $\overline{\mathcal{P}}_1 = \mathsf{P}(\overline{\mathcal{R}}_1 > \epsilon/15)$. It suffices to find bounds to $\widetilde{\mathcal{P}}_1$, $\overline{\mathcal{P}}_1$.

8. *Bounding the probability* $\widetilde{\mathcal{P}}_1$. Let $z_{l,i} = |w_{in}||J_{i,p}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}})|$ and write $\widetilde{\mathcal{R}}_1$ as the maximum of $|n^{-1/2}\sum_{i=1}^{n}(z_{l,i} - \mathsf{E}_{i-1}z_{l,i})|$ over index $l = k_{\overline{m}}$ with $L = K_{\overline{m}} = 2^{\overline{m}}$. The difference of indicators can be bounded by one, so it holds for the moment condition that $\mathsf{E}z_{l,i}^4 \le \mathsf{E}|w_{in}|^4 u_i^{4p} = \mathsf{E}|w_{in}|^4\mathsf{E}u_i^{4p} < \infty$. This is because u_i is independent of $w_{in} \in \mathcal{F}_{i-1}$ and due to Assumption 3.5.1(ia, va) that $\mathsf{E}u_i^{4p}, \mathsf{E}|w_{in}|^4 < \infty$. Consider $1 \le q \le s$ and $s = 2$. The inequality of (3.7.5) shows uniformly in $l = k_{\overline{m}}$ that

$$\mathsf{E}_{i-1}z_{l,i}^{2^q} = |w_{in}|^{2^q}\mathsf{E}_{i-1}J_{i,p}^{2^q}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}}) < (1+|w_{in}|^{2^s})\mathsf{d}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}}).$$

As $\mathsf{d}(c_{k_{\overline{m}}-1,\overline{m}},c_{k_{\overline{m}},\overline{m}}) = H_s/K_{\overline{m}}$ and Assumption 3.5.1(va) $n^{-1}\mathsf{E}\sum_{i=1}^{n}(1+|w_{in}|^{2^s}) = \mathrm{O}(1)$, then we have $D = 2^{-\overline{m}}H_s$ such that

$$\mathsf{E}D_q = \mathsf{E}\max_{1\le l\le L}\sum_{i=1}^{n}\mathsf{E}_{i-1}z_{l,i}^{2^q} < 2^{-\overline{m}}H_s\mathsf{E}\sum_{i=1}^{n}(1+|w_{in}|^{2^s}) \le Dn.$$

Apply Lemma 3.7.5 with $\kappa = \epsilon/15$ to get

$$\widetilde{\mathcal{P}}_1 \le \frac{15(2^{\overline{m}}+1)\theta^3 H_s}{2^{\overline{m}}\epsilon n} + \frac{15\theta H_s}{2^{\overline{m}}\epsilon} + 2^2 2^{\overline{m}}\exp(-\frac{\epsilon\theta}{210}) \le \frac{C\theta^3}{\epsilon n} + \frac{C\theta}{2^{\overline{m}}\epsilon} + C2^{\overline{m}}\exp(-\frac{\epsilon\theta}{210}).$$

Choose $\theta = 210\epsilon^{-1}(\log 2^{2\overline{m}} + \log\phi^{-1})$. Use the inequality $(x+y)^3 \le C(x^3+y^3)$ and then $\epsilon^{-2} \le \phi^{-(1-\nu)/4}$ and $n^{-1}\epsilon^2 \le C2^{-2\overline{m}}\phi^{-(1-\nu)/2}$ implied by bounds in item 1, 2 to obtain

$$\widetilde{\mathcal{P}}_{1,1} = \frac{C\theta^3}{\epsilon n} \le \frac{C}{\epsilon^4 n}(\overline{m}^3 + \log^3\frac{1}{\phi}) \le C2^{-2\overline{m}}\phi^{-(1-\nu)/2}\phi^{-3(1-\nu)/4}(\overline{m}^3 + \log^3\frac{1}{\phi}).$$

Since $2^{-\overline{m}/2} \le 2^{-\underline{m}/2} \le C\phi^{(1-\nu)/2}$ implied by the bound in item 3, and the functions $m^3 2^{-m/2}$ and $\phi^{(1-\nu)/2}\log^3\phi^{-1}$ are bounded for $m \ge 1$ and $0 < \phi < 1$, we have

$$\widetilde{\mathcal{P}}_{1,1} \le C(2^{-\overline{m}/2}\phi^{-(1-\nu)/2})^3(\overline{m}^3 2^{-\overline{m}/2} + 2^{-\overline{m}/2}\phi^{-(1-\nu)/2}\phi^{(1-\nu)/2}\log^3\frac{1}{\phi})\phi^{(1-\nu)/4} \le C\phi^{(1-\nu)/4}.$$

By $\epsilon^{-2} \le \phi^{-(1-\nu)/4}$ given in item 1, it holds

$$\widetilde{\mathcal{P}}_{1,2} = \frac{C\theta}{2^{\overline{m}}\epsilon} \le \frac{C}{2^{\overline{m}}\epsilon^2}(\overline{m} + \log\frac{1}{\phi}) \le C2^{-\overline{m}}\phi^{-(1-\nu)/4}(\overline{m} + \log\frac{1}{\phi}).$$

Rewrite this bound and argue as for $\widetilde{\mathcal{P}}_{1,1}$ to get

$$\widetilde{\mathcal{P}}_{1,2} \le C2^{-\overline{m}/2}\phi^{-(1-\nu)/2}(\overline{m}2^{-\overline{m}/2} + 2^{-\overline{m}/2}\phi^{-(1-\nu)/2}\phi^{(1-\nu)/2}\log\frac{1}{\phi})\phi^{(1-\nu)/4} \le C\phi^{(1-\nu)/4}.$$

For $0 < \phi < 1$ and $0 < \nu < 1$ it holds $\phi \le \phi^{(1-\nu)/4}$ so that

$$\widetilde{\mathcal{P}}_{1,3} = C2^{\overline{m}}\exp(-\frac{\epsilon\theta}{210}) = C2^{-\overline{m}}\phi \le C\phi^{(1-\nu)/4}.$$

In summary, $\widetilde{\mathcal{P}}_1 \le \widetilde{\mathcal{P}}_{1,1} + \widetilde{\mathcal{P}}_{1,2} + \widetilde{\mathcal{P}}_{1,3} \le C\phi^{(1-\nu)/4}$.

9. *Bounding the probability* $\overline{\mathcal{P}}_1$. The inequality of (3.7.5) shows that uniformly in $k_{\overline{m}}$ we have $\mathsf{E}_{i-1}|J_{i,p}(c_{k_{\overline{m}}-1,\overline{m}}, c_{k_{\overline{m}},\overline{m}})| < \mathsf{d}(c_{k_{\overline{m}}-1,\overline{m}}, c_{k_{\overline{m}},\overline{m}}) = H_s/K_{\overline{m}}$. Then by $K_{\overline{m}} = 2^{\overline{m}}$ and Assumption 3.5.1(va) it follows $\mathsf{E}\overline{\mathcal{R}}_1 < n^{-1/2}\mathsf{E}\sum_{i=1}^n |w_{in}|H_s/K_{\overline{m}} \le Cn^{1/2}2^{-\overline{m}}H_s$. By Markov inequality and the bound $2^{-\overline{m}}H_s \le n^{-1/2}\epsilon\phi^{(1-\nu)/4}$ in item 2, we get

$$\overline{\mathcal{P}}_1 \le \frac{15\mathsf{E}\overline{\mathcal{R}}_1}{\epsilon} \le \frac{Cn^{1/2}2^{-\overline{m}}H_s}{\epsilon} \le C\phi^{(1-\nu)/4}.$$

10. *Bounding the probability* $\mathcal{P}_1$. Combine results in item $7-9$ to get $\mathcal{P}_1 \le C\phi^{(1-\nu)/4}$.

11. *Bounding the probability* $\mathcal{P}_2$. This is similar as to show $\mathcal{P}_1 \le C\phi^{(1-\nu)/4}$. Thus the same argument can be made to demonstrate $\mathcal{P}_2 \le C\phi^{(1-\nu)/4}$ through item $7-10$.

12. *Bounding the probability* $\mathcal{P}_3$. Notice $\mathcal{R}_3 \le \mathcal{R}_1$ so $\mathcal{P}_3 \le \mathcal{P}_1 \le C\phi^{(1-\nu)/4}$.

13. *Decompose the term* $\mathcal{R}_4$. From partition in item $4, 5$, $\overline{c}_{m-1} = c_{k_{m-1},m-1}$ equals either $\overline{c}_m = c_{k_m,m}$ or $c_{k_m+1,m}$, then $\overline{c}_m$, $\overline{c}_{m-1}$ are at most 1 step apart in the m-grid so that $R(\overline{c}_m, \overline{c}_{m-1})$ is either 0 or $R(c_{k_m,m}, c_{k_m+1,m})$. Then we get

$$|R(\overline{c}_m, \overline{c}_{m-1})| \le |R(c_{k_m,m}, c_{k_m+1,m})| \le \max_{1\le k_m\le K_m} |R(c_{k_m-1,m}, c_{k_m,m})|.$$

Note that the right-hand side on the last inequality does not depend on ψ, $\psi^\dagger$ so

$$\mathcal{R}_4 \le \sum_{m=\underline{m}+1}^{\overline{m}} \max_{1\le k_m\le K_m} |R(c_{k_m-1,m}, c_{k_m,m})|.$$

14. *Bounding the probability* $\mathcal{P}_4$. Notice first that

$$\sum_{m=\underline{m}+1}^{\overline{m}} 2^{(\underline{m}-m)/4} < \sum_{j=1}^{\infty} 2^{-j/4} = \frac{1}{2^{1/4}-1} < 6,$$

so $\sum_{m=\underline{m}+1}^{\overline{m}} 2^{(\underline{m}-m)/4}\epsilon/30 < \epsilon/5$. Then $\mathcal{R}_4 > \epsilon/5$ implies

$$\sum_{m=\underline{m}+1}^{\overline{m}} \max_{1\le k_m\le K_m} |R(c_{k_m-1,m}, c_{k_m,m})| > \sum_{m=\underline{m}+1}^{\overline{m}} \frac{2^{(\underline{m}-m)/4}\epsilon}{30}.$$

It therefore holds

$$\mathcal{P}_4 \le \mathsf{P}[\bigcup_{m=\underline{m}+1}^{\overline{m}} \{\max_{1\le k_m\le K_m} |R(c_{k_m-1,m}, c_{k_m,m})| > \frac{2^{(\underline{m}-m)/4}\epsilon}{30}\}].$$

Then by Boole's inequality, we have

$$\mathcal{P}_4 \le \sum_{m=\underline{m}+1}^{\overline{m}} \mathsf{P}\{\max_{1\le k_m\le K_m} |R(c_{k_m-1,m}, c_{k_m,m})| > \frac{2^{(\underline{m}-m)/4}\epsilon}{30}\}.$$

Let $z_{l,i} = w_{in}J_{i,p}(c_{k_m-1,m}, c_{k_m,m})$ and write $R(c_{k_m-1,m}, c_{k_m,m})$ as $n^{-1/2}\sum_{i=1}^n (z_{l,i} - \mathsf{E}_{i-1}z_{l,i})$, where l represents the index k_m with $L = K_m = 2^m$. The moment condition $\mathsf{E}z_{l,i}^4 < \infty$ holds due to Assumption 3.5.1(ia, va). Consider $1 \le q \le s$ and $s = 2$. The inequality of (3.7.5) shows uniformly in $l = k_m$

$$\mathsf{E}_{i-1}z_{l,i}^{2^q} = |w_{in}|^{2^q}\mathsf{E}_{i-1}J_{i,p}^{2^q}(c_{k_m-1,m}, c_{k_m,m}) < (1 + |w_{in}|^{2^s})\mathsf{d}(c_{k_m-1,m}, c_{k_m,m}).$$

Since $\mathsf{d}(c_{k_m-1,m}, c_{k_m,m}) = H_s/K_m$ and Assumption 3.5.1(va), we have $D = 2^{-m}H_s$ so that

$$\mathsf{E}D_q = \mathsf{E}\max_{1\le l\le L}\sum_{i=1}^n \mathsf{E}_{i-1}z_{l,i}^{2^q} < 2^{-m}H_s\mathsf{E}\sum_{i=1}^n (1 + |w_{in}|^{2^s}) \le Dn.$$

Apply Lemma 3.7.5 with $\kappa = 2^{(\underline{m}-m)/4}\epsilon/30$ to get

$$\mathcal{P}_4 \le \sum_{m=\underline{m}+1}^{\overline{m}} \{\frac{C\theta_m^3}{2^{(\underline{m}-m)/4}\epsilon n} + \frac{C\theta_m}{2^m 2^{(\underline{m}-m)/4}\epsilon} + C2^m \exp(-\frac{2^{(\underline{m}-m)/4}\epsilon\theta_m}{420})\}.$$

Choose $\theta_m = 420\epsilon^{-1}2^{(m-\underline{m})/4}(\log 4^{m-\underline{m}} + \log\phi^{-1})$. For the first term above, use the inequality $(x+y)^3 \le C(x^3+y^3)$ to get

$$\mathcal{P}_{4,1} = \sum_{m=\underline{m}+1}^{\overline{m}} \frac{C\theta_m^3}{2^{(\underline{m}-m)/4}\epsilon n} \le \sum_{m=\underline{m}+1}^{\overline{m}} \frac{C}{2^{\underline{m}-m}\epsilon^4 n}\{(m-\underline{m})^3 + \log^3\frac{1}{\phi}\}.$$

Since $\epsilon^{-2} \le \phi^{-(1-\nu)/4}$ and $n^{-1}\epsilon^2 \le C2^{-2\overline{m}}\phi^{-(1-\nu)/2}$ implied by bounds in item 1, 2, then

$$\mathcal{P}_{4,1} \le \sum_{m=\underline{m}+1}^{\overline{m}} C2^{-2\overline{m}-\underline{m}+m}\phi^{-5(1-\nu)/4}\{(m-\underline{m})^3 + \log^3\frac{1}{\phi}\}.$$

Rewrite $2^{-2\overline{m}-\underline{m}+m} = 2^{3(m-\overline{m})/2-(m-\underline{m})/2-\overline{m}/2-3\underline{m}/2}$ so

$$\mathcal{P}_{4,1} \leq \sum_{m=\underline{m}+1}^{\overline{m}} C(2^{-\underline{m}/2}\phi^{-(1-\nu)/2})^3 \frac{2^{3(m-\overline{m})/2}}{2^{(m-\underline{m})/2}}\{\frac{(m-\underline{m})^3}{2^{\overline{m}/2}} + \frac{\phi^{(1-\nu)/2}}{2^{\overline{m}/2}\phi^{(1-\nu)/2}}\log^3\frac{1}{\phi}\}\phi^{(1-\nu)/4}.$$

Using $2^{-\overline{m}/2} \leq 2^{-\underline{m}/2} \leq C\phi^{(1-\nu)/2}$ from item 3 and that geometric sums are finite, argue as for $\widetilde{\mathcal{P}}_{1,1}$ in item 8 to obtain $\mathcal{P}_{4,1} \leq C\phi^{(1-\nu)/4}$. Then the second term satisfies

$$\mathcal{P}_{4,2} = \sum_{m=\underline{m}+1}^{\overline{m}} \frac{C\theta_m}{2^m 2^{(\underline{m}-m)/4}\epsilon} \leq \sum_{m=\underline{m}+1}^{\overline{m}} \frac{C}{2^{(\underline{m}+m)/2}\epsilon^2}\{(m-\underline{m}) + \log\frac{1}{\phi}\}.$$

Use $\epsilon^{-2} \leq \phi^{-(1-\nu)/4}$ from item 1 to get

$$\mathcal{P}_{4,2} \leq \sum_{m=\underline{m}+1}^{\overline{m}} C2^{-(\underline{m}+m)/2}\phi^{-(1-\nu)/4}\{(m-\underline{m}) + \log\frac{1}{\phi}\}.$$

Rewrite $2^{-(\underline{m}+m)/2} = 2^{-(m-\underline{m})/2}2^{-\underline{m}}$ so

$$\mathcal{P}_{4,2} \leq \sum_{m=\underline{m}+1}^{\overline{m}} C2^{-\underline{m}/2}\phi^{-(1-\nu)/2}2^{-(m-\underline{m})/2}(\frac{m-\underline{m}}{2^{\underline{m}/2}} + \frac{\phi^{(1-\nu)/2}}{2^{\underline{m}/2}\phi^{(1-\nu)/2}}\log\frac{1}{\phi})\phi^{(1-\nu)/4}.$$

Since $2^{-\underline{m}/2} \leq C\phi^{(1-\nu)/2}$ from item 3, argue as for $\mathcal{P}_{4,1}$ to show $\mathcal{P}_{4,2} \leq C\phi^{(1-\nu)/4}$. Then the third term satisfies

$$\mathcal{P}_{4,3} = \sum_{m=\underline{m}+1}^{\overline{m}} C2^m \exp(-\frac{2^{(\underline{m}-m)/4}\epsilon\theta_m}{420}) = \sum_{m=\underline{m}+1}^{\overline{m}} C2^{-(m-\underline{m})}2^{\underline{m}}\phi.$$

As $2^{\underline{m}} \leq C\phi^{\nu-1}$ implied by the bound in item 3, $\mathcal{P}_{4,3} \leq \sum_{m=\underline{m}+1}^{\overline{m}} C2^{-(m-\underline{m})}\phi^{\nu} \leq C\phi^{\nu}$. In summary, $\mathcal{P}_4 \leq \mathcal{P}_{4,1} + \mathcal{P}_{4,2} + \mathcal{P}_{4,3} \leq C\phi^{(1-\nu)/4} + C\phi^{\nu}$.

15. *Bounding the probability* $\mathcal{P}_5$. Apply the same argument as for $\mathcal{P}_4$ through item 13, 14 to show $\mathcal{P}_5 \leq C\phi^{(1-\nu)/4} + C\phi^{\nu}$.

16. *Tightness*. Combine the above result to obtain $\mathcal{P} \leq C\phi^{(1-\nu)/4} + C\phi^{\nu}$. For a given $0 < \epsilon < 1$ and a large n, the only constraint to $\phi \in (0,1)$ is $\phi^{(1-\nu)/4} \leq \epsilon^2$. Thus we have $\mathcal{P} \to 0$ as $\phi \downarrow 0$. ■

3.8.4 Uniform n-convergence of empirical process

While the above is to show $n^{1/2}$-convergence result, we then prove n-convergence for a particular type of empirical process with the weight $n^{1/2}z_{in} = z_i$ and mark r_i. We first demonstrate uniform convergence for the indicator process, then incorporate the weights and marks by Hölder inequality.

Theorem 3.8.6. *Let $c_\psi = \mathsf{F}_u^{-1}(\psi)$. Suppose Assumption 3.5.1 holds with $\nu = 1$ and $s \geq 2$ satisfying (3.5.4). Then for any $B > 0$ and as $n \to \infty$*

$$\sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B}\frac{1}{n}\sum_{i=1}^{n}|1_{(u_i\leq\sigma c_\psi+n^{-1/2}ac_\psi+z_{in}'\Pi b+n^{-1/2}r_i'b)}-1_{(u_i\leq\sigma c_\psi)}| = \mathrm{O_P}(n^{-1/4-\eta}).$$

Proof of Theorem 3.8.6. Denote $\mathcal{R}_n' = \sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B} R_n'(a,b,c_\psi)$, where $g_i^{a,b,c_\psi} = 1_{(u_i\leq\sigma c_\psi+n^{-1/2}ac_\psi+z_{in}'\Pi b+n^{-1/2}r_i'b)} - 1_{(u_i\leq\sigma c_\psi)}$ and $R_n'(a,b,c_\psi) = n^{-1}\sum_{i=1}^n |g_i^{a,b,c_\psi}|$, then we show $\mathcal{R}_n' = \mathrm{O_P}(n^{-1/4-\eta})$ as $n\to\infty$. We construct the martingale $R_{n,1}'(a,b,c_\psi)$ by adding and subtracting a compensator $R_{n,2}'(a,b,c_\psi)$ to $R_n'(a,b,c_\psi)$, therefore we have $R_n'(a,b,c_\psi) = R_{n,1}'(a,b,c_\psi) + R_{n,2}'(a,b,c_\psi)$ where

$$R_{n,1}'(a,b,c_\psi) = \frac{1}{n}\sum_{i=1}^n\{|g_i^{a,b,c_\psi}| - \mathsf{E}_{i-1}|g_i^{a,b,c_\psi}|\}, \qquad R_{n,2}'(a,b,c_\psi) = \frac{1}{n}\sum_{i=1}^n \mathsf{E}_{i-1}|g_i^{a,b,c_\psi}|.$$

Triangle inequality gives $\mathcal{R}_n' \leq \mathcal{R}_{n,1}' + \mathcal{R}_{n,2}'$ where

$$\mathcal{R}_{n,1}' = \sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B}|R_{n,1}'(a,b,c_\psi)|, \qquad \mathcal{R}_{n,2}' = \sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B}R_{n,2}'(a,b,c_\psi).$$

It then suffices to prove $\mathcal{R}_{n,1}' = \mathrm{o_P}(n^{-1/2})$ and $\mathcal{R}_{n,2}' = \mathrm{O_P}(n^{-1/4-\eta})$. Set $w_{in} = 1$, $p = 0$ and adjust the rate n^{-1} instead of $n^{-1/2}$ to the empirical process, then use the argument in Theorem 3.5.1 to show $\mathcal{R}_{n,1}' = \mathrm{o_P}(n^{-1/2})$, while $\mathcal{R}_{n,2}' = \mathrm{O_P}(n^{-1/4-\eta})$ can be demonstrated by the reasoning of Theorem 3.5.2. ■

Proof of Theorem 3.5.4. Let $g_i^{a,b,c_\psi} = 1_{(u_i\leq\sigma c_\psi+n^{-1/2}ac_\psi+z_{in}'\Pi b+n^{-1/2}r_i'b)} - 1_{(u_i\leq\sigma c_\psi)}$ and $R_n^{z,r\prime}(a,b,c_\psi) = n^{-1/2}\sum_{i=1}^n z_{in}r_i'g_i^{a,b,c_\psi}$, then we want to show $\mathcal{R}_n^{z,r\prime} = \mathrm{o_P}(1)$ as $n\to\infty$ where $\mathcal{R}_n^{z,r\prime} = \sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B}|R_n^{z,r\prime}(a,b,c_\psi)|$. By the triangle inequality we have $|R_n^{z,r\prime}(a,b,c_\psi)| \leq n^{-1}\sum_{i=1}^n |n^{1/2}z_{in}||r_i||g_i^{a,b,c_\psi}|$. Then apply Hölder inequality to get

$$|R_n^{z,r\prime}(a,b,c_\psi)| \leq (n^{-1}\sum_{i=1}^n |n^{1/2}z_{in}|^2|r_i|^2)^{1/2}(n^{-1}\sum_{i=1}^n |g_i^{a,b,c_\psi}|^2)^{1/2}.$$

Notice $|g_i^{a,b,c_\psi}|^2 = |g_i^{a,b,c_\psi}|$ and let $\mathcal{R}_n' = \sup_{0\leq\psi\leq 1}\sup_{|a|,|b|\leq n^{1/4-\eta}B}R_n'(a,b,c_\psi)$ where $R_n'(a,b,c_\psi) = n^{-1}\sum_{i=1}^n|g_i^{a,b,c_\psi}|$ so $\mathcal{R}_n^{z,r\prime} \leq (n^{-1}\sum_{i=1}^n|n^{1/2}z_{in}|^2|r_i|^2)^{1/2}(\mathcal{R}_n')^{1/2}$. Theorem 3.8.6 shows $\mathcal{R}_n' = \mathrm{O_P}(n^{-1/4-\eta})$ so $(\mathcal{R}_n')^{1/2} = \mathrm{O_P}(n^{-1/8-\eta/2}) = \mathrm{o_P}(1)$. Then it suffices to demonstrate $n^{-1}\sum_{i=1}^n|n^{1/2}z_{in}|^2|r_i|^2 = \mathrm{O_P}(1)$. This is shown by Markov inequality using independence of z_i and r_i and Assumption 3.5.1(iia, va) that $\mathsf{E}|n^{1/2}z_{in}|^2, \mathsf{E}|r_i|^2 < \infty$. ■

3.8.5 Results for empirical processes of absolute residuals

Empirical processes with absolute indicators can be expressed by the difference of two one-sided processes, thus we prove Theorem 3.5.5, 3.5.6, 3.5.7, 3.5.8 by using the one-sided process results in Theorem 3.5.1, 3.5.2, 3.5.3, 3.5.4. The absolute empirical process results are given under more restrictive Assumption 3.3.1, so we first investigate the lemma concerning the relationship between Assumption 3.3.1 and 3.5.1.

Lemma 3.8.1. *Suppose* w_{in} *is either of* 1, $n^{1/2}z_{in} = z_i$, $nz_{in}z'_{in} = z_iz'_i$ *and* p *is either of* 0, 1, 2. *Then Assumption 3.3.1*$(ia, iia, iiia, ivb, ivc)$ *implies Assumption 3.5.1 with* $s \geq 2$ *satisfying (3.5.4).*

Proof of Theorem 3.8.1. Assumption 3.3.1(ia) shows Assumption 3.5.1(ia, ic), while Assumption 3.5.1(ib) further needs continuous differentiability of the marginal density f_u, see discussion in Johansen and Nielsen (2016a, Remark 4.1c). Assumption 3.3.1(iia, ivb) is the same as Assumption 3.5.1(iia, iva), and Assumption 3.3.1$(iiia)$ deduces Assumption 3.5.1$(iiia)$ with continuous differentiability of the conditional density $\mathsf{f}_{u|r}$. At last Assumption 3.3.1(ivc) implies Assumption 3.5.1(va) and (vb) by Markov inequality. ■

Proof of Theorem 3.5.5. The term of interest is $\mathcal{G} = \mathbb{G}^{w,p}_{u,n}(a, b, c_\psi) - \mathbb{G}^{w,p}_{u,n}(0, 0, c_\psi)$. Our focus is on the absolute quantile $c_\psi = \mathsf{G}_u^{-1}(\psi) > 0$ rather than the one-sided quantile $c_{\psi^*} = \mathsf{F}_u^{-1}(\psi^*) \in \mathbb{R}$. Note $|u_i|/\sigma \sim \mathsf{G}_u$ and $u_i/\sigma \sim \mathsf{F}_u$. Since

$$\begin{aligned}&1_{(|u_i - z'_{in}\Pi b - n^{-1/2}r'_i b| \leq \sigma c_\psi + n^{-1/2}ac_\psi)}\\ =&1_{(u_i \leq \sigma c_\psi + n^{-1/2}ac_\psi + z'_{in}\Pi b + n^{-1/2}r'_i b)} - 1_{(u_i \leq -\sigma c_\psi - n^{-1/2}ac_\psi + z'_{in}\Pi b + n^{-1/2}r'_i b)}\end{aligned}$$

and by (3.5.3), (3.5.7), we have $\mathbb{G}^{w,p}_{u,n}(a, b, c_\psi) = \mathbb{F}^{w,p}_{u,n}(a, b, c_\psi) - \lim_{c^\dagger_\psi \downarrow c_\psi} \mathbb{F}^{w,p}_{u,n}(a, b, -c^\dagger_\psi)$ for any $c_\psi > 0$. By this and the triangle inequality, then for any $c_\psi = \mathsf{G}_u^{-1}(\psi) > 0$,

$$|\mathcal{G}| \leq |\mathbb{F}^{w,p}_{u,n}(a, b, c_\psi) - \mathbb{F}^{w,p}_{u,n}(0, 0, c_\psi)| + \lim_{c^\dagger_\psi \downarrow c_\psi} |\mathbb{F}^{w,p}_{u,n}(a, b, -c^\dagger_\psi) - \mathbb{F}^{w,p}_{u,n}(0, 0, -c^\dagger_\psi)|.$$

These vanish uniformly in ψ, a, b by Theorem 3.5.1 using Assumption 3.5.1 with $\nu = 1$ and $s \geq 2$ such that (3.5.4) holds. Lemmma 3.8.1 shows Assumption 3.3.1$(ia, iia, iiia, ivb, ivc)$ suffices. ■

Proof of Theorem 3.5.6. Argue as in the proof of Theorem 3.5.5 but using Theorem 3.5.2 instead of Theorem 3.5.1. Due to symmetry of f_u, approximate the compensator

by

$$\mathcal{B}_{\mathsf{G}_u,n}(a,b,c_\psi) = \sigma^{p-1}c_\psi^p \mathsf{f}_u(c_\psi) n^{-1/2} \sum_{i=1}^{n} w_{in}[\{1+(-1)^p\}n^{-1/2}ac_\psi$$
$$+ n^{-1/2}\{\xi_{c_\psi} - (-1)^p \xi_{-c_\psi}\}'\Sigma^{1/2}b + \{1-(-1)^p\}z_{in}'\Pi b]. \quad \blacksquare$$

Proof of Theorem 3.5.7. The same argument as Theorem 3.5.5 applies while the tightness result in Theorem 3.5.3 is used in this proof instead. ■

Proof of Theorem 3.5.8. Argue along the lines of the proof in Theorem 3.5.5, however replace Theorem 3.5.1 by Theorem 3.5.4. ■

3.9 Appendix C: Proofs of the main results

We first present some preliminary results for stochastic expansions of two stage least squares statistics. Asymptotics are shown for Algorithm 3.2.1, then weak convergence for Algorithm 3.2.2 and 3.2.3. Finally, a new type of Hausman test is established for outlier robustness checks.

3.9.1 Auxillary results for product moments

The robust estimators in Algorithm 3.2.1 are two stage least squares estimators for selected observations. Outliers are detected through the absolute value of estimated structural errors u_i adjusted by its scale σ in (3.2.1). Introduce the indicators in the empirical processes in §4 as

$$v_i^{a,b,c} = 1_{(|u_i - z_{in}'\Pi b - n^{-1/2}r_i'b| \le \sigma c + n^{-1/2}ac)}, \tag{3.9.1}$$

so given the cut-off c the indicators (3.2.9) for selecting non-outlying observations have the relationship

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \le \widehat{\sigma}_c^{(m)} c)} = 1_{(|u_i - z_{in}'\Pi\widehat{b}_c^{(m)} - n^{-1/2}r_i'\widehat{b}_c^{(m)}| \le \sigma c + n^{-1/2}\widehat{a}_c^{(m)} c)} = v_i^{\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c}, \tag{3.9.2}$$

where $\widehat{b}_c^{(m)} = n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$, $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are the estimation errors for β, σ. Thus in Algorithm 3.2.1 the least squares statistics in the estimators (3.2.10), (3.2.11), (3.2.12) for Π, β, σ are expressed by the weighted and marked empirical processes analyzed in section §3.5. The following results describe the asymptotic behaviour of the corresponding product moments.

Lemma 3.9.1. *Suppose Assumption 3.3.1(ia, iia, iiia, ivb, ivc) holds. Then we have*

$$
\begin{aligned}
n^{-1/2}\sum_{i=1}^{n} v_i^{a,b,c} &= n^{-1/2}\sum_{i=1}^{n} 1_{(|u_i|\le \sigma c)} + \mathsf{f}_u(c)\frac{2ac+\zeta_c^{-\prime}\Sigma^{1/2}b}{\sigma} + R_v(a,b,c),\\
n^{-1/2}\sum_{i=1}^{n} u_i^2 v_i^{a,b,c} &= n^{-1/2}\sum_{i=1}^{n} u_i^2 1_{(|u_i|\le \sigma c)} + \sigma^2 c^2 \mathsf{f}_u(c)\frac{2ac+\zeta_c^{-\prime}\Sigma^{1/2}b}{\sigma} + R_{vuu}(a,b,c),\\
\sum_{i=1}^{n} z_{in} u_i v_i^{a,b,c} &= \sum_{i=1}^{n} z_{in} u_i 1_{(|u_i|\le \sigma c)} + c\mathsf{f}_u(c)\sum_{i=1}^{n} z_{in}(n^{-1/2}\zeta_c^{+\prime}\Sigma^{1/2} + 2z_{in}'\Pi)b\\
&\quad + R_{vzu}(a,b,c),\\
n^{1/2}\sum_{i=1}^{n} z_{in} z_{in}' v_i^{a,b,c} &= n^{1/2}\sum_{i=1}^{n} z_{in} z_{in}' 1_{(|u_i|\le \sigma c)} + \mathsf{f}_u(c) M_{zz,n}\frac{2ac+\zeta_c^{-\prime}\Sigma^{1/2}b}{\sigma} + R_{vzz}(a,b,c).
\end{aligned}
$$

Let $R(a,b,c) = |R_v(a,b,c)| + |R_{vuu}(a,b,c)| + |R_{vzu}(a,b,c)| + |R_{vzz}(a,b,c)|$. *Then for any* $B > 0$ *and as* $n \to \infty$

$$
\sup_{0<c<\infty}\sup_{|a|,|b|\le n^{1/4-\eta}B} |R(a,b,c)| = \mathrm{o}_{\mathsf{P}}(1).
$$

Proof of Lemma 3.9.1. The terms of interest are

$$
\mathcal{M}_n = n^{-1/2}\sum_{i=1}^{n} w_{in} u_i^p v_i^{a,b,c}, \quad v_i^{a,b,c} = 1_{(|u_i - z_{in}'\Pi b - n^{-1/2} r_i' b| \le \sigma c + n^{-1/2} ac)}.
$$

1. *Decompose* $\mathcal{M}_n$. Write $\mathcal{M}_n = \mathcal{M}_{n,1} + \mathcal{M}_{n,2} + \mathcal{M}_{n,3}$, where

$$
\begin{aligned}
\mathcal{M}_{n,1} &= n^{-1/2}\sum_{i=1}^{n} w_{in} u_i^p 1_{(|u_i|\le\sigma c)}, \quad \mathcal{M}_{n,2} = n^{-1/2}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1} u_i^p\{v_i^{a,b,c} - 1_{(|u_i|\le\sigma c)}\},\\
\mathcal{M}_{n,3} &= n^{-1/2}\sum_{i=1}^{n} w_{in} u_i^p\{v_i^{a,b,c} - 1_{(|u_i|\le\sigma c)}\} - n^{-1/2}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1} u_i^p\{v_i^{a,b,c} - 1_{(|u_i|\le\sigma c)}\}.
\end{aligned}
$$

Therefore, the first term in stochastic expansion is $\mathcal{M}_{n,1}$. We will linearize $\mathcal{M}_{n,2}$ to obtain the second term, and argue that $\mathcal{M}_{n,3}$ is small in probability.

2. *Linearize* $\mathcal{M}_{n,2}$. Note $\mathcal{M}_{n,2} = n^{1/2}\{\overline{\mathsf{G}}_{u,n}^{w,p}(a,b,c) - \overline{\mathsf{G}}_{u,n}^{w,p}(0,0,c)\}$, see (3.5.6). Apply Theorem 3.5.6 by Assumption 3.3.1$(ia, iia, iiia, ivc)$ to get $\mathcal{M}_{n,2} = \mathcal{B}_{\mathsf{G}_u,n}(a,b,c) + \mathrm{O}_{\mathsf{P}}(n^{-2\eta})$. Then $\mathcal{B}_{\mathsf{G}_u,n}(a,b,c)$ reduces as desired. Note $0 < \eta \le 1/4$ so $\mathcal{M}_{n,2} = \mathcal{B}_{\mathsf{G}_u,n}(a,b,c) + \mathrm{o}_{\mathsf{P}}(1)$ uniformly in $0 < c < \infty$ and $|a|, |b| \le n^{1/4-\eta}B$.

3. *Bounding* $\mathcal{M}_{n,3}$. Note $\mathcal{M}_{n,3} = \mathbb{G}_{u,n}^{w,p}(a,b,c) - \mathbb{G}_{u,n}^{w,p}(0,0,c)$, see (3.5.7). By Assumption 3.3.1$(ia, iia, iiia, ivb, ivc)$, Theorem 3.5.5 shows $\mathcal{M}_{n,3} = \mathrm{o}_{\mathsf{P}}(1)$ uniformly in a, b, c. ■

While the above results are asymptotic expansions with $n^{1/2}$ rate for two stage least squares statistics, stochastic approximation with rate n are also required to demonstrate the main asymptotic results for the estimators and gauge. Thus we establish some n-convergence expansions in the following.

Lemma 3.9.2. *Suppose Assumption 3.3.1$(ia, iia, iiia, iv)$ holds. Then we have*

$$n^{-1}\sum_{i=1}^{n} v_i^{a,b,c} = n^{-1}\sum_{i=1}^{n} 1_{(|u_i|\leq \sigma c)} + R'_v(a,b,c),$$
$$\sum_{i=1}^{n} z_{in}z'_{in}v_i^{a,b,c} = \sum_{i=1}^{n} z_{in}z'_{in}1_{(|u_i|\leq \sigma c)} + R'_{vzz}(a,b,c),$$

where for any $B > 0$ and as $n \to \infty$

$$\sup_{0<c<\infty}\sup_{|a|,|b|\leq n^{1/4-\eta}B} |R'_v(a,b,c)| + |R'_{vzz}(a,b,c)| = \mathrm{o_P}(1).$$

Proof of Lemma 3.9.2. Adjust the first and fourth items in Lemma 3.9.1 by the rate $n^{-1/2}$ and then observe the second terms in the expansions. We find $\sup_{c\in\mathbb{R}_+} c\mathsf{f}_u(c) < \infty$, $\sup_{c\in\mathbb{R}_+} |\zeta_c^-|\mathsf{f}_u(c) < \infty$ by Assumption 3.3.1(ia), and $n^{-1/2}|a|, n^{-1/2}|b| \leq n^{-1/4-\eta}B$, while $M_{zz,n} \xrightarrow{\mathsf{P}} M_{zz}$ by Assumption 3.3.1(iva) and M_{zz}, Σ, σ are treated as constant. Notice that ζ_c^- can be expressed as the function of c, for instance under normality (3.2.5) it holds $\zeta_c^- = 2\Omega c$. Indeed the second terms in expansions are $\mathrm{O_P}(n^{-1/4-\eta})$ whereas the third terms are $\mathrm{o_P}(n^{-1/2})$ so combine them to show the remainder terms are $\mathrm{o_P}(1)$. ■

Lemma 3.9.3. *Suppose Assumption 3.3.1$(ia, iia, iiia, ivb, ivc)$ holds. Then we have*

$$n^{-1/2}\sum_{i=1}^{n} z_{in}r'_i v_i^{a,b,c} = n^{-1/2}\sum_{i=1}^{n} z_{in}r'_i 1_{(|u_i|\leq \sigma c)} + R'_{vzr}(a,b,c),$$

where for any $B > 0$ and as $n \to \infty$

$$\sup_{0<c<\infty}\sup_{|a|,|b|\leq n^{1/4-\eta}B} |R'_{vzr}(a,b,c)| = \mathrm{o_P}(1).$$

Proof of Lemma 3.9.3. Decompose the term of interest as

$$n^{-1/2}\sum_{i=1}^{n} z_{in}r'_i v_i^{a,b,c} = n^{-1/2}\sum_{i=1}^{n} z_{in}r'_i 1_{(|u_i|\leq \sigma c)} + n^{-1/2}\sum_{i=1}^{n} z_{in}r'_i\{v_i^{a,b,c} - 1_{(|u_i|\leq \sigma c)}\}.$$

Then Theorem 3.5.8 shows the second term in the above is $\mathrm{o_P}(1)$ uniformly in a, b, c. ■

3.9.2 Results for Algorithm 3.2.1

Since the indicators $v_{i,c}^{(m)}$ in 2SLS statistics equals to $v_i^{\widehat{a}_c^{(m)},\widehat{b}_c^{(m)},c}$, see (3.9.2), all processes $M_n(\widehat{a},\widehat{b},c)$ we considered depend on estimation errors $\widehat{b}=n^{1/2}(\widehat{\beta}-\beta)$, $\widehat{a}=n^{1/2}(\widehat{\sigma}-\sigma)$. If $\widehat{a},\widehat{b}$ are bounded by a compact set in probability, then $M_n(\widehat{a},\widehat{b},c)$ can be studied by analyzing the behaviour of $M_n(a,b,c)$ uniformly in $a,b\in\Theta$ for a compact set Θ. This is due to the lemma below.

Lemma 3.9.4. *If for any $\epsilon>0$ there exists a compact set Θ so $\mathsf{P}(\widehat{a},\widehat{b}\in\Theta^c)<\epsilon$ as $n\to\infty$. Then we have $\mathsf{P}\{M_n(\widehat{a},\widehat{b},c)>\epsilon\}\leq\mathsf{P}\{\sup_{a,b\in\Theta}|M_n(a,b,c)|>\epsilon\}+\epsilon$.*

Proof of Lemma 3.9.4. Let $\mathcal{A}=\{|M_n(\widehat{a},\widehat{b},c)|>\epsilon\},\mathcal{B}=(\widehat{a},\widehat{b}\in\Theta)$. Then Boole's inequality shows $\mathsf{P}(\mathcal{A})\leq\mathsf{P}(\mathcal{A}\cap\mathcal{B})+\mathsf{P}(\mathcal{B}^c)$ as $\mathcal{A}=\mathcal{A}\cap(\mathcal{B}\cup\mathcal{B}^c)=(\mathcal{A}\cap\mathcal{B})\cup(\mathcal{A}\cap\mathcal{B}^c)$. The probability $\mathsf{P}(\mathcal{A}\cap\mathcal{B})$ is bounded by $\mathsf{P}\{\sup_{a,b\in\Theta}|M_n(a,b,c)|>\epsilon\}$ while $\mathsf{P}(\mathcal{B}^c)<\epsilon$ by assumption. ■

First demonstrate consistency of $\widehat{\Pi}_c^{(m+1)}$ uniformly in $c\in[c_+,\infty)$ for a finite number $c_+>0$ given that $\widehat{b}_c^{(m)}=n^{1/2}(\widehat{\beta}_c^{(m)}-\beta)$, $\widehat{a}_c^{(m)}=n^{1/2}(\widehat{\sigma}_c^{(m)}-\sigma)$ are bounded in probability for any $m\in[0,\infty)$. The lemma below explores the uniform limiting behaviour of least squares statistics in the first stage regression (3.2.2).

Lemma 3.9.5. *Let $c_\psi=\mathsf{G}_u^{-1}(\psi)$. Suppose Assumption 3.3.1(iia,ivc) holds. Then as $n\to\infty$*

$$\sup_{0\leq\psi\leq1}|n^{-1/2}\sum_{i=1}^n z_{in}r_i'1_{(|u_i|\leq\sigma c_\psi)}|=\mathsf{o_P}(1).$$

Proof of Lemma 3.9.5. First observe that ψ is varying in the compact set $[0,1]$. Moreover $z_ir_i'1_{(|u_i|\leq\sigma c_\psi)}$ is continuous in $\psi\in[0,1]$ with probability 1 since the only discontinuous point for c_ψ is at the realized value of u_i/σ while u_i/σ is assumed to have the continuous density f_u by Assumption 3.3.1. Then $|z_ir_i'1_{(|u_i|\leq\sigma c_\psi)}|\leq|z_i||r_i|$ and $\mathsf{E}|z_i||r_i|<\infty$ since z_i is independent of r_i,u_i by Assumption 3.3.1 and $\mathsf{E}|z_i|,\mathsf{E}|r_i|<\infty$ by Assumption 3.3.1(iia,ivc). Therefore apply Uniform Law of Large Numbers, see Theorem 2 in Jennrich (1969), to show

$$n^{-1}\sum_{i=1}^n z_ir_i'1_{(|u_i|\leq\sigma c_\psi)}\xrightarrow{\mathsf{P}}\mathsf{E}z_ir_i'1_{(|u_i|\leq\sigma c_\psi)}=\mathsf{E}z_i\mathsf{E}r_i'1_{(|u_i|\leq\sigma c_\psi)}=0_{d_z\times d_x}$$

uniformly in ψ. The last equality follows $\mathsf{E}z_i=0_{d_z}$ and $\mathsf{E}|r_i'1_{(|u_i|\leq\sigma c_\psi)}|\leq\mathsf{E}|r_i|<\infty$. ■

Remark 3.9.1. *Notice in general that* $\mathsf{E}r_i'1_{(|u_i|\leq\sigma c_\psi)} \neq 0_{d_x}'$ *except when* $\psi = 1$ *and* $c_\psi = \infty$ *so* $\mathsf{E}r_i'1_{(|u_i|\leq\sigma c_\psi)} = \mathsf{E}r_i' = 0_{d_x}'$. *Write the expectation as an integral*

$$\mathsf{E}r_i'1_{(|u_i|\leq\sigma c_\psi)} = \int_{x\in\mathbb{R}^{d_x}}\int_{-c_\psi}^{c_\psi} x'\mathsf{f}_{u,r}(y,x)dy(dx)\Sigma^{1/2}.$$

The above expected value equals to $0_{d_x}'$ *if for any* $y \in \mathbb{R}, x \in \mathbb{R}^{d_x}$ *the density* $\mathsf{f}_{u,r}$ *satisfies* $\mathsf{f}_{u,r}(y,x) = -\mathsf{f}_{u,r}(y,-x)$ *or* $\mathsf{f}_{u,r}(y,x) = -\mathsf{f}_{u,r}(-y,x)$ *unless* $\psi = 1$. *However, symmetricity does not hold for the joint density in its second argument, for instance, under the normality assumption (3.2.5) when* $\Omega \neq 0_{d_x}$, *while obviously* $\mathsf{f}_{u,r}$ *is not an odd function.*

Proof of Theorem 3.3.1. For any $m \in [0,\infty)$ the $m+1$ step estimator for Π defined in (3.2.10) satisfies

$$\widehat{\Pi}_c^{(m+1)} - \Pi = (\sum_{i=1}^n z_{in}z_{in}'v_{i,c}^{(m)})^{-1}(n^{-1/2}\sum_{i=1}^n z_{in}r_i'v_{i,c}^{(m)}). \tag{3.9.3}$$

First notice $v_{i,c}^{(m)} = v_i^{\widehat{a}_c^{(m)},\widehat{b}_c^{(m)},c}$, see (3.9.2). Suppose $|\widehat{b}_c^{(m)}| + |\widehat{a}_c^{(m)}| = \mathsf{O_P}(1)$ and by Assumption 3.3.1$(ia, iia, iiia, iv)$, Lemma 3.9.2 and Lemma 3.9.3 show asymptotic expansions for least squares statistics for non-outlying observations. Lemma 3.9.4 allows to replace deterministic terms b, a by stochastic estimation errors $\widehat{b}_c^{(m)}, \widehat{a}_c^{(m)}$. Substitute these expansions into (3.9.3) to obtain

$$\widehat{\Pi}_c^{(m+1)} - \Pi = (\sum_{i=1}^n z_{in}z_{in}'1_{(|u_i|\leq\sigma c)})^{-1}(n^{-1/2}\sum_{i=1}^n z_{in}r_i'1_{(|u_i|\leq\sigma c)}) + R_\Pi'(\widehat{a}_c^{(m)},\widehat{b}_c^{(m)},c),$$

where $R_\Pi'(a,b,c)$ vanishes uniformly in $c_+ \leq c < \infty$ and $|a|, |b| \leq B$. Check the moment condition $\mathsf{E}|z_iz_i'1_{(|u_i|\leq\sigma c)}| \leq \mathsf{E}|z_i|^2 < \infty$ by Assumption 3.3.1(ivc), Law of Large Numbers shows for any $c \in [c_+,\infty)$

$$n^{-1}\sum_{i=1}^n z_iz_i'1_{(|u_i|\leq\sigma c)} \xrightarrow{\mathsf{P}} \mathsf{E}z_iz_i'1_{(|u_i|\leq\sigma c)} = \mathsf{E}z_iz_i'\mathsf{E}1_{(|u_i|\leq\sigma c)} = M_{zz}\psi,$$

where the first equality is because z_i is independent of u_i. Theorem 3.5.7 demonstrates tightness of the above sequence so $\sum_{i=1}^n z_{in}z_{in}'1_{(|u_i|\leq\sigma c)}$ converges in probability to $M_{zz}\psi$ uniformly in $c \in [c_+,\infty)$. A key to this is that c is bounded away from 0 and $M_{zz} > 0$ by Assumption 3.3.1(iva) so $\psi, M_{zz}\psi > 0$. Thus further by Lemma 3.9.5 due to Assumption 3.3.1(iia, ivc) and by Slutsky's theorem uniform consistency then follows. ■

Then we prove the theorem about stochastic expansion of the updated estimator in Algorithm 3.2.1.

Proof of Theorem 3.3.2. First notice that by substituting the Π estimator (3.2.10) the inverse term in the updated β estimator (3.2.11) is expressed as

$$\begin{aligned}\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_i z_i' v_{i,c}^{(m)}\widehat{\Pi}_c^{(m+1)} &= \widehat{\Pi}_c^{(m+1)\prime}(\sum_{i=1}^n z_i z_i' v_{i,c}^{(m)})(\sum_{i=1}^n z_i z_i' v_{i,c}^{(m)})^{-1}(\sum_{i=1}^n z_i x_i' v_{i,c}^{(m)}) \\ &= \widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_i x_i' v_{i,c}^{(m)}. \end{aligned} \tag{3.9.4}$$

Apply (3.9.4) and (3.2.1) into the $m+1$ step β estimator (3.2.11) so for any $m \in [0,\infty)$

$$\widehat{\beta}_c^{(m+1)} = \beta + (\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_i x_i' v_{i,c}^{(m)})^{-1}(\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_i u_i v_{i,c}^{(m)}).$$

Use (3.9.4) again and adjust the above equation by the rate $n^{1/2}$ to get

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) = (\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_{in} z_{in}' v_{i,c}^{(m)}\widehat{\Pi}_c^{(m+1)})^{-1}(\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_{in} u_i v_{i,c}^{(m)}). \tag{3.9.5}$$

Suppose $|\widehat{b}_c^{(m)}| + |\widehat{a}_c^{(m)}| = \mathsf{O_P}(1)$, Theorem 3.3.1 by Assumption 3.3.1$(ia, iia, iiia, iv)$ shows $\widehat{\Pi}_c^{(m+1)} \xrightarrow{\mathsf{P}} \Pi$ uniformly in $c \in [c_+, \infty)$. Apply the same argument in the proof of Theorem 3.3.1, so by Assumption 3.3.1$(ia, iia, iiia, iv)$ it allows to substitute expansions in Lemma 3.9.1 and 3.9.2 into (3.9.5). Then by Slutsky's theorem we have

$$\begin{aligned}\widehat{b}_c^{(m+1)} = (\sum_{i=1}^n \tilde{x}_{in}\tilde{x}_{in}' 1_{(|u_i|\le\sigma c)})^{-1}\{c\mathsf{f}_u(c)\sum_{i=1}^n \tilde{x}_{in}(n^{-1/2}\zeta_c^{+\prime}\Sigma^{1/2} + 2\tilde{x}_{in}')\widehat{b}_c^{(m)} \\ + \sum_{i=1}^n \tilde{x}_{in}u_i 1_{(|u_i|\le\sigma c)}\} + R_\beta(\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c),\end{aligned}$$

where $R_\beta(a,b,c)$ vanishes uniformly in $c_+ \le c < \infty$ and $|a|, |b| \le B$. Explore the inverse term in the above equation

$$\sum_{i=1}^n \tilde{x}_{in}\tilde{x}_{in}' 1_{(|u_i|\le\sigma c)} = \sum_{i=1}^n \tilde{x}_{in}\tilde{x}_{in}'\psi + \sum_{i=1}^n \tilde{x}_{in}\tilde{x}_{in}'(1_{(|u_i|\le\sigma c)} - \psi) = M_{\tilde{x}\tilde{x},n}\psi + \mathsf{o_P}(1),$$

where the second term converges in probability to zero uniformly in c by Law of Large Numbers and tightness property shown in Theorem 3.5.7, see the similar argument in the proof of Theorem 3.3.1. Note $n^{-1/2}\sum_{i=1}^n \tilde{x}_{in} \xrightarrow{\mathsf{P}} \mathsf{E}\tilde{x}_i = \Pi'\mathsf{E}z_i = 0_{d_x}$ as $\mathsf{E}z_i = 0_{d_z}$. Then explore the first term in $\widehat{b}_c^{(m+1)}$ it follows

$$(M_{\tilde{x}\tilde{x},n}\psi)^{-1}c\mathsf{f}_u(c)n^{-1/2}\sum_{i=1}^n \tilde{x}_{in}\zeta_c^{+\prime}\Sigma^{1/2}\widehat{b}_c^{(m)} \xrightarrow{\mathsf{P}} (M_{\tilde{x}\tilde{x}}\psi)^{-1}c\mathsf{f}_u(c)0_{d_x}\zeta_c^{+\prime}\Sigma^{1/2}\widehat{b}_c^{(m)} = 0_{d_x},$$

uniformly in $c_+ \leq c < \infty$. Uniformity requires $\sup_{c\in\mathbb{R}_+} c|\zeta_c^+|\mathsf{f}_u(c) < \infty$ by Assumption 3.3.1(ia) and $\widehat{b}_c^{(m)} = \mathrm{O}_\mathsf{P}(1)$. Also notice $\zeta_c^+ = 0_{d_x}$ if $\mathsf{E}x_iu_i = 0_{d_x}$ and then $\mathsf{E}r_iu_i = 0_{d_x}$ or under normality assumption (3.2.5) for $(u_i/\sigma, \Sigma^{-1/2}r_i)$ so the above term also vanishes in these two situations. Then we obtain

$$\widehat{b}_c^{(m+1)} = \frac{2c\mathsf{f}_u(c)}{\psi}\widehat{b}_c^{(m)} + (M_{\tilde{x}\tilde{x},n}\psi)^{-1}\sum_{i=1}^n \tilde{x}_{in}u_i1_{(|u_i|\leq\sigma c)} + R_\beta(\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c),$$

where $R_\beta(a,b,c)$ vanishes uniformly. A key to above derivation is that c is bounded away from 0 and $M_{\tilde{x}\tilde{x}} > 0$ by Assumption 3.3.1(iva) so $\psi, M_{\tilde{x}\tilde{x}}\psi > 0$.

Secondly for any $m \in [0,\infty)$ we get an expression for $(\widehat{\sigma}_c^{(m+1)})^2$ in (3.2.12). By Taylor expansion, first note that

$$n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = \frac{1}{2\sigma}n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\} + n^{-1/2}\mathrm{O}[n\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\}^2].$$

Then apply arguments as above to get

$$\begin{aligned}\widehat{a}_c^{(m+1)} = \frac{c(c^2-\varsigma_c^2)\mathsf{f}_u(c)}{\tau_2^c}\widehat{a}_c^{(m)} + \frac{\sigma}{2\tau_2^c}n^{-1/2}\sum_{i=1}^n(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)} \\ + \frac{(c^2-\varsigma_c^2)\mathsf{f}_u(c)\zeta_c^{-\prime}\Sigma^{1/2}}{2\tau_2^c}\widehat{b}_c^{(m)} + R_\sigma(\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c),\end{aligned}$$

where the remainder term $R_\sigma(a,b,c)$ also vanishes uniformly. Notice for analysis of $\widehat{a}_c^{(m+1)}$ we further require the moment condition $\mathsf{E}|x_iu_i|^2 < \infty$ and $\mathsf{E}|x_ix_i'|^2 < \infty$. Since x_i is correlated with u_i in the IVs context and so apply Cauchy-Schwarz inequality to the above product moment, then $\mathsf{E}|x_iu_i|^2 = \mathsf{E}|x_i|^2|u_i|^2 \leq (\mathsf{E}|x_i|^4)^{1/2}(\mathsf{E}|u_i|^4)^{1/2}$. Also note $\mathsf{E}|x_ix_i'|^2 \leq \mathsf{E}|x_i|^4$. Thus we need $\mathsf{E}|u_i|^4 < \infty$, $\mathsf{E}|r_i|^4 < \infty$, $\mathsf{E}|z_i|^4 < \infty$ by Assumption 3.3.1(ia, iia, ivc). ■

Remark 3.9.2. *For any $m \in [0,\infty)$ Algorithm 3.2.1 applies indicators $v_{i,c}^{(m)}$ in (3.2.9) to choose non-outlying observations for both the first stage regression (3.2.2) and structural equation (3.2.1), then updated estimators $\widehat{\Pi}_c^{(m+1)}$, $\widehat{\beta}_c^{(m+1)}$, $(\widehat{\sigma}_c^{(m+1)})^2$ are obtained based on the same sub-sample selected, see (3.2.10), (3.2.11), (3.2.12). The equality (3.9.4) comes from the fact that the sub-sample chosen for non-outlying observations is the same for (3.2.2) and (3.2.1), while it is significant for proving consistency of β estimator. Otherwise without (3.9.4), we can only get*

$$\begin{aligned}\widehat{\beta}_c^{(m+1)} = (\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_iz_i'v_{i,c}^{(m)}\widehat{\Pi}_c^{(m+1)})^{-1}(\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_ix_i'v_{i,c}^{(m)})\beta \\ + (\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_iz_i'v_{i,c}^{(m)}\widehat{\Pi}_c^{(m+1)})^{-1}(\widehat{\Pi}_c^{(m+1)\prime}\sum_{i=1}^n z_iu_iv_{i,c}^{(m)}),\end{aligned}$$

and the first product term before β does not reduce to the identity matrix so form the bias term, whereas the second product moment is treated by the same argument in the proof of Theorem 3.3.2. Thus the estimator for β is inconsistent if two different sub-samples are used for running the first stage and structural equation.

Remark 3.9.3. *The factor ς_c^2 in (3.2.8) needs to be adjusted in the variance estimator (3.2.12) to ensure the consistent property. For any $m \in [0,\infty)$ observe the above 1-step stochastic expansion of variance estimator $(\widehat{\sigma}_c^{(m+1)})^2$ multiplied by the rate $n^{-1/2}$*

$$\widehat{\sigma}_c^{(m+1)} - \sigma = \frac{c(c^2-\varsigma_c^2)\mathsf{f}_u(c)}{\tau_2^c}(\widehat{\sigma}_c^{(m)} - \sigma) + \frac{\sigma}{2\tau_2^c} n^{-1}\sum_{i=1}^n (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i|\leq \sigma c)} \\ + \frac{(c^2-\varsigma_c^2)\mathsf{f}_u(c)\zeta_c^{-\prime}\Sigma^{1/2}}{2\tau_2^c}(\widehat{\beta}_c^{(m)} - \beta) + \mathrm{o_P}(n^{-1/2}).$$

Since assume tightness of m step estimator, consistency holds such that $\widehat{\sigma}_c^{(m)} - \sigma$, $\widehat{\beta}_c^{(m)} - \beta$ are $\mathrm{o_P}(1)$. Furthermore, coefficients before them are bounded uniformly in $c \in [c_+,\infty)$ by Assumption 3.3.1(ia). Then the first and third terms in the above expansion converges to zero in probability. Thus uniform consistency of variance estimator $(\widehat{\sigma}_c^{(m+1)})^2$ requires the second kernel term vanishes. Bias correction factor ς_c^2 in fact appears after minus sign in the bracket in the kernel term, which lead to $n^{-1}\sum_{i=1}^n (u_i^2/\sigma^2 - \varsigma_c^2) 1_{(|u_i|\leq \sigma c)} \xrightarrow{\mathsf{P}} 0$ uniformly in c by Uniform Law of Large Numbers, see the argument in the proof of Lemma 3.9.5.

The next proof is to establish the tightness result.

Proof of Theorem 3.3.3. As Assumption 3.3.1$(ia, iia, iiia, iv)$, Theorem 3.3.2 gives for any $m \in [0,\infty)$

$$\widehat{\theta}_c^{(m+1)} = \Gamma_c \widehat{\theta}_c^{(m)} + \Lambda_c + R_\theta(\widehat{\theta}_c^{(m)}, c), \tag{3.9.6}$$

where the remainder term satisfies $\sup_{c_+\leq c<\infty}\sup_{|\theta|\leq B}|R_\theta(\theta,c)| = \mathrm{o_P}(1)$ and

$$\widehat{\theta}_c^{(m)} = \begin{pmatrix} \widehat{b}_c^{(m)} \\ \widehat{a}_c^{(m)} \end{pmatrix} = \begin{Bmatrix} n^{1/2}(\widehat{\beta}_c^{(m)} - \beta) \\ n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) \end{Bmatrix}, \quad \Gamma_c = \begin{Bmatrix} \frac{2c\mathsf{f}_u(c)}{\psi} I_{d_x} & 0_{d_x} \\ \frac{(c^2-\varsigma_c^2)\mathsf{f}_u(c)\zeta_c^{-\prime}\Sigma^{1/2}}{2\tau_2^c} & \frac{c(c^2-\varsigma_c^2)\mathsf{f}_u(c)}{\tau_2^c} \end{Bmatrix}, \tag{3.9.7}$$

$$\Lambda_c = \begin{Bmatrix} (M_{\tilde{x}\tilde{x},n}\psi)^{-1} & 0_{d_x} \\ 0'_{d_x} & \frac{\sigma}{2\tau_2^c} \end{Bmatrix} \sum_{i=1}^n \begin{Bmatrix} \tilde{x}_{in} u_i \\ n^{-1/2}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2) \end{Bmatrix} 1_{(|u_i|\leq \sigma c)}. \tag{3.9.8}$$

Apply the autoregressive equation (3.9.6) recursively to obtain the representation

$$\widehat{\theta}_c^{(m+1)} = \Gamma_c^{m+1}\widehat{\theta}_c^{(0)} + \sum_{l=0}^m \Gamma_c^l \{\Lambda_c + R_\theta(\widehat{\theta}_c^{(m-l)}, c)\}. \tag{3.9.9}$$

Note the spectral norm of a matrix is compatible with Euclidean norm of a vector so $|Mx| \leq |M||x|$, and use the triangle inequality to get

$$|\widehat{\theta}_c^{(m+1)}| \leq |\Gamma_c^{m+1}||\widehat{\theta}_c^{(0)}| + \{|\Lambda_c| + \max_{0\leq l\leq m} |R_\theta(\widehat{\theta}_c^{(l)}, c)|\} \sum_{l=0}^{m} |\Gamma_c^l|.$$

Notice Γ_c has d_x number of eigenvalues $2c\mathsf{f}_u(c)/\psi$ and one eigenvalue $c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)/\tau_2^c$, and (3.3.1) in Remark 3.3.2 shows both of them are strictly bounded by one, so we have

$$\sup_{c_+\leq c<\infty} \max |\text{eigen}(\Gamma_c)| < 1. \tag{3.9.10}$$

The fact, that the spectral radius of autoregressive coefficient matrix Γ_c is bounded by one as shown in (3.9.10), is essential to build tightness and fixed point results for the iterative system (3.9.6). Then because of (3.9.10) two equalities follow

$$\sum_{l=0}^{m} \Gamma_c^l = (I_{d_x+1} - \Gamma_c)^{-1}(I_{d_x+1} - \Gamma_c^{m+1}) = (I_{d_x+1} - \Gamma_c^{m+1})(I_{d_x+1} - \Gamma_c)^{-1}. \tag{3.9.11}$$

When $m \to \infty$, see 5.4(3b) in Lütkepohl (1996), it holds

$$\Gamma_c^{m+1} \to 0_{(d_x+1)\times(d_x+1)}, \qquad \sum_{l=0}^{\infty} \Gamma_c^l = (I_{d_x+1} - \Gamma_c)^{-1}. \tag{3.9.12}$$

See Varga (2000, Theorem 3.4), Gelfand's formula gives

$$\lim_{m\to\infty} |\Gamma_c^m|^{1/m} = \max |\text{eigen}(\Gamma_c)|. \tag{3.9.13}$$

Then (3.9.13) implies for some ω such that $\sup_{c_+\leq c<\infty} \max |\text{eigen}(\Gamma_c)| < \omega < 1$ there exists $m_0 > 0$ so for all $m > m_0$ we have

$$\sup_{c_+\leq c<\infty} |\Gamma_c^m| < \omega^m < 1. \tag{3.9.14}$$

This with the equality in (3.9.12) implies for some $1 < B_0 < \infty$

$$\sup_{0\leq m<\infty} \sup_{c_+\leq c<\infty} |\Gamma_c^m| < B_0, \quad \sup_{c_+\leq c<\infty} |(I_{d_x+1} - \Gamma_c)^{-1}| \leq \sum_{l=0}^{\infty} \sup_{c_+\leq c<\infty} |\Gamma_c^l| < B_0. \tag{3.9.15}$$

Use (3.9.15) to show for all $m \in [0, \infty)$

$$|\widehat{\theta}_c^{(m+1)}| < B_0\{|\widehat{\theta}_c^{(0)}| + |\Lambda_c| + \max_{0\leq l\leq m} |R_\theta(\widehat{\theta}_c^{(l)}, c)|\}. \tag{3.9.16}$$

Assumption 3.3.1(v) with $\eta = 1/4$ guarantees boundedness of $\widehat{\theta}_c^{(0)}$, while the kernel Λ_c is tight by Theorem 3.5.7 using Assumption 3.3.1(ia, ivc). Thus, for all $\epsilon, \delta > 0$ there exist $n_0, \Theta_0 > 0$ so that for all $n > n_0$ the set

$$\mathcal{A}_n = \{B_0 \sup_{c_+\leq c<\infty} (|\widehat{\theta}_c^{(0)}| + |\Lambda_c|) \leq \Theta_0/3, B_0 \sup_{c_+\leq c<\infty} \sup_{|\theta|\leq\Theta_0} |R_\theta(\theta, c)| < \delta/2\} \tag{3.9.17}$$

has probability larger than $1-\epsilon$.

Mathematical induction over m is used to show $\sup_{0\leq m<\infty}\sup_{c_+\leq c<\infty}|\widehat{\theta}_c^{(m)}|\leq\Theta_0$ on the set $\mathcal{A}_n$. For $m=0$ as induction starts, $\sup_{c_+\leq c<\infty}|\widehat{\theta}_c^{(0)}|\leq B_0^{-1}\Theta_0/3<\Theta_0$ holds since $B_0>1$. The induction assumption is that $\sup_{0\leq l\leq m}\sup_{c_+\leq c<\infty}|\widehat{\theta}_c^{(l)}|\leq\Theta_0$. This implies $B_0\max_{0\leq l\leq m}|R_\theta(\widehat{\theta}_c^{(l)},c)|<\delta/2$, and then the bound in (3.9.16) becomes $\sup_{c_+\leq c<\infty}|\widehat{\theta}_c^{(m+1)}|<2\Theta_0/3+\delta/2<\Theta_0$ so that $\sup_{0\leq l\leq m+1}\sup_{c_+\leq c<\infty}|\widehat{\theta}_c^{(l)}|\leq\Theta_0$. ■

Proof of Corollary 3.3.1. Tightness of initial estimators for structural parameters β, σ^2 in (3.2.1) implicitly assumes consistency of the initial estimator for location parameter Π in the first stage regression (3.2.2). Combine Theorem 3.3.3 and 3.3.1, then consistency of $\widehat{\Pi}_c^{(m)}$ follows uniformly in $m\in[0,\infty)$ and $c\in[c_+,\infty)$. ■

With tightness results in Theorem 3.3.3, one-step expansion in Theorem 3.3.2 can be applied recursively to obtain Theorem 3.3.4.

Proof of Theorem 3.3.4. By Assumption 3.3.1, Theorem 3.3.2 shows the one-step expansion (3.9.6). Then apply it recursively to obtain (3.9.9) so for any $m\in[0,\infty)$

$$\widehat{\theta}_c^{(m+1)}=\Gamma_c^{m+1}\widehat{\theta}_c^{(0)}+\sum_{l=0}^{m}\Gamma_c^l\Lambda_c+\sum_{l=0}^{m}\Gamma_c^lR_\theta(\widehat{\theta}_c^{(m-l)},c),$$

where $\sup_{c_+\leq c<\infty}\sup_{|\theta|\leq B}|R_\theta(\theta,c)|=\mathsf{o_P}(1)$. As the spectral radius of Γ_c is bounded by one, see (3.9.10), then (3.9.15) shows for $1<B_0<\infty$

$$\sup_{0\leq m<\infty}\sup_{c_+\leq c<\infty}|\Gamma_c^m|<B_0,\quad \sup_{c_+\leq c<\infty}|\sum_{l=0}^{m}\Gamma_c^l|\leq\sup_{c_+\leq c<\infty}\sum_{l=0}^{m}|\Gamma_c^l|\leq\sum_{l=0}^{\infty}\sup_{c_+\leq c<\infty}|\Gamma_c^l|<B_0.$$

Further with tightness $\sup_{0\leq m<\infty}\sup_{c_+\leq c<\infty}|\widehat{\theta}_c^{(m)}|=\mathsf{O_P}(1)$ shown in Theorem 3.3.3, the third term in the representation of $\widehat{\theta}_c^{(m+1)}$ vanishes, then apply the second equality in (3.9.11) so uniformly in $c\in[c_+,\infty)$ we have for any $m\in[0,\infty)$

$$\widehat{\theta}_c^{(m+1)}=\Gamma_c^{m+1}\widehat{\theta}_c^{(0)}+(I_{d_x+1}-\Gamma_c^{m+1})(I_{d_x+1}-\Gamma_c)^{-1}\Lambda_c+\mathsf{o_P}(1). \tag{3.9.18}$$

See Γ_c in (3.9.7), by matrix manipulations we get

$$\Gamma_c^{m+1}=\begin{pmatrix}\varrho_{\beta\beta,c}^{(m+1)}I_{d_x} & 0_{d_x}\\ \varrho_{\sigma\beta,c}^{(m+1)\prime} & \varrho_{\sigma\sigma,c}^{(m+1)}\end{pmatrix},(I_{d_x+1}-\Gamma_c^{m+1})(I_{d_x+1}-\Gamma_c)^{-1}=\begin{pmatrix}\psi\varrho_{\beta\tilde{x}u,c}^{(m+1)}I_{d_x} & 0_{d_x}\\ \psi\varrho_{\sigma\tilde{x}u,c}^{(m+1)\prime} & \frac{2\tau_2^c}{\sigma}\varrho_{\sigma uu,c}^{(m+1)}\end{pmatrix},$$

see expressions for $\varrho_{\beta\beta,c}^{(m+1)}$, $\varrho_{\beta\tilde{x}u,c}^{(m+1)}$, $\varrho_{\sigma\sigma,c}^{(m+1)}$, $\varrho_{\sigma uu,c}^{(m+1)}$, $\varrho_{\sigma\beta,c}^{(m+1)}$, $\varrho_{\sigma\tilde{x}u,c}^{(m+1)}$ in Theorem 3.3.4. Substitute these and $\widehat{\theta}_c^{(0)}$, Λ_c into (3.9.18), see $\widehat{\theta}_c^{(0)}$ in (3.9.7) and Λ_c in (3.9.8), then the expression for $\widehat{\theta}_c^{(m+1)}$ is established. ■

The next step is to prove the fixed point theorem when iteration step becomes sufficiently large.

Proof of Theorem 3.3.5. Since the spectral radius of Γ_c is strictly smaller than one, see (3.9.10), when $m \to \infty$ we can apply (3.9.12) to the recursive representation (3.9.18) shown in Theorem 3.3.4 by Assumption 3.3.1, then as $n \to \infty$ and uniformly in $c \in [c_+, \infty)$ we have the fixed point

$$\widehat{\theta}_c^{(*)} = \widehat{\theta}_c^{(\infty)} = (I_{d_x+1} - \Gamma_c)^{-1}\Lambda_c. \tag{3.9.19}$$

See Γ_c in (3.9.7), we get

$$(I_{d_x+1} - \Gamma_c)^{-1} = \begin{bmatrix} \frac{\psi}{\psi - 2c\mathsf{f}_u(c)} I_{d_x} & 0_{d_x} \\ \frac{\psi(c^2-\varsigma_c^2)\mathsf{f}_u(c)\zeta_c^{-\prime}\Sigma^{1/2}}{2\{\psi-2c\mathsf{f}_u(c)\}\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}} & \frac{\tau_2^c}{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)} \end{bmatrix}.$$

Then substitute this and Λ_c into (3.9.19), see Λ_c in (3.9.8), so the expression for the fixed point $\widehat{\theta}_c^{(*)}$ is demonstrated.

Substitute (3.9.9) and (3.9.19) into the deviation $\widehat{\Delta}_c^{(m+1)} = \widehat{\theta}_c^{(m+1)} - \widehat{\theta}_c^{(*)}$, then apply the second equality in (3.9.12) so uniformly in $c \in [c_+, \infty)$ we have for any $m \in [0, \infty)$

$$\widehat{\Delta}_c^{(m+1)} = \Gamma_c^{m+1}\{\widehat{\theta}_c^{(0)} - (I_{d_x+1} - \Gamma_c)^{-1}\Lambda_c\} + \sum_{l=0}^{m} \Gamma_c^l R_\theta(\widehat{\theta}_c^{(m-l)}, c).$$

To bound $\widehat{\Delta}_c^{(m+1)}$, use the triangle inequality and $|Mx| \leq |M||x|$ to get

$$|\widehat{\Delta}_c^{(m+1)}| \leq |\Gamma_c^{m+1}|\{|\widehat{\theta}_c^{(0)}| + |(I_{d_x+1} - \Gamma_c)^{-1}||\Lambda_c|\} + \max_{0\leq l\leq m} |R_\theta(\widehat{\theta}_c^{(l)}, c)| \sum_{l=0}^{m} |\Gamma_c^l|.$$

Now consider the large m such that $m > m_0$ then (3.9.14) can be used, so along with the second inequality in (3.9.15) it follows

$$|\widehat{\Delta}_c^{(m+1)}| < \omega^{m+1}(|\widehat{\theta}_c^{(0)}| + B_0|\Lambda_c|) + B_0 \max_{0\leq l\leq m} |R_\theta(\widehat{\theta}_c^{(l)}, c)|.$$

On the set $\mathcal{A}_n$ as in (3.9.17), Theorem 3.3.3 shows $\sup_{0\leq m<\infty}\sup_{c_+\leq c<\infty}|\widehat{\theta}_c^{(m)}| \leq \Theta_0$, so

$$|\widehat{\Delta}_c^{(m+1)}| < \omega^{m+1}(B_0^{-1}\Theta_0/3 + \Theta_0/3) + \delta/2 < \omega^{m+1}\Theta_0 + \delta/2,$$

where the second inequality follows by $1 < B_0 < \infty$. As $0 < \omega < 1$, ω^{m+1} declines exponentially so m_0 can be chosen sufficiently large that for $m > m_0$ then $\omega^{m+1}\Theta_0 < \delta/2$. Thus for $n > n_0$, $m > m_0$ we have $\mathsf{P}(\sup_{c_+\leq c<\infty}|\widehat{\Delta}_c^{(m+1)}| < \delta) > 1 - \epsilon$. ■

3.9.3 Results for Algorithm 3.2.2 and 3.2.3

We first present a lemma to describe the limiting distribution of kernels.

Lemma 3.9.6. *Suppose Assumption 3.3.1(ia, iva, ivc) holds. Then as $n \to \infty$ we have for any $c \in [c_+, \infty)$*

$$\sum_{i=1}^{n} \begin{pmatrix} \tilde{x}_{in} u_i \\ \tilde{x}_{in} u_i 1_{(|u_i| \le \sigma c)} \end{pmatrix} \xrightarrow{\mathsf{D}} \mathsf{N}\left\{ \begin{pmatrix} 0_{d_x} \\ 0_{d_x} \end{pmatrix}, \sigma^2 \tau_2^c \begin{pmatrix} \frac{1}{\tau_2^c} M_{\tilde{x}\tilde{x}} & M_{\tilde{x}\tilde{x}} \\ M_{\tilde{x}\tilde{x}} & M_{\tilde{x}\tilde{x}} \end{pmatrix} \right\}.$$

Proof of Theorem 3.9.6. As $\tilde{x}_i = \Pi' z_i$ and z_i is independent of u_i by Assumption 3.3.1. Then $\mathsf{E}\tilde{x}_i u_i = 0_{d_x}$ and $\mathsf{Var}(\tilde{x}_i u_i) = \mathsf{E}\tilde{x}_i \tilde{x}_i' u_i^2 = \mathsf{E}u_i^2 \mathsf{E}\tilde{x}_i \tilde{x}_i' = \sigma^2 M_{\tilde{x}\tilde{x}}$. Furthermore, we have $\mathsf{E}\tilde{x}_i u_i 1_{(|u_i| \le \sigma c)} = \mathsf{E}\tilde{x}_i \mathsf{E}u_i 1_{(|u_i| \le \sigma c)} = 0_{d_x}$ and

$$\mathsf{Var}(\tilde{x}_i u_i 1_{(|u_i| \le \sigma c)}) = \mathsf{E}\tilde{x}_i \tilde{x}_i' u_i^2 1_{(|u_i| \le \sigma c)} = \mathsf{E}u_i^2 1_{(|u_i| \le \sigma c)} \mathsf{E}\tilde{x}_i \tilde{x}_i' = \sigma^2 \tau_2^c M_{\tilde{x}\tilde{x}}.$$

Finally we calculate $\mathsf{Cov}(\tilde{x}_i u_i, \tilde{x}_i u_i 1_{(|u_i| \le \sigma c)}) = \mathsf{E}\tilde{x}_i \tilde{x}_i' u_i^2 1_{(|u_i| \le \sigma c)} = \sigma^2 \tau_2^c M_{\tilde{x}\tilde{x}}$. Note that $\tilde{x}_{in} = n^{-1/2} \tilde{x}_i$, so apply CLT to obtain for any $c \in [c_+, \infty)$

$$n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} \tilde{x}_i u_i \\ \tilde{x}_i u_i 1_{(|u_i| \le \sigma c)} \end{pmatrix} \xrightarrow{\mathsf{D}} \mathsf{N}\left\{ \begin{pmatrix} 0_{d_x} \\ 0_{d_x} \end{pmatrix}, \begin{pmatrix} \sigma^2 M_{\tilde{x}\tilde{x}} & \sigma^2 \tau_2^c M_{\tilde{x}\tilde{x}} \\ \sigma^2 \tau_2^c M_{\tilde{x}\tilde{x}} & \sigma^2 \tau_2^c M_{\tilde{x}\tilde{x}} \end{pmatrix} \right\}. \qquad \blacksquare$$

The next step is to provide the expansion and limiting distribution of the full sample two stage least squares estimator.

Proof of Lemma 3.3.1. Substitute (3.2.2) into the location estimator for Π to obtain

$$\widetilde{\Pi} = \Pi + M_{zz,n}^{-1} n^{-1/2} \sum_{i=1}^{n} z_{in} r_i' \xrightarrow{\mathsf{P}} \Pi + M_{zz}^{-1} \mathsf{E} z_i r_i' = \Pi + M_{zz}^{-1} 0_{d_z \times d_x} = \Pi,$$

where the probability limit follows by Assumption 3.3.1(iva) that $M_{zz,n} \xrightarrow{\mathsf{P}} M_{zz} > 0$, LLN, Slutsky's theorem, and $\mathsf{E} z_i r_i' = 0_{d_z \times d_x}$. Notice LLN requires moment conditions $\mathsf{E}|r_i|, \mathsf{E}|z_i| < \infty$ by Assumption 3.3.1(iia, ivc). We have $\widetilde{\Pi}' \sum_{i=1}^{n} z_i z_i' \widetilde{\Pi} = \widetilde{\Pi}' \sum_{i=1}^{n} z_i x_i'$, see (3.9.4), then apply this and structural equation (3.2.1) to get

$$n^{1/2}(\widetilde{\beta} - \beta) = (\widetilde{\Pi}' M_{zz,n} \widetilde{\Pi})^{-1} (\widetilde{\Pi}' \sum_{i=1}^{n} z_{in} u_i) = M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + \mathsf{o}_{\mathsf{P}}(1),$$

where the second equality is due to consistency of $\widetilde{\Pi}$. Thus the limiting distribution follows by Lemma 3.9.6 and Assumption 3.3.1(iva) that $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}} > 0$. $\blacksquare$

Then weak convergence results are proved for Algorithm 3.2.2, 3.2.3.

Proof of Theorem 3.3.6. Algorithm 3.2.2 chooses the full sample two stage least squares as the initial estimator $\widehat{\beta}_c^{(0)}$ in Algorithm 3.2.1. Thus substitute the asymptotic expansion of $n^{1/2}(\widehat{\beta}_c^{(0)} - \beta) = n^{1/2}(\widetilde{\beta} - \beta)$ in Lemma 3.3.1 into the recursive equation in Theorem 3.3.4, then uniformly in $c \in [c_+, \infty)$ we have for any $m \in [0, \infty)$

$$\mathbb{G}_n^{(m+1)}(c) = \begin{pmatrix} \varrho_{\beta\beta,c}^{(m+1)} M_{\tilde{x}\tilde{x},n}^{-1} \\ \varrho_{\beta\tilde{x}u,c}^{(m+1)} M_{\tilde{x}\tilde{x},n}^{-1} \end{pmatrix}' \sum_{i=1}^{n} \begin{pmatrix} \tilde{x}_{in} u_i \\ \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} \end{pmatrix} + \mathsf{o}_{\mathsf{P}}(1),$$

where expressions of $\varrho_{\beta\beta,c}^{(m+1)}$, $\varrho_{\beta\tilde{x}u,c}^{(m+1)}$ are shown in Theorem 3.3.4. Then $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}} > 0$, Lemma 3.9.6, and Slutsky's theorem show for any $c \in [c_+, \infty)$

$$\mathbb{G}_n^{(m+1)}(c) \xrightarrow{\mathsf{D}} \begin{pmatrix} \varrho_{\beta\beta,c}^{(m+1)} M_{\tilde{x}\tilde{x}}^{-1} \\ \varrho_{\beta\tilde{x}u,c}^{(m+1)} M_{\tilde{x}\tilde{x}}^{-1} \end{pmatrix}' \mathsf{N}\left\{ \begin{pmatrix} 0_{d_x} \\ 0_{d_x} \end{pmatrix}, \sigma^2 \tau_2^c \begin{pmatrix} \frac{1}{\tau_2^c} M_{\tilde{x}\tilde{x}} & M_{\tilde{x}\tilde{x}} \\ M_{\tilde{x}\tilde{x}} & M_{\tilde{x}\tilde{x}} \end{pmatrix} \right\}.$$

Since a transformation of multivariate normal is still normal, it follows

$$\mathbb{G}_n^{(m+1)}(c) \xrightarrow{\mathsf{D}} \mathsf{N}[0_{d_x}, \{(\varrho_{\beta\beta,c}^{(m+1)})^2 + 2\tau_2^c \varrho_{\beta\beta,c}^{(m+1)} \varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c (\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\} \sigma^2 M_{\tilde{x}\tilde{x}}^{-1}].$$

Theorem 3.3.3 demonstrates the process $\mathbb{G}_n^{(m+1)}$ is tight for any $m \in [0, \infty)$ thus

$$\mathbb{G}_n^{(m+1)} \rightsquigarrow \mathbb{G}^{(m+1)},$$

where the weak limit $\mathbb{G}^{(m+1)}$ is a zero mean Gaussian process with the variance

$$\mathsf{Var}\{\mathbb{G}^{(m+1)}(c)\} = \{(\varrho_{\beta\beta,c}^{(m+1)})^2 + 2\tau_2^c \varrho_{\beta\beta,c}^{(m+1)} \varrho_{\beta\tilde{x}u,c}^{(m+1)} + \tau_2^c (\varrho_{\beta\tilde{x}u,c}^{(m+1)})^2\} \sigma^2 M_{\tilde{x}\tilde{x}}^{-1}. \qquad \blacksquare$$

Proof of Corollary 3.3.2. This is a special case of Theorem 3.3.6 when $m = 0$ such that $\varrho_{\beta\beta,c}^{(1)} = 2c\mathsf{f}_u(c)/\psi$, $\varrho_{\beta\tilde{x}u,c}^{(1)} = \psi^{-1}$, see (3.3.2). $\blacksquare$

Proof of Theorem 3.3.7. The first step updated estimator for Π is expressed as

$$\begin{aligned} \widehat{\Pi}_c^{(1)} - \Pi = & \Big(\sum_{j=1,2} \frac{n_{3-j}}{n} \sum_{i \in \mathcal{I}_{3-j}} z_{in_{3-j}} z_{in_{3-j}}' 1_{(|y_i - x_i' \widehat{\beta}_j| \leq \widehat{\sigma}_j c)}\Big)^{-1} \\ & \Big(\sum_{j=1,2} \frac{n_{3-j}}{n} n_{3-j}^{-1/2} \sum_{i \in \mathcal{I}_{3-j}} z_{in_{3-j}} r_i' 1_{(|y_i - x_i' \widehat{\beta}_j| \leq \widehat{\sigma}_j c)}\Big). \end{aligned}$$

Assumption 3.3.1(ia, iia, iva, ivc) holds for each of sub-sample $\mathcal{I}_j$, so two stage least squares $\widehat{\beta}_j, \widehat{\sigma}_j^2$ are tight for $j = 1, 2$, see Lemma 3.3.1. Argue as Theorem 3.3.1 to obtain

$$\widehat{\Pi}_c^{(1)} - \Pi = \Big(\sum_{i=1}^{n} z_{in} z_{in}' 1_{(|u_i| \leq \sigma c)}\Big)^{-1} \Big(n^{-1/2} \sum_{i=1}^{n} z_{in} r_i' 1_{(|u_i| \leq \sigma c)}\Big) + \mathsf{o}_{\mathsf{P}}(1),$$

uniformly in $c \in [c_+, \infty)$. Therefore uniform consistency of $\widehat{\Pi}_c^{(1)}$ immediately follows, see proof of Theorem 3.3.1.

The updated estimator for β is expressed as

$$n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = (\widehat{\Pi}_c^{(1)\prime} \sum_{j=1,2} \frac{n_{3-j}}{n} \sum_{i \in \mathcal{I}_{3-j}} z_{in_{3-j}} z'_{in_{3-j}} 1_{(|y_i - x_i'\widehat{\beta}_j| \leq \widehat{\sigma}_j c)} \widehat{\Pi}_c^{(1)})^{-1}$$
$$(\widehat{\Pi}_c^{(1)\prime} \sum_{j=1,2} (\frac{n_{3-j}}{n})^{1/2} \sum_{i \in \mathcal{I}_{3-j}} z_{in_{3-j}} u_i 1_{(|y_i - x_i'\widehat{\beta}_j| \leq \widehat{\sigma}_j c)}).$$

Argue along the lines of Theorem 3.3.2, then it follows

$$n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = (M_{\tilde{x}\tilde{x},n}\psi)^{-1}\{2c\mathsf{f}_u(c) \sum_{j=1,2} (\frac{n_{3-j}}{n})^{1/2} (\frac{n_{3-j}}{n_j})^{1/2} M_{\tilde{x}\tilde{x},n_{3-j}}^{\mathcal{I}_{3-j}} n_j^{1/2}(\widehat{\beta}_j - \beta)$$
$$+ \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)}\} + \mathsf{o}_{\mathsf{P}}(1),$$

uniformly in $c \in [c_+, \infty)$ and where we denote $M_{\tilde{x}\tilde{x},n_j}^{\mathcal{I}_j} = \sum_{i \in \mathcal{I}_j} \tilde{x}_{in_j} \tilde{x}'_{in_j}$ for $j = 1, 2$. Again Assumption 3.3.1(ia, iia, iva, ivc) holds for each sub-sample $\mathcal{I}_j$, so Lemma 3.3.1 shows $n_j^{1/2}(\widehat{\beta}_j - \beta) = (M_{\tilde{x}\tilde{x},n_j}^{\mathcal{I}_j})^{-1} \sum_{i \in \mathcal{I}_j} \tilde{x}_{in_j} u_i + \mathsf{o}_{\mathsf{P}}(1)$ and notice $M_{\tilde{x}\tilde{x},n_j}^{\mathcal{I}_j} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}}$ as $n \to \infty$ such that $n_j \to \infty$ for each $j = 1, 2$. Thus we have

$$n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = \frac{2c\mathsf{f}_u(c)}{\psi} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{j=1,2} \frac{n_{3-j}}{n_j} \sum_{i \in \mathcal{I}_j} \tilde{x}_{in} u_i + (M_{\tilde{x}\tilde{x},n}\psi)^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} + \mathsf{o}_{\mathsf{P}}(1),$$

and if assume $n_1 = n_2 = n/2$ it further holds

$$n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = \frac{2c\mathsf{f}_u(c)}{\psi} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + (M_{\tilde{x}\tilde{x},n}\psi)^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} + \mathsf{o}_{\mathsf{P}}(1).$$

We find the expansion of $n^{1/2}(\widehat{\beta}_c^{(1)} - \beta)$ for Algorithm 3.2.3 is the same as Algorithm 3.2.2, see Corollary 3.3.2, therefore two algorithms have the identical asymptotics even they apply different initial estimates when running Algorithm 3.2.1. ■

Finally we establish weak convergence for the fixed point of Algorithm 3.2.1, 3.2.2, 3.2.3.

Proof of Theorem 3.3.8. Consider Algorithm 3.2.1, 3.2.2, 3.2.3. Let $m \to \infty$ then Theorem 3.3.5 shows the fixed point for $c \in [c_+, \infty)$

$$\mathbb{G}_n^{(*)}(c) = \frac{1}{\psi - 2c\mathsf{f}_u(c)} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)}.$$

As $n \to \infty$ then $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}} > 0$ and Lemma 3.9.6 demonstrate finite dimensional convergence

$$\mathbb{G}_n^{(*)}(c) \xrightarrow{\mathsf{D}} \mathsf{N}[0_{d_x}, \frac{\tau_2^c}{\{\psi - 2c\mathsf{f}_u(c)\}^2}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}].$$

Tightness shown in Theorem 3.3.3 further establishes weak convergence

$$\mathbb{G}_n^{(*)} \rightsquigarrow \mathbb{G}^{(*)},$$

where $\mathbb{G}^{(*)}$ is a zero mean Gaussian process with variance

$$\mathsf{Var}\{\mathbb{G}^{(*)}(c)\} = \frac{\tau_2^c}{\{\psi - 2c\mathsf{f}_u(c)\}^2}\sigma^2 M_{\tilde{x}\tilde{x}}^{-1}. \qquad \blacksquare$$

3.9.4 Results for a new type of Hausman test

We first prove the stochastic expansion and weak limit of the difference processes between the robustified and full sample 2SLS.

Proof of Theorem 3.3.9. Notice that $\mathbb{H}_n^{(m+1)}(c)$ can be rewritten as

$$\mathbb{H}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) - n^{1/2}(\widetilde{\beta} - \beta).$$

Substitute expansions of $n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta)$ in Theorem 3.3.6, 3.3.7 and $n^{1/2}(\widetilde{\beta} - \beta)$ in Lemma 3.3.1 into the above term, then the expansion is attained for $\mathbb{H}_n^{(m+1)}(c)$ for any $c \in [c_+, \infty)$ and $m \in [0, \infty)$. Argue along Proof of Theorem 3.3.6 but replace $\varrho_{\beta\beta,c}^{(m+1)}$ by $\varrho_{\beta\beta,c}^{(m+1)} - 1$ to obtain the weak Gaussian limit $\mathbb{H}^{(m+1)}$ of processes $\mathbb{H}_n^{(m+1)}$. $\blacksquare$

The next two proofs establish corollaries on a new Hausman type test for performing outlier robustness checks.

Proof of Corollary 3.3.3. Gaussian weak limit in Theorem 3.3.9 immediately implies the finite dimensional convergence result as normal distribution. The limiting χ^2 distribution then follows for the proposed test statistics with the degree of freedom d_x as the dimension of the structural parameter β. $\blacksquare$

Proof of Corollary 3.3.4. Set $m = 0, \infty$ and substitute the terms $\varrho_{\beta\beta,c}^{(1)} = 2c\mathsf{f}_u(c)/\psi$, $\varrho_{\beta\tilde{x}u,c}^{(1)} = \psi^{-1}$, $\varrho_{\beta\beta,c}^{(\infty)} = 0$, $\varrho_{\beta\tilde{x}u,c}^{(\infty)} = \{\psi - 2c\mathsf{f}_u(c)\}^{-1}$ from (3.3.2), (3.3.3) into $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ given by (3.3.9), then our proposed tests follow. $\blacksquare$

Conclusion

This thesis established the asymptotic theory of the iterated one-step Huber-skip M-estimators in the (IVs) regression equation. Outliers are ubiquitous in economic data and very likely to distort estimation, statistical inference, and forecasting. In applied economics, there is a recurrent critique questioning whether outliers may have driven the empirical results. The iterated one-step Huber-skip M-estimators help to alleviate the pernicious effects of unmodelled outliers and provide a statistical tool to perform outlier robustness checks.

The asymptotic theory was firstly derived in the linear regression equation, where the cut-off value chosen for the iterated one-step Huber-skip M-estimators can be varying with sample size and iteration step. The first ingredient is to show the tightness property of the iterated estimator even with the varying cut-off. Since the contraction mapping holds for the one-step stochastic expansion of the updated estimator in terms of the original one, the algorithm converges to the fixed point upon through infinite iterations. Lastly, the Poisson exceedance theory was established for the gauge that arises in the situation where the cut-off is set to allow the fixed number of outliers regardless of the sample size. The theory is valid for regressors with any form of time series correlation, whether they are stationary, deterministic trends, or unit roots.

Secondly, using the distributional results on the gauge, we proposed two sets of testing procedures for the null hypothesis of no outliers based on the iterated one-step Huber-skip M-estimators: first, a set of tests on the proportion and count of detected outliers; second, a set of scaling tests. Tests on the proportion and count have power against the number of outliers, while scaling tests have power against both the number and magnitude of outliers. These tests perform well with size close to nominal levels and high power against both magnitude and number of outliers.

Lastly, this thesis turned into the IVs regression, to which the asymptotic theory of the iterated one-step Huber-skip estimators was extended. As compared to the ordinary estimator the Huber-skip type is less efficient under the null, though robust to outliers

under alternatives, so it is of our first interest to compare the relative efficiency between these two estimates. We then suggested the adjusted inferential procedure on the structural parameters robust to outliers. At last, a Hausman type test was constructed to perform outlier robustness checks, enabling us to compare between two estimates (the robust and ordinary) in a formal statistical way.

To analyse the Huber-skip type estimator asymptotically in the (IVs) regression, this thesis generalized the theory for a new class of the weighted and marked empirical processes of residuals, which is also of independent interest in applied probability. Our applications show the utility and versatility of the proposed theory. In the second chapter, we re-visited the studies by Mankiw, Romer, and Weil (1992) and Pretis (2019), estimating the Solow growth model and impact of carbon tax on CO2 emissions respectively. The thrid chapter re-examined the outlier robustness analysis in Acemoglu et al. (2017), empirically addressing the debate on the performance of democratic institutions.

There remains a fruitful area for future research, including those discussed at the end of each individual chapters. For example, the robust device in this thesis relies on the distributional assumption on the errors, since outliers are defined and detected relative to this reference distribution, also from which the cut-off is selected. However, in practice, the assumed distribution might not hold in the true DGP, say the cut-off 1.96 is selected from normal for 5% level but the errors in fact follow the t-distribution. Thus, given the cut-off it would be of interest to propose an inference robust to arbitrary distributions of errors. Clustered samples are increasingly popular in applied economics and are related to panel data. Likewise, it deserves to propose inference robust to arbitrary forms of within-cluster correlation. An exploration would be applying the bootstrap to perform inference on the structural parameters. The bootstrapping method could have finite sample improvements on the one hand when these assumptions hold, and could still work on the other hand.

References

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70, 91-117.

Acemoglu, D., Johnson, S., and Robinson, J.A. (2001). The colonial origins of comparative development: an empirical investigation. *American Economic Review* 91, 1369-1401.

Acemoglu, D., Johnson, S., and Robinson, J.A. (2012). The colonial origins of comparative development: an empirical investigation: reply. *American Economic Review* 102, 3077-3110.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J.A. (2017). Democracy does cause growth. *Journal of Political Economy*, forthcoming.

Adichie, J.N. (1967). Estimates of regression parameters based on rank tests. *Annals of Mathematical Statistics* 38, 894-904.

Adler, R.J. and Taylor, J.E. (2009). *Random Fields and Geometry*. New York: Springer.

Ahumada, H. and Cornejo, M. (2016). Forecasting food prices: The case of corn, soybeans and wheat. *International Journal of Forecasting* 32, 838-848.

Albouy, D.Y. (2012). The colonial origins of comparative development: an empirical investigation: comment. *American Economic Review* 102, 3059-3076.

Anundsen, A.K. (2015). Econometric regime shifts and the US subprime bubble. *Journal of Applied Econometrics* 30, 145-169.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies* 58, 277-297.

Auerbach, A.J., Hassett, K.A., and Oliner, S.D. (1994). Reassessing the social returns to equipment investment. *Quarterly Journal of Economics* 109, 789-802.

Bazinas, V. and Nielsen, B. (2015). Causal transmission in reduced-form models. *Working Paper*.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 1165-1188.

Bercu, B. and Touati, A. (2008). Exponential inequalities for self-normalized martingales with applications. *Annals of Applied Probability* 18, 1848-1869.

Berenguer-Rico, V., Johansen, S., and Nielsen, B. (2019a). The analysis of marked and weighted empirical processes of estimated residuals. *Working Paper*.

Berenguer-Rico, V., Johansen, S., and Nielsen, B. (2019b). Uniform consistency of marked and weighted empirical distribution of residuals. *Working Paper*.

Berenguer-Rico, V. and Nielsen, B. (2018). Marked and weighted empirical processes of residuals with applications to robust regressions. *Working Paper*.

Berenguer-Rico, V. and Wilms, I. (2018). White heteroscedasticty testing after outlier removal. *Working Paper*.

Bickel, P.J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association* 70, 428-434.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Castle, J.L., Doornik, J.A., Hendry, D.F., and Pretis, F. (2015). Detecting location shifts by step-indicator saturation during model selection. *Econometrics* 3, 240-264.

Castle, J.L. and Hendry, D.F. (2009). The long-run determinants of UK wages, 1860–2004. *Journal of Macroeconomics* 31, 5-28.

Castle, J.L. and Hendry, D.F. (2012). Automatic selection for non-linear models. In *System Identification, Environmental Modelling, and Control System Design* 229-250. London: Springer.

Cavaliere, G. and Georgiev, I. (2013). Exploiting infinite variance through dummy variables in nonstationary autoregression. *Econometric Theory* 29, 1162-1195.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 15.

Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica* 73, 245-261.

Chhaochharia, V. and Grinstein, Y. (2009). CEO compensation and board structure. *Journal of Finance* 64, 231-261.

Clements, M. and Hendry, D. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Croux, C. and Wilms, I. (2016). Discussion of 'Asymptotic theory of outlier detection algorithms for linear time series regression models'. *Scandinavian Journal of Statistics* 43, 353-356.

De Long, J.B. and Summers, L.H. (1991). Equipment investment and economic growth. *Quarterly Journal of Economics* 106, 445-502.

De Long, J.B. and Summers, L.H. (1994). Equipment investment and economic growth: reply. *Quarterly Journal of Economics* 109, 803-807.

Dollinger, M.B. and Staudte, R.G. (1991). Influence functions of iteratively re-weighted least squares estimators. *Journal of the American Statistical Association* 86, 709-716.

Donsker, M.D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics* 23, 277-281.

Doob, J.L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics* 20, 393-403.

Doornik, J.A. (2009). Autometrics. In Castle, J.L. and Shephard, N. (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honor of David F. Hendry*, pp. 88-121. Oxford: Oxford University Press.

Doornik, J.A. and Hendry, D.F. (2013). *PcGive 14*. London: Timberlake Consultants.

Doornik, J.A. and Hendry, D.F. (2016). Outliers and model selection: Discussion of the Paper by Søren Johansen and Bent Nielsen. *Scandinavian Journal of Statistics* 43, 360-365.

Dreger, C. and Wolters, J. (2014). Money demand and the role of monetary indicators in forecasting euro area inflation. *International Journal of Forecasting* 30, 303-312.

Durbin, J. (1954). Errors in variables. *Revue de l'institut International de Statistique* 22, 23-32.

Engle, R.F., Hendry, D.F., and Richard, J.F. (1983). Exogeneity. *Econometrica* 51, 277-304.

Engler, E. and Nielsen, B. (2009). The empirical process of autoregressive residuals. *Econometrics Journal* 12, 367-381.

Ericsson, N.R. (2017). How biased are US government forecasts of the federal debt? *International Journal of Forecasting* 33, 543-559.

Fabrizio, K.R., Rose, N.L., and Wolfram, C.D. (2007). Do markets reduce costs? Assessing the impact of regulatory restructuring on US electric generation efficiency. *American Economic Review* 97, 1250-1277.

Freedman, D.A. (1975). On tail probabilities for martingales. *Annals of Probability* 3, 100-118.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* 11, 1-21.

Guthrie, K., Sokolowsky, J., and Wan, K.M. (2012). CEO compensation and board structure revisited. *Journal of Finance* 67, 1149-1168.

Hahn, J., Hausman, J.A., and Kuersteiner, G.M. (2001). Bias corrected instrumental variables estimation for dynamic panel models with fixed effects. *Working Paper*.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics*. New York: Wiley.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029-1054.

Hansen, L.P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* 14, 262-280.

Hausman, J.A. (1978). Specification tests in econometrics. *Econometrica* 46, 1251-1271.

Hausman, J.A. and Taylor, W.E. (1981). A generalized specification test. *Economics Letters* 8, 239-245.

Hendry, D.F. (1999). An econometric analysis of US food expenditure, 1931-1989. In Magnus, J.R. and Morgan, M.S. (eds.) *Methodology and Tacit Knowledge: Two Experiments in Econometrics*, pp. 341-361. New York: Wiley.

Hendry, D.F. (2011). Revisiting UK consumers' expenditure: cointegration, breaks and robust forecasts. *Applied Financial Economics* 21, 19-32.

Hendry, D.F. and Doornik, J.A. (2014). *Empirical Model Discovery and Theory Evaluation*. Cambridge MA: MIT Press.

Hendry, D.F., Johansen, S., and Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics* 23, 317-335. Erratum 337-339.

Hendry, D.F. and Mizon, G.E. (2011). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, 3.

Hendry, D.F. and Pretis, F. (2013). Anthropogenic influences on atmospheric CO2. *Handbook on Energy and Climate Change* 287.

Hendry, D.F. and Santos, C. (2010). An automatic test of super exogeneity. In Bollerslev, T., Russell, J.R., and Watson, M.W. (eds.) *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, pp. 164-193. Oxford: Oxford University Press.

Herndon, T., Ash, M., and Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics* 38, 257-279.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review* 22, 85-126.

Holly, A. (1982). A remark on Hausman's specification test. *Econometrica* 50, 749-759.

Honoré, B.E. and Hu, L. (2004). On the performance of some robust instrumental variables estimators. *Journal of Business and Economic Statistics* 22, 30-39.

Honoré, B.E. and Powell, J.L. (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics* 64, 241-278.

Hoover, K.D. and Perez, S.J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal* 2, 167-191.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73-101.

Imbens, G.W. and Angrist, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica* 62, 467-475.

Imbens, G.W. and Newey, W.K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77, 1481-1512.

Jennrich, R.I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics* 40, 633-643.

Jiao, X. (2018). Uniform convergence of marked empirical processes in location and scale model. *Working Paper*.

Jiao, X. and Nielsen, B. (2017). Asymptotic analysis of iterated 1-step Huber-skip M-estimators with varying cut-offs. In Antoch, J., Jurečková, J., Maciak, M., and Pešta, M. (eds.) *Analytical Methods in Statistics: AMISTAT, Prague, November 2015*. Springer Proceedings in Mathematics and Statistics 193, 23-52.

Jiao, X. and Pretis, F. (2018). Testing the presence of outliers. *Working Paper*.

Johansen, S. and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, J.L. and Shephard, N. (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honor of David F. Hendry*, pp. 1-36. Oxford: Oxford University Press.

Johansen, S. and Nielsen, B. (2013). Asymptotic theory for iterated one-step Huber-skip estimators. *Econometrics* 1, 53-70.

Johansen, S. and Nielsen, B. (2016a). Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli* 22, 1131-1183.

Johansen, S. and Nielsen, B. (2016b). Asymptotic theory of outlier detection algorithms for linear time series regression models (with discussion). *Scandinavian Journal of Statistics* 43, 321-381.

Johansen, S. and Nielsen, B. (2018). Boundedness of M-estimators for multiple linear regression in time series. *Econometric Theory*, forthcoming.

Jurečková, J., Sen, P.K., and Picek, J. (2013). *Methodology in Robust and Nonparametric Statistics*. Boca Raton, FL: CRC Press.

Kaji, T. (2018). Switching to the new norm: from heuristics to formal tests using integrable empirical processes. *Working Paper*.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica* 46. 33-50.

Kolmogorov, A.N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale Dell'Instituto Italiano Degli Attuari* 4, 83-91.

Koul, H.L. (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models*, 2nd edition. New York: Springer.

Koul, H.L. and Ling, S. (2006). Fitting an error distribution in some heteroscedastic time series models. *Annals of Statistics* 34, 994-1012.

Koul, H.L. and Ossiander, M. (1994). Weak convergence of randomly weighted dependent residual empiricals with applications to autoregression. *Annals of Statistics* 22, 540-562.

Lehmann, E.L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics* 37, 1137-1153.

Li, G. (1985). Robust regression. In Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (eds.) *Exploring Data Tables, Trends, and Shapes*, pp. 281-343. New York: Wiley.

Lütkepohl, H. (1996). *Handbook of Matrices*. Chichester: Wiley.

Mankiw, N.G., Romer, D., and Weil, D.N. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107, 407-437.

Marczak, M. and Proietti, T. (2016). Outlier detection in structural time series models: The indicator saturation approach. *International Journal of Forecasting* 32, 180-202.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. Chicester: Wiley.

Maronna, R.A. and Yohai, V.J. (1997). Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference* 65, 233-244.

Martinez, A.B.(2015). How good are US government forecasts of the federal debt? *International Journal of Forecasting* 31, 312-324.

Martinez, A.B. (2019). A false sense of security: the impact of forecast uncertainty on hurricane damages. *Working Paper*.

Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica* 49, 1417-1426.

Nielsen, B. and Qian, M. (2018). Asymptotic properties of the gauge of step-indicator saturation. *Working Paper*.

Nielsen, H.B. (2004). Cointegration analysis in the presence of outliers. *Econometrics Journal* 7, 249-271.

Nymoen, R. and Sparrman, V. (2015). Equilibrium unemployment dynamics in a panel of OECD countries. *Oxford Bulletin of Economics and Statistics* 77, 164-190.

Prescott, P. (1975). An approximate test for outliers in linear models. *Technometrics* 17, 129-132.

Pretis, F. (2019). Does a carbon tax reduce CO2 emissions? Evidence from British Columbia. *Working Paper*.

Pretis, F., Mann, M.L., and Kaufmann, R.K. (2015). Testing competing models of the temperature hiatus: assessing the effects of conditioning variables and temporal uncertainties through sample-wide break detection. *Climatic Change* 131, 705-718.

Pretis, F., Reade, J., and Sucarrat, G. (2018). Automated General-to-Specific (GETS) regression modeling and indicator saturation methods for the detection of outliers and structural breaks. *Journal of Statistical Software* 86.

Pretis, F., Schwarz, M., Tang, K., Haustein, K., and Allen, M.R. (2018). Uncertain impacts on economic growth when stabilizing global temperatures at 1.5 C or 2 C warming. *Philosophical Transactions of the Royal Society* A, 376.

Reinhart, C.M. and Rogoff, K.S. (2010). Growth in a time of debt. *American Economic Review* 100, 573-578.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Ruppert, D. and Carroll, R.J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association* 75, 828-838.

Sarkar, T.K. (1969). Some lower bounds of reliability. *Stanford University Calif Department of Statistics*. STANFORD UNIV CALIF DEPT OF STATISTICS.

Sarkar, S.K. and Chang, C.K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 92, 1601-1608.

Seeger, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics* 10, 586-593.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73, 751-754.

Smirnov, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathèmatique de l'Université de Moscou* 2(2).

Srivastava, M.S. and von Rosen, D. (1998). Outliers in multivariate regression models. *Journal of Multivariate Analysis* 65, 195-208.

Stillwagon, J.R. (2016). Non-linear exchange rate relationships: An automated model selection approach with indicator saturation. *The North American Journal of Economics and Finance* 37, 84-109.

Temple, J.R. (1998). Robustness tests of the augmented Solow model. *Journal of Applied Econometrics* 13, 361-375.

Tietjen, G.L., Moore, R., and Beckman, R. (1973). Testing for a single outlier in simple linear regression. *Technometrics* 15, 717-721.

Toda, A.A. and Walsh, K.J. (2015). The double power law in consumption and implications for testing euler equations. *Journal of Political Economy* 123, 1177-1200.

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

Varga, R.S. (2000). *Matrix Iterative Analysis*. Berlin: Springer.

Welsh, A.H. and Ronchetti, E. (2002). A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference* 103, 287-310.

Wu, D.M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 41, 733.

Wüthrich, K. (2018). A comparison of two quantile models with endogeneity. *Journal of Business and Economic Statistics*, forthcoming.

Young, A. (2017). Channelling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *Working Paper*.

Young, A. (2018). Consistency without inference: instrumental variables in practical application. *Working Paper*.