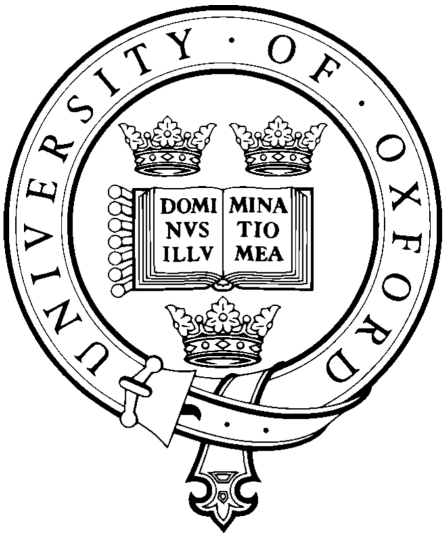
Essays on Outlier and Break Detection



Jonas Kai Kurle

Department of Economics and Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Economics

Trinity Term 2024

Supervisor This thesis was supervised by Vanessa Berenguer-Rico.

Word Count This thesis contains approximately 85,084 words. The word count was calculated as the number of words on a typical page with only text (page 2, 356 words) multiplied by the total number of pages excluding the cover, dedication and acknowledgements, table of contents, list of figures, list of tables, and references.

Statement of Authorship Chapter 1 is joint work with Xiyu Jiao. Chapters 2 and 3 are single-authored.

Acknowledgement of Funding I gratefully acknowledge funding from the Department of Economics (Two-Year Departmental Doctoral Bursary and George Webb Medley Fund), Balliol College (Simon Linnett Scholarship and various Graduate Project Grants), the Forethought Foundation (Global Priorities Fellowship), and the Evangelisches Studienwerk e.V. (Promotionsförderung). I am also thankful for generous financial support from Climate Econometrics.

In loving memory of my mother, Claudia.

To my father, Jürgen, and my sister, Julia.

Acknowledgements

First and foremost, I want to thank my supervisor, Vanessa Berenguer-Rico, for fostering my intellectual and mathematical rigour, patiently helping me refine my writing style, and guiding me to become a better researcher more generally. Vanessa was always understanding and encouraging when I faced challenges of various kinds, and offered me valuable career advice.

I am also very grateful to Jennifer Castle and David Hendry for their significant contributions to my third chapter in this thesis. This chapter builds on work from my MPhil thesis, which was supervised by David Hendry and Rick van der Ploeg. Both Jennie and David were extremely generous with their time, providing invaluable feedback and greatly enhancing the quality of my research.

My sincere thanks go to my co-author and friend, Xiyu Jiao, who was my econometrics class teacher during the MPhil and who invited me to collaborate on a research project at the beginning of my DPhil studies, which subsequently evolved into the first chapter of this thesis.

I am very thankful to the wider econometrics research group for interesting discussions and helpful comments about my work. I want to especially thank Sophocles Mavroeidis, Bent Nielsen and Martin Weidner for their feedback on the first two chapters of this thesis.

I gratefully acknowledge the financial support from the Department of Economics (Two-Year Departmental Doctoral Bursary and George Webb Medley Fund), Balliol College (Simon Linnett Scholarship and various Graduate Project Grants), the Forethought Foundation (Global Priorities Fellowship), and the Evangelisches Studienwerk e.V. (Promotionsförderung). I am also thankful for generous financial support from Climate Econometrics, which enabled my attendance at several conferences, provided me with office space, and fostered my overall well-being during my studies.

I would like to thank the Nuffield College IT Team for providing me access to their Compute Server. Similarly, Oxford University's Advanced Research Computing (ARC) allowed me to use their compute clusters for my computationally intensive simulations. Their support team was always available to assist and taught me about high-performance computing, job scheduling, and efficient resource management.

I am very grateful to my colleagues, friends, and fellow DPhil students both at the Department of Economics and Climate Econometrics for their camaraderie and moral support. I would like to especially thank Lisa Martin for our productive co-working sessions and open conversations; Ryan Rafaty for including me in a policy-oriented project with the World Bank; and Michael Wulfsohn for countless interesting conversations and shared experiences during our DPhil journey.

During the most challenging times, including periods of personal loss and health challenges, I was met with immense understanding, kindness, and support from my friends, both old and new. I want to especially thank Michael Göpper for being a close friend since the MPhil, always available to help and discuss my problems; Ebba Mark for being a compassionate listener and providing excellent advice; Fulvia Marotta for our open and honest conversations about personal experiences we both relate to; Markus Schmidt for being a lifelong friend since high school, always ready to help and listen; Angela Wenham for looking out for me and ensuring that I take care of myself; Peter Nitsche-Whitfield for being there when I needed him the most; and Michael Wulfsohn for his support and advice during challenging times.

I want to express my heartfelt gratitude to my family, who have provided constant support throughout my life. I owe everything to my parents, Claudia and Jürgen, for their endless love and encouragement. My mother's remarkable strength and my father's steady support have been crucial throughout this experience. Their support and dedication have provided me with immense comfort and stability.

I am also very grateful to my sister, Julia, for her loving support and belief in me. A special, heartfelt thanks goes to my partner, Gigi, whose love, patience, and understanding have been a reliable source of strength and motivation through the many challenges that we have faced together.

Contents

Introduction **1**

1 An Asymptotic Study of the False Outlier Detection Rate in Robust Two-Stage Least Squares Regressions **5**

1.1 Introduction . . . 6
1.2 Model and Robust Methods . . . 11
1.2.1 Model . . . 11
1.2.2 Algorithms Robust to Outliers . . . 15
1.2.3 False Outlier Detection Rate (FODR) . . . 19
1.3 The Main Results . . . 21
1.3.1 Assumptions . . . 21
1.3.2 Asymptotic Approximation to the FODR . . . 23
1.3.2.1 Tightness . . . 25
1.3.2.2 Gaussian Approximation . . . 27
1.3.2.3 Poisson Approximation . . . 32
1.4 Simulation Study . . . 34
1.4.1 Evaluation of the Gaussian Approximation . . . 35
1.4.1.1 Distributional Comparison . . . 36
1.4.1.2 Mean Comparison . . . 38
1.4.1.3 Variance and Covariance Comparison . . . 39
1.4.2 Evaluation of the Poisson Approximation . . . 41
1.5 Conclusion . . . 43

Appendices of Chapter 1 **45**

Appendix 1.A LLN and CLT for Empirical Processes . . . 45
Appendix 1.B Asymptotics of Iterated Estimators . . . 50
Appendix 1.C Main Results for the FODR and Their Proofs . . . 61
Appendix 1.D Additional Simulations . . . 79
1.D.1 Performance of Saturated 2SLS with Unequal Splits . . . 79
1.D.2 Evaluation of the Convergent Iterations . . . 81
Appendix 1.E Robustness of Algorithms . . . 82
1.E.1 Intuition for Differences in Robustness . . . 83
1.E.2 Monte Carlo Illustration . . . 84

2 Testing for Outliers in Robust Two-Stage Least Squares Regressions **89**

2.1 Introduction . . . 90
2.2 Classical 2SLS Regression Model and Notation . . . 93

2.3 Test Statistics . . . 96
2.3.1 Outlier Share and Outlier Count . . . 96
2.3.2 Stylised Illustration . . . 98
2.3.3 Overview Over Tests . . . 100
2.3.3.1 Proportion Test . . . 101
2.3.3.2 Count Test . . . 101
2.3.3.3 Multiple Proportion and Count Test . . . 102
2.3.3.4 Sum Test . . . 103
2.3.3.5 Supremum Test . . . 104
2.4 Outlier Detection Algorithms . . . 105
2.5 Asymptotic Theory . . . 110
2.5.1 Assumptions . . . 110
2.5.2 Asymptotic Distributions of the Test Statistics . . . 111
2.5.2.1 Proportion Test . . . 111
2.5.2.2 Count Test . . . 113
2.5.2.3 Multiple Proportion and Count Test . . . 114
2.5.2.4 Sum Test . . . 115
2.5.2.5 Supremum Test . . . 116
2.6 Simulation Study . . . 118
2.6.1 Size . . . 118
2.6.1.1 DGP . . . 119
2.6.1.2 Proportion Test . . . 119
2.6.1.3 Count Test . . . 121
2.6.1.4 Multiple Proportion and Count Test . . . 122
2.6.1.5 Sum and Supremum Test . . . 125
2.6.2 Power Epsilon-Contamination . . . 128
2.6.2.1 DGP . . . 128
2.6.2.2 Proportion Test . . . 131
2.6.2.3 Count Test . . . 133
2.6.2.4 Multiple Proportion and Count Test . . . 133
2.6.2.5 Sum Test . . . 137
2.6.2.6 Supremum Test . . . 137
2.6.3 Power LTS-Contamination . . . 140
2.6.3.1 DGP . . . 141
2.6.3.2 Proportion and Count Test . . . 141
2.6.3.3 Sum and Supremum Test . . . 144
2.7 Discussion . . . 144
2.8 Empirical Illustration . . . 148
2.9 Conclusion . . . 152

Appendices of Chapter 2 **155**
Appendix 2.A Simplifications Under Normality . . . 155
Appendix 2.B Proofs of Asymptotics of Test Statistics . . . 160
Appendix 2.C Comparison to Normality Tests . . . 164

3 A Monte Carlo Study of Super Saturation for Detecting Outliers and Location Shifts **168**
3.1 Introduction . . . 169
3.2 Super Saturation: Overview and Estimation . . . 175
3.3 New Performance Metrics for Super Saturation . . . 178
3.3.1 Existing Performance Metrics for Automatic Model Selection . 179
3.3.2 Shortcomings of Existing Metrics for Super Saturation . . . 180
3.3.3 Adapted, Phenomenon-Based Performance Metrics . . . 181
3.4 Performance in a Location-Scale Regression Model . . . 183
3.4.1 DGP without Outliers or Location Shifts . . . 183
3.4.1.1 Asymptotic Theory for the Gauge of IIS and SIS . . 184
3.4.1.2 Descriptive Statistics of the Gauge . . . 186
3.4.1.3 Response Surface of the Gauge . . . 191
3.4.1.4 Guidance for Controlling the Gauge . . . 202
3.4.2 DGP with Outliers . . . 204
3.4.3 DGP with Location Shifts . . . 208
3.4.4 DGP with Both Location Shift and Outliers . . . 213
3.5 Performance Prior to Model Selection in a Regression Model with Covariates . . . 215
3.5.1 DGP without Outliers or Location Shifts . . . 218
3.5.1.1 Cost of Search versus Cost of Inference . . . 218
3.5.1.2 Additional Cost of Search of Super Saturation . . . 221
3.5.2 DGP with Location Shift and Outliers in Regressand . . . 222
3.5.3 DGP with Location Shift in Relevant Regressor and Outliers in Regressand . . . 224
3.5.4 DGP with Location Shift in Omitted Relevant Regressor and Outliers in Regressand . . . 226
3.5.5 DGP with Location Shift in Irrelevant Regressor and Outliers in Regressand . . . 228
3.6 Conclusion . . . 230

Appendices of Chapter 3 **233**
Appendix 3.A Regressors in Response Surface Modelling . . . 233
Appendix 3.B Response Surfaces for All Simulations . . . 234

Conclusion **237**

References **239**

List of Figures

1.3.1 Asymptotic Variance of the FODR 31
1.4.1 Evaluation of the Distribution of the FODR 37
1.4.2 Evaluation of the Asymptotic Mean of the FODR 38
1.4.3 Evaluation of the Asymptotic Variance of the FODR 40
1.4.4 Evaluation of the Distribution of the Number of Detected Outliers 42
1.D.1 Evaluation of the FODR: Saturated 2SLS, $m = 0$ 80
1.D.2 Evaluation of the Convergent Iterations 81
1.E.1 Comparison of RMSE 85
1.E.2 Comparison of Gauge and Potency 88

2.3.1 Illustration of Testing Idea 99
2.6.1 Size of Proportion Test, $\alpha = 0.05$ 120
2.6.2 Size of Count Test, $\alpha = 0.05$ 122
2.6.3 Size of Multiple Proportion Test, $\alpha = 0.05$ 123
2.6.4 Size of Multiple Count Test, $\alpha = 0.05$ 124
2.6.5 Size of Sum Test, $\alpha = 0.05$ 126
2.6.6 Size of Supremum Test, $\alpha = 0.05$ 127
2.6.7 Power of Proportion Test, $\alpha = 0.05$ 132
2.6.8 Power of Count Test, $\alpha = 0.05$ 134
2.6.9 Power of Multiple Proportion Test, $\alpha = 0.05$ 135
2.6.10 Power of Multiple Count Test, $\alpha = 0.05$ 136
2.6.11 Power of Sum Test, $\alpha = 0.05$ 138
2.6.12 Power of Supremum Test, $\alpha = 0.05$ 139
2.6.13 Illustration of LTS-Contamination 142
2.6.14 Power of the Tests, $\alpha = 0.05$ 143
2.8.1 Density Comparison of Standardised Residuals 151
2.C.1 Power Comparison, $\alpha = 0.05$ 166

3.3.1 Illustration of Ambiguity of Indicator Relevance 181
3.4.1 Effect of Diagnostic Tests and Backtesting on the Gauge 187
3.4.2 Histogram of the Gauge, $\gamma = 0.05$ 190
3.4.3 Correlation Between the Impulse and Step Gauge 191
3.4.4 Response Surfaces for the Mean of the Gauge 197
3.4.5 Response Surfaces for the Variance of the Gauge 200
3.4.6 Stylised Illustrations of the DGPs with Outliers 204
3.4.7 Stylised Illustrations of the DGPs with Location Shifts 208
3.4.8 Stylised Illustration of a DGP with Location Shift and Outliers . 214

3.B.1 Response Surfaces for the Mean of the Gauge 235
3.B.2 Response Surfaces for the Variance of the Gauge 236

List of Tables

1.2.1 Number of Errors Committed When Testing m Null Hypotheses . 20
1.3.1 Probability of Detecting at Most x Outliers 34
1.4.1 Evaluation of the Asymptotic Covariance of the FODR 41
1.4.2 Cut-off Values for Poisson Approximation 42

2.3.1 Overview Over Tests 105
2.6.1 Contamination Scenarios and Values of γ_c^1 129
2.8.1 Comparison of Full-Sample and Robust Estimation Results 150

3.4.1 Response Surface for the Mean of the Impulse Gauge 194
3.4.2 Response Surface for the Mean of the Step Gauge 196
3.4.3 Response Surface for the Variance of the Impulse Gauge 199
3.4.4 Response Surface for the Variance of the Step Gauge 201
3.4.5 Single Outlier 205
3.4.6 Two Outliers, Equal and Opposite 207
3.4.7 Many Outliers 207
3.4.8 Single Permanent Shift 209
3.4.9 Single Permanent Shift Tolerating Misspecified Timing 210
3.4.10 Single Temporary Shift 211
3.4.11 Single Temporary Shift Across Both Halves 212
3.4.12 Two Equal Temporary Shifts 213
3.4.13 Two Opposite Temporary Shifts 214
3.4.14 Single Shift and Multiple Outliers 215
3.5.1 Overview Over DGPs 216
3.5.2 Overview of Terms Included in the GUMs 217
3.5.3 Comparison of Reduction, No Outliers and No Mean Shifts 219
3.5.4 Comparison of Reduction, Outliers and Mean Shift in y 223
3.5.5 Comparison of Reduction, Outliers in y and Mean Shift in x_1 . . 225
3.5.6 Comparison of Reduction, Outliers in y and Mean Shift in x_3 but Omitted in Estimation 227
3.5.7 Comparison of Reduction, Outliers in y and Mean Shift in x_6 . . 229

Abstract

The study of outliers and structural breaks in econometric models is crucial because of their potential to distort statistical inference, model selection, and forecasting. This thesis investigates algorithms that are used in applied research to detect outliers and structural breaks in econometric models. It is organised into two parts.

The first part considers outlier detection algorithms in two-stage least squares regression models for cross-sectional data. These algorithms classify observations with standardised residuals beyond a certain cut-off value as outliers and re-estimate the model using the remaining observations. Chapter 1 analyses the false outlier detection rate of these algorithms under the null hypothesis of no outliers in the data, providing an asymptotic theory for the share and number of falsely detected outliers in the sample. These theoretical insights provide guidance on how to set the cut-off value.

Chapter 2 builds on the results of Chapter 1 by proposing statistical tests for the presence of outliers in the sample, whose asymptotic distribution is derived under the additional assumption of normally distributed errors for the non-outlying observations. Monte Carlo simulations assess and compare the performance of the different tests, showing that the size can be well controlled and that the power of the tests increases both in the outlier frequency and outlier magnitude. The methods and tests analysed in Chapters 1 and 2 are implemented in the R package *robust2sls*, providing a useful tool for empirical researchers.

The second part shifts focus to time series regression models estimated by ordinary least squares, examining the performance of super saturation, a method that is used to detect both outliers and location shifts. Chapter 3 uses Monte Carlo simulations to assess the performance of this method, offering recommendations for setting its tuning parameters to reach high detection rates of outliers and location shifts while limiting the risk of spurious discoveries.

Overall, this thesis advances the field by providing new theoretical and practical insights into outlier and break detection, both for cross-sectional and time series econometric models.

Introduction

Undetected outliers and structural breaks can significantly distort econometric models and yield misleading results if not properly addressed. First, outliers and structural breaks can distort coefficient estimates in regression models, especially when the influence function of their estimators is unbounded (Hampel et al. 1986; Zhelonkin, Genton and Ronchetti 2012). Second, they can adversely affect statistical inference. For example, the presence of outliers can affect diagnostic testing, such as testing for normality (Berenguer-Rico and Nielsen 2023b) or heteroskedasticity (Berenguer-Rico and Wilms 2021) of the errors. Similarly, they can impact the first-stage F-test for instrument relevance in two-stage least squares regressions (Young 2022) and the Anderson-Rubin test for inference under weak instruments (Klooster and Zhelonkin 2024). Undetected structural breaks can impact unit root testing in time series regressions (Hendry and Neal 1991), potentially leading to wrong conclusions about the dynamics of the time series. Third, location shifts can cause systematic forecast failure, especially when the location shift occurs towards the end of the sample (Clements and Hendry 1998). Fourth, outliers and location shifts can induce non-constancy in under-specified models and impact model selection by causing spurious correlations with irrelevant variables (Castle and Hendry 2014, 2019).

Outliers are often defined relative to both a statistical model and a reference distribution. An observation may appear as an outlier in one model but not in another. For example, an observation may appear as an outlier if an important variable is omitted or when the relationship is modelled linearly while the true relationship is non-linear. Similarly, outliers are often defined relative to a reference distribution, as in Huber's (1964) ϵ-contamination framework, in which observations are generated from a mixture of a "good" reference distribution and a contaminating distribution. A realisation that seems extreme or unlikely under one reference

distribution may be relatively common under another. In the outlier detection algorithms considered in this thesis, the model-dependence and reference-dependence of outliers are combined because outliers are identified by comparing the magnitude of the standardised residuals from a statistical model to a reference distribution.

Similarly, location shifts can be model-dependent. For example, the exclusion of a relevant variable that experiences a location shift induces a shift in the intercept of such an under-specified model.

The causes of outliers and location shifts in real-world data are manifold. They may represent genuine anomalies but can also be due to human recording errors or inaccuracies in measurement, such that the unusual values are in fact incorrect. Location shifts are longer-lasting shifts in the mean of a variable. Therefore, they more likely stem from lasting changes in measurement techniques, shifts in policy and regulations, or the adoption of new technologies that shift the mean of a random variable.

The reason for an outlier or location shift may also determine the appropriate response to its detection. Barnett (1978) categorises four different ways of handling outliers. The first possibility is to limit the influence of outliers in general by relying on robust estimators (see for example Hampel et al. 1986, for a popular approach in the robust statistics literature). Second, the outlier may be incorporated to yield a homogeneous sample by replacing the reference distribution with one that can describe all data points. Third, the identification of outliers may lead to further investigation into their causes and subsequent revisions of the statistical model. Finally, the outliers can be discarded, while the remaining observations are treated as having been generated by the original model or reference distribution.

Given the challenges posed by undetected outliers and location shifts, this thesis aims to enhance our understanding of methods that are commonly used to detect and address these phenomena. The thesis is split into two parts. The first part analyses outlier detection algorithms and proposes tests for the presence of outliers in two-stage least squares regression models for cross-sectional data. The second part shifts the focus to outlier and structural break detection in time series regression

models estimated by ordinary least squares.

Chapter 1 introduces the outlier detection heuristics, which classify observations with standardised residuals beyond a certain cut-off value as outliers and re-estimate the model using the remaining observations. After discussing the choice of initial estimator and the possibility of iterating the procedure, we formalise the procedures in the form of algorithms. The algorithms have a positive probability of falsely classifying an observation as an outlier, conceptualised as the false outlier detection rate (FODR), which is analysed asymptotically under the null hypothesis of no outliers using empirical process results derived in Jiao (2022). We first show tightness of the FODR, uniformly both in the cut-off value and the iteration step of the algorithm. Next, we establish the weak convergence of the share of falsely detected outliers to a Gaussian process and provide a Poisson approximation for the number of falsely detected outliers. A simulation study assesses the finite sample performance of the asymptotic results.

Chapter 2 uses the asymptotic theory derived in Chapter 1 to propose statistical tests for the presence of outliers in the sample. Based on the algorithms discussed in the previous chapter, the idea is to test whether the share or number of detected outliers is significantly higher than the expected false detections under the null hypothesis of no outliers in the sample. Following Jiao and Pretis (2022), we distinguish two types of test statistics. Simple tests consider the frequency of detected outliers relative to a single cut-off value, while scaling tests run the algorithms for a range of cut-off values and thereby also incorporate the magnitude of outliers. The asymptotic distribution of the test statistics is derived assuming normality of the errors. Extensive Monte Carlo simulations are used to assess the size and power properties of the different tests. For the contamination scenarios under the alternative, we implement outliers using Huber's (1964) ϵ-contamination framework and leverage point outliers using the LTS contamination framework introduced in Berenguer-Rico and Nielsen (2023a). Based on the simulation results as well as theoretical considerations, we discuss the advantages and disadvantages of the different test statistics. Finally, we illustrate the usage of the tests by re-examining the effect of schooling

on earnings in the seminal paper by Angrist and Krueger (1991).

Chapter 3 uses Monte Carlo simulations to assess the performance of super saturation. Super saturation is a method that transforms the detection of outliers and location shifts into a model selection problem by searching over sets of impulse indicators that can capture an outlier (Hendry, Johansen and Santos 2008) and sets of step indicators that can capture a location shift (Castle et al. 2015). This thesis focuses on the evaluation of super saturation as implemented in the general-to-specific selection algorithm Autometrics (Doornik 2009) and provides recommendations for applied researchers on how to set its tuning parameters. First, we adapt the classical evaluation measures of gauge and potency that measure the false retention of irrelevant and the correct retention of relevant indicators, respectively, to allow for the possibility of different combinations of indicators modelling a single outlier or location shift. Second, we simulate the gauge under the null for many different choices of tuning parameters and model their impact on the gauge by estimating response surfaces. These response surfaces provide valuable insights for researchers on how to set the tuning parameters. Next, we study the correct detection rate (potency) in data generating processes that contain outliers and location shifts. Finally, we assess the impact of super saturation on subsequent selection over other covariates included in the model.

CHAPTER 1

An Asymptotic Study of the False Outlier Detection Rate in Robust Two-Stage Least Squares Regressions*

with **Xiyu Jiao**

Abstract Many empirical research papers in economics provide robustness checks to show that their results are not driven by a subset of the sample that they analyse. There has been a recent resurgence of interest in the outlier robustness of two-stage least squares (2SLS) regression models, which can be very prone to outliers. A common practice to perform outlier robustness checks is to first run 2SLS on the full sample and classify observations with standardised residuals beyond a chosen cut-off value as outliers. Then, the 2SLS estimator is applied to the remaining subsample of non-outlying observations. This procedure may be iterated until the classification of outliers does not change any more. However, the above trimmed 2SLS has a positive probability of finding outliers even when the data generating process contains none. To analyse the performance of this heuristic approach, this chapter formalises the heuristic approach, proposes various improvements to it, and studies the concept of false outlier detection rate (FODR) asymptotically using empirical processes techniques. The established asymptotic theory provides guidance for setting the cut-off value. Moreover, the theory serves as the basis for the construction of tests for the overall presence of outliers in Chapter 2.

∗. We are grateful to participants at the Oxford Econometrics Lunch Seminar and at the workshops and conferences CFE-CMStatistics 2020, RES Annual Meeting 2021, EcoSta 2021, ESEM 2021, Oxford Encounters in Econometric Theory 2022, and IAAE 2022 for helpful comments and discussions. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (Richards 2015).

1.1 Introduction

A frequent concern in applied economics is that key empirical findings may be driven by a tiny set of outliers. Researchers should therefore confirm that their results are robust to outliers in order to increase confidence in their validity. Recent studies have highlighted the sensitivity of empirical results to outliers, especially in the instrumental variables setting. Young (2022) analyses 1309 IV regressions in 30 papers and finds that the results are very sensitive to removing just one or two observations from the sample. The author also finds that outliers seem to strongly affect the first stage weak instrument F-tests, thereby undermining their credibility. Similarly, Klooster and Zhelonkin (2024) show that the Anderson-Rubin test for inference under weak instruments (Anderson and Rubin 1949) is not robust to outliers and provide a robust alternative. Broderick, Giordano and Meager (2023) develop a metric related to influence functions called "Approximate Maximum Inference Perturbation" to find small subsets of the data that strongly affect regression estimates, including IV regressions. Like Young (2022), they also find that some empirical results are strongly driven by a small share of observations. This literature demonstrates how prone IV regressions are to outliers and either suggests robust methods or ways to determine whether results are driven by a small number or share of observations.

In practice, the most common approach to conduct outlier analysis is a heuristic approach based on the size of regression residuals (see for example Acemoglu, Johnson and Robinson 2001, 2012; Albouy 2012; Acemoglu et al. 2019). To confirm the validity of their estimates of the impact of institutions on economic growth, they first run ordinary two-stage least squares (2SLS) on the full sample and classify observations with standardised residuals beyond a chosen cut-off value as outliers. Subsequently, 2SLS is re-calculated based on the non-outlying observations. We refer to the heuristic approach as *Trimmed 2SLS* following the trimmed least squares terminology introduced by Ruppert and Carroll (1980). In this chapter, we analyse the trimmed 2SLS and several extensions. To achieve higher outlier robustness, we

suggest three improvements to the trimmed 2SLS: (i) iterating the procedure, (ii) choosing a different initial estimator to classify observations as outliers, and (iii) bias-correcting the variance estimator in the iterated procedure.

Most empirical studies stop after the initial iteration, which classifies observations as outliers using the full-sample estimate. In contrast, we suggest iterating the procedure until a fixed point is reached. We call the procedure that starts with the full-sample estimates and is then iterated *Robustified 2SLS*, following Johansen and Nielsen (2009) who call the analogous procedure in the least squares setting robustified least squares. Johansen and Nielsen (2013) and Jiao (2022) show that the fixed point of the iterated procedure has the same first order asymptotic expansion as the Huber-skip M-estimator in the classical and the IVs regression setting, respectively. Thus, repeated application of the procedure may provide better estimates and classifications, yielding higher resilience to outliers.

As a second potential improvement, we suggest using a more robust initial estimator. Since the initial full-sample estimator of Robustified 2SLS is not robust to outliers itself, it may not be a good starting point for the classification, especially in presence of high leverage points. We therefore also consider an alternative procedure, in which the sample is partitioned into two subsamples and the estimates of one subsample are used to classify outliers in the second one. This procedure is especially suitable for situations in which we suspect outliers to be concentrated in a certain part of the sample. We call this procedure *Saturated 2SLS* as it is a simplified version of impulse indicator saturation (IIS) (Hendry 1999; Hendry, Johansen and Santos 2008), which runs multiple block searches for outliers across many subsets of the sample. For analytical tractability, we focus on the stylised version with two subsamples as an approximation to IIS. IIS has been widely used in the empirical literature to detect outliers, for example for modelling wages (Castle and Hendry 2009) and food expenditure (Hendry and Mizon 2011), evaluating debt forecasts (Ericsson 2017), and estimating economic impacts of climate change (Pretis et al. 2018).

Third, we suggest using the bias-correction factor of the variance that is derived in Jiao (2022). The described algorithms all have a positive probability of declaring

observations with large residuals as outliers even when the true data generating process (DGP) contains none, which causes the usual 2SLS estimator of the variance to be downward biased. A lack of correction not only distorts inference on the parameters if there are no true outliers in the DGP, but it also causes the iterated detection procedure — as suggested in the first improvement — to falsely detect more outliers in subsequent iterations.

All of the above procedures can be conceptualised as algorithms with the cut-off value and the iteration step as tuning parameters that the researcher must specify. In practice, this decision seems to be relatively ad-hoc. For example, Acemoglu et al. (2019) choose a cut-off value of 1.96, presumably with the 97.5%-quantile of the normal distribution as a reference in mind. Previous research took the choice of the cut-off value as given. Kaji (2018) and Jiao (2022) propose Hausman-type tests to compare the full-sample to the trimmed 2SLS estimates to assess the influence of outliers given a specific cut-off value. We aim to provide guidance on how to choose the cut-off value depending on the researcher's tolerance for false outlier detections. To that end, we analyse the algorithms by introducing the concept of the *false outlier detection rate* (FODR) and establishing its asymptotic theory. The FODR is the share of observations in the sample that have been falsely declared as outliers, which is formally defined under the null of no outliers as

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

where $v_{i,c}^{(m)}$ is an indicator variable that takes on the value 0 if observation i is classified as an outlier and 1 otherwise. The indices c and m represent the cut-off value and the iteration step of the algorithm, respectively. The FODR captures the rate of type I error and has previously been named *gauge* in the context of variable selection (Castle, Doornik and Hendry 2011) and outlier detection. For example, Johansen and Nielsen (2016b) study the gauge of various outlier robust estimators based on ordinary least squares in a time series setting. We use the term false *outlier detection* rate to emphasise that we analyse the performance of *outlier detection* instead of variable selection procedures.

We conduct our analysis of the FODR under the null hypothesis of no outliers. Under the alternative, there are many contamination scenarios one could consider but given the outlier detection procedure that we describe and analyse, we suggest Huber's (1964) ϵ-contamination framework in order to fix ideas. In this setup, with probability $1-\epsilon$ the structural error is drawn from some well-behaved distribution f_u (as defined in our assumptions), while with probability ϵ it is drawn from some contaminating distribution f_u^{cont}, which produces values that are extreme and unlikely compared to the well-behaved reference distribution.

The Robustified and Saturated 2SLS along with their FODRs involve weighted and marked empirical processes in IVs regressions. These processes have been studied recently by Jiao (2022) using a martingale decomposition argument with the main results summarised in Lemma 1.A.1, which is about the tightness of the empirical processes and their first order expansions. Equipped with the empirical processes theory, we can establish the asymptotic theory of the FODR by analysing it as a process in the cut-off value c. The cut-off value c is the quantile associated with a certain probability γ_c of falsely classifying an observation as an outlier. The first result shows the tightness of the FODR uniformly in the cut-off value c and iteration step m and for all our analysed algorithms. It implies that the FODR converges in probability to γ_c. This consistency result also holds uniformly in the cut-off value c and iteration step m. It is worth highlighting that for the iterated procedure ($m \geq 1$), if the variance were not bias-corrected, the FODR would instead converge to some $\gamma_c^{bias} > \gamma_c$.

Second, we study the weak convergence of the FODR for each algorithm separately. We start by analysing the Robustified 2SLS procedure. We show that the FODR as a sequence of processes in the cut-off value c converges to a Gaussian process as the sample size goes to infinity. The second moments of the Gaussian process depend both on the cut-off value c and the iteration step m of the algorithm. The variance of the Gaussian process increases in each iteration step but quickly levels off. Next, we turn to the Saturated 2SLS algorithm. In its general form, the sample can be split into any two subsamples — including unequal splits. For this case,

we analyse the weak convergence of the FODR in its non-iterated version. In the special case where the sample is split in half, we provide results also for the iterated version. We show that in this case, the procedure yields a FODR whose weak limit is identical to that of Robustified 2SLS. Hence, it follows the same Gaussian process in the cut-off value c and for every iteration m.

Our asymptotic theory provides guidance for applied researchers on how to set the cut-off value such that it matches their tolerance for false detections. Currently, researchers choose the cut-off value corresponding to some quantile of a reference distribution, assuming that it results in their intuitively expected false detection rate. For example, using the 97.5%-quantile of a symmetric reference distribution, one would expect to falsely classify 5% of the sample as outliers. We confirm this procedure for $m = 0$, show that a naive iterated procedure without bias correction would yield over-detection, and provide a corrected procedure that enables the researcher to correctly target their false detection rate also for $m \geq 1$. The Gaussian approximation allows the researcher to control the share of falsely detected outliers and quantifies the uncertainty around it. Similarly, we provide guidance for setting the cut-off value when the number instead of the share of outliers is targeted. This is done by providing a Poisson exceedance theory for the expected number of falsely detected outliers. Ultimately, our guidance addresses the robustness-efficiency trade-off that applied researchers face: choosing a smaller cut-off value classifies more observations as outliers thereby increasing robustness but at the same time losing more observations for the estimation, including observations that the algorithms incorrectly classify as outliers.

In the wider literature on robust methods underlying the Huber-skip estimator, Hendry and Johansen (2015) apply IIS for outlier detection while retaining theory relevant variables and selecting over other candidate variables. Johansen and Nielsen (2016a, 2019) and Jiao and Nielsen (2017) analyse L-type and M-type estimators. Berenguer-Rico and Nielsen (2023b) and Berenguer-Rico and Wilms (2021) analyse diagnostic tests on residuals for normality and heteroskedasticity after outlier removal. Jiao and Pretis (2022) test whether the number of outliers is different

from the expected value when no outliers are present. Jiao, Pretis and Schwarz (2024) provide tests whether the estimated coefficients are distorted due to outliers. However, this existing work is restricted to the ordinary least squares setting. This chapter considers instrumental variables regressions and derives the asymptotic theory for the FODR of Huber-skip type estimators. The established theory will be used in Chapter 2 to devise statistical tests for the presence of outliers in the sample, allowing applied researchers to directly test whether a robust but potentially costly procedure is necessary.

This chapter is organised as follows. Section 1.2 introduces the two-stage least squares regression setup and our notation. We then define the algorithms robust to outliers before introducing the concept of the false outlier detection rate and how it relates to existing concepts, such as the gauge, false discovery rate, and family-wise error rate. Section 1.3 lists the assumptions required for the derivation of the asymptotic theory of the FODR. We then present the Gaussian and Poisson approximations for the different algorithms in several theorems. Section 1.4 evaluates Monte Carlo simulations to verify the asymptotic distribution results and to assess their finite sample performance. All the proofs as well as additional simulation results can be found in the Appendices.

1.2 Model and Robust Methods

We first describe the instrumental variables (IVs) regression model along with some notation. Then, we review a class of robust methods related to the iterated one-step Huber-skip M-estimators but in the instrumental variables setup. Finally, this section introduces the concept of FODR to control the false outlier detection rate of the robust IV methods, for which this chapter derives an asymptotic theory.

1.2.1 Model

The exposition follows Jiao (2022). Suppose that we have independently and identically distributed cross-sectional data $\{(y_i, x_i, z_i)\}_{i=1}^n$, where y_i is univariate and x_i, z_i

are multivariate with dimensions d_x and d_z, respectively.

Assume the data $\{(y_i, x_i)\}_{i=1}^n$ satisfy the structural equation

$$y_i = x_i'\beta + u_i, \quad i = 1, 2, \ldots, n. \tag{1.2.1}$$

The structural errors $\{u_i\}_{i=1}^n$ are univariate iid random variables with scale σ so that u_i/σ has 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 potentially endogenous, i.e. $\mathsf{E}x_i u_i \neq 0_{d_x}$, where 0_{d_x} denotes a column vector of zeros of length d_x. To be more specific $x_i = (x_{1i}', x_{2i}')'$, where x_{1i} is the set of exogenous variables with dimension d_{x_1} such that $\mathsf{E}x_{1i}u_i = 0_{d_{x_1}}$, while x_{2i} corresponds to the set of endogenous variables with dimension d_{x_2} such that $\mathsf{E}x_{2i}u_i \neq 0_{d_{x_2}}$. Instruments z_i are thus required for estimating the parameter vector $\beta \in \mathbb{R}^{d_x}$ consistently. We partition the instruments $z_i = (z_{1i}', z_{2i}')' = (x_{1i}', z_{2i}')'$, where z_{1i} coincides with the set of exogenous regressors x_{1i} and z_{2i} is the set of excluded instruments with dimension d_{z_2}. The total number of regressors and instruments is each given by $d_x = d_{x_1} + d_{x_2}$ and $d_z = d_{x_1} + d_{z_2}$, respectively. The first stage regression for $\{(x_i, z_i)\}_{i=1}^n$ is

$$x_i' = z_i'\Pi + r_i', \quad i = 1, 2, \ldots, n, \tag{1.2.2}$$

where the coefficient matrix Π of the first stage can be expressed as the following partitioned matrix

$$\Pi = \begin{pmatrix} I_{d_{x_1}} & \Pi_1 \\ 0_{d_{z_2} \times d_{x_1}} & \Pi_2 \end{pmatrix}, \tag{1.2.3}$$

with Π_1 having dimension $d_{x_1} \times d_{x_2}$ and Π_2 dimension $d_{z_2} \times d_{x_2}$. The first stage errors $\{r_i\}_{i=1}^n$ are d_x-variate iid random vectors that can be partitioned into two subvectors: (i) the d_{x_1}-dimensional errors r_{1i} corresponding to the exogenous regressors x_{1i} and equalling exactly $0_{d_{x_1}}$; and (ii) the d_{x_2}-dimensional errors r_{2i} corresponding to the endogenous regressors x_{2i}. Hence, $r_i = (r_{1i}', r_{2i}')' = (0_{d_{x_1}}', r_{2i}')'$. Furthermore, assume r_i has mean 0_{d_x} and dispersion matrix $\Sigma \in \mathbb{R}^{d_x \times d_x}$. Notice that Σ is symmetric but positive semi-definite instead of strictly positive definite. We can explicitly show the structure of Σ and define Σ^k, for example for $k \in \{-1, -1/2, 1/2\}$, in terms of the symmetric and positive definite dispersion matrix $\Sigma_2 \in \mathbb{R}^{d_{x_2} \times d_{x_2}}$ of r_{2i}.

Specifically, we note that

$$\Sigma = \begin{pmatrix} 0_{d_{x_1} \times d_{x_1}} & 0_{d_{x_1} \times d_{x_2}} \\ 0_{d_{x_2} \times d_{x_1}} & \Sigma_2 \end{pmatrix}, \qquad \Sigma^k = \begin{pmatrix} 0_{d_{x_1} \times d_{x_1}} & 0_{d_{x_1} \times d_{x_2}} \\ 0_{d_{x_2} \times d_{x_1}} & \Sigma_2^k \end{pmatrix}. \tag{1.2.4}$$

Then, the scaled errors $\Sigma^{-1/2} r_i = \{0'_{d_{x_1}}, (\Sigma_2^{-1/2} r_{2i})'\}'$ follow 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 dispersion matrix

$$I^*_{d_x} = \begin{pmatrix} 0_{d_{x_1} \times d_{x_1}} & 0_{d_{x_1} \times d_{x_2}} \\ 0_{d_{x_2} \times d_{x_1}} & I_{d_{x_2}} \end{pmatrix}. \tag{1.2.5}$$

Although $I^*_{d_x}$ is not the identity matrix, we still have $\Sigma^{1/2}\Sigma^{-1/2} r_i = I^*_{d_x} r_i = r_i$, or explicitly

$$\begin{pmatrix} 0_{d_{x_1}} \\ \Sigma_2^{1/2}\Sigma_2^{-1/2} r_{2i} \end{pmatrix} = \begin{pmatrix} 0_{d_{x_1}} \\ I_{d_{x_2}} r_{2i} \end{pmatrix} = \begin{pmatrix} 0_{d_{x_1}} \\ r_{2i} \end{pmatrix},$$

which will be frequently used in our analysis for normalizing the first stage errors r_i. The parameter Π lies in $\mathbb{R}^{d_z \times d_x}$ and we assume that the order condition ($d_x \leq d_z$) is satisfied, meaning that the number of regressors is weakly smaller than the number of instruments. Furthermore, to identify the structural parameters β, the instruments z_i need to fulfil two conditions: they must be valid, i.e. $\mathsf{E} z_i u_i = 0_{d_z}$, and informative, i.e. the following rank conditions must hold: $\operatorname{rank} \Pi = d_x$ and $\operatorname{rank} \mathsf{E} z_i z_i' = d_z$. The instruments z_i are assumed to be orthogonal to the errors r_i in the first stage equation (1.2.2) such that $\mathsf{E} z_i r_i' = 0_{d_z \times d_x}$.

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 that the joint distribution $\mathsf{f}_{u,r}, \mathsf{F}_{u,r}$ also depends on the covariance between the scaled structural and first stage errors, $\Omega = \mathsf{Cov}(u_i/\sigma, \Sigma^{-1/2} r_i) = \mathsf{E}(\Sigma^{-1/2} r_i u_i/\sigma) \in \mathbb{R}^{d_x}$. For simplicity, Ω is suppressed in the notation of the joint density. We can write Ω explicitly as

$$\Omega = \mathsf{Cov}\left\{ u_i/\sigma, \begin{pmatrix} 0_{d_{x_1}} \\ \Sigma_2^{-1/2} r_{2i} \end{pmatrix} \right\} = \mathsf{E}\left\{ \begin{pmatrix} 0_{d_{x_1}} \\ \Sigma_2^{-1/2} r_{2i} \end{pmatrix} u_i/\sigma \right\} = \begin{pmatrix} 0_{d_{x_1}} \\ \Omega_2 \end{pmatrix},$$

where $\Omega_2 \in \mathbb{R}^{d_{x_2}}$ is the covariance between u_i/σ and $\Sigma_2^{-1/2} r_{2i}$. Furthermore, we decompose the joint density into the conditional and marginal ones such that we have $\mathsf{f}_{u,r}(y, x) = \mathsf{f}_{u|r}(y|x)\mathsf{f}_r(x) = \mathsf{f}_{r|u}(x|y)\mathsf{f}_u(y)$. The conditional expected value is

denoted by

$$\xi_y = \mathsf{E}(\Sigma^{-1/2} r_i | u_i/\sigma = y) = \int_{\mathbb{R}^{d_x}} x \mathsf{f}_{r|u}(x|y)(dx). \tag{1.2.6}$$

Note that $\xi_y \in \mathbb{R}^{d_x}$ is related to the covariance Ω between u_i/σ and $\Sigma^{-1/2} r_i$ and its first d_{x_1} elements are zero. Introduce the notation

$$\zeta_y^+ = \xi_y + \xi_{-y}, \qquad \zeta_y^- = \xi_y - \xi_{-y}. \tag{1.2.7}$$

Let the absolute error $|u_i|/\sigma$ in (1.2.1) have 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 $\mathsf{g}_u(y) = 2\mathsf{f}_u(y)$ for $y > 0$. Define $\psi_c = \mathsf{G}_u(c)$, so the probability of exceeding the cut-off c is $\gamma_c = 1 - \psi_c$. Supposing the k-th moment of the density f_u exists, we introduce the moments

$$\tau_k = \int_{-\infty}^{\infty} y^k \mathsf{f}_u(y) dy, \qquad \tau_k^c = \int_{-c}^{c} y^k \mathsf{f}_u(y) dy. \tag{1.2.8}$$

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

$$\varsigma_c^2 = \frac{\tau_2^c}{\psi_c} = \frac{\int_{-c}^{c} y^2 \mathsf{f}_u(y) dy}{\mathsf{P}(|u_i| \leq \sigma c)}. \tag{1.2.9}$$

This will be used as a bias correction factor for the variance estimator of σ^2 computed from the selected non-outlying sample. Lastly, we introduce the truncated covariance

$$\omega_c = \mathsf{E}(\Sigma^{-1/2} r_i \frac{u_i}{\sigma} 1_{(|u_i| \leq \sigma c)}) = \iint_{y \in \mathbb{R}, x \in \mathbb{R}^{d_x}} x y 1_{(|y| \leq c)} \mathsf{f}_{u,r}(y, x) dy (dx), \tag{1.2.10}$$

which can be written in terms of the truncated conditional expected value ξ_y and yields

$$\omega_c = \int_{-c}^{c} \{ \int_{x \in \mathbb{R}^{d_x}} x \mathsf{f}_{r|u}(x|y)(dx) \} y \mathsf{f}_u(y) dy = \int_{-c}^{c} \xi_y y \mathsf{f}_u(y) dy. \tag{1.2.11}$$

Remark 1.2.1. *Consider the normality case where the scaled errors follow*

$$\begin{pmatrix} u_i/\sigma \\ \Sigma^{-1/2} r_i \end{pmatrix} \overset{\mathsf{D}}{=} \mathsf{N} \left\{ \begin{pmatrix} 0 \\ 0_{d_x} \end{pmatrix}, \begin{pmatrix} 1 & \Omega' \\ \Omega & I_{d_x}^* \end{pmatrix} \right\}, \tag{1.2.12}$$

$$\begin{pmatrix} u_i/\sigma \\ 0_{d_{x_1}} \\ \Sigma_2^{-1/2} r_{2i} \end{pmatrix} \overset{\mathsf{D}}{=} \mathsf{N} \left\{ \begin{pmatrix} 0 \\ 0_{d_{x_1}} \\ 0_{d_{x_2}} \end{pmatrix}, \begin{pmatrix} 1 & 0'_{d_{x_1}} & \Omega_2' \\ 0_{d_{x_1}} & 0_{d_{x_1} \times d_{x_1}} & 0_{d_{x_1} \times d_{x_2}} \\ \Omega_2 & 0_{d_{x_2} \times d_{x_1}} & I_{d_{x_2}} \end{pmatrix} \right\}.$$

As the conditional and marginal distribution of the multivariate normal are still normally distributed, the conditional random vector $\Sigma^{-1/2} r_i | u_i/\sigma \sim \mathsf{f}_{r|u}$ *follows the normal distribution* $\mathsf{N}(u_i\Omega/\sigma, I^*_{d_x} - \Omega\Omega')$ *and the random variable* $u_i/\sigma \sim \mathsf{f}_u$ *follows a standard normal distribution* $\mathsf{N}(0,1)$*. Thus, the conditional expectation in* (1.2.6) *has the form* $\xi_y = y\Omega$ *and* $\xi_{-y} = -y\Omega = -\xi_y$ *such that* $\zeta_y^+ = 0_{d_x}$ *and* $\zeta_y^- = 2\xi_y = 2y\Omega$*. Furthermore, the truncated covariance in* (1.2.10) *and* (1.2.11) *is* $\omega_c = \tau_2^c\Omega$*. Finally, we can show* $\tau_2^c = \psi_c - 2c\mathsf{f}_u(c)$*,* $\tau_4 = 3$*, and* $\tau_4^c = 3\psi_c - 2c(c^2+3)\mathsf{f}_u(c)$*.*

For matrices M, choose the spectral norm $|M| = \max\{\text{eigen}(M'M)\}^{1/2}$ so that for vectors x, the notation $|x|$ is the Euclidean norm. The spectral norm is compatible with respect to the Euclidean norm such that $|Mx| \leq |M||x|$. Notice (dM) represents the exterior product of all elements in the matrix M to describe the multivariate integral, see chapter 2 in Muirhead (1982). For instance, consider a vector $x \in \mathbb{R}^k$, then (dx) is the exterior product $dx_1\, dx_2 \cdots dx_k$ of some measures on $\mathbb{R}^k$.

1.2.2 Algorithms Robust to Outliers

The Trimmed 2SLS and its variations are extensively used in empirical studies in economics as a heuristic approach to conduct outlier robustness analysis. Such methods are all based on the same principle: First, choose a cut-off value c with respect to a (implicitly assumed) reference distribution. Residuals beyond this cut-off value are treated as unlikely draws. In practice, some initial estimates of β and σ^2 are used to estimate the error by calculating the standardised residuals of each observation. Observations with standardised residuals beyond the chosen cut-off value are then classified as outliers. Subsequently, the 2SLS regression is re-estimated on the subsample of non-outlying observations.

For example, Acemoglu et al. (2019) perform the trimmed 2SLS to demonstrate outlier robustness of their empirical findings that democracy causes economic growth. They initially compute the full-sample 2SLS estimates to calculate structural residuals, choose the cut-off value 1.96, and remove observations with standardised residuals beyond that value as outliers. Presumably, their choice of cut-off value $c = 1.96$ has implicitly been made with the the normal distribution as a

reference in mind.

In principle, the heuristic version of trimmed 2SLS that is used in the literature can be improved in three aspects according to Jiao (2022). First, the trimmed 2SLS has a positive probability to remove observations with large errors from the sample even when the model contains no outliers. The variance of the structural errors is thus underestimated in this scenario. To consistently estimate σ^2, we then need to introduce the bias correction factor ς_c^2 given in equation (1.2.9) and multiply the ordinary variance estimator by its inverse. Second, instead of starting with the full-sample 2SLS estimator, we can apply a different and possibly more robust initial estimator to classify and trim outlying observations. One specific example is to start with the so-called Saturated 2SLS, which will be introduced later in this subsection. Third, the trimmed 2SLS estimator is in fact a one-step method updated from the full-sample 2SLS estimator. It can be further iterated until converging to a fixed point, which Jiao (2022) shows to have the same first order asymptotics as the Huber-skip regression[1]. Thus, we suggest iterating this one-step procedure which results in an estimator that is more resistant to outliers. The iterated procedure actually mimics the iterated one-step Huber-skip M-estimator in the IVs regression setting. We present Algorithm 1.I next, which incorporates these three improvements.

1. The Huber (1964) skip regression has the criterion function

$$\rho(t) = \begin{cases} \frac{t^2}{2}, & \text{if } |t| \leq c, \\ \frac{c^2}{2}, & \text{otherwise,} \end{cases}$$

as opposed to the Huber regression with the criterion

$$\rho(t) = \begin{cases} \frac{t^2}{2}, & \text{if } |t| \leq c, \\ c|t| - c^2/2, & \text{otherwise,} \end{cases}$$

see Hampel et al. (1986) (p. 104).

Algorithm 1.*I* (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. Compute the indicator variables for selecting non-outlying observations

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \leq \widehat{\sigma}_c^{(m)} c)}. \tag{1.2.13}$$

3. Compute the updated least squares estimator for Π

$$\widehat{\Pi}_c^{(m+1)} = (\sum_{i=1}^n z_i z_i' v_{i,c}^{(m)})^{-1} (\sum_{i=1}^n z_i x_i' v_{i,c}^{(m)}). \tag{1.2.14}$$

4. Compute the updated 2SLS estimators for β *and* σ^2

$$\widehat{\beta}_c^{(m+1)} = (\widehat{\Pi}_c^{(m+1)\prime} \sum_{i=1}^n z_i z_i' v_{i,c}^{(m)} \widehat{\Pi}_c^{(m+1)})^{-1} (\widehat{\Pi}_c^{(m+1)\prime} \sum_{i=1}^n z_i y_i v_{i,c}^{(m)}), \tag{1.2.15}$$

$$(\widehat{\sigma}_c^{(m+1)})^2 = \varsigma_c^{-2} (\sum_{i=1}^n v_{i,c}^{(m)})^{-1} \{\sum_{i=1}^n (y_i - x_i'\widehat{\beta}_c^{(m+1)})^2 v_{i,c}^{(m)}\}. \tag{1.2.16}$$

5. Let $m = m + 1$ *and repeat steps 2 - 5 until the desired number of iterations is reached.*

Algorithm 1.I does not explicitly specify how the initial estimators in step 1 are obtained. Next, we specify two particular initial estimators. We start by describing the simplest and most commonly used procedure, based on the full-sample 2SLS as used in Acemoglu et al. (2019) for outlier analysis. First, calculate 2SLS on the full sample and use the estimates to classify observations with residuals beyond a chosen cut-off as outliers. Subsequently, re-calculate 2SLS on the "clean" subsample of non-outlying observations. In analogy to the least squares literature, we call this algorithm Robustified 2SLS. It is defined formally in Algorithm 1.R.

Algorithm 1.*R* (Robustified 2SLS). *Choose a cut-off* $c > 0$.

1.1. Compute the least squares estimator for Π *based on the whole sample*

$$\widehat{\Pi}_c^{(0)} = (\sum_{i=1}^n z_i z_i')^{-1} (\sum_{i=1}^n z_i x_i'). \tag{1.2.17}$$

1.2. Compute the 2SLS estimators for β *and* σ^2 *based on the whole sample*

$$\widehat{\beta}_c^{(0)} = (\widehat{\Pi}_c^{(0)\prime} \sum_{i=1}^n z_i z_i' \widehat{\Pi}_c^{(0)})^{-1} (\widehat{\Pi}_c^{(0)\prime} \sum_{i=1}^n z_i y_i), \qquad (\widehat{\sigma}_c^{(0)})^2 = \frac{1}{n} \sum_{i=1}^n (y_i - x_i'\widehat{\beta}_c^{(0)})^2. \tag{1.2.18}$$

1.3. Select $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ *given in equation* (1.2.18) *as initial estimators and let* $m = 0$.

2. Follow the steps 2 - 5 in Algorithm 1.I.

The initial estimator of Algorithm 1.R is not robust itself, especially not with respect to high leverage points. One possible solution is to instead use a possibly more robust initial estimator. Here, we will consider Saturated 2SLS described in Algorithm 1.S. The procedure is as follows. We divide the full sample into two subsamples and calculate the 2SLS estimates for each of them. We then use the estimates from each subsample to calculate standardised residuals and subsequently detect outliers in the other subsample. This procedure yields two subsamples of non-outlying observations, which are then combined to a joint subsample on which the updated 2SLS estimates are computed.

Algorithm 1.*S* (Saturated 2SLS). *Choose a cut-off* $c > 0$.

1.1. Split the full sample into two sets $\mathcal{I}_j$ *of* n_j *observations,* $j = 1, 2$, *where* $n_1 + n_2 = n$.

1.2. Compute the least squares estimators for Π *for each subsample* $\mathcal{I}_j$, $j = 1, 2$

$$\widehat{\Pi}_j = (\sum_{i \in \mathcal{I}_j} z_i z_i')^{-1} (\sum_{i \in \mathcal{I}_j} z_i x_i'). \tag{1.2.19}$$

1.3. Compute the 2SLS estimators for β *and* σ^2 *for each subsample* $\mathcal{I}_j$, $j = 1, 2$

$$\widehat{\beta}_j = (\widehat{\Pi}_j' \sum_{i \in \mathcal{I}_j} z_i z_i' \widehat{\Pi}_j)^{-1} (\widehat{\Pi}_j' \sum_{i \in \mathcal{I}_j} z_i y_i), \qquad \widehat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} (y_i - x_i' \widehat{\beta}_j)^2. \tag{1.2.20}$$

1.4. Compute the indicator variables for selecting non-outlying observations

$$v_{i,c}^{(0)} = 1_{(i \in \mathcal{I}_1)} 1_{(|y_i - x_i' \widehat{\beta}_2| \leq \widehat{\sigma}_2 c)} + 1_{(i \in \mathcal{I}_2)} 1_{(|y_i - x_i' \widehat{\beta}_1| \leq \widehat{\sigma}_1 c)}. \tag{1.2.21}$$

1.5. Compute $\widehat{\Pi}_c^{(1)}$, $\widehat{\beta}_c^{(1)}$, $(\widehat{\sigma}_c^{(1)})^2$ *using equations* (1.2.14), (1.2.15), (1.2.16) *with* $m = 0$ *and let* $m = 1$.

2. Follow the steps 2 - 5 in Algorithm 1.I.

Note that this procedure will be particularly effective if outliers are concentrated in a certain subset of the whole sample. Saturated 2SLS is a simplified version of impulse indicator saturation (IIS), which was originally introduced in Hendry

(1999) when analysing US food expenditure time series data. IIS runs multiple block searches for outliers across many subsets of the sample. For analytical tractability, we focus on the stylised version presented in Algorithm 1.S with only two subsamples as an approximation to IIS.

We do not analyse and compare the robustness properties of Algorithms 1.R and 1.S formally. A formal evaluation is left for future research but Appendix 1.E explains in more detail why Algorithm 1.S is expected to be more robust in certain settings, supported by small, preliminary Monte Carlo simulations.

1.2.3 False Outlier Detection Rate (FODR)

The algorithms presented in Section 1.2.2 all have a positive probability to classify observations as outliers and hence remove them from the sample even when the true DGP contains no outliers. To evaluate the performance of the algorithms, the concept of the false outlier detection rate (FODR) is introduced. It is the sample proportion of falsely classified outliers and is defined as follows

$$\widehat{\gamma}_c^{(m)} = \frac{1}{n}\sum_{i=1}^{n}(1 - v_{i,c}^{(m)}). \tag{1.2.22}$$

The FODR is a measure to conceptualise the rate of type I error and to provide a way of choosing the cut-off value c, which is a tuning parameter in the algorithms. The purpose of this chapter is to derive an asymptotic theory for the FODR of the algorithms described in Section 1.2.2. Chapter 2 builds on this asymptotic theory to devise statistical tests for the presence of outliers in IVs regressions.

The FODR can be shown to coincide with the *gauge* in a special case of model selection, namely impulse indicator saturation (IIS), see Hendry (1999) who analyses US food expenditure between 1931 and 1989 by including sets of impulse indicators (dummies for each observation) in the time series regression, which were able to account for outliers. This procedure was later developed more formally into impulse indicator saturation in Hendry, Johansen and Santos (2008). The concept of the gauge was originally introduced in Hendry and Santos (2010) and Castle, Doornik and Hendry (2011) and more generally refers to the share of falsely retained regressors in a selection procedure, i.e. regressors that were selected but do

not belong into the DGP. Retaining an impulse indicator is equivalent to classifying this observation as an outlier. It follows that in the context of impulse indicator saturation when there are no outliers in the data generating process, the gauge equals the FODR. Johansen and Nielsen (2016b) study the gauge of various outlier robust estimators, including impulse indicator saturation, in an ordinary least squares setting.

Another well-known concept for measuring the degree of false reject decisions is the *false discovery rate* (FDR), which was introduced in Benjamini and Hochberg (1995). The FDR is related to but yet distinct from the FODR. Benjamini and Hochberg (1995) define the FDR as the expected value of the share of falsely rejected hypotheses to the total number of rejected hypotheses. Expressed in terms of the conceptual schema presented in Table 1.2.1, the FDR equals $\mathsf{E}(V/R)$. Since we analyse outlier detection under the null hypothesis that there are no outliers, we have $m = m_0$ and every rejected hypothesis is falsely rejected. Therefore, in our context the share V/R is either 1 (if at least one hypothesis is rejected) or 0 (if no rejection). As already noted in Benjamini and Hochberg (1995), when all hypotheses are true, the FDR equals the *family-wise error rate* (FWER), which is the probability of at least one false rejection, $\mathsf{P}(V \geq 1)$. This is distinct from the share of false rejections.

The population FODR is a more natural concept to evaluate the performance of outlier detection since it measures the expected share of falsely rejected hypotheses to the total number of true hypotheses, i.e. $\mathsf{E}(V/m_0)$. In our context, this translates into the share of falsely classified outliers, which is our object of interest (see Johansen and Nielsen 2016b).

The FODR is equal to a different concept, however, when only true hypotheses are tested as in our setting. In terms of Table 1.2.1, this means that $m_0 = m$. The *per*

Table 1.2.1: Number of Errors Committed When Testing m Null Hypotheses

	not rejected	rejected	total
true null hypotheses	U	V	m_0
false null hypotheses	T	S	$m - m_0$
total	$m - R$	R	m

comparison error rate (PCER) is defined as the expected value of the share of falsely rejected hypotheses to the total number of hypotheses, i.e. $\mathsf{E}(V/m)$. When only true hypotheses are tested, the PCER hence coincides with the FODR, $\mathsf{E}(V/m_0)$.

1.3 The Main Results

This section studies the FODR of the algorithms presented in Section 1.2.2 using empirical processes techniques and presents the main theoretical results. We start by listing the assumptions that are required for our asymptotic analysis. Then, we establish two types of asymptotic theory for the FODR: Gaussian and Poisson approximations.

1.3.1 Assumptions

We list assumptions sufficient for our asymptotic analysis of the FODR of the algorithms described in Section 1.2.2. Moment conditions are required for the structural errors u_i, first stage errors r_i, and instruments z_i.

We focus on the cross-sectional setup where the data $\{(y_i, x_i, z_i)\}_{i=1}^n$ are assumed to be iid. Thus, the instruments z_i can be normalised by $n^{-1/2}$ using the notation $z_{in} = n^{-1/2} z_i$. We then 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 the Law of Large Numbers (LLN) assuming $\mathsf{E}|z_i|^2 < \infty$. Let the infeasible (ideal) fitted values from the first stage be denoted by $\tilde{x}_i = \Pi' z_i$ while its feasible counterpart is attained by estimating Π consistently. The normalised $\tilde{x}_i$ is $\tilde{x}_{in} = n^{-1/2} \tilde{x}_i = \Pi' z_{in}$ 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 the rank condition for identification holds such that $\operatorname{rank} \Pi = d_x$ and $M_{zz} > 0$, then we also have $M_{\tilde{x}\tilde{x}} > 0$.

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

Assumption 1.3.1. *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 density* $\mathsf{f}_u(y)$ *is symmetric and satisfies for* $y \in \mathbb{R}$

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

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

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

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

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

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

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

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

  (*b*) $\sup_{y \in \mathbb{R}} (|\xi_y| + |\xi_{-y}|) y^2 \mathsf{f}_u(y) < \infty$*;*

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

While these assumptions may appear abstract, the conditions (i, ii, iii, iv) are satisfied in a range of situations. In particular, (ia) is satisfied by the normal and t-distribution; see Example 3.1 in Johansen and Nielsen (2016a), while (ib, ic) as well as (iia, iii) hold for the normal distribution, see Remark 2 in Johansen and Nielsen (2016b). The first stage errors r_i have finite fourth moment by (iia), thus condition

(iib) holds by Boole's and Markov's inequalities. Condition (iv) holds for stationary instruments; see Example 3.2 in Johansen and Nielsen (2016a). Condition (v) allows the standardised estimation errors to diverge at a rate of $n^{1/4-\eta}$ rather than being bounded in probability. As a special case, $\eta = 1/4$ can be chosen for estimators with standard convergence rates. Assumptions (ib, ic, iib, ivb) are only used to establish the Poisson exceedance theory.

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 of moments s for the density $\mathsf{f}_{u,r}$. In standard situations, the normalised estimator is bounded such that $\eta = 1/4$ and the moment condition reduces to $s = 2$. Otherwise, when the initial estimator diverges at rate η, the number of moments s grows linearly with the dimension of the regressors, d_x. This would be relevant for the $n^{1/3}$-consistent least median of squares estimator by Rousseeuw (1984).

1.3.2 Asymptotic Approximation to the FODR

The purpose of this subsection is to establish the asymptotic theory of the proportion $\widehat{\gamma}_c$ (FODR) and the number $n\widehat{\gamma}_c$ of falsely discovered outliers. This helps us understand the performance of Algorithms $1.I$, $1.R$, and $1.S$. Furthermore, it provides guidance for setting the cut-off value c to control the expected share or number of falsely detected outliers.

Section 1.3.2.2 approximates the FODR $\widehat{\gamma}_c$ as a stochastic process with varying cut-off values $c \in [c_+, \infty)$ by a Gaussian weak limit as $n \to \infty$, where $c_+ > 0$ is taken to be any small finite, positive number to ensure that we do not classify all observations as outliers. Section 1.3.2.3 approximates the number of false discoveries $n\widehat{\gamma}_c$ by a Poisson distribution as $c \to \infty$ when $n \to \infty$ in order to keep the expected number of outliers fixed.

The main asymptotic results are structured as follows: five main theorems are presented in Section 1.3 along with three further results in Appendix 1.C. First of all, Theorem 1.3.1 shows the tightness property for all our algorithms. This implies that

the FODR approaches the chosen level of significance as the sample size becomes large regardless of the choice of cut-off value and iteration step.

Next, we focus on the non-iterated version of Robustified 2SLS described in Algorithm 1.R, which is most common in empirical studies. Theorem 1.3.2 ($R^{(0)}$) builds up a Gaussian approximation to the FODR for the initial step of Algorithm 1.R.

Theorem 1.3.3 ($S^{(0)}$) then considers Saturated 2SLS described in Algorithm 1.S and establishes a Gaussian approximation to its FODR for the initial step.

Following our suggestions to iterate the procedures, we explore the fixed point of our algorithms and construct a Gaussian approximation to the FODR in Theorem 1.3.4 ($R^{(*)}$ and split-half $S^{(*)}$).

Finally, Theorem 1.3.5 establishes a Poisson approximation to the number of falsely discovered outliers at any iteration step, which holds for all our presented algorithms.

To make our analysis complete, the chapter demonstrates three further important results in Appendix 1.C. First, Theorem 1.C.1 ($R^{(1)}$) investigates the first iteration of Algorithm 1.R and derives a Gaussian weak limit for its FODR. Second, Theorem 1.C.2 ($R^{(m)}$ for $m \geq 1$) generalises the result to any iteration of Algorithm 1.R. Third, Theorem 1.C.3 (split-half $S^{(m)}$ for $m \geq 0$) demonstrates asymptotic equivalence between the FODR of Algorithm 1.S when the sample is split in half to that of Algorithm 1.R. This result holds for all iterations $m \geq 0$.

All auxiliary lemmas and proofs are provided in the Appendices. Appendix 1.A presents LLNs and Central Limit Theorems (CLTs) for a class of weighted and marked empirical processes that form the basis of the derivations of the main theorems. Appendix 1.B studies the asymptotic properties of the iterated estimators of (β, σ^2), which are necessary for establishing the theory of the FODR in Appendix 1.C.

1.3.2.1 Tightness

First, we show that the FODR of Algorithm 1.I approaches the chosen level of significance as the sample size becomes large uniformly in the cut-off value c and iteration step m. The result also applies to Algorithms 1.R and 1.S because their initial estimators are known and tight. In order for this result to hold for $m \geq 1$, the bias correction of the variance in our suggested iterated procedure (see equation (1.2.16)) is essential. After stating our result, we will explore the consequences of not using the bias correction factor.

Theorem 1.3.1. *Consider Algorithm 1.I. Suppose Assumption 1.3.1*(ia, iia, iii, iv, v) *holds with* $\eta = 1/4$. *Then, as* $n \to \infty$

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

The proof requires an asymptotic expansion of the FODR in terms of the iterated (β, σ^2)-estimators, which is presented in Lemma 1.C.1. The result then follows in combination with the tightness result of those iterated (β, σ^2)-estimators.

This tightness result implies the uniform consistency property of the FODR such that $\widehat{\gamma}_c^{(m)} \xrightarrow{\mathsf{P}} \gamma_c$ uniformly in the cut-off value c and iteration step m as $n \to \infty$. This establishes an asymptotic relationship between the unique tuning parameter c of Algorithm 1.I and the selection probability of an individual outlier. Thus, it can be used for calibration of the algorithm and provides a rationale for choosing c to match the desired population FODR, γ_c. That is to say, we first specify our tolerance for the false positive rate expressed by γ_c, which results in a selection quantile c given the reference distribution f_u. For example, consider a standard normal distribution for the reference distribution f_u and a sample size of $n = 100$. When we tolerate in expectation one falsely detected outlier, we then want to set the FODR to $\gamma_c = 1\%$ and thus choose $c = 2.58$.

Importance of Bias-Correcting

Suppose we used a naive iterated procedure analogous to Algorithm 1.I but without the bias correction factor ς_c^{-2} for the variance estimator in equation (1.2.16). For

illustrative purposes, we focus on the first iteration step $m = 1$. The biased variance estimator is therefore given in terms of the unbiased estimator as $(\widehat{\sigma}^{(1)}_{c,bias})^2 = \varsigma_c^2(\widehat{\sigma}^{(1)}_c)^2$. Its probability limit is $(\widehat{\sigma}^{(1)}_{c,bias})^2 \xrightarrow{\mathsf{P}} \varsigma_c^2\sigma^2 < \sigma^2$, such that we would underestimate the true variance of the structural error. The reason is that in the initial step $m = 0$, we classified those observations with large residuals as outliers. The subsequent estimation ($m = 1$) on the subsample of non-outlying observations therefore underestimates the variance.

This in turn impacts the new classification of outliers based on the updated estimates $\widehat{\beta}^{(1)}_c$ and $\widehat{\sigma}^{(1)}_{c,bias}$. An observation is classified as an outlier if its standardised residual is larger than the cut-off value, which now becomes $1_{(|\widehat{u}^{(1)}_i/\widehat{\sigma}^{(1)}_{c,bias}|>c)}$. This can equivalently be written in terms of the unbiased variance estimator as $1_{(|\widehat{u}^{(1)}_i/\widehat{\sigma}^{(1)}_c|>\varsigma_c c)}$. It shows that using the naive, biased variance estimator with cut-off value c is equivalent to using the bias-corrected variance estimator with some alternative cut-off value $c' := \varsigma_c c < c$. Since the effective cut-off value c' is smaller than our intended cut-off value c, we find ourselves making more false classifications as outliers than intended. Using the same arguments as in the proof of Theorem 1.3.1, it can be shown formally that $\widehat{\gamma}^{(1)}_{c,bias} \xrightarrow{\mathsf{P}} \gamma_{c'} > \gamma_c$. The bias can be substantial and is worse for larger values of the target FODR γ_c (smaller cut-off values c). For example, under normality and targeting $\gamma_c = 0.05$, the biased procedure will actually produce $\widehat{\gamma}^{(1)}_{c,bias} \xrightarrow{\mathsf{P}} \gamma_{c'} \approx 0.09$. In conclusion, using a naive iterated procedure that does not bias correct the variance will lead to (possibly substantial) over-detection of outliers, giving the researcher a false sense of control of the false outlier detection rate and potentially false conclusions about the presence of outliers in their sample.

Returning to our suggested algorithms with bias correction and their tightness result, we next consider the FODR with varying cut-offs $c \in [c_+, \infty)$. In this case, the normalised FODR $n^{1/2}(\widehat{\gamma}^{(0)}_c - \gamma_c)$ calculated at the initial step $m = 0$ of Algorithms $1.R$ or $1.S$ becomes a sequence of stochastic processes. In the coming section, the tightness result is required to establish its weak convergence theory. In addition, tightness is a key ingredient of our argument to show the fixed point $\widehat{\gamma}^{(*)}_c$ of the algorithms as the iteration step increases, $m \to \infty$.

1.3.2.2 Gaussian Approximation

Since Algorithms $1.R$ and $1.S$ satisfy the assumptions of Theorem 1.3.1, we can further derive their weak convergence by establishing their finite dimensional convergence. We start by demonstrating the Gaussian approximation to the FODR at the initial step of Algorithms $1.R$ and $1.S$.

Theorem 1.3.2. *Consider Algorithm 1.R. Suppose Assumption 1.3.1*(ia, iia, iii, iv) *holds. Let* $\mathbb{G}_n^{(0)}(c) = n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c)$ *for* $c \in [c_+, \infty)$ *denote a sequence of processes of the FODR at the initial step* $m = 0$*. Then, as* $n \to \infty$

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

where $\mathsf{GP}(c; 0, K^{(0)})$ *for* $c \in [c_+, \infty)$ *refers to the Gaussian process with expected value* $\mathsf{E}\{\mathsf{GP}(t; 0, K^{(0)})\} = 0$ *and covariance* $K_{st}^{(0)} = \mathsf{Cov}\{\mathsf{GP}(s; 0, K^{(0)}), \mathsf{GP}(t; 0, K^{(0)})\}$ *for any* $c_+ \leq s \leq t < \infty$ *such that*

$$\begin{aligned} K_{st}^{(0)} = {} & \gamma_t(1-\gamma_s) - s\mathsf{f}_u(s)(1-\gamma_t-\tau_2^t) - t\mathsf{f}_u(t)(1-\gamma_s-\tau_2^s) + st\mathsf{f}_u(s)\mathsf{f}_u(t)(\tau_4-1) \\ & + \mathsf{f}_u(s)\mathsf{f}_u(t)(\zeta_s^- - 2s\Omega)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}(\zeta_t^- - 2t\Omega). \end{aligned}$$

Robustified 2SLS starts with the full-sample 2SLS estimator and Lemma 1.C.1 expands its FODR at the initial iteration in terms of the full-sample 2SLS. We finish the proof by studying the asymptotic properties of the 2SLS estimators, especially the σ-estimator in Lemma 1.B.1.

To compare the result to the corresponding one in the classical regression without endogeneity, assume that $\mathsf{E}x_i u_i = 0_{d_x}$ such that $\mathsf{E}r_i u_i = 0_{d_x}$ and $\zeta_c^- = \Omega = 0_{d_x}$. Then the fifth term $\mathsf{f}_u(s)\mathsf{f}_u(t)(\zeta_s^- - 2s\Omega)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}(\zeta_t^- - 2t\Omega)$ in the asymptotic covariance $K_{st}^{(0)}$ becomes zero, and the above convergence result reduces to that shown in the classical setting, see Theorem 6 in Jiao and Pretis (2022) as well as Theorem 7 in Johansen and Nielsen (2016b). Recall that both Ω and $\zeta_c^- = \xi_c - \xi_{-c}$ with $\xi_c = \mathsf{E}(\Sigma^{-1/2}r_i | u_i/\sigma = c)$ and $\Omega = \mathsf{E}(\Sigma^{-1/2}r_i u_i/\sigma)$ are related measures of the level of endogeneity in the regression, which appear in the expression of the IV setting.

Moreover, under normality of the errors, see Remark 1.2.1, we have $\zeta_c^- = 2c\Omega$, and the asymptotic result of the FODR in the IV setting also coincides with that in the classical regression even with endogeneity when $\mathsf{E}x_i u_i \neq 0_{d_x}$. Next, we turn to Algorithm 1.S and establish the weak convergence result for its FODR at the initial step $m = 0$ for any split of subsamples, including unequal splits.

Theorem 1.3.3. *Consider Algorithm 1.S. Suppose that the sample is split into a partition, $\mathcal{I}_1$ and $\mathcal{I}_2$, with number of observations $n_1 = \text{int}[\alpha n]$ and $n_2 = n - n_1$ where $\alpha \in (0,1)$ represents the share of observations in subsample $\mathcal{I}_1$. Suppose Assumption 1.3.1(ia, iia, iii, iv) holds for each subsample. Let $\mathbb{G}_n^{(0)}(c) = n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c)$ for $c \in [c_+, \infty)$ denote a sequence of processes of the FODR at step $m = 0$. Then, as $n \to \infty$*

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

where $\mathsf{GP}(c; 0, K^{(0)})$ for $c \in [c_+, \infty)$ refers to the Gaussian process with expected value $\mathsf{E}\{\mathsf{GP}(t; 0, K^{(0)})\} = 0$ and covariance $K_{st}^{(0)} = \mathsf{Cov}\{\mathsf{GP}(s; 0, K^{(0)}), \mathsf{GP}(t; 0, K^{(0)})\}$ for any $c_+ \leq s \leq t < \infty$ such that

$$\begin{aligned} K_{st}^{(0)} &= \gamma_t(1-\gamma_s) - s\mathsf{f}_u(s)(1-\gamma_t-\tau_2^t) - t\mathsf{f}_u(t)(1-\gamma_s-\tau_2^s) \\ &\quad + \frac{\alpha^3+(1-\alpha)^3}{\alpha(1-\alpha)} st\mathsf{f}_u(s)\mathsf{f}_u(t)(\tau_4-1) \\ &\quad + \frac{\alpha^3+(1-\alpha)^3}{\alpha(1-\alpha)} \mathsf{f}_u(s)\mathsf{f}_u(t)(\zeta_s^- - 2s\Omega)'\Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}(\zeta_t^- - 2t\Omega). \end{aligned}$$

Saturated 2SLS splits the whole sample into two subsamples and uses 2SLS estimates calculated on one subsample to detect outliers in the other subsample. In Lemma 1.C.3, we derive an expansion of the FODR in terms of the two 2SLS estimators that are computed on each subsample. Applying a similar argument as in the proof of Theorem 1.3.2 ($R^{(0)}$) completes the proof of the above weak convergence result.

The FODR at the initial step of Algorithm 1.S has a similar asymptotic covariance of the Gaussian weak limit as that of Algorithm 1.R in Theorem 1.3.2 ($R^{(0)}$). The two algorithms share the same first three terms but the fourth and fifth terms for Algorithm 1.S are multiplied by a factor that depends on α, which represents how the sample is split. Splitting the sample in half, i.e. $\alpha = 1/2$, we find that the

factor becomes 1 and it immediately follows that the split-half version of Algorithm 1.S has the same limiting Gaussian process as Algorithm 1.R.

The weak convergence results for the Gaussian approximation to the iterated FODR $\widehat{\gamma}_c^{(m)}$, $m \in [1, \infty)$ become much more complicated for both Algorithms 1.R and 1.S. Thus, we present those results in Appendix 1.C and illustrate them in the coming subsection by plotting the asymptotic variances of $\widehat{\gamma}_c^{(m)}$ for a variety of cut-off values c and iteration steps m. Theorem 1.C.1 ($R^{(1)}$) considers the first iteration $m = 1$ of Algorithm 1.R and establishes a Gaussian approximation to $\widehat{\gamma}_c^{(1)}$. Theorem 1.C.2 ($R^{(m)}$ for $m \geq 1$) extends the results to $\widehat{\gamma}_c^{(m)}$ for any iteration step $m \in [1, \infty)$. The proof requires an expansion of $\widehat{\gamma}_c^{(m)}$ for $m \in [1, \infty)$ in terms of the initial estimators $(\widehat{\beta}_c^{(0)}, \widehat{\sigma}_c^{(0)})$. This is shown in Lemma 1.C.2, which builds on a one-step asymptotic relationship between consecutive estimators of (β, σ^2) across two consecutive iterations, stated in Lemma 1.B.2. We then return to the split-half version ($\alpha = 1/2$) of Algorithm 1.S and demonstrate its asymptotic equivalence to Algorithm 1.R. The FODRs of the two algorithms then have the same Gaussian weak limit at any iteration step. This result is presented in Theorem 1.C.3 (split-half $S^{(m)}$ for $m \geq 0$) with the proof involving stochastic expansions of the iterated estimators of (β, σ^2) when starting with an equal sample split, provided in Lemma 1.B.3.

In most empirical studies, the FODR $\widehat{\gamma}_c^{(0)}$ calculated at the initial iteration $m = 0$ of Algorithm 1.R is of special interest. Jiao (2022) suggests iterating the algorithm until finding a fixed point, which has the same first order asymptotics as the Huber-skip regression and thus results in an estimator that is more resistant to outliers. Therefore, we now study the FODR $\widehat{\gamma}_c^{(m)}$ when the procedure is iterated infinitely as described in Algorithm 1.I, i.e. as $m \to \infty$.

Theorem 1.3.4. *Consider Algorithm 1.I started from Algorithm 1.R or 1.S in its split-half version. Suppose further that Assumption 1.3.1*(ia, iia, iii, iv) *holds. Then, for all* $\epsilon, \delta > 0$ *a pair* $m_0, n_0 > 0$ *exists, such that for all* $m > m_0$ *and* $n > n_0$

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

Moreover, let $\mathbb{G}_n^{(*)}(c) = n^{1/2}(\widehat{\gamma}_c^{(*)} - \gamma_c)$ *for* $c \in [c_+, \infty)$ *denote a sequence of processes of the corresponding FODR at the fixed point when* $m \to \infty$*. As* $n \to \infty$*, we obtain*

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

where $\mathsf{GP}(c; 0, K^{(*)})$ *for* $c \in [c_+, \infty)$ *refers to the Gaussian process with expected value* $\mathsf{E}\{\mathsf{GP}(t; 0, K^{(*)})\} = 0$ *and covariance* $K_{st}^{(*)} = \mathsf{Cov}\{\mathsf{GP}(s; 0, K^{(*)}), \mathsf{GP}(t; 0, K^{(*)})\}$ *for any* $c_+ \leq s \leq t < \infty$ *such that*

$$\begin{aligned} K_{st}^{(*)} &= \gamma_t(1-\gamma_s) + \frac{\psi_s t \mathsf{f}_u(t)(\varsigma_s^2 - \varsigma_t^2)}{\tau_2^t - t\mathsf{f}_u(t)(t^2 - \varsigma_t^2)} + \frac{s\mathsf{f}_u(s)(\tau_4^s - \tau_2^s\varsigma_s^2)}{\tau_2^s - s\mathsf{f}_u(s)(s^2 - \varsigma_s^2)} \frac{t\mathsf{f}_u(t)}{\tau_2^t - t\mathsf{f}_u(t)(t^2 - \varsigma_t^2)} \\ &+ \frac{\tau_2^s \mathsf{f}_u(s)}{\{\psi_s - 2s\mathsf{f}_u(s)\}\{\tau_2^s - s\mathsf{f}_u(s)(s^2 - \varsigma_s^2)\}} \frac{\mathsf{f}_u(t)}{\{\psi_t - 2t\mathsf{f}_u(t)\}\{\tau_2^t - t\mathsf{f}_u(t)(t^2 - \varsigma_t^2)\}} \\ &\times (\tau_2^s \zeta_s^- - 2s\omega_s)' \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \Sigma^{1/2} (\tau_2^t \zeta_t^- - 2t\omega_t). \end{aligned}$$

Algorithm 1.R and the split-half version of Algorithm 1.S converge in probability to the same fixed point when the iteration becomes sufficiently large despite starting with different initial estimators. The argument uses the Gaussian approximation results for the FODRs $\widehat{\gamma}_c^{(m)}$ of Algorithms 1.R and 1.S and letting $m \to \infty$ in Theorems 1.C.2 ($R^{(m)}$ for $m \geq 1$) and 1.C.3 (split-half $S^{(m)}$ for $m \geq 0$) completes the proof.

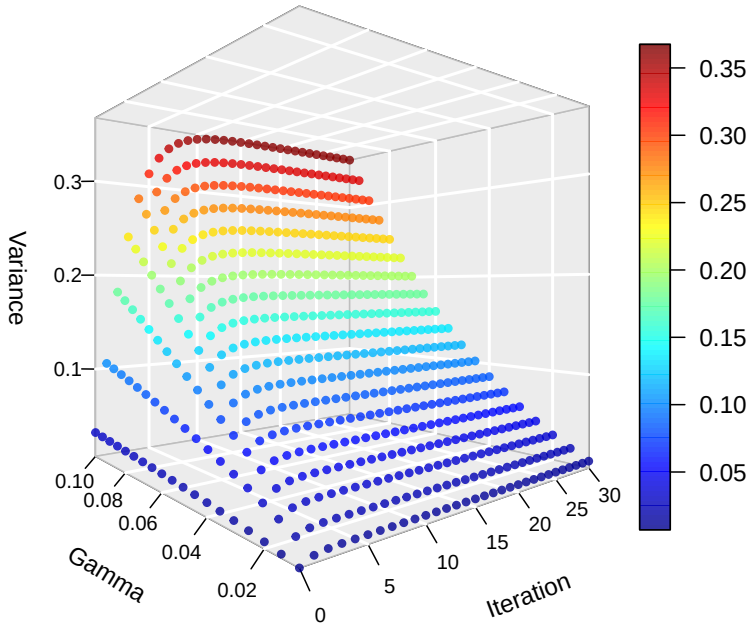
The expression of the Gaussian weak limit for $\widehat{\gamma}_c^{(m)}$ is very complicated for general m, however it simplifies for the fixed point when $m \to \infty$. This is similar to the results derived by Johansen and Nielsen (2016b) in the classical regression setting. As before, when considering a setting with exogenous regressors $\mathsf{E}x_i u_i = 0_{d_x}$, the terms related to the degree of endogeneity ζ_c^-, Ω, and $\omega_c = \mathsf{E}(\Sigma^{-1/2} r_i u_i 1_{(|u_i| \leq \sigma c)}/\sigma)$ become zero such that the fourth term in the asymptotic covariance becomes zero and the covariance reduces to that shown in the classical setting, see Theorem 7 in Jiao and Pretis (2022). Again, the IV result also coincides with classical regressions under normality of the errors because then $\omega_c = \tau_2^c \Omega$, as shown in Remark 1.2.1.

Graphical Illustration of Gaussian Approximation Results

We now illustrate and discuss the effect of both tuning parameters, the cut-off value c and the iteration step m, on the asymptotic variance of the Gaussian approximation of the FODR. In general, the formulas for the asymptotic variance of the FODR depend on certain moments of the errors: Σ, Ω, ζ_c^-, and ω_c. Under joint normality of the errors as in Remark 1.2.1, the formulas for the asymptotic variance simplify such that they only depend on c and m.

The following illustration refers to Theorems 1.3.2 ($R^{(0)}$), 1.3.4 ($R^{(*)}$ and split-half $S^{(*)}$), 1.C.2 ($R^{(m)}$ for $m \geq 1$), and 1.C.3 (split-half $S^{(m)}$ for $m \geq 0$). Figure 1.3.1 shows the asymptotic variance of the FODR for different values of the iteration step (m) and of the probability of exceeding the cut-off (γ_c), which is equivalent to varying the cut-off level (c). The asymptotic variance increases both as the number of iterations increases and as the cut-off decreases.

Figure 1.3.1: Asymptotic Variance of the FODR



Plot of the asymptotic variance of the FODR $\widehat{\gamma}_c^{(m)}$ for different values of iteration m ranging from 0 to 30 and of the probability of exceeding the cut-off γ_c ranging from 0.01 to 0.1 (under normality corresponding to cut-off c ranging from 2.58 to 1.64).

Decreasing the cut-off means that in expectation a larger share of observations is classified as outliers. This leads to a more robust estimator when the sample is contaminated but is costly under the null when there are no actual outliers in the sample, leading to less efficient estimates. This represents an efficiency-robustness trade-off.

For a fixed probability γ_c of exceeding the cut-off c, the asymptotic variance of the FODR $\widehat{\gamma}_c^{(m)}$ increases in the number of iterations m while its probability limit remains at γ_c. Thus, when the sample is not contaminated we expect the same share of observations to be classified as outliers for different iterations. However, the variability in the share of classified outliers from previous iterations seems to persist and even amplify as the number of iterations increases. This loss of precision when estimating the FODR when there are no actual outliers may be compensated by higher robustness under contamination. Increasing the number of iterations of the outlier detection algorithm may result in a more robust estimator. For example, the initial full-sample estimator of Algorithm 1.R is not robust itself while the iterated procedure might provide higher power to detect outliers in a contaminated sample. The iterated procedures described in Algorithm 1.I are similar to the iterated one-step Huber-skip estimator in the classical regression setting.

Figure 1.3.1 also shows that the value of the asymptotic variance levels off relatively quickly as the number of iterations m increases. This happens faster the higher the cut-off value c is. We do not show the asymptotic variance of the fixed point FODR given in Theorem 1.3.4 ($R^{(*)}$ and $S^{(*)}$) because it is virtually the same as that of iteration step $m = 30$. The maximum absolute difference between $\widehat{\gamma}_c^{(*)}$ and $\widehat{\gamma}_c^{(30)}$ across the presented range of c is less than 4.4×10^{-7}.

1.3.2.3 Poisson Approximation

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

$$n\mathsf{P}(|u_i| > \sigma c_n) = \lambda. \tag{1.3.1}$$

Notice that $c_n \to \infty$ as $n \to \infty$. Define $v_{i,c_n}^{(m)}$, $\widehat{\beta}_{c_n}^{(m+1)}$, and $(\widehat{\sigma}_{c_n}^{(m+1)})^2$ by replacing c with c_n in expressions (1.2.13), (1.2.15), and (1.2.16). The corresponding FODR is

$$\widehat{\gamma}_{c_n}^{(m)} = \frac{1}{n}\sum_{i=1}^{n}(1 - v_{i,c_n}^{(m)}) = \frac{1}{n}\sum_{i=1}^{n} 1_{(|y_i - x_i'\widehat{\beta}_{c_n}^{(m)}| > \widehat{\sigma}_{c_n}^{(m)} c_n)}. \tag{1.3.2}$$

Making use of the tightness property of the iterated β, σ^2 estimators, one can find two indicators that bound $1_{(|y_i - x_i'\widehat{\beta}_{c_n}^{(m)}| > \widehat{\sigma}_{c_n}^{(m)} c_n)}$ and do not depend on the estimators $\widehat{\beta}_{c_n}^{(m)}, \widehat{\sigma}_{c_n}^{(m)}$. By exploring these upper and lower bounds, the following Poisson limiting theorem arises. The idea underlying the proof is similar to Theorem 8 in Johansen and Nielsen (2016b).

Theorem 1.3.5. *Consider Algorithm 1.I. Let c_n be defined as in* (1.3.1)*. Suppose Assumption 1.3.1 holds with $\eta = 1/4$. Then, uniformly in $m \in [0, \infty)$ and as $n \to \infty$, the FODR in* (1.3.2) *satisfies*

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

Table 1.3.1 shows example cut-off values for sample sizes of $n = 100$ and $n = 200$, assuming that u_i/σ follows a standard normal distribution. For a given expected number of detected outliers λ, the cut-off in equation (1.3.1) depends on the sample size to satisfy $c_n = \Phi^{-1}\{1 - \lambda/(2n)\}$. 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 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 of finding at most 1 and 2 outliers are 0.74 and 0.92, respectively. This means it regularly finds two outliers when there are none in the true DGP.

The Poisson approximation is uniform in the iteration step m. Thus, it is more robust than the Gaussian approximation, which has a different variance for different iteration steps m. Algorithms 1.R and 1.S are special versions of Algorithm 1.I with different starting points. Their initial estimates do not depend on the cut-off, and thus satisfy the tightness property. Therefore, the above Poisson approximation theorem also applies to both of these algorithms.

Table 1.3.1: Probability of Detecting at Most x Outliers

			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				

The probability of detecting at most x outliers approximated by a Poisson distribution for a given expected value λ, and the varying cut-off $c_n = \Phi^{-1}\{1-\lambda/(2n)\}$ in order to fix the expected number of detected outliers λ for $n = 100, 200$.

1.4 Simulation Study

In this section, we verify the asymptotic distribution results derived in the previous section and check their finite sample performance using extensive Monte Carlo simulations. The simulations are based on the *robust2sls* R package (Kurle 2022b), which is the companion R package to this chapter that implements all the algorithms presented in Section 1.2.2, the results in Section 1.3, and the corrected inference on and testing of the location parameters derived in Jiao (2022).[2]

We use a DGP where the structural equation contains an intercept and one endogenous regressor, which is instrumented by one excluded and informative instrument, such that we are in the just-identified case. Formally, the structural equation is given as

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \quad i = 1, 2, \ldots, n,$$

where $x_{1i} = 1$ for all i and $\beta = (\beta_1, \beta_2)' = (2, 4)'$. The first stage projection is

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \pi_1 & \pi_2 \end{pmatrix} \begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix} + \begin{pmatrix} 0 \\ r_{2i} \end{pmatrix}, \quad i = 1, 2, \ldots, n.$$

We choose $\pi_1 = 0$ and $\pi_2 = 1$ such that Π is a 2×2 identity matrix I_2. Both the structural error u_i and the first stage error r_{2i} are taken to have variance 1, i.e. $\sigma^2 = 1$ and $\Sigma_2 = 1$. The covariance between the two errors, which is the source of endogeneity, is varied across five different settings: $\Omega_2 \in \{3/4, 1/2, 1/3, 1/6, 0\}$,

2. The simulations were performed using R Statistical Software (v4.1.2; R Core Team 2022) in RStudio (Posit Team 2023). We used the tidyverse R packages (Wickham et al. 2019) to evaluate and visualise the simulation results.

where x_{2i} is exogenous in the final setting. The excluded instrument z_{2i} is also mean zero with variance one, i.e. $\mu_{z_2} = 0$ and $\sigma^2_{z_2} = 1$. To ensure the validity of the instrument, $\mathsf{E}z_{2i}u_i = 0$, the errors and z_{2i} are in practice drawn from a multivariate normal distribution

$$\begin{pmatrix} u_i \\ r_{2i} \\ z_{2i} \end{pmatrix} \overset{\mathsf{D}}{=} \mathsf{N}\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \Omega_2 & 0 \\ \Omega_2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\}.$$

The sample size is varied widely from 100 to 1 million in order to both assess the finite sample performance and to check the correctness of the asymptotic theory, i.e. $n \in \{100, 1000, 10{,}000, 100{,}000, 1{,}000{,}000\}$. We set $W = 50{,}000$ replications for our Monte Carlo experiments.

1.4.1 Evaluation of the Gaussian Approximation

In this subsection, we assess the Gaussian approximation theory for the Robustified 2SLS Algorithm 1.R and the split-half version of the Saturated 2SLS Algorithm 1.S. Results for Saturated 2SLS with unequal splits can be found in Appendix 1.D.1.

We vary γ_c, the probability of drawing an error that lies beyond the cut-off value c. We choose $\gamma_c \in \{0.01, 0.05\}$, which corresponds to $c \in \{2.58, 1.96\}$ under normality of the reference distribution f_u.

In addition, we vary the iteration step $m \in \{0, 1, 5, *\}$, where the asterisk $*$ denotes the fixed point, in which the algorithms are iterated until convergence. This ensures that the selection of non-outlying observations would not change if iterated further. Distance between iterations is measured by the L_2 norm between subsequent β-estimates and convergence is reached when $\|\widehat{\beta}_c^{(m)} - \widehat{\beta}_c^{(m-1)}\|_2 = 0$. Since we cannot guarantee that this condition is reached in a reasonable amount of iterations and thus to reduce the computational cost, we add a second stopping criterion that halts the algorithm after 50 iterations if the previous condition is not reached. This potential limitation is negligible in practice. Across all fixed point simulations, only 0.0011% of replications were stopped at $m = 50$ instead of converging earlier. The maximum share of such replications within a single Monte Carlo experiment

was 0.016%. Hence, the effect of iterations that have not fully converged is negligible and the presented simulations for the fixed point can be interpreted as such. Appendix 1.D.2 presents more detailed results about the convergent iterations, including histograms of the convergent iterations and their average.

Overall, the simulations cover the results in Theorem 1.3.2 ($R^{(0)}$), Theorem 1.3.4 ($R^{(*)}$ and split-half $S^{(*)}$), Theorem 1.C.1 ($R^{(1)}$), Theorem 1.C.2 ($R^{(m)}$ for $m \geq 1$), and Theorem 1.C.3 (split-half $S^{(m)}$ for $m \geq 0$). The asymptotic theory for the Gaussian approximation presented in these various theorems is of the form

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) \xrightarrow{\mathsf{D}} \mathsf{N}(0, K_{cc}^{(m)}),$$

where we focus on the finite dimensional convergence result and where $K_{cc}^{(m)}$ represents the asymptotic variance. It can be re-arranged to obtain the approximation

$$\widehat{\gamma}_c^{(m)} \overset{a}{\sim} \mathsf{N}(\gamma_c, \frac{K_{cc}^{(m)}}{n}),$$

which is then evaluated by the simulation study in this subsection.

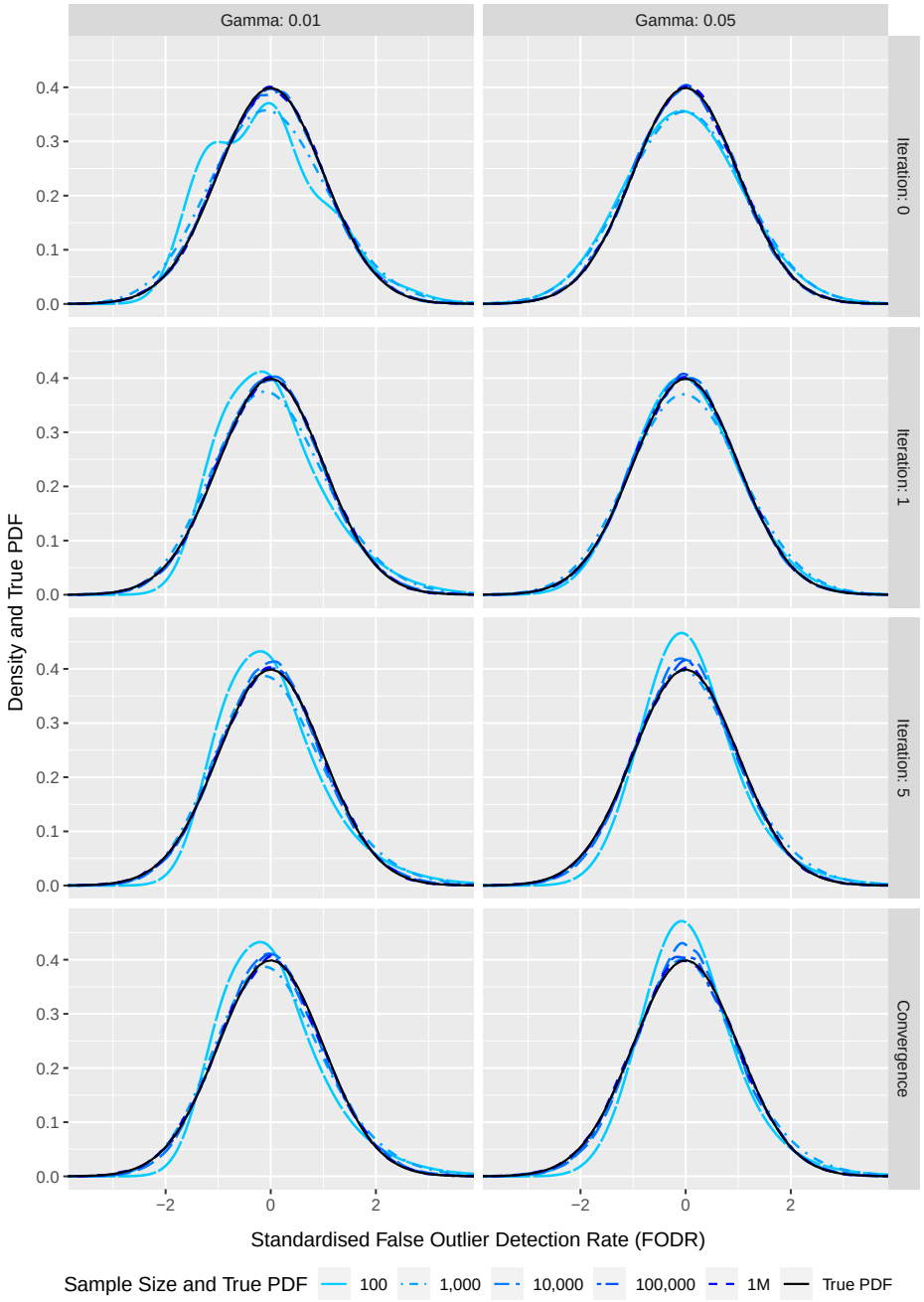
1.4.1.1 Distributional Comparison

In this subsection, we present the distribution of the simulated FODR and compare it to the normal distribution. For brevity, Figure 1.4.1 only includes the Robustified 2SLS Algorithm 1.R and the DGP with the highest degree of endogeneity, $\Omega_2 = 0.75$. After confirming the Gaussian limit distribution, the next two subsections analyse a wider variety of settings, focusing on the mean and variance since the first two moments suffice to characterise the normal distribution.

Figure 1.4.1 plots the probability density function of the standard normal distribution and smoothed kernel density estimates of the standardised distribution of the FODR, $n^{1/2}(\widehat{\gamma}_{c,w}^{(m)} - \gamma_c)/(K_{cc}^{(m)})^{1/2}$, where the subscript $w = 1, 2, \ldots, W$ refers to the replication of the simulation and $W = 50{,}000$.

The graphs show that the standardised FODR converges to a Gaussian distribution and that the approximation is very close for a sample size of $n = 10{,}000$ and larger. For small sample sizes, the density estimates of the simulations are less smooth and slightly skewed due to integer issues in finite samples. This is because

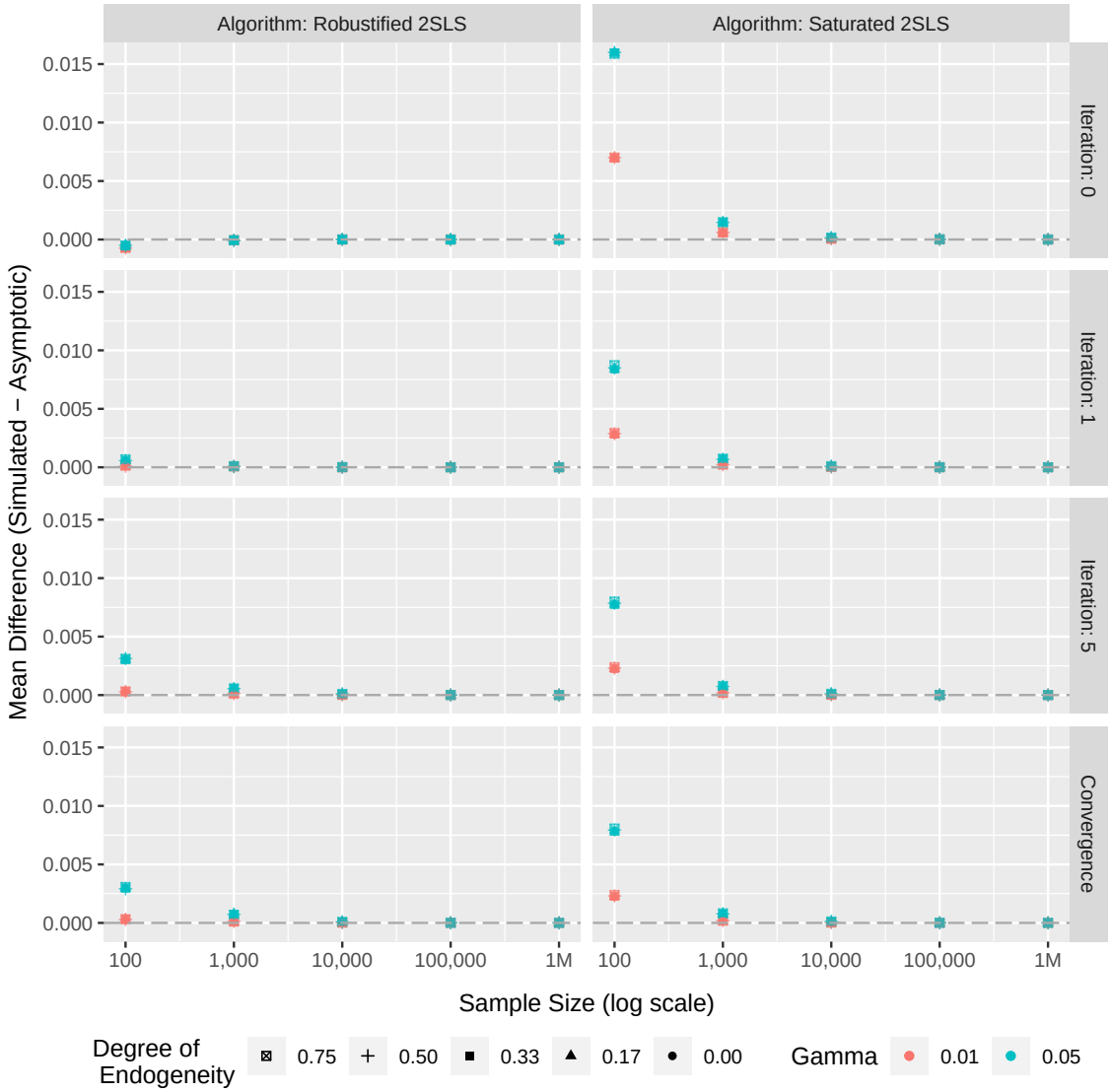
Figure 1.4.1: Evaluation of the Distribution of the FODR



Comparison of the theoretical distribution of the standardised FODR (black) with the simulated distributions of the FODR (different shades of blue and different line types) for the Robustified 2SLS Algorithm 1.R. Several settings are varied: the iteration step m, the cut-off c (or equally γ_c), and the sample size n. All Monte Carlo simulations used $W = 50{,}000$ replications.

the FODR is a share of the sample size and as such cannot take on as many different values in small samples as in large samples.

Figure 1.4.2: Evaluation of the Asymptotic Mean of the FODR



Comparison of the theoretical asymptotic mean of the FODR with the simulated mean of the FODR. The y-axis shows the difference between simulated and asymptotic mean such that a value of 0 means that they coincide. Several settings are varied: the algorithm (initial estimator), the iteration step m, the degree of endogeneity Ω_2 in the DGP, the cut-off c (or equally γ_c), and the sample size n. All Monte Carlo simulations used $W = 50{,}000$ replications.

1.4.1.2 Mean Comparison

Figure 1.4.2 evaluates the difference between the simulated mean of the FODR and the theoretical mean. The simulated mean is defined as the average FODR across simulations, $\bar{\widehat{\gamma}}_c^{(m)} = \sum_{w=1}^{W} \widehat{\gamma}_{c,w}^{(m)}/W$ and Figure 1.4.2 displays $\bar{\widehat{\gamma}}_c^{(m)} - \gamma_c$. Values close to zero confirm the theory. Across all settings, we find that the degree of endogeneity has virtually no effect on the performance.

Figure 1.4.2 shows that the asymptotic theory with respect to the mean works well for both algorithms. For large samples, the difference between simulated and theoretical mean is close to zero, supporting the theoretical result that $\widehat{\gamma}_c^{(m)} \xrightarrow{\mathsf{P}} \gamma_c$.

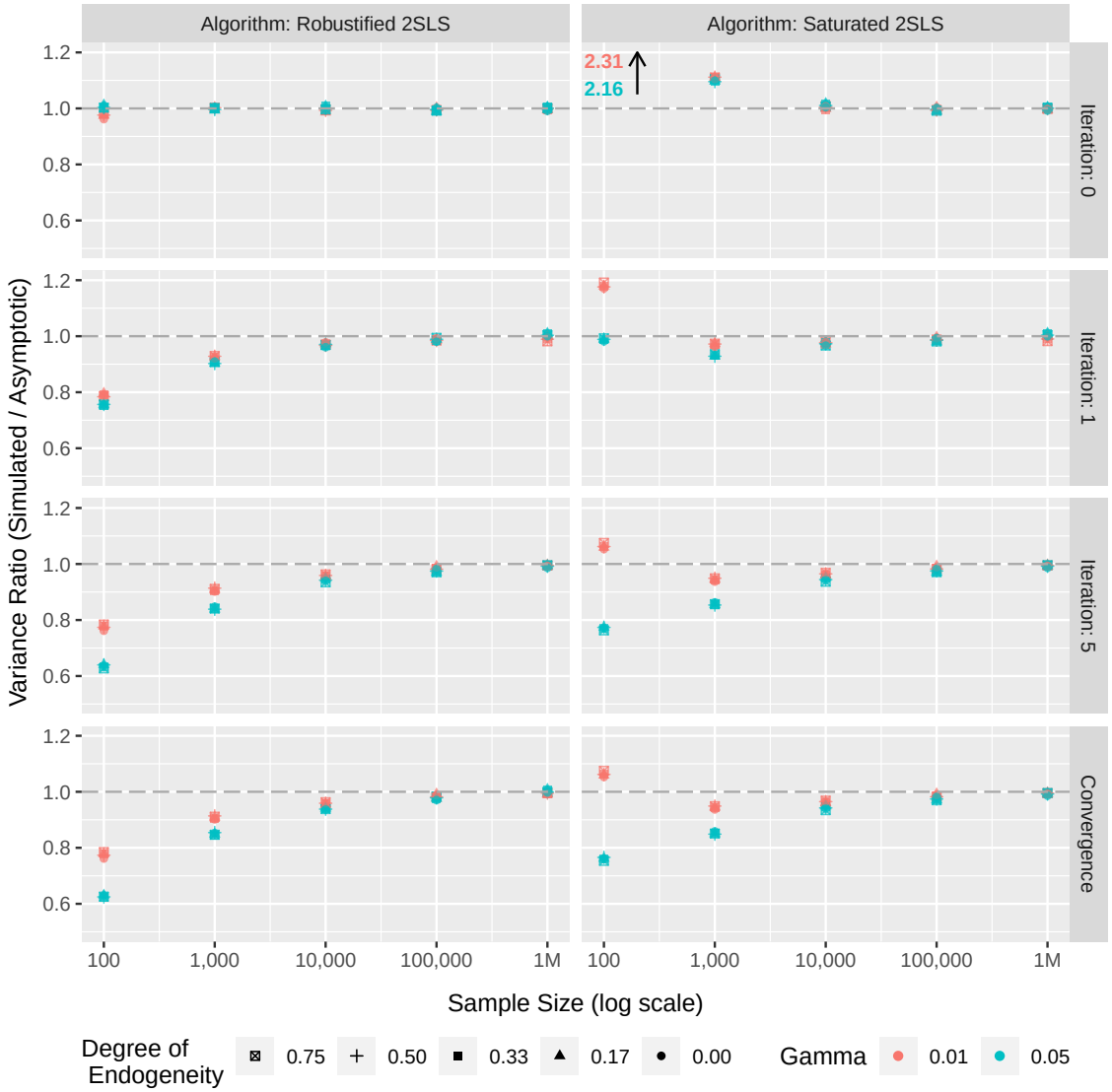
In terms of finite sample performance, the Robustified 2SLS Algorithm 1.R only shows major deviations from the theoretical value of zero for a small sample size of 100 and the looser significance level of 0.05. This deviation seems to be larger for higher iterations but shrinks quickly as the sample size increases. For the split-half version of the Saturated 2SLS Algorithm 1.S, we find the same pattern of larger deviations from zero for looser significance levels. For $n = 100$, the deviation is much more pronounced than in the case of Robustified 2SLS. However, it now decreases not only for larger sample sizes but also for higher iterations. Robustified and Saturated 2SLS only differ in their initial step while the iterated part of the procedure is identical, potentially explaining why they become closer. Even for large iterations, however, the finite sample bias of Saturated 2SLS remains larger than that of Robustified 2SLS.

1.4.1.3 Variance and Covariance Comparison

Figure 1.4.3 evaluates the second part of the asymptotic theory — the asymptotic variance. We estimate the variance of the FODR using the Monte Carlo simulations by calculating $\bar{K}_{cc}^{(m)} = n \times \sum_{w=1}^{W} (\widehat{\gamma}_{c,w}^{(m)} - \bar{\widehat{\gamma}}_c^{(m)})^2/(W-1)$. We then take the ratio of the simulated variance and the asymptotic variance, $\bar{K}_{cc}^{(m)}/K_{cc}^{(m)}$. A value close to unity therefore indicates that the asymptotic variance formula is correct. As for the mean, we find that the degree of endogeneity does not matter for the performance of the asymptotic theory.

Figure 1.4.3 indicates that the asymptotic variance formula is correct because the ratio is close to 1 for large sample sizes across all settings. In general, the approximation seems to hold well for sample sizes 10,000 and larger. However, unlike the asymptotic mean, the performance in finite samples is considerably worse except for Robustified 2SLS and $m = 0$. For Robustified 2SLS, the approximation is again better for the tighter significance level of $\gamma_c = 0.01$.

Figure 1.4.3: Evaluation of the Asymptotic Variance of the FODR



Comparison of the theoretical asymptotic variance of the FODR with the simulated variance of the FODR. The y-axis shows the ratio of simulated to asymptotic variance such that a value of 1 means that they coincide. Several settings are varied: the algorithm (initial estimator), the iteration step m, the degree of endogeneity Ω_2 in the DGP, the cut-off c (or equally γ_c), and the sample size n. All Monte Carlo simulations used $W = 50{,}000$ replications.

The results are more erratic for the Saturated 2SLS Algorithm 1.S. There seems to be an upward finite sample bias for lower iterations of the algorithm. At the same time, there seems to exist a downward finite sample bias that is larger for looser significance levels. These biases seem to cancel to some degree, such that for low iterations, a looser significance level yields a variance of the FODR closer to the asymptotic variance than a stricter significance level. For higher iterations,

however, it is the other way around and tighter significance levels perform closer to the asymptotic theory. Since these biases disappear as the sample size grows, the choice of significance level does not matter for large values of n.

Table 1.4.1: Evaluation of the Asymptotic Covariance of the FODR

	sample size n				
iteration m	100	1000	10,000	100,000	1M
0	1.152	1.038	1.068	0.903	1.033
1	0.751	0.902	0.994	0.961	1.003
5	0.59	0.831	0.961	0.961	0.996
convergence	0.581	0.824	0.943	0.991	0.999

Comparison of the theoretical asymptotic covariance of the FODR with the simulated covariance of the FODR across the pair $(\gamma_s, \gamma_t) = (0.05, 0.01)$ for the DGP with $\Omega_2 = 3/4$ and the Robustified 2SLS Algorithm $1.R$. All Monte Carlo simulations used $W = 50{,}000$ replications.

Finally, we analyse the covariance result of the Gaussian process for the Robustified 2SLS Algorithm $1.R$ by taking the ratio of the simulated covariance $\bar{K}_{st}^{(m)}$ and the theoretical covariance $K_{st}^{(m)}$, where $\bar{K}_{st}^{(m)} = n \times \sum_{w=1}^{W} (\widehat{\gamma}_{s,w}^{(m)} - \bar{\widehat{\gamma}}_{s}^{(m)})(\widehat{\gamma}_{t,w}^{(m)} - \bar{\widehat{\gamma}}_{t}^{(m)})/(W-1)$. The covariance is calculated across the pair $(\gamma_s, \gamma_t) = (0.05, 0.01)$ or $(s,t) = (1.96, 2.58)$.

Table 1.4.1 shows a similar pattern as for the variance: the finite sample performance is worse for higher iterations but is already close to the Gaussian approximation for low iterations and relatively small sample sizes of $n = 1000$ and larger. For all iterations, the simulated covariance becomes close to the theoretical covariance as the sample size increases, supporting the theoretical results.

1.4.2 Evaluation of the Poisson Approximation

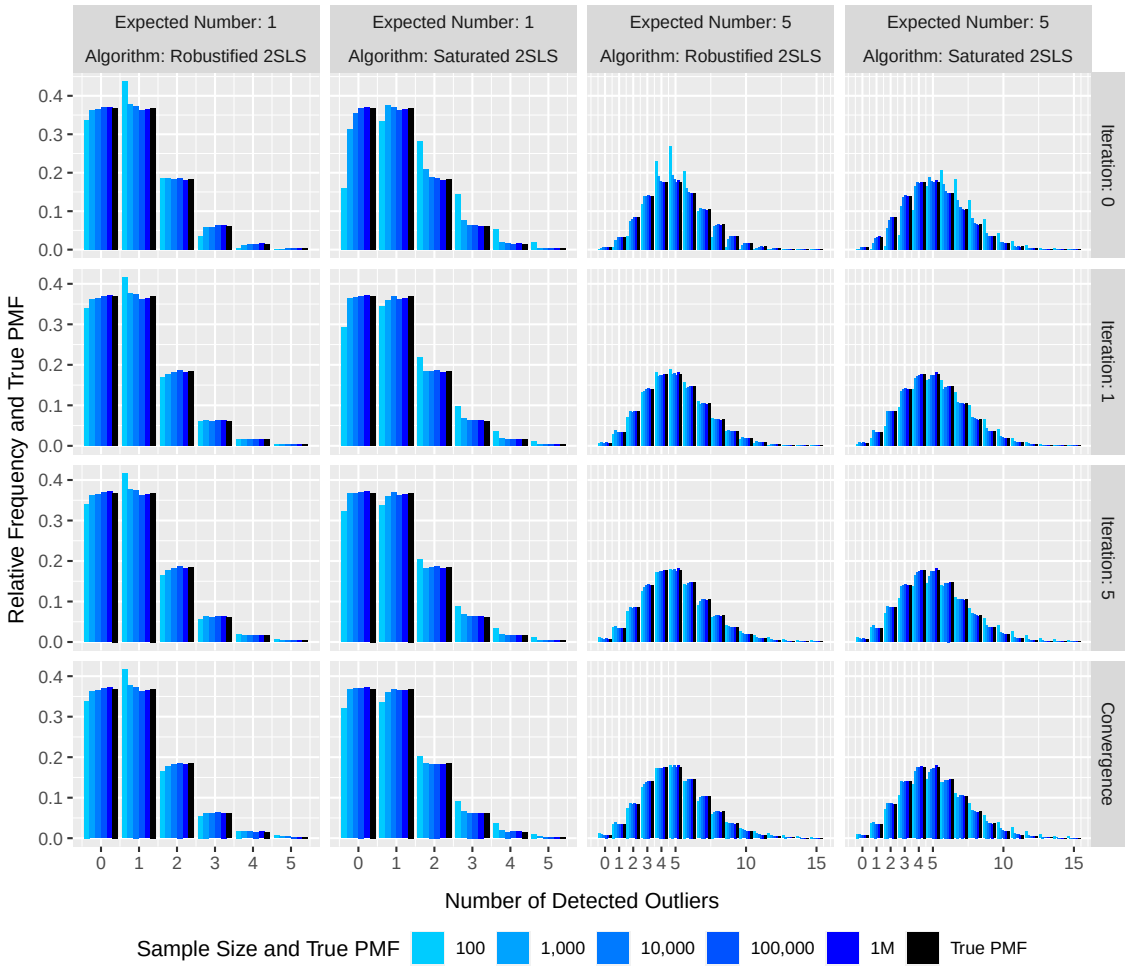
We now analyse the Poisson approximation for the expected number of detected outliers. For simplicity, we focus on the DGP with the highest degree of endogeneity, $\Omega_2 = 3/4$, but vary the initial estimator, the number of iterations, and the sample size as before. Theorem 1.3.1 of the Poisson approximation fixes the expected number of detected outliers instead of the share. In the simulations, we choose the expected number of detected outliers as $\lambda \in \{1, 5\}$. Keeping the expected num-

ber of detected outliers constant requires the cut-off to progressively increase as the sample size increases, see Table 1.4.2.

Table 1.4.2: Cut-off Values for Poisson Approximation

	sample size n				
expected value λ	100	1000	10,000	100,000	1M
1	2.58	3.29	3.89	4.42	4.89
5	1.96	2.81	3.48	4.06	4.56

Figure 1.4.4: Evaluation of the Distribution of the Number of Detected Outliers



Comparison of the theoretical distribution of the number of detected outliers (black) with the simulated distributions (different shades of blue). The y-axis shows the true probability mass of the Poisson distribution and the simulated relative frequencies, respectively. Several settings are varied: the algorithm (initial estimator), the iteration step m, the expected number of detected outliers λ, and the sample size n. All Monte Carlo simulations used $W = 50{,}000$ replications.

Figure 1.4.4 displays the relative frequencies of the number of detected outliers across various simulations and the probability mass function (PMF) of the Poisson

distribution. The simulated frequencies are calculated as $\bar{\mathsf{f}}(x)^{(m)}_{\lambda_{c_n}} = \sum_{s=1}^{S} 1_{(n\widehat{\gamma}^{(m)}_{c,s}=x)}/S$ for $x = 0, 1, 2, \ldots$. For $\lambda = 1$ and $\lambda = 5$, we do not show any relative frequencies for outcomes larger than 5 or 15, respectively. The sum of the probabilities of such larger outcomes is less than 1% in each case, both for the simulated frequencies and the true probabilities.

The simulations confirm the Poisson theory with the relative frequencies for large sample sizes being virtually the same as the true probability mass function. The finite sample frequencies are already close to the theoretical probabilities for sample sizes of $n = 1000$ and larger. For $n = 100$, the approximation performs better when the algorithms are iterated, especially for $\lambda = 5$. Overall, Figure 1.4.4 strongly supports Theorem 1.3.1 and shows good finite sample performance.

1.5 Conclusion

Outliers and their influence on parameter estimates are a frequent concern in empirical research. To confirm that empirical results are not driven by a small set of outliers, applied researchers often use a heuristic procedure that classifies observations with standardised residuals beyond a certain cut-off value as outliers and re-estimates the model on the subset of non-outlying observations.

We first analyse the common practice in the empirical literature of starting the procedure with full-sample estimates, which we term Robustified 2SLS (Algorithm 1.R). The drawback of this method is that the full-sample estimates are not robust to outliers themselves. We therefore suggest a more robust procedure called Saturated 2SLS (Algorithm 1.S), which partitions the whole sample into two subsamples and uses the estimates from one subsample to classify observations as outliers in the other subsample. Furthermore, we suggest iterating the procedures (Algorithm 1.I) in order to achieve higher robustness, which mimics the iterated one-step Huber-skip M-estimator in the instrumental variables regression setting.

Even when the true DGP has no outliers, these procedures have a positive probability of falsely classifying observations as outliers. We analyse this false outlier

detection rate (FODR), prove its tightness, and show that the bias correction implemented in all our suggested procedures is required to properly control the FODR. With a naive, biased procedure, the researcher would instead be unable to properly control the FODR by detecting too many outliers. In order to quantify the uncertainty around the FODR, we derive a Gaussian approximation for the expected share of falsely classified outliers and a Poisson approximation for the expected number of falsely classified outliers. Overall, our asymptotic theory can serve as guidance for applied researchers to control the expected share or number of falsely classified outliers by selecting an appropriate cut-off value. Extensive Monte Carlo simulations support the asymptotic theory and show their finite sample performance. The algorithms and the results of this chapter are readily available in the open-source R package *robust2sls* (Kurle 2022b).

Chapter 2 applies the asymptotic theory of the FODR derived in this chapter to devise statistical tests for the presence of outliers in instrumental variables regressions. Such tests address the robustness-efficiency trade-off and help to evaluate whether the gain in robustness exceeds the corresponding loss in efficiency due to losing parts of the data when applying the algorithms analysed in this chapter. In addition, the theory in the present chapter is restricted to the iid setting but could be extended to time series and panel data in future research.

Appendices of Chapter 1

1.A LLN and CLT for Empirical Processes

Appendix 1.A provides some asymptotic results for empirical processes along with several lemmas, which are frequently used in the proofs of the Gaussian and Poisson approximation theory of the FODR in Section 1.3. Appendix 1.B then focuses on the iterated estimators of (β, σ^2) and their asymptotic properties. The iterated estimators appear in Algorithm 1.I, which nests Algorithms 1.R and 1.S when starting with the full-sample or split-sample estimates, respectively. Since the estimators of (β, σ^2) in iteration m determine the classification of outliers in the next step $m+1$, the results shown in Appendix 1.B are sufficient to demonstrate the main theory for the FODR. The proofs of the asymptotics of the FODR are presented in Appendix 1.C.

The whole argument of the chapter requires an empirical process theory recently developed by Jiao (2022). The first lemma provided below shows first-order asymptotic expansions of weighted and marked empirical processes in IVs regressions, which form the basis for analysing the FODR.

Lemma 1.A.1. *(Jiao (2022), Lemma C.1) Suppose Assumption 1.3.1(ia, iia, iii, iv) holds. Let w_{in} be any of $1, n^{1/2}z_{in}, nz_{in}z_{in}'$ and p any of $0, 1, 2$. Then, we have an expansion*

$$n^{-1/2}\sum_{i=1}^{n} w_{in}u_i^p 1_{(|u_i - z_{in}'\Pi b - n^{-1/2}r_i'b| \le \sigma c + n^{-1/2}ac)} = n^{-1/2}\sum_{i=1}^{n} w_{in}u_i^p 1_{(|u_i|\le \sigma c)} + \mathcal{B}_n(a,b,c) + R(a,b,c),$$

where the bias term is expressed as

$$\mathcal{B}_n(a,b,c) = \sigma^{p-1}c^p \mathsf{f}_u(c) n^{-1/2}\sum_{i=1}^{n} w_{in}[\{1+(-1)^p\}n^{-1/2}ac + n^{-1/2}\{\xi_c - (-1)^p\xi_{-c}\}'\Sigma^{1/2}b + \{1-(-1)^p\}z_{in}'\Pi b].$$

For any $B > 0$ and as $n \to \infty$, the remainder term satisfies

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

In addition, the normalised process $n^{-1/2}\sum_{i=1}^n w_{in}(u_i^p 1_{(|u_i|\leq\sigma c)} - \mathsf{E}_{i-1}u_i^p 1_{(|u_i|\leq\sigma c)})$ is tight in $c \in \mathbb{R}_+$, where $\mathsf{E}_{i-1}(\cdot) = \mathsf{E}(\cdot|\mathcal{F}_{i-1})$ and $\mathsf{E}_{i-1}u_i^p 1_{(|u_i|\leq\sigma c)} = \mathsf{E}u_i^p 1_{(|u_i|\leq\sigma c)} = \sigma^p \tau_p^c$.

The lemma above provides $n^{1/2}$-convergence results for a class of weighted and marked empirical processes. Recall that $\xi_c = \mathsf{E}(\Sigma^{-1/2}r_i|u_i/\sigma = c)$, which appears in the bias term $\mathcal{B}_n(a,b,c)$, is a measure for the level of endogeneity in the regression. The next corollary to Lemma 1.A.1 presents the implied n-consistency results for the same class of empirical processes.

Corollary 1.A.1. *Suppose Assumption 1.3.1(ia, iia, iii, iv) holds. Let w_{in} be any of 1, $n^{1/2}z_{in}$, $nz_{in}z_{in}'$ and p any of $0, 1, 2$. Then, we have an expansion*

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

For any $B > 0$ and as $n \to \infty$, the remainder term satisfies

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

Proof of Corollary 1.A.1. Use Assumption 1.3.1(ia, iia, iii, iv) to apply Lemma 1.A.1 and divide its empirical processes by the rate $n^{-1/2}$ such that

$$n^{-1}\sum_{i=1}^n w_{in}u_i^p 1_{(|u_i - z_{in}'\Pi b - n^{-1/2}r_i'b|\leq\sigma c + n^{-1/2}ac)} = n^{-1}\sum_{i=1}^n w_{in}u_i^p 1_{(|u_i|\leq\sigma c)} + n^{-1/2}\mathcal{B}_n(a,b,c) + n^{-1/2}R(a,b,c).$$

Consider the cases where w_{in} is either of 1, $n^{1/2}z_{in}$, or $nz_{in}z_{in}'$ and p either of 0, 1, or 2. We then find $\sup_{0<c<\infty}\sup_{|a|,|b|\leq n^{1/4-\eta}B}|\mathcal{B}_n(a,b,c)| = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$ such that $n^{-1/2}\mathcal{B}_n(a,b,c)$ is $\mathrm{O}_{\mathsf{P}}(n^{-1/4-\eta})$, since $\sup_{0<c<\infty}|c^3\mathsf{f}_u(c)| < \infty$ implied by Assumption 1.3.1(ia), $\sup_{0<c<\infty}|(|\xi_c| + |\xi_{-c}|)c^2\mathsf{f}_u(c)| < \infty$ by Assumption 1.3.1$(iiib)$, and $n^{-1}\sum_{i=1}^n z_i$ and $n^{-1}\sum_{i=1}^n z_i z_i'$ are $\mathrm{O}_{\mathsf{P}}(1)$ by the LLN and Assumption 1.3.1(iva, ivc). Next, Lemma 1.A.1 shows $\sup_{0<c<\infty}\sup_{|a|,|b|\leq n^{1/4-\eta}B}|n^{-1/2}R(a,b,c)| = \mathsf{o}_{\mathsf{P}}(n^{-1/2})$. Thus, the last two terms as remainder terms are $\mathsf{o}_{\mathsf{P}}(1)$, and combining the three terms ends the proof. ■

For our derivations of the asymptotics of the FODR, we also need n-consistency results for a different class of weighted and marked empirical processes not covered by Corollary 1.A.1, where combinations of weights and marks can be $n^{1/2}z_{in}r_i'$, r_ir_i', and r_iu_i. The next lemma demonstrates n-consistency results for these three empirical processes.

Lemma 1.A.2. *Suppose Assumption 1.3.1(ia, iia, iii, ivb, ivc) holds. Let combinations of weights and marks $w_{in}m_i$ be any of $n^{1/2}z_{in}r_i'$, r_ir_i', or r_iu_i. Then, we have an expansion*

$$n^{-1}\sum_{i=1}^{n} w_{in}m_i 1_{(|u_i - z_{in}'\Pi b - n^{-1/2}r_i'b| \leq \sigma c + n^{-1/2}ac)} = n^{-1}\sum_{i=1}^{n} w_{in}m_i 1_{(|u_i|\leq \sigma c)} + R^{w,m}(a,b,c).$$

For any $B > 0$ and as $n \to \infty$, the remainder term satisfies

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

Proof of Lemma 1.A.2. Let $g_i^{a,b,c} = 1_{(|u_i - z_{in}'\Pi b - n^{-1/2}r_i'b| \leq \sigma c + n^{-1/2}ac)} - 1_{(|u_i|\leq \sigma c)}$ such that $R^{w,m}(a,b,c) = n^{-1}\sum_{i=1}^{n} w_{in}m_i g_i^{a,b,c}$, then the aim is to show $\mathcal{R}^{w,m} = \mathrm{o_P}(1)$, where $\mathcal{R}^{w,m} = \sup_{0<c<\infty}\sup_{|a|,|b|\leq n^{1/4-\eta}B}|R^{w,m}(a,b,c)|$. Using the triangle inequality and consistency property of the matrix norm, we have the following inequality $|R^{w,m}(a,b,c)| \leq n^{-1}\sum_{i=1}^{n}|w_{in}||m_i||g_i^{a,b,c}|$. Next, Hölder inequality gives

$$|R^{w,m}(a,b,c)| \leq (n^{-1}\sum_{i=1}^{n}|w_{in}|^2|m_i|^2)^{1/2}(n^{-1}\sum_{i=1}^{n}|g_i^{a,b,c}|^2)^{1/2}.$$

Notice that $|g_i^{a,b,c}|^2 = |g_i^{a,b,c}|$, and let $\mathcal{R}^g = \sup_{0<c<\infty}\sup_{|a|,|b|\leq n^{1/4-\eta}B} R^g(a,b,c)$, where $R^g(a,b,c) = n^{-1}\sum_{i=1}^{n}|g_i^{a,b,c}|$, thus $\mathcal{R}^{w,m} \leq (n^{-1}\sum_{i=1}^{n}|w_{in}|^2|m_i|^2)^{1/2}(\mathcal{R}^g)^{1/2}$. Apply Theorem B.6, together with the argument used in Theorem 5.5, in Jiao (2022) to get $\mathcal{R}^g = \mathrm{O_P}(n^{-1/4-\eta})$ such that $(\mathcal{R}^g)^{1/2} = \mathrm{O_P}(n^{-1/8-\eta/2}) = \mathrm{o_P}(1)$, which requires Assumption 1.3.1(ia, iia, iii, ivb, ivc). Finally, it suffices to demonstrate the condition $n^{-1}\sum_{i=1}^{n}|w_{in}|^2|m_i|^2 = \mathrm{O_P}(1)$. Consider the cases where w_{in} is any of 1 or $n^{1/2}z_{in}$ and m_i any of r_i', r_ir_i' or r_iu_i. When $w_{in} = n^{1/2}z_{in}$ and $m_i = r_i'$, the boundedness can be shown by the LLN, independence of z_i and r_i, and Assumption 1.3.1(iia, ivc) that $\mathsf{E}|n^{1/2}z_{in}|^2, \mathsf{E}|r_i|^2 < \infty$. When $w_{in} = 1$ and $m_i = r_ir_i'$, we further require $\mathsf{E}|r_i|^4 < \infty$ by Assumption 1.3.1(iia). When $w_{in} = 1$ and $m_i =$

$r_i u_i$, Cauchy-Schwarz inequality shows $\mathsf{E}|r_i u_i|^2 = \mathsf{E}|r_i|^2|u_i|^2 \leq (\mathsf{E}|r_i|^4)^{1/2}(\mathsf{E}|u_i|^4)^{1/2}$ so that $\mathsf{E}|u_i|^4 < \infty$ is further needed and given by Assumption 1.3.1(ia). ■

Furthermore, we present an application of the CLT for the normalised first terms in the expansions of the empirical processes in Lemma 1.A.1. These terms appear in the expansion of the FODR, so their limiting distribution will be used to derive the Gaussian approximation results.

Lemma 1.A.3. *Suppose Assumption 1.3.1(ia, iva, ivc) holds. Then, as $n \to \infty$ and for any $c \in [c_+, \infty)$*

$$n^{-1/2}\sum_{i=1}^{n}\begin{Bmatrix} 1_{(|u_i|>\sigma c)} - \gamma_c \\ \frac{u_i^2}{\sigma^2} - 1 \\ (\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)} \end{Bmatrix} \xrightarrow{\mathsf{D}} \mathsf{N}\left[\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{Bmatrix} \gamma_c(1-\gamma_c) & 1-\gamma_c-\tau_2^c & 0 \\ 1-\gamma_c-\tau_2^c & \tau_4-1 & \tau_4^c - \frac{(\tau_2^c)^2}{1-\gamma_c} \\ 0 & \tau_4^c - \frac{(\tau_2^c)^2}{1-\gamma_c} & \tau_4^c - \frac{(\tau_2^c)^2}{1-\gamma_c} \end{Bmatrix}\right].$$

In addition, we have another vector which is asymptotically independent of the above also converging to a normal distribution

$$\sum_{i=1}^{n}\begin{pmatrix} \tilde{x}_{in}u_i \\ \tilde{x}_{in}u_i 1_{(|u_i|\leq\sigma c)} \end{pmatrix} \xrightarrow{\mathsf{D}} \mathsf{N}\left\{\begin{pmatrix}0_{d_x}\\0_{d_x}\end{pmatrix}, \sigma^2\tau_2^c\begin{pmatrix} \frac{1}{\tau_2^c}M_{\tilde{x}\tilde{x}} & M_{\tilde{x}\tilde{x}} \\ M_{\tilde{x}\tilde{x}} & M_{\tilde{x}\tilde{x}} \end{pmatrix}\right\}.$$

Proof of Lemma 1.A.3. We initially establish the limiting distribution of the first kernel vector. We can see that $\mathsf{E}(1_{(|u_i|>\sigma c)} - \gamma_c) = 0$ and $\mathsf{E}(u_i^2/\sigma^2 - 1) = 0$. Notice that $\varsigma_c^2 = \tau_2^c/\psi_c$, so we also have $\mathsf{E}(u_i^2/\sigma^2 - \varsigma_c^2)1_{(|u_i|\leq\sigma c)} = \tau_2^c - \varsigma_c^2\psi_c = 0$. The variances are $\mathsf{Var}(1_{(|u_i|>\sigma c)} - \gamma_c) = \mathsf{Var}(1_{(|u_i|>\sigma c)}) = \gamma_c(1-\gamma_c)$ and $\mathsf{Var}(u_i^2/\sigma^2 - 1) = \mathsf{Var}(u_i^2/\sigma^2) = \tau_4 - 1$. Furthermore, we have

$$\mathsf{Var}\{(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)}\} = \mathsf{E}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)^2 1_{(|u_i|\leq\sigma c)} = \mathsf{E}\{\frac{u_i^4}{\sigma^4} - 2\varsigma_c^2\frac{u_i^2}{\sigma^2} + (\varsigma_c^2)^2\}1_{(|u_i|\leq\sigma c)}.$$

With $\psi_c = 1-\gamma_c$, $\mathsf{Var}\{(u_i^2/\sigma^2 - \varsigma_c^2)1_{(|u_i|\leq\sigma c)}\} = \tau_4^c - 2\varsigma_c^2\tau_2^c + (\varsigma_c^2)^2\psi_c = \tau_4^c - (\tau_2^c)^2/(1-\gamma_c)$. We next compute the covariance

$$\mathsf{Cov}(1_{(|u_i|>\sigma c)} - \gamma_c, \frac{u_i^2}{\sigma^2} - 1) = \mathsf{Cov}(1_{(|u_i|>\sigma c)}, \frac{u_i^2}{\sigma^2}) = \mathsf{E}\frac{u_i^2}{\sigma^2}1_{(|u_i|>\sigma c)} - \mathsf{E}1_{(|u_i|>\sigma c)}\mathsf{E}\frac{u_i^2}{\sigma^2}.$$

As $\mathsf{E}u_i^2 1_{(|u_i|>\sigma c)}/\sigma^2 = 1 - \tau_2^c$, then $\mathsf{Cov}(1_{(|u_i|>\sigma c)} - \gamma_c, u_i^2/\sigma^2 - 1) = 1 - \gamma_c - \tau_2^c$. In addition,

$$\mathsf{Cov}\{1_{(|u_i|>\sigma c)} - \gamma_c, (\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)}\} = \mathsf{Cov}\{1_{(|u_i|>\sigma c)}, (\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)}\} = 0,$$

where the last equality is due to $1_{(|u_i|>\sigma c)}1_{(|u_i|\leq\sigma c)} = 0$. The last covariance is

$$\mathsf{Cov}\{\frac{u_i^2}{\sigma^2}-1, (\frac{u_i^2}{\sigma^2}-\varsigma_c^2)1_{(|u_i|\leq\sigma c)}\} = \mathsf{Cov}\{\frac{u_i^2}{\sigma^2}, (\frac{u_i^2}{\sigma^2}-\varsigma_c^2)1_{(|u_i|\leq\sigma c)}\} = \mathsf{E}\frac{u_i^2}{\sigma^2}(\frac{u_i^2}{\sigma^2}-\varsigma_c^2)1_{(|u_i|\leq\sigma c)}.$$

Then, $\mathsf{Cov}\{u_i^2/\sigma^2-1, (u_i^2/\sigma^2-\varsigma_c^2)1_{(|u_i|\leq\sigma c)}\} = \tau_4^c - \varsigma_c^2\tau_2^c = \tau_4^c - (\tau_2^c)^2/(1-\gamma_c)$. The limiting distribution of the first kernel vector is thus given by applying the CLT, which requires Assumption 1.3.1(ia) that $\mathsf{E}u_i^4 < \infty$.

The next step is to establish the limiting distribution of the second kernel vector. As z_i and $\tilde{x}_i = \Pi' z_i$ are independent of u_i and f_u is symmetric due to Assumption 1.3.1(ia), it follows that $\mathsf{E}\tilde{x}_i u_i = 0_{d_x}$ and $\mathsf{E}\tilde{x}_i u_i 1_{(|u_i|\leq\sigma c)} = \mathsf{E}\tilde{x}_i \mathsf{E}u_i 1_{(|u_i|\leq\sigma c)} = 0_{d_x}$. In general, the symmetry of f_u implies that all odd moments, truncated or not, are zero. Next, we compute the variance $\mathsf{Var}(\tilde{x}_i u_i) = \mathsf{E}\tilde{x}_i\tilde{x}_i' u_i^2 = \mathsf{E}u_i^2 \mathsf{E}\tilde{x}_i\tilde{x}_i' = \sigma^2 M_{\tilde{x}\tilde{x}}$. Furthermore,

$$\mathsf{Var}(\tilde{x}_i u_i 1_{(|u_i|\leq\sigma c)}) = \mathsf{E}\tilde{x}_i\tilde{x}_i' u_i^2 1_{(|u_i|\leq\sigma c)} = \mathsf{E}u_i^2 1_{(|u_i|\leq\sigma c)} \mathsf{E}\tilde{x}_i\tilde{x}_i' = \sigma^2\tau_2^c M_{\tilde{x}\tilde{x}}.$$

The covariance is $\mathsf{Cov}(\tilde{x}_i u_i, \tilde{x}_i u_i 1_{(|u_i|\leq\sigma c)}) = \mathsf{E}\tilde{x}_i\tilde{x}_i' u_i^2 1_{(|u_i|\leq\sigma c)} = \sigma^2\tau_2^c M_{\tilde{x}\tilde{x}}$. Note that $\tilde{x}_{in} = n^{-1/2}\tilde{x}_i$, so the limiting distribution of the second kernel vector is given by applying the CLT, which further requires Assumption 1.3.1(iva, ivc) that $M_{zz} > 0$ and $\mathsf{E}|z_i|^2 < \infty$.

Finally, we prove independence between the two kernel vectors. As $\tilde{x}_{in} = n^{-1/2}\tilde{x}_i$, we compute the covariance

$$\mathsf{Cov}(1_{(|u_i|>\sigma c)} - \gamma_c, \tilde{x}_i u_i) = \mathsf{Cov}(1_{(|u_i|>\sigma c)}, \tilde{x}_i u_i) = \mathsf{E}\tilde{x}_i u_i 1_{(|u_i|>\sigma c)} = 0_{d_x},$$

where the last equality follows from $\mathsf{E}\tilde{x}_i u_i 1_{(|u_i|>\sigma c)} = \mathsf{E}\tilde{x}_i \mathsf{E}u_i 1_{(|u_i|>\sigma c)}$ as well as $\mathsf{E}u_i 1_{(|u_i|>\sigma c)} = 0$. As $1_{(|u_i|\leq\sigma c)}1_{(|u_i|>\sigma c)} = 0$, the next covariance is

$$\mathsf{Cov}(1_{(|u_i|>\sigma c)} - \gamma_c, \tilde{x}_i u_i 1_{(|u_i|\leq\sigma c)}) = \mathsf{Cov}(1_{(|u_i|>\sigma c)}, \tilde{x}_i u_i 1_{(|u_i|\leq\sigma c)}) = 0_{d_x}.$$

In addition, as $\mathsf{E}u_i^3/\sigma^3 = 0$, we have

$$\mathsf{Cov}(\frac{u_i^2}{\sigma^2} - 1, \tilde{x}_i u_i) = \mathsf{Cov}(\frac{u_i^2}{\sigma^2}, \tilde{x}_i u_i) = \sigma\mathsf{E}\tilde{x}_i \frac{u_i^3}{\sigma^3} = \sigma\mathsf{E}\tilde{x}_i \mathsf{E}\frac{u_i^3}{\sigma^3} = 0_{d_x}.$$

In the same way, since $\mathsf{E}u_i^3 1_{(|u_i|\leq\sigma c)}/\sigma^3 = 0$, it follows that

$$\mathsf{Cov}(\frac{u_i^2}{\sigma^2} - 1, \tilde{x}_i u_i 1_{(|u_i|\leq\sigma c)}) = \mathsf{Cov}(\frac{u_i^2}{\sigma^2}, \tilde{x}_i u_i 1_{(|u_i|\leq\sigma c)}) = \sigma\mathsf{E}\tilde{x}_i \frac{u_i^3}{\sigma^3} 1_{(|u_i|\leq\sigma c)} = 0_{d_x}.$$

In addition, use $\mathsf{E}u_i 1_{(|u_i|\leq\sigma c)}/\sigma = 0$ to show that

$$\mathsf{Cov}\{(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)}, \tilde{x}_i u_i\} = \sigma\mathsf{E}\tilde{x}_i(\frac{u_i^3}{\sigma^3} - \varsigma_c^2\frac{u_i}{\sigma})1_{(|u_i|\leq\sigma c)} = 0_{d_x}.$$

By the same argument as above, the last covariance is

$$\mathsf{Cov}\{(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)}, \tilde{x}_i u_i 1_{(|u_i|\leq\sigma c)}\} = \sigma\mathsf{E}\tilde{x}_i(\frac{u_i^3}{\sigma^3} - \varsigma_c^2\frac{u_i}{\sigma})1_{(|u_i|\leq\sigma c)} = 0_{d_x}.$$

Therefore, the two kernel vectors are uncorrelated and thus independent of each other, since uncorrelatedness implies independence under joint normality. ■

To establish the Gaussian approximation of the FODR, we rely on expansions of the iterated estimators of (β, σ^2), derived in the next Appendix 1.B. We also need to obtain an expansion of $n^{1/2}(\widehat{\sigma} - \sigma)$ from $n^{1/2}(\widehat{\sigma}^2 - \sigma^2)$, for which a variation of the delta method is required. It is presented as the final lemma in the current section.

Lemma 1.A.4. *Let $\{X_n\}$ be a sequence of random variables and θ be a deterministic parameter. Assume that a univariate function g has the first and second derivatives $\dot{g}, \ddot{g}$. Then, we have*

$$n^{1/2}\{g(X_n) - g(\theta)\} = \dot{g}(\theta)n^{1/2}(X_n - \theta) + \frac{1}{2}n^{-1/2}\ddot{g}(\bar{\theta})\{n^{1/2}(X_n - \theta)\}^2,$$

where $|\bar{\theta} - \theta| \leq |X_n - \theta|$.

Proof of Lemma 1.A.4. Approximate g around the point θ by a linear function using the Taylor expansion and evaluate the approximation at the point X_n, then

$$g(X_n) = g(\theta) + \dot{g}(\theta)(X_n - \theta) + \frac{1}{2}\ddot{g}(\bar{\theta})(X_n - \theta)^2,$$

where $|\bar{\theta} - \theta| \leq |X_n - \theta|$. Rearranging the above immediately gives the expansion shown in the lemma. ■

1.B Asymptotics of Iterated Estimators

The analysis of the FODR relies on the analysis of the (β, σ^2)-estimators. Therefore, we now investigate the iterated estimators of (β, σ^2) produced by Algorithm 1.*I*. The

statistical properties of the β-estimator have been established in Jiao (2022), so the current section focuses on demonstrating the asymptotics of the σ-estimator.

The section is comprised of three lemmas that show stochastic expansions of the iterated estimators of (β, σ^2). Algorithm 1.R starts with the full-sample 2SLS estimates, so the first lemma provides an expansion of the full-sample 2SLS estimators $(\widetilde{\beta}, \widetilde{\sigma})$. Then, to investigate the iterative procedure of Algorithm 1.I, we build up a one-step expansion of the updated estimators in step $m+1$ in terms of the ones in the previous iterated step, m. With these two lemmas, we can establish the asymptotic properties of the (β, σ^2)-estimators in any iteration step. The final lemma explores the split-half version of Algorithm 1.S and demonstrates its asymptotic equivalence to Algorithm 1.R.

Lemma 1.B.1. *Consider the full-sample two-stage least squares estimators*

$$\widetilde{\beta} = (\widetilde{\Pi}' \sum_{i=1}^{n} z_i z_i' \widetilde{\Pi})^{-1} (\widetilde{\Pi}' \sum_{i=1}^{n} z_i y_i), \quad \textit{where} \quad \widetilde{\Pi} = (\sum_{i=1}^{n} z_i z_i')^{-1} (\sum_{i=1}^{n} z_i x_i'),$$
$$\widetilde{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (y_i - x_i' \widetilde{\beta})^2.$$

Suppose Assumption 1.3.1(ia, iia, iva, ivc) *holds. Then, as* $n \to \infty$

$$n^{1/2}(\widetilde{\beta} - \beta) = M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + \mathrm{O}_{\mathsf{P}}(n^{-1/2}),$$
$$n^{1/2}(\widetilde{\sigma} - \sigma) = \frac{\sigma}{2} n^{-1/2} \sum_{i=1}^{n} (\frac{u_i^2}{\sigma^2} - 1) - \Omega' \Sigma^{1/2} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + \mathrm{O}_{\mathsf{P}}(n^{-1/2}).$$

Proof of Lemma 1.B.1. We initially show $\widetilde{\Pi} = \Pi + \mathrm{O}_{\mathsf{P}}(n^{-1/2})$. Substitute the first stage equation (1.2.2) into $\widetilde{\Pi}$ to obtain

$$n^{1/2}(\widetilde{\Pi} - \Pi) = M_{zz,n}^{-1} \sum_{i=1}^{n} z_{in} r_i' = \mathrm{O}_{\mathsf{P}}(1),$$

where the second equality follows by Assumption 1.3.1(iva) that $M_{zz,n} \xrightarrow{\mathsf{P}} M_{zz} > 0$, applying the CLT to $\sum_{i=1}^{n} z_{in} r_i'$, and Slutsky's theorem. The CLT requires $\mathsf{E} z_i r_i' = 0_{d_z \times d_x}$ and moment conditions $\mathsf{E}|r_i|^2, \mathsf{E}|z_i|^2 < \infty$ by Assumption 1.3.1(iia, ivc).

The next step is to establish the expansion of $\widetilde{\beta}$. Combining

$$\widetilde{\Pi}' \sum_{i=1}^{n} z_i z_i' \widetilde{\Pi} = \widetilde{\Pi}' (\sum_{i=1}^{n} z_i z_i') (\sum_{i=1}^{n} z_i z_i')^{-1} (\sum_{i=1}^{n} z_i x_i') = \widetilde{\Pi}' \sum_{i=1}^{n} z_i x_i',$$

and the structural equation (1.2.1), we can rearrange $\widetilde{\beta}$ to obtain

$$n^{1/2}(\widetilde{\beta}-\beta) = (\widetilde{\Pi}' M_{zz,n} \widetilde{\Pi})^{-1} (\widetilde{\Pi}' \sum_{i=1}^{n} z_{in} u_i) = M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + \mathrm{O_P}(n^{-1/2}),$$

where the second equality follows by substituting $\widetilde{\Pi} = \Pi + \mathrm{O_P}(n^{-1/2})$, Assumption 1.3.1$(iva)$ that $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}} > 0$, applying the CLT to $\sum_{i=1}^{n} z_{in} u_i$, and Slutsky's theorem. The CLT requires $\mathsf{E} z_i u_i = 0_{d_z}$ and moment conditions $\mathsf{E}|u_i|^2, \mathsf{E}|z_i|^2 < \infty$ by Assumption 1.3.1(ia, ivc). It immediately follows from the above expansion that $n^{1/2}(\widetilde{\beta}-\beta) = \mathrm{O_P}(1)$.

Finally, we build up the expansion of $\widetilde{\sigma}$. Substitute the structural equation (1.2.1) and rearrange $\widetilde{\sigma}^2$ to obtain

$$\begin{aligned} n^{1/2}(\widetilde{\sigma}^2 - \sigma^2) = \sigma^2 n^{-1/2} \sum_{i=1}^{n} (\frac{u_i^2}{\sigma^2} - 1) - 2n^{-1/2} (n^{-1/2} \sum_{i=1}^{n} x_i u_i)' n^{1/2} (\widetilde{\beta} - \beta) \\ + n^{-1/2} n^{1/2} (\widetilde{\beta} - \beta)' (n^{-1} \sum_{i=1}^{n} x_i x_i') n^{1/2} (\widetilde{\beta} - \beta). \end{aligned}$$

The first term above, $\mathcal{T}_1$, is already as required in the expansion, so our focus is on the second and third terms $\mathcal{T}_2, \mathcal{T}_3$.

We first analyse the second term $\mathcal{T}_2$. Substitute the first stage equation (1.2.2) and rearrange the term to get

$$\mathcal{T}_2 = -2n^{-1/2} \{ \sum_{i=1}^{n} \tilde{x}_{in} u_i + n^{-1/2} \sum_{i=1}^{n} (r_i u_i - \mathsf{E} r_i u_i) + n^{1/2} \mathsf{E} r_i u_i \}' n^{1/2} (\widetilde{\beta} - \beta).$$

The CLT shows $\sum_{i=1}^{n} \tilde{x}_{in} u_i = \mathrm{O_P}(1)$ and $n^{-1/2} \sum_{i=1}^{n} (r_i u_i - \mathsf{E} r_i u_i) = \mathrm{O_P}(1)$, which requires $\mathsf{E}|r_i u_i|^2 < \infty$ and which is implied by the Cauchy–Schwarz inequality $\mathsf{E}|r_i u_i|^2 \leq (\mathsf{E}|r_i|^4 \mathsf{E}|u_i|^4)^{1/2}$ and Assumption 1.3.1(ia, iia) that $\mathsf{E}|u_i|^4, \mathsf{E}|r_i|^4 < \infty$. Together with $n^{1/2}(\widetilde{\beta} - \beta) = \mathrm{O_P}(1)$,

$$\mathcal{T}_2 = -2(\mathsf{E} r_i u_i)' n^{1/2} (\widetilde{\beta} - \beta) + \mathrm{O_P}(n^{-1/2}).$$

As $\mathsf{E}(\Sigma^{-1/2} r_i u_i / \sigma) = \Omega$, we have $\mathsf{E} r_i u_i = \sigma \Sigma^{1/2} \Omega$. Notice $\Sigma^{1/2} \Sigma^{-1/2} = I_{d_x}^*$ and $I_{d_x}^* r_i = r_i$ even though $I_{d_x}^*$ is not the identity matrix. Substituting the above $\mathsf{E} r_i u_i$ and the expansion of $n^{1/2}(\widetilde{\beta} - \beta)$, we obtain the second term in the expansion of $n^{1/2}(\widetilde{\sigma}^2 - \sigma^2)$.

The last step is to demonstrate

$$\mathcal{T}_3 = n^{-1/2}n^{1/2}(\widetilde{\beta}-\beta)'(n^{-1}\sum_{i=1}^{n}x_ix_i')n^{1/2}(\widetilde{\beta}-\beta) = \mathrm{O_P}(n^{-1/2}).$$

Plugging in the first stage equation (1.2.2), it follows that

$$n^{-1}\sum_{i=1}^{n}x_ix_i' = M_{\tilde{x}\tilde{x},n} + n^{-1/2}\sum_{i=1}^{n}\tilde{x}_{in}r_i' + n^{-1/2}\sum_{i=1}^{n}r_i\tilde{x}_{in}' + n^{-1}\sum_{i=1}^{n}r_ir_i'.$$

The LLN can demonstrate that $n^{-1/2}\sum_{i=1}^{n}\tilde{x}_{in}r_i' = \mathrm{o_P}(1)$, $n^{-1/2}\sum_{i=1}^{n}r_i\tilde{x}_{in}' = \mathrm{o_P}(1)$, and $n^{-1}\sum_{i=1}^{n}r_ir_i' = \Sigma + \mathrm{o_P}(1)$, which requires $\mathsf{E}z_ir_i' = 0_{d_z\times d_x}$, $\mathsf{E}r_iz_i' = 0_{d_x\times d_z}$, $\mathsf{E}r_ir_i' = \Sigma \geq 0$, and $\mathsf{E}|r_i|^2, \mathsf{E}|z_i| < \infty$, all implied by Assumption 1.3.1(iia, ivc). With Assumption 1.3.1(iva) that $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}} > 0$, we have $n^{-1}\sum_{i=1}^{n}x_ix_i' = M_{\tilde{x}\tilde{x}} + \Sigma + \mathrm{o_P}(1) = \mathrm{O_P}(1)$. Combine this with $n^{1/2}(\widetilde{\beta}-\beta) = \mathrm{O_P}(1)$ to achieve $\mathcal{T}_3 = \mathrm{O_P}(n^{-1/2})$.

Hence, we have $n^{1/2}(\widetilde{\sigma}^2-\sigma^2) = \mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3$, where $\mathcal{T}_1 = \sigma^2n^{-1/2}\sum_{i=1}^{n}(u_i^2/\sigma^2-1)$, $\mathcal{T}_2 = -2\sigma\Omega'\Sigma^{1/2}M_{\tilde{x}\tilde{x},n}^{-1}\sum_{i=1}^{n}\tilde{x}_{in}u_i + \mathrm{O_P}(n^{-1/2})$, and $\mathcal{T}_3 = \mathrm{O_P}(n^{-1/2})$. It immediately follows that $n^{1/2}(\widetilde{\sigma}^2-\sigma^2) = \mathrm{O_P}(1)$ from its expansion. To attain the expansion of $\widetilde{\sigma}$ from $\widetilde{\sigma}^2$, write $n^{1/2}(\widetilde{\sigma}-\sigma) = n^{1/2}\{(\widetilde{\sigma}^2)^{1/2}-(\sigma^2)^{1/2}\}$ and let $g(x) = x^{1/2}$, $X_n = \widetilde{\sigma}^2$, $\theta = \sigma^2$. Then, apply Lemma 1.A.4 to obtain

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

Note that $\dot{g}(x) = x^{-1/2}/2$ and $\ddot{g}(x) = -x^{-3/2}/4$. Since $n^{1/2}(\widetilde{\sigma}^2-\sigma^2) = \mathrm{O_P}(1)$, it follows that $n^{1/2}(\widetilde{\sigma}-\sigma) = (2\sigma)^{-1}n^{1/2}(\widetilde{\sigma}^2-\sigma^2) + \mathrm{O_P}(n^{-1/2})$. Inserting the expansion of $n^{1/2}(\widetilde{\sigma}^2-\sigma^2)$ ends the proof. ■

We just showed the expansion of the full-sample 2SLS, which is used as the initial estimator of Algorithm 1.R. To understand the iterated estimators of Algorithm 1.R in any iteration step, the next lemma considers Algorithm 1.I and provides a one-step expansion of the updated estimators in step $m+1$ in terms of the previous step m, the kernel terms, and some small probability terms. The indicators (1.2.13) for selecting non-outlying observations can be expressed as

$$v_{i,c}^{(m)} = 1_{(|y_i-x_i'\widehat{\beta}_c^{(m)}|\leq\widehat{\sigma}_c^{(m)}c)} = 1_{(|u_i-z_{in}'\Pi\widehat{b}_c^{(m)}-n^{-1/2}r_i'\widehat{b}_c^{(m)}|\leq\sigma c+n^{-1/2}\widehat{a}_c^{(m)}c)}, \tag{1.B.1}$$

where $\widehat{b}_c^{(m)} = n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$ and $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are the estimation errors of the m-step estimators of (β, σ^2). The weighted and marked empirical processes of Appendix 1.A are involved in the updated estimators of (β, σ^2) (1.2.15, 1.2.16) and the FODR (1.2.22), which will become clear in the following proofs.

Lemma 1.B.2. *Consider Algorithm 1.I. Suppose Assumption 1.3.1(ia, iia, iii, iv) holds, and that $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are $\mathrm{O_P}(1)$ for any $m \in [0, \infty)$. Then, as $n \to \infty$ and uniformly in $c \in [c_+, \infty)$*

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

Proof of Lemma 1.B.2. The proof uses empirical processes theory presented in Lemma 1.A.1, Corollary 1.A.1, and Lemma 1.A.2. The one-step expansion of the updated beta estimator is shown in the Proof of Theorem 3.2 in Jiao (2022).

Here we concentrate on demonstrating the one-step expansion of the updated sigma estimator. Start from the expression (1.2.16) of $(\widehat{\sigma}_c^{(m+1)})^2$ and substitute the structural equation (1.2.1) to obtain

$$
\begin{aligned}
n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\} &= \varsigma_c^{-2} (n^{-1} \sum_{i=1}^n v_{i,c}^{(m)})^{-1} \{\sigma^2 n^{-1/2} \sum_{i=1}^n (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) v_{i,c}^{(m)} && \text{(1.B.2)} \\
&\quad - 2(n^{-1} \sum_{i=1}^n x_i u_i v_{i,c}^{(m)})' n^{1/2} (\widehat{\beta}_c^{(m+1)} - \beta) && \text{(1.B.3)} \\
&\quad + n^{-1/2} n^{1/2} (\widehat{\beta}_c^{(m+1)} - \beta)' (n^{-1} \sum_{i=1}^n x_i x_i' v_{i,c}^{(m)}) n^{1/2} (\widehat{\beta}_c^{(m+1)} - \beta)\}. && \text{(1.B.4)}
\end{aligned}
$$

We first find the limit of the denominator. Note that $\widehat{b}_c^{(m)} = n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$ and $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are the estimation errors of the original m-step beta and sigma estimators so that $v_{i,c}^{(m)} = 1_{(|u_i - z_{in}' \Pi \widehat{b}_c^{(m)} - n^{-1/2} r_i' \widehat{b}_c^{(m)}| \leq \sigma c + n^{-1/2} \widehat{a}_c^{(m)} c)}$, see (1.B.1). By Assumption 1.3.1(ia, iia, iii, iv) and the fact that $|\widehat{b}_c^{(m)}| + |\widehat{a}_c^{(m)}| = \mathrm{O_P}(1)$, we

apply Corollary 1.A.1 to show $n^{-1}\sum_{i=1}^n v_{i,c}^{(m)} = n^{-1}\sum_{i=1}^n 1_{(|u_i|\leq\sigma c)} + \mathsf{o}_\mathsf{P}(1)$, such that we have

$$\varsigma_c^{-2}(n^{-1}\sum_{i=1}^n v_{i,c}^{(m)})^{-1} = (\frac{\tau_2^c}{\psi_c})^{-1}\psi_c^{-1} + \mathsf{o}_\mathsf{P}(1) = \frac{1}{\tau_2^c} + \mathsf{o}_\mathsf{P}(1)$$

uniformly in the cut-offs c. We apply the LLN, the CMT, and the fact that $\psi_c > 0$ as $c > c_+$ to obtain the above, where $c_+ > 0$ is the lower bound of the cut-offs. The next step is to derive the limits of the numerators (1.B.2), (1.B.3), and (1.B.4).

Analysing (1.B.2) requires Lemma 1.A.1, which shows the two expansions

$$n^{-1/2}\sum_{i=1}^n v_{i,c}^{(m)} = n^{-1/2}\sum_{i=1}^n 1_{(|u_i|\leq\sigma c)} + \frac{\mathsf{f}_u(c)}{\sigma}(2c\widehat{a}_c^{(m)} + \zeta_c^{-\prime}\Sigma^{1/2}\widehat{b}_c^{(m)}) + \mathsf{o}_\mathsf{P}(1),$$
$$n^{-1/2}\sum_{i=1}^n u_i^2 v_{i,c}^{(m)} = n^{-1/2}\sum_{i=1}^n u_i^2 1_{(|u_i|\leq\sigma c)} + \sigma c^2\mathsf{f}_u(c)(2c\widehat{a}_c^{(m)} + \zeta_c^{-\prime}\Sigma^{1/2}\widehat{b}_c^{(m)}) + \mathsf{o}_\mathsf{P}(1).$$

Substitute the above in (1.B.2) to obtain

$$\sigma^2 n^{-1/2}\sum_{i=1}^n(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)v_{i,c}^{(m)} = 2\sigma\{c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\widehat{a}_c^{(m)} + \frac{\sigma}{2}n^{-1/2}\sum_{i=1}^n(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)}$$
$$+ \frac{(c^2-\varsigma_c^2)\mathsf{f}_u(c)}{2}\zeta_c^{-\prime}\Sigma^{1/2}\widehat{b}_c^{(m)}\} + \mathsf{o}_\mathsf{P}(1).$$

Uniformity in c further requires $\sup_{c_+\leq c<\infty}|c(c^2-\varsigma_c^2)\mathsf{f}_u(c)| < \infty$ implied by Assumption 1.3.1(ia), $\sup_{c_+\leq c<\infty}|(c^2-\varsigma_c^2)\mathsf{f}_u(c)\zeta_c^-| < \infty$ implied by Assumption 1.3.1($iiib$), and tightness of $n^{-1/2}\sum_{i=1}^n(u_i^2/\sigma^2 - \varsigma_c^2)1_{(|u_i|\leq\sigma c)}$ shown in Lemma 1.A.1.

Substitute the first stage equation (1.2.2) into (1.B.3) so that

$$-2(n^{-1}\sum_{i=1}^n x_i u_i v_{i,c}^{(m)})'\widehat{b}_c^{(m+1)} = -2(\Pi' n^{-1}\sum_{i=1}^n z_i u_i v_{i,c}^{(m)} + n^{-1}\sum_{i=1}^n r_i u_i v_{i,c}^{(m)})'\widehat{b}_c^{(m+1)}.$$

Corollary 1.A.1 shows $n^{-1}\sum_{i=1}^n z_i u_i v_{i,c}^{(m)} = n^{-1}\sum_{i=1}^n z_i u_i 1_{(|u_i|\leq\sigma c)} + \mathsf{o}_\mathsf{P}(1)$ and Lemma 1.A.2 shows $n^{-1}\sum_{i=1}^n r_i u_i v_{i,c}^{(m)} = n^{-1}\sum_{i=1}^n r_i u_i 1_{(|u_i|\leq\sigma c)} + \mathsf{o}_\mathsf{P}(1)$. Use the expansion of the beta estimator $\widehat{b}_c^{(m+1)} = \{2c\mathsf{f}_u(c)/\psi_c\}\widehat{b}_c^{(m)} + (\psi_c M_{\tilde{x}\tilde{x},n})^{-1}\sum_{i=1}^n \tilde{x}_{in}u_i 1_{(|u_i|\leq\sigma c)} + \mathsf{o}_\mathsf{P}(1)$ and apply the LLN and the CMT to get

$$-2(n^{-1}\sum_{i=1}^n x_i u_i v_{i,c}^{(m)})'\widehat{b}_c^{(m+1)} = -\frac{4\sigma c\mathsf{f}_u(c)}{\psi_c}\omega_c'\Sigma^{1/2}\widehat{b}_c^{(m)}$$
$$- \frac{2\sigma}{\psi_c}\omega_c'\Sigma^{1/2}M_{\tilde{x}\tilde{x},n}^{-1}\sum_{i=1}^n \tilde{x}_{in}u_i 1_{(|u_i|\leq\sigma c)} + \mathsf{o}_\mathsf{P}(1).$$

In order to derive the above, we also use $\mathsf{E}z_i u_i 1_{(|u_i|\leq\sigma c)} = \mathsf{E}z_i\mathsf{E}u_i 1_{(|u_i|\leq\sigma c)} = 0_{d_z\times 1}$ and $\mathsf{E}r_i u_i 1_{(|u_i|\leq\sigma c)} = \sigma\Sigma^{1/2}\omega_c$. The uniformity argument in c here requires boundedness

of $\widehat{b}_c^{(m+1)}$ and ω_c. It immediately follows from the beta expansion that $\widehat{b}_c^{(m+1)}$ is $\mathrm{O}_\mathsf{P}(1)$ uniformly in c due to $\sup_{c_+\leq c<\infty}|2c\mathsf{f}_u(c)/\psi_c| < 1$ implied by Assumption 1.3.1(ia), $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}} > 0$ implied by Assumption 1.3.1(iva), and tightness of $\sum_{i=1}^n \tilde{x}_{in}u_i 1_{(|u_i|\leq\sigma c)}$ shown in Lemma 1.A.1. Since $\mathsf{E}(\Sigma^{-1/2}r_iu_i/\sigma) = \Omega$ and ω_c is the truncated covariance, boundedness of ω_c follows directly.

To analyse (1.B.4), substitute the first stage equation (1.2.2) again such that

$$n^{-1/2}\widehat{b}_c^{(m+1)\prime}n^{-1}\sum_{i=1}^n x_ix_i'v_{i,c}^{(m)}\widehat{b}_c^{(m+1)} = n^{-1/2}\widehat{b}_c^{(m+1)\prime}(\Pi' n^{-1}\sum_{i=1}^n z_iz_i'v_{i,c}^{(m)}\Pi + n^{-1}\sum_{i=1}^n r_ir_i'v_{i,c}^{(m)} + \Pi' n^{-1}\sum_{i=1}^n z_ir_i'v_{i,c}^{(m)} + n^{-1}\sum_{i=1}^n r_iz_i'v_{i,c}^{(m)}\Pi)\widehat{b}_c^{(m+1)}.$$

Again, Corollary 1.A.1 shows $n^{-1}\sum_{i=1}^n z_iz_i'v_{i,c}^{(m)} = n^{-1}\sum_{i=1}^n z_iz_i'1_{(|u_i|\leq\sigma c)} + \mathrm{o}_\mathsf{P}(1)$ and Lemma 1.A.2 shows two other expansions $n^{-1}\sum_{i=1}^n r_ir_i'v_{i,c}^{(m)} = n^{-1}\sum_{i=1}^n r_ir_i'1_{(|u_i|\leq\sigma c)} + \mathrm{o}_\mathsf{P}(1)$ and $n^{-1}\sum_{i=1}^n z_ir_i'v_{i,c}^{(m)} = n^{-1}\sum_{i=1}^n z_ir_i'1_{(|u_i|\leq\sigma c)} + \mathrm{o}_\mathsf{P}(1)$. Apply the LLN and explore the three different expected values $\mathsf{E}z_iz_i'1_{(|u_i|\leq\sigma c)} = \mathsf{E}z_iz_i'\mathsf{E}1_{(|u_i|\leq\sigma c)} = \psi_cM_{zz}$, $\mathsf{E}r_ir_i'1_{(|u_i|\leq\sigma c)} \leq \mathsf{E}r_ir_i' = \Sigma$, and $\mathsf{E}z_ir_i'1_{(|u_i|\leq\sigma c)} = \mathsf{E}z_i\mathsf{E}r_i'1_{(|u_i|\leq\sigma c)} = \mathsf{E}z_i0_{1\times d_x} = 0_{d_z\times d_x}$. Thus, all terms in the brackets are bounded in probability. Again, $\widehat{b}_c^{(m+1)}$ is $\mathrm{O}_\mathsf{P}(1)$ so that it holds uniformly in c that

$$n^{-1/2}\widehat{b}_c^{(m+1)\prime}n^{-1}\sum_{i=1}^n x_ix_i'v_{i,c}^{(m)}\widehat{b}_c^{(m+1)} = \mathrm{O}_\mathsf{P}(n^{-1/2}) = \mathrm{o}_\mathsf{P}(1).$$

In summary, combine the three terms (1.B.2), (1.B.3), (1.B.4) in the numerator divided by the denominator to obtain the expansion of $n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\}$. It immediately follows from the established expansion that $n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\}$ is $\mathrm{O}_\mathsf{P}(1)$ uniformly in c. Using Lemma 1.A.4 in the same way as in the proof of Lemma 1.B.1 shows

$$n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = (2\sigma)^{-1}n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\} + \mathrm{O}_\mathsf{P}(n^{-1/2}).$$

Finally, inserting the expansion of $n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\}$ ends the proof. ■

In the proof of Lemma 1.B.2, the following three expected values $\mathsf{E}r_iu_i1_{(|u_i|\leq\sigma c)}$, $\mathsf{E}r_ir_i'1_{(|u_i|\leq\sigma c)}$, and $\mathsf{E}r_i1_{(|u_i|\leq\sigma c)}$ need to be analysed. The remark below explores these three integrals.

Remark 1.B.1. *As shown in (1.2.10) and (1.2.11),*

$$\omega_c = \mathsf{E}\Sigma^{-1/2} r_i \frac{u_i}{\sigma} 1_{(|u_i| \leq \sigma c)} = \int_{x \in \mathbb{R}^{d_x}} \int_{-c}^{c} xy \mathsf{f}_{u,r}(y,x) dy(dx).$$

Let $g_1(x,y) = xy\mathsf{f}_{u,r}(y,x)$*, then* $g_1(-x,-y) = g_1(x,y)$ *if* $\mathsf{f}_{u,r}(-y,-x) = \mathsf{f}_{u,r}(y,x)$*. Thus,* $\omega_c = \int_{x \in \mathbb{R}^{d_x}} \int_{-c}^{c} g_1(x,y) dy(dx) \neq 0_{d_x}$ *as long as* c *is bounded away from* 0*, since the integral of an even function on the symmetric domain is not zero. The Remark 1.2.1 also shows* $\omega_c = \tau_2^c \Omega$ *under normality.*

Next, we analyse

$$\begin{aligned} \mathsf{E}\Sigma^{-1/2} r_i r_i' \Sigma^{-1/2} 1_{(|u_i| \leq \sigma c)} &= \iint_{y \in \mathbb{R}, x \in \mathbb{R}^{d_x}} x x' 1_{(|y| \leq c)} \mathsf{f}_{u,r}(y,x) dy(dx) \\ &= \int_{x \in \mathbb{R}^{d_x}} \int_{-c}^{c} x x' \mathsf{f}_{u,r}(y,x) dy(dx). \end{aligned}$$

Let $g_2(x,y) = xx'\mathsf{f}_{u,r}(y,x)$*, then* g_2 *is even and positive semi-definite so that the above integral is positive definite.*

The third expected value is

$$\mathsf{E}\Sigma^{-1/2} r_i 1_{(|u_i| \leq \sigma c)} = \iint_{y \in \mathbb{R}, x \in \mathbb{R}^{d_x}} x 1_{(|y| \leq c)} \mathsf{f}_{u,r}(y,x) dy(dx) = \int_{x \in \mathbb{R}^{d_x}} \int_{-c}^{c} x \mathsf{f}_{u,r}(y,x) dy(dx).$$

Let $g_3(x,y) = x\mathsf{f}_{u,r}(y,x)$*, then* g_3 *is an odd function as* $g_3(-x,-y) = -g_3(x,y)$*. Thus, the integral* $\mathsf{E}\Sigma^{-1/2} r_i 1_{(|u_i| \leq \sigma c)} = \int_{x \in \mathbb{R}^{d_x}} \int_{-c}^{c} g_3(x,y) dy(dx) = 0_{d_x}$*.*

It can be shown that the asymptotic expansions of the iterated (β, σ^2)-estimators of Algorithm 1.R and the split-half version of Algorithm 1.S coincide. Theorem 3.7 in Jiao (2022) shows asymptotic equality of the β-expansion. In the final lemma of this section, we also show asymptotic equality of the σ-expansion.

Lemma 1.B.3. *Consider the split-half version of Algorithm 1.S, where the sample is partitioned into two halves,* $\mathcal{I}_1$ *and* $\mathcal{I}_2$*, with number of observations* $n_1 = \text{int}[n/2]$ *and* $n_2 = n - n_1$*. Suppose Assumption 1.3.1*(ia, iia, iii, iv) *holds for each half* $\mathcal{I}_1$*,* $\mathcal{I}_2$*. Then, as* $n \to \infty$ *and for any* $m \in [0, \infty)$*, the variance estimator* $n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma)$ *has the same asymptotic expansion as that of Algorithm 1.R, uniformly in* $c \in [c_+, \infty)$*.*

Proof of Lemma 1.B.3. Since Algorithm 1.R and the split-half version of Algorithm 1.S only differ in the choice of their initial estimators but proceed in the

same way thereafter, it is sufficient to show that the asymptotic expansions of their first step updated variance estimators coincide uniformly in the cut-off value c. We first consider Algorithm 1.R and obtain its first step σ-expansion. Lemma 1.B.2 provides the expansion of $\widehat{\sigma}_c^{(1)}$ in terms of $\widehat{\sigma}_c^{(0)}$, $\widehat{\beta}_c^{(0)}$, and the kernel terms. The expansions of the full-sample 2SLS estimators $(\widehat{\beta}_c^{(0)}, \widehat{\sigma}_c^{(0)}) = (\widetilde{\beta}, \widetilde{\sigma})$ are shown in Lemma 1.B.1, such that the expansion of $\widehat{\sigma}_c^{(1)}$ immediately follows by combining these two lemmas, resulting in

$$\begin{aligned} n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma) = {} & \frac{\sigma c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)}{2\tau_2^c} n^{-1/2} \sum_{i=1}^n (\frac{u_i^2}{\sigma^2} - 1) + \frac{\sigma}{2\tau_2^c} n^{-1/2} \sum_{i=1}^n (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i| \leq \sigma c)} \\ & + \frac{\mathsf{f}_u(c)}{\tau_2^c} \{ \frac{c^2 - \varsigma_c^2}{2} \zeta_c^- - \frac{2c}{\psi_c} \omega_c - c(c^2 - \varsigma_c^2)\Omega \}' \Sigma^{1/2} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^n \tilde{x}_{in} u_i \\ & - \frac{1}{\psi_c \tau_2^c} \omega_c' \Sigma^{1/2} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^n \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} + \mathsf{o}_\mathsf{P}(1), \end{aligned}$$

uniformly in $c \in [c_+, \infty)$.

For the remainder of the proof, we use the split-half version of Algorithm 1.S with $n_1 = \text{int}[n/2]$ and $n_2 = n - n_1$. Our goal is to derive its expansion of $n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma)$ and then show its asymptotic equivalence to that of Algorithm 1.R. We start with the definition of $(\widehat{\sigma}_c^{(1)})^2$ in (1.2.16). Substituting the structural equation (1.2.1) and re-arranging yields

$$\begin{aligned} n^{1/2}\{(\widehat{\sigma}_c^{(1)})^2 - \sigma^2\} = {} & \varsigma_c^{-2} (n^{-1} \sum_{i=1}^n v_{i,c}^{(0)})^{-1} \{ \sigma^2 n^{-1/2} \sum_{i=1}^n (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) v_{i,c}^{(0)} \\ & - 2(n^{-1} \sum_{i=1}^n x_i u_i v_{i,c}^{(0)})' n^{1/2} (\widehat{\beta}_c^{(1)} - \beta) \\ & + n^{-1/2} n^{1/2} (\widehat{\beta}_c^{(1)} - \beta)' (n^{-1} \sum_{i=1}^n x_i x_i' v_{i,c}^{(0)}) n^{1/2} (\widehat{\beta}_c^{(1)} - \beta) \}. \end{aligned}$$

Remember that $v_{i,c}^{(0)} = 1_{(i \in \mathcal{I}_1)} 1_{(|y_i - x_i' \widehat{\beta}_2| \leq \widehat{\sigma}_2 c)} + 1_{(i \in \mathcal{I}_2)} 1_{(|y_i - x_i' \widehat{\beta}_1| \leq \widehat{\sigma}_1 c)}$, so we have

$$\sum_{i=1}^n v_{i,c}^{(0)} = \sum_{j=1}^2 \sum_{i \in \mathcal{I}_j} 1_{(|y_i - x_i' \widehat{\beta}_{3-j}| \leq \widehat{\sigma}_{3-j} c)}.$$

Let us first analyse the denominator. Since Assumption 1.3.1(ia, iia, iva, ivc) holds for each subsample $\mathcal{I}_1$ and $\mathcal{I}_2$, we can apply the argument of Lemma 1.B.1 to each subsample and obtain tightness of $\widehat{\beta}_j$ and $\widehat{\sigma}_j^2$ for $j = 1, 2$. Combining this with

Assumption 1.3.1(ia, iia, iii, iv), we can apply Corollary 1.A.1 to show

$$n^{-1}\sum_{i=1}^{n} v_{i,c}^{(0)} = \sum_{j=1}^{2} \frac{n_j}{n} n_j^{-1} \sum_{i\in\mathcal{I}_j} 1_{(|y_i - x_i'\widehat{\beta}_{3-j}|\leq \widehat{\sigma}_{3-j}c)} = \sum_{j=1}^{2} \frac{n_j}{n} n_j^{-1} \sum_{i\in\mathcal{I}_j} 1_{(|u_i|\leq\sigma c)} + \mathsf{o_P}(1),$$

so that $n^{-1}\sum_{i=1}^{n} v_{i,c}^{(0)} = n^{-1}\sum_{i=1}^{n} 1_{(|u_i|\leq\sigma c)} + \mathsf{o_P}(1)$ and the denominator can thus be expressed as $\varsigma_c^{-2}(n^{-1}\sum_{i=1}^{n} v_{i,c}^{(0)})^{-1} = 1/\tau_2^c + \mathsf{o_P}(1)$ uniformly in $c \in [c_+, \infty)$. Next, we analyse the first term in the numerator, to which we apply Lemma 1.A.1 using Assumption 1.3.1(ia, iia, iii, iv) and the tightness of both subsample estimators:

$$\begin{aligned}\sigma^2 n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)v_{i,c}^{(0)} &= \sigma^2 n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)} \\ &+ \sigma(c^2 - \varsigma_c^2)\mathsf{f}_u(c)\sum_{j=1}^{2}(\frac{n_j}{n})^{1/2}(\frac{n_j}{n_{3-j}})^{1/2}\{\zeta_c^{-\prime}\Sigma^{1/2} n_{3-j}^{1/2}(\widehat{\beta}_{3-j} - \beta) \\ &+ 2cn_{3-j}^{1/2}(\widehat{\sigma}_{3-j} - \sigma)\} + \mathsf{o_P}(1).\end{aligned}$$

We replace $n_{3-j}^{1/2}(\widehat{\beta}_{3-j} - \beta)$ and $n_{3-j}^{1/2}(\widehat{\sigma}_{3-j} - \sigma)$ for $j = 1, 2$ by the expansions in Lemma 1.B.1 applied to each subsample. It follows that

$$\begin{aligned}\sigma^2 n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)v_{i,c}^{(0)} &= \sigma^2 n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)} \\ &+ \sigma(c^2 - \varsigma_c^2)\mathsf{f}_u(c)n^{-1/2}\sum_{j=1}^{2}\frac{n_j}{n_{3-j}}\{\sigma c\sum_{i\in\mathcal{I}_{3-j}}(\frac{u_i^2}{\sigma^2} - 1) \\ &+ (\zeta_c^- - 2c\Omega)'\Sigma^{1/2}(M_{\tilde{x}\tilde{x},n_{3-j}}^{\mathcal{I}_{3-j}})^{-1}\sum_{i\in\mathcal{I}_{3-j}}\tilde{x}_i u_i\} + \mathsf{o_P}(1).\end{aligned}$$

Note that as $n \to \infty$ then $n_j/n_{3-j} \to 1$ and $M_{\tilde{x}\tilde{x},n_{3-j}}^{\mathcal{I}_{3-j}} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}}$ for $j = 1, 2$. Thus, the first term in the numerator has the expansion

$$\begin{aligned}\sigma^2 n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)v_{i,c}^{(0)} &= \sigma^2 c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - 1) \\ &+ \sigma^2 n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)} \\ &+ \sigma(c^2 - \varsigma_c^2)\mathsf{f}_u(c)(\zeta_c^- - 2c\Omega)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\sum_{i=1}^{n}\tilde{x}_{in}u_i + \mathsf{o_P}(1).\end{aligned}$$

To show uniformity in c, we also need to apply Assumption 1.3.1$(ia, iiib)$ such that $\sup_{c_+\leq c<\infty}|c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)| < \infty$ and $\sup_{c_+\leq c<\infty}|(c^2 - \varsigma_c^2)\mathsf{f}_u(c)\zeta_c^-| < \infty$. Next, we analyse the second term in the numerator and substitute the first stage equation

$x_i = \Pi' z_i + r_i$ into it to get

$$-2(n^{-1}\sum_{i=1}^{n} x_i u_i v_{i,c}^{(0)})' n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = -2(\Pi' n^{-1}\sum_{i=1}^{n} z_i u_i v_{i,c}^{(0)} + n^{-1}\sum_{i=1}^{n} r_i u_i v_{i,c}^{(0)})' n^{1/2}(\widehat{\beta}_c^{(1)} - \beta).$$

Arguing along the lines of the denominator, we have

$$-2(n^{-1}\sum_{i=1}^{n} x_i u_i v_{i,c}^{(0)})' n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = -2\{\Pi' n^{-1}\sum_{i=1}^{n} z_i u_i 1_{(|u_i|\leq \sigma c)} + n^{-1}\sum_{i=1}^{n} r_i u_i 1_{(|u_i|\leq \sigma c)} + \mathsf{o_P}(1)\}' n^{1/2}(\widehat{\beta}_c^{(1)} - \beta).$$

In addition to Corollary 1.A.1 that deals with the term $z_i u_i$, we also apply Lemma 1.A.2 to handle the empirical process containing the term $r_i u_i$, which requires Assumption 1.3.1(ia, iia, iii, ivb, ivc). Moreover, with Assumption 1.3.1(ia, iia, iii, iv), we use Theorem 3.7 in Jiao (2022), which shows the tightness of $n^{1/2}(\widehat{\beta}_c^{(1)} - \beta)$ and its expansion, to obtain

$$n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = \frac{2c\mathsf{f}_u(c)}{\psi_c} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + (\psi_c M_{\tilde{x}\tilde{x},n})^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i|\leq \sigma c)} + \mathsf{o_P}(1).$$

Moreover, by the LLN and using the two expected values $\mathsf{E} z_i u_i 1_{(|u_i|\leq \sigma c)} = 0_{d_z}$ and $\mathsf{E} r_i u_i 1_{(|u_i|\leq \sigma c)} = \sigma \Sigma^{1/2} \omega_c$, we derive

$$-2(n^{-1}\sum_{i=1}^{n} x_i u_i v_{i,c}^{(0)})' n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = \frac{-4\sigma c\mathsf{f}_u(c)}{\psi_c} \omega_c' \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i - \frac{2\sigma}{\psi_c} \omega_c' \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i|\leq \sigma c)} + \mathsf{o_P}(1).$$

Note that we also need $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}}$ to obtain the above expansion and the whole argument holds uniformly in c. Using a similar argument for the second numerator term and also following the arguments of the proof of Lemma 1.B.2 — particularly the part analysing the third numerator term (1.B.4) of $\widehat{\sigma}_c^{(m+1)}$ — we can then show

$$n^{-1/2} n^{1/2}(\widehat{\beta}_c^{(1)} - \beta)'(n^{-1}\sum_{i=1}^{n} x_i x_i' v_{i,c}^{(0)}) n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) = \mathsf{O_P}(n^{-1/2}) = \mathsf{o_P}(1),$$

uniformly in c. Having analysed the denominator and all three terms in the numerator, we can combine their expressions to get the expansion of $n^{1/2}\{(\widehat{\sigma}_c^{(1)})^2 - \sigma^2\}$.

The final step is to apply Lemma 1.A.4 to derive the expansion of $n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma)$ so that

$$\begin{aligned} n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma) = {} & \frac{\sigma c(c^2 - \varsigma_c^2)\mathsf{f}_u(c)}{2\tau_2^c} n^{-1/2} \sum_{i=1}^{n} (\frac{u_i^2}{\sigma^2} - 1) + \frac{\sigma}{2\tau_2^c} n^{-1/2} \sum_{i=1}^{n} (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i| \leq \sigma c)} \\ & + \frac{\mathsf{f}_u(c)}{\tau_2^c} \{ \frac{c^2 - \varsigma_c^2}{2} \zeta_c^- - \frac{2c}{\psi_c} \omega_c - c(c^2 - \varsigma_c^2)\Omega \}' \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i \\ & - \frac{1}{\psi_c \tau_2^c} \omega_c' \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} + \mathsf{o}_{\mathsf{P}}(1), \end{aligned}$$

uniformly in c. We again used $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}}$ and immediately find the equivalence of the expansions of $n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma)$ between Algorithm 1.R and the split-half version of Algorithm 1.S. ■

1.C Main Results for the FODR and Their Proofs

In the main text, we provide two types of asymptotic approximations: (i) a normal asymptotic theory when the proportion of detected outliers is controlled by choosing a fixed cut-off value c as n increases and (ii) a Poisson exceedance theory when the number of detected outliers is controlled by increasing the cut-off values c_n as n increases.

For the Gaussian approximation, we first show the stochastic expansion and tightness for the FODR of Algorithm 1.I. We then consider Algorithm 1.R and derive the Gaussian weak limits for $n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)$ with drifting cut-off values c for the initial step $m = 0$, the first iteration $m = 1$, and the general iterated step $m \in [1, \infty)$. For Algorithm 1.S, we first derive the expansion of the FODR and derive its Gaussian weak limit for any split — including unequal splits — for the initial step $m = 0$. Then, restricted to the split-half version of Algorithm 1.S, we demonstrate that in this case the FODR has the same Gaussian limit as that of Algorithm 1.R for any iteration step $m \in [0, \infty)$. Then, we return to the iterated versions of the algorithms and show the fixed point of the FODR when the procedure is iterated infinitely, as described in Algorithm 1.I. The final proof establishes the Poisson approximation to the number $n\widehat{\gamma}_c^{(m)}$ of falsely detected outliers at any

iteration $m \in [0, \infty)$ and for all Algorithms $1.I$, $1.R$, and $1.S$.

The first lemma in this section derives the stochastic expansion of the FODR $\widehat{\gamma}_c^{(m)}$ of Algorithm $1.I$ in terms of the m-step estimation errors $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$ and $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ by applying the empirical process theory shown in Lemma 1.A.1.

Lemma 1.C.1. *Consider Algorithm 1.I. Suppose Assumption 1.3.1*(ia, iia, iii, iv) *holds, and that* $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ *are* $\mathrm{O_P}(1)$. *Then, for any* $m \in [0, \infty)$ *and uniformly in* $c \in [c_+, \infty)$ *we have as* $n \to \infty$

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

Proof of Lemma 1.C.1. The definition of the FODR in equation (1.2.22) gives

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

Recall that $\widehat{b}_c^{(m)} = n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$ and $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are the estimation errors for β and σ of algorithm step m, so $v_{i,c}^{(m)} = 1_{(|u_i - z_{in}' \Pi \widehat{b}_c^{(m)} - n^{-1/2} r_i' \widehat{b}_c^{(m)}| \leq \sigma c + n^{-1/2} \widehat{a}_c^{(m)} c)}$, see (1.B.1). By Assumption 1.3.1(ia, iia, iii, iv) and the fact that $|\widehat{b}_c^{(m)}| + |\widehat{a}_c^{(m)}| = \mathrm{O_P}(1)$, we apply the expansion of the empirical processes to the FODR, where we choose $w_{in} = 1$ and $p = 0$ in Lemma 1.A.1. Then, for any $0 \leq m < \infty$

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

where the remainder term $R(a, b, c)$ vanishes in probability uniformly in $c \in [c_+, \infty)$ and $|a|, |b| \leq B$. Remember the notation $\zeta_c^- = \xi_c - \xi_{-c}$ in (1.2.7) to arrive at the final expression. ■

The expansion of the FODR in the setup of instrumental variables is almost identical to that in the classical regression, see Lemma A.4 in Jiao and Pretis (2022). The expansion in the IVs regression includes one more term related to the estimation error of β, $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$, which is new here and further complicates our asymptotic analysis. Based on Lemma 1.C.1, we prove the tightness property of the FODR of Algorithm $1.I$, shown in Theorem 1.3.1 in Section 1.3.

Proof of Theorem 1.3.1. Immediately following from Lemma 1.C.1, it holds that

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}(1_{(|u_i|>\sigma c)} - \gamma_c) - \frac{\mathsf{f}_u(c)}{\sigma}\zeta_c^{-\prime}\Sigma^{1/2}n^{1/2}(\widehat{\beta}_c^{(m)} - \beta) - \frac{2c\mathsf{f}_u(c)}{\sigma}n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) + \mathrm{o}_{\mathsf{P}}(1),$$

where, uniformly in $m \in [0, \infty)$ and $c \in [c_+, \infty)$, the first term is $\mathrm{O}_{\mathsf{P}}(1)$ due to the tightness result of the empirical processes shown in Lemma 1.A.1, and the second and third terms are $\mathrm{O}_{\mathsf{P}}(1)$ since $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta) = \mathrm{O}_{\mathsf{P}}(1)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) = \mathrm{O}_{\mathsf{P}}(1)$ as shown by Theorem 3.3 in Jiao (2022) and $\sup_{c_+\leq c<\infty}|\zeta_c^{-}\mathsf{f}_u(c)| < \infty$, $\sup_{c_+\leq c<\infty}|c\mathsf{f}_u(c)| < \infty$ by Assumption 1.3.1$(ia, iiib)$. Theorem 3.3 in Jiao (2022) further needs Assumption 1.3.1(v) with $\eta = 1/4$ and we assume $\sigma > 0$ in our analysis. Therefore, the FODR adjusted by the rate $n^{1/2}$ is bounded in probability uniformly in m and c. ■

Equipped with Lemmas 1.A.3, 1.B.1, and 1.C.1, we are able to establish the Gaussian approximation result in Theorem 1.3.2 ($R^{(0)}$) for the FODR calculated at the initial step of Algorithm 1.R.

Proof of Theorem 1.3.2. The proof is divided into two parts. First, we establish the finite dimensional convergence of the FODR process. This result is then combined with the tightness of the FODR process in Theorem 1.3.1 to obtain the weak convergence.

Robustified 2SLS described in Algorithm 1.R starts with the full-sample 2SLS estimators such that $(\widehat{\beta}_c^{(0)}, \widehat{\sigma}_c^{(0)}) = (\widetilde{\beta}, \widetilde{\sigma})$. Thus, by Assumption 1.3.1(ia, iia, iva, ivc), Lemma 1.B.1 shows that $n^{1/2}(\widehat{\beta}_c^{(0)} - \beta)$ and $n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma)$ are $\mathrm{O}_{\mathsf{P}}(1)$. Together with Assumption 1.3.1(ia, iia, iii, iv), Lemma 1.C.1 gives the expansion of the FODR at step $m = 0$, so uniformly in $c \in [c_+, \infty)$ we have

$$n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}(1_{(|u_i|>\sigma c)} - \gamma_c) - \frac{\mathsf{f}_u(c)}{\sigma}\zeta_c^{-\prime}\Sigma^{1/2}n^{1/2}(\widehat{\beta}_c^{(0)} - \beta) - \frac{2c\mathsf{f}_u(c)}{\sigma}n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) + \mathrm{o}_{\mathsf{P}}(1).$$

Next, insert the expansions of the initial estimation errors from Lemma 1.B.1 and

rearrange the whole expression to obtain

$$\mathbb{G}_n^{(0)}(c) = \begin{Bmatrix} 1 \\ -c\mathsf{f}_u(c) \\ 0 \end{Bmatrix}' n^{-1/2} \sum_{i=1}^{n} \begin{Bmatrix} 1_{(|u_i|>\sigma c)} - \gamma_c \\ \frac{u_i^2}{\sigma^2} - 1 \\ (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i|\leq \sigma c)} \end{Bmatrix}$$
$$- \frac{\mathsf{f}_u(c)}{\sigma} \begin{pmatrix} \zeta_c^- - 2c\Omega \\ 0_{d_x} \end{pmatrix}' \begin{pmatrix} \Sigma^{1/2} M_{\tilde{x}\tilde{x},n}^{-1} & 0_{d_x \times d_x} \\ 0_{d_x \times d_x} & \Sigma^{1/2} M_{\tilde{x}\tilde{x},n}^{-1} \end{pmatrix} \sum_{i=1}^{n} \begin{pmatrix} \tilde{x}_{in} u_i \\ \tilde{x}_{in} u_i 1_{(|u_i|\leq \sigma c)} \end{pmatrix} + o_{\mathsf{P}}(1).$$

Note that to guarantee the uniformity argument over $c \in [c_+, \infty)$, we also apply $\sup_{c_+ \leq c < \infty} |\zeta_c^- \mathsf{f}_u(c)| < \infty$ and $\sup_{c_+ \leq c < \infty} |c\mathsf{f}_u(c)| < \infty$ by Assumption 1.3.1$(ia, iiib)$. To derive the limiting distribution of $\mathbb{G}_n^{(0)}(c)$, we need to apply $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}} > 0$ by Assumption 1.3.1(iva), Lemma 1.A.3 by Assumption 1.3.1(ia, iva, ivc), Slutsky's theorem, and the fact that a linear transformation of a multivariate normal distribution is still normal. Thus, for any $c \in [c_+, \infty)$ and as $n \to \infty$

$$\mathbb{G}_n^{(0)}(c) \xrightarrow{\mathsf{D}} \mathsf{N}(0, K_{cc}^{(0)}),$$

where

$$K_{cc}^{(0)} = \gamma_c(1-\gamma_c) - 2c\mathsf{f}_u(c)(1 - \gamma_c - \tau_2^c) + c^2 \mathsf{f}_u^2(c)(\tau_4 - 1)$$
$$+ \mathsf{f}_u^2(c)(\zeta_c^- - 2c\Omega)' \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \Sigma^{1/2} (\zeta_c^- - 2c\Omega).$$

A similar exercise concludes the proof of the finite dimensional convergence of $\mathbb{G}_n^{(0)}$ by using the multivariate CLT.

Theorem 1.3.1 shows the tightness of $\mathbb{G}_n^{(0)}$ by Assumption 1.3.1(ia, iia, iii, iv) and the fact that the full-sample 2SLS as the initial estimator is tight. Thus, combining the finite dimensional convergence and tightness, we can now establish the weak convergence

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

see Billingsley (1968). Finally, the covariance of the limiting Gaussian process between two cut-off values s and t needs to be calculated, where $c_+ \leq s \leq t < \infty$. Replace c by s and t, respectively, in the expansion of $\mathbb{G}_n^{(0)}(c)$ and apply similar reasoning as in Lemma 1.A.3 to compute their asymptotic covariance

$$K_{st}^{(0)} = \gamma_t(1-\gamma_s) - s\mathsf{f}_u(s)(1 - \gamma_t - \tau_2^t) - t\mathsf{f}_u(t)(1 - \gamma_s - \tau_2^s) + st\mathsf{f}_u(s)\mathsf{f}_u(t)(\tau_4 - 1)$$
$$+ \mathsf{f}_u(s)\mathsf{f}_u(t)(\zeta_s^- - 2s\Omega)' \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \Sigma^{1/2} (\zeta_t^- - 2t\Omega).$$

The above calculation further requires the following four indicator equalities

$$1_{(|u_i|>\sigma s)}1_{(|u_i|\leq\sigma t)} = 1_{(\sigma s<|u_i|\leq\sigma t)}, \qquad 1_{(|u_i|\leq\sigma s)}1_{(|u_i|\leq\sigma t)} = 1_{(|u_i|\leq\sigma s)},$$

$$1_{(|u_i|\leq\sigma s)}1_{(|u_i|>\sigma t)} = 0, \qquad 1_{(|u_i|>\sigma s)}1_{(|u_i|>\sigma t)} = 1_{(|u_i|>\sigma t)},$$

for $s \leq t$. ■

Next, we show the Gaussian approximation result for the FODR calculated at the first iteration step of Algorithm $1.R$.

Theorem 1.C.1. *Consider Algorithm 1.R. Suppose Assumption 1.3.1(ia, iia, iii, iv) holds. Let $\mathbb{G}_n^{(1)}(c) = n^{1/2}(\widehat{\gamma}_c^{(1)} - \gamma_c)$ for $c \in [c_+, \infty)$ denote a sequence of processes of the FODR at the first step $m = 1$. Then, as $n \to \infty$*

$$\mathbb{G}_n^{(1)} \rightsquigarrow \mathsf{GP}(0, K^{(1)}),$$

where $\mathsf{GP}(c; 0, K^{(1)})$ for $c \in [c_+, \infty)$ refers to the Gaussian process with expected value $\mathsf{E}\{\mathsf{GP}(t; 0, K^{(1)})\} = 0$ and covariance $K_{st}^{(1)} = \mathsf{Cov}\{\mathsf{GP}(s; 0, K^{(1)}), \mathsf{GP}(t; 0, K^{(1)})\}$ for any $c_+ \leq s \leq t < \infty$ such that

$$\begin{aligned}
K_{st}^{(1)} &= \gamma_t(1-\gamma_s) - \frac{s^2\mathsf{f}_u^2(s)(s^2-\varsigma_s^2)}{\tau_2^s}(1-\gamma_t-\tau_2^t) - \frac{t^2\mathsf{f}_u^2(t)(t^2-\varsigma_t^2)}{\tau_2^t}(1-\gamma_s-\tau_2^s) \\
&+ \frac{s^2\mathsf{f}_u^2(s)(s^2-\varsigma_s^2)}{\tau_2^s}\frac{t^2\mathsf{f}_u^2(t)(t^2-\varsigma_t^2)}{\tau_2^t}(\tau_4-1) + \frac{t\mathsf{f}_u(t)}{\tau_2^t}\frac{s\mathsf{f}_u(s)(\tau_4^s-\tau_2^s\varsigma_s^2)}{\tau_2^s} \\
&+ \frac{s\mathsf{f}_u(s)(\tau_4^s-\tau_2^s\varsigma_s^2)}{\tau_2^s}\frac{t^2\mathsf{f}_u^2(t)(t^2-\varsigma_t^2)}{\tau_2^t} + \frac{t\mathsf{f}_u(t)(\tau_4^t-\tau_2^t\varsigma_t^2)}{\tau_2^t}\frac{s^2\mathsf{f}_u^2(s)(s^2-\varsigma_s^2)}{\tau_2^s} \\
&- t\mathsf{f}_u(t)(\frac{\psi_s}{\psi_t} - \frac{\tau_2^s}{\tau_2^t}) + \varsigma_s^2\mathsf{f}_u(s)\frac{\mathsf{f}_u(t)}{\psi_t}(\zeta_s^- - \frac{2s}{\tau_2^s}\omega_s)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}(\zeta_t^- - \frac{2t}{\tau_2^t}\omega_t) \\
&+ \varsigma_s^2\mathsf{f}_u(s)\frac{2t\mathsf{f}_u^2(t)}{\tau_2^t}(\zeta_s^- - \frac{2s}{\tau_2^s}\omega_s)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}\{\frac{t^2+\varsigma_t^2}{2}\zeta_t^- - \frac{2t}{\psi_t}\omega_t - t(t^2-\varsigma_t^2)\Omega\} \\
&+ \varsigma_t^2\mathsf{f}_u(t)\frac{2s\mathsf{f}_u^2(s)}{\tau_2^s}(\zeta_t^- - \frac{2t}{\tau_2^t}\omega_t)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}\{\frac{s^2+\varsigma_s^2}{2}\zeta_s^- - \frac{2s}{\psi_s}\omega_s - s(s^2-\varsigma_s^2)\Omega\} \\
&+ \frac{2s\mathsf{f}_u^2(s)}{\tau_2^s}\frac{2t\mathsf{f}_u^2(t)}{\tau_2^t}\{\frac{s^2+\varsigma_s^2}{2}\zeta_s^- - \frac{2s}{\psi_s}\omega_s - s(s^2-\varsigma_s^2)\Omega\}'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2} \\
&\times \{\frac{t^2+\varsigma_t^2}{2}\zeta_t^- - \frac{2t}{\psi_t}\omega_t - t(t^2-\varsigma_t^2)\Omega\}.
\end{aligned}$$

Proof of Theorem 1.C.1. The proof is analogous to that of Theorem 1.3.2 ($R^{(0)}$), which first establishes the finite dimensional convergence of the first step FODR process and proceeds to demonstrate the weak convergence.

Algorithm 1.R starts with the full-sample 2SLS estimator $(\widehat{\beta}_c^{(0)}, \widehat{\sigma}_c^{(0)}) = (\widetilde{\beta}, \widetilde{\sigma})$, which are tight. Thus, together with Assumption 1.3.1(ia, iia, iii, iv), Lemma 1.B.2 shows that $n^{1/2}(\widehat{\beta}_c^{(1)} - \beta)$ and $n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma)$ are $\mathrm{O_P}(1)$ uniformly in $c \in [c_+, \infty)$. Using this tightness condition, Lemma 1.C.1 then gives the expansion of the FODR at step $m = 1$ so that

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^{(1)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}(1_{(|u_i|>\sigma c)} - \gamma_c) - \frac{\mathsf{f}_u(c)}{\sigma}\zeta_c^{-\prime}\Sigma^{1/2}n^{1/2}(\widehat{\beta}_c^{(1)} - \beta) \\ - \frac{2c\mathsf{f}_u(c)}{\sigma}n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma) + \mathrm{o_P}(1). \end{aligned}$$

Plugging in the expansions of $n^{1/2}(\widehat{\beta}_c^{(1)} - \beta)$ and $n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma)$ from Lemma 1.B.2, we have

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^{(1)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}(1_{(|u_i|>\sigma c)} - \gamma_c) - \frac{c\mathsf{f}_u(c)}{\tau_2^c}n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\le\sigma c)} \\ - \frac{\mathsf{f}_u(c)}{\sigma\psi_c}(\zeta_c^- - \frac{2c}{\tau_2^c}\omega_c)'\Sigma^{1/2}M_{\tilde{x}\tilde{x},n}^{-1}\sum_{i=1}^{n}\tilde{x}_{in}u_i1_{(|u_i|\le\sigma c)} \\ - \frac{2c\mathsf{f}_u^2(c)}{\sigma\tau_2^c}(\frac{c^2+\varsigma_c^2}{2}\zeta_c^- - \frac{2c}{\psi_c}\omega_c)'\Sigma^{1/2}n^{1/2}(\widehat{\beta}_c^{(0)} - \beta) \\ - \frac{2c^2\mathsf{f}_u^2(c)(c^2-\varsigma_c^2)}{\sigma\tau_2^c}n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) + \mathrm{o_P}(1). \end{aligned}$$

Next, insert the expansions of the initial estimation errors from Lemma 1.B.1 and collect terms to obtain

$$\begin{aligned} \mathbb{G}_n^{(1)}(c) = \begin{Bmatrix} 1 \\ -\frac{c^2\mathsf{f}_u^2(c)(c^2-\varsigma_c^2)}{\tau_2^c} \\ -\frac{c\mathsf{f}_u(c)}{\tau_2^c} \end{Bmatrix}' n^{-1/2}\sum_{i=1}^{n}\begin{Bmatrix} 1_{(|u_i|>\sigma c)} - \gamma_c \\ \frac{u_i^2}{\sigma^2} - 1 \\ (\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\le\sigma c)} \end{Bmatrix} \\ - \frac{\mathsf{f}_u(c)}{\sigma\psi_c}\begin{bmatrix} \frac{2c\mathsf{f}_u(c)}{\varsigma_c^2}\{\frac{c^2+\varsigma_c^2}{2}\zeta_c^- - \frac{2c}{\psi_c}\omega_c - c(c^2-\varsigma_c^2)\Omega\} \\ \zeta_c^- - \frac{2c}{\tau_2^c}\omega_c \end{bmatrix}' \begin{pmatrix} \Sigma^{1/2}M_{\tilde{x}\tilde{x},n}^{-1} & 0_{d_x\times d_x} \\ 0_{d_x\times d_x} & \Sigma^{1/2}M_{\tilde{x}\tilde{x},n}^{-1} \end{pmatrix} \\ \times \sum_{i=1}^{n}\begin{pmatrix} \tilde{x}_{in}u_i \\ \tilde{x}_{in}u_i1_{(|u_i|\le\sigma c)} \end{pmatrix} + \mathrm{o_P}(1). \end{aligned}$$

We apply Lemma 1.A.3 to derive the limiting distribution of $\mathbb{G}_n^{(1)}(c)$ so that for any $c \in [c_+, \infty)$ and as $n \to \infty$

$$\mathbb{G}_n^{(1)}(c) \xrightarrow{\mathsf{D}} \mathsf{N}(0, K_{cc}^{(1)}),$$

where

$$
\begin{aligned}
K_{cc}^{(1)} = {} & \gamma_c(1-\gamma_c) - \frac{2c^2\mathsf{f}_u^2(c)(c^2-\varsigma_c^2)}{\tau_2^c}(1-\gamma_c-\tau_2^c) + \{\frac{c^2\mathsf{f}_u^2(c)(c^2-\varsigma_c^2)}{\tau_2^c}\}^2(\tau_4-1) \\
& + \frac{2c^3\mathsf{f}_u^3(c)(c^2-\varsigma_c^2)}{(\tau_2^c)^2}(\tau_4^c-\tau_2^c\varsigma_c^2) + \{\frac{c\mathsf{f}_u(c)}{\tau_2^c}\}^2(\tau_4^c-\tau_2^c\varsigma_c^2) \\
& + \frac{\varsigma_c^2\mathsf{f}_u^2(c)}{\psi_c}(\zeta_c^- - \frac{2c}{\tau_2^c}\omega_c)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}(\zeta_c^- - \frac{2c}{\tau_2^c}\omega_c) \\
& + \frac{4c\mathsf{f}_u^3(c)}{\psi_c}(\zeta_c^- - \frac{2c}{\tau_2^c}\omega_c)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}\{\frac{c^2+\varsigma_c^2}{2}\zeta_c^- - \frac{2c}{\psi_c}\omega_c - c(c^2-\varsigma_c^2)\Omega\} \\
& + \{\frac{2c\mathsf{f}_u^2(c)}{\tau_2^c}\}^2\{\frac{c^2+\varsigma_c^2}{2}\zeta_c^- - \frac{2c}{\psi_c}\omega_c - c(c^2-\varsigma_c^2)\Omega\}'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2} \\
& \times \{\frac{c^2+\varsigma_c^2}{2}\zeta_c^- - \frac{2c}{\psi_c}\omega_c - c(c^2-\varsigma_c^2)\Omega\}.
\end{aligned}
$$

A similar exercise using the multivariate CLT concludes the proof of the finite dimensional convergence of $\mathbb{G}_n^{(1)}$.

Theorem 1.3.1 shows the tightness of $\mathbb{G}_n^{(1)}$, which now allows us to establish the weak convergence

$$\mathbb{G}_n^{(1)} \rightsquigarrow \mathsf{GP}(0, K^{(1)}).$$

The final step is to calculate the covariance $K_{st}^{(1)}$ of the limiting Gaussian process between two cut-off values s and t, where $c_+ \leq s \leq t < \infty$. Replace c by s and t, respectively, in the expansion of $\mathbb{G}_n^{(1)}(c)$ and apply similar reasoning as in Lemma 1.A.3 and Theorem 1.3.2 ($R^{(0)}$) to compute the asymptotic covariance $K_{st}^{(1)}$. ■

Before being able to analyse the FODR $\widehat{\gamma}_c^{(m)}$ of Algorithm 1.R at a general iteration step $m \in [1, \infty)$, we must return to Algorithm 1.I and demonstrate an expansion of the FODR $\widehat{\gamma}_c^{(m)}$ in terms of the initial estimation errors $n^{1/2}(\widehat{\beta}_c^{(0)} - \beta)$ and $n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma)$. The proof requires the expansions of the estimation errors $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$ and $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ at iteration steps $m \in [1, \infty)$ in terms of those at the initial step $m = 0$, which has been established in Theorem 3.4 in Jiao (2022).

Lemma 1.C.2. *Consider Algorithm 1.I. Suppose Assumption 1.3.1*(ia, iia, iii, iv, v) *holds with* $\eta = 1/4$. *Then, as* $n \to \infty$ *and uniformly in* $c \in [c_+, \infty)$ *we can express the expansion of the FODR at step* $m \in [1, \infty)$ *in terms of the initial estimation*

errors

$$
\begin{aligned}
n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}(1_{(|u_i|>\sigma c)} - \gamma_c) - \frac{\mathsf{f}_u(c)}{\sigma}[\{\varrho_{\beta\beta,c}^{(m)} + \frac{c\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}(c^2-\varsigma_c^2)}{\tau_2^c}\}\zeta_c^- \\
- \frac{4c^2\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}}{\psi_c\tau_2^c}\omega_c]'\Sigma^{1/2}n^{1/2}(\widehat{\beta}_c^{(0)} - \beta) - \frac{2c\mathsf{f}_u(c)}{\sigma}\varrho_{\sigma\sigma,c}^{(m)}n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) \\
- \frac{\mathsf{f}_u(c)}{\sigma}[\{\varrho_{\beta\tilde{x}u,c}^{(m)} + c\varrho_{\sigma\tilde{x}u,c}^{(m)}(c^2-\varsigma_c^2)\}\zeta_c^- - \frac{2c(2c\varrho_{\sigma\tilde{x}u,c}^{(m)} + \varrho_{\sigma uu,c}^{(m)})}{\psi_c}\omega_c]'\Sigma^{1/2} \\
\times M_{\tilde{x}\tilde{x},n}^{-1}\sum_{i=1}^{n}\tilde{x}_{in}u_i1_{(|u_i|\leq\sigma c)} - c\mathsf{f}_u(c)\varrho_{\sigma uu,c}^{(m)}n^{-1/2}\sum_{i=1}^{n}(\frac{u_i^2}{\sigma^2} - \varsigma_c^2)1_{(|u_i|\leq\sigma c)} \\
+ \mathsf{o}_{\mathsf{P}}(1),
\end{aligned}
$$

where the coefficients have expressions

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

Proof of Lemma 1.C.2. First, Theorem 3.3 in Jiao (2022) shows for Algorithm 1.*I* that $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta) = \mathsf{O}_{\mathsf{P}}(1)$ and $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) = \mathsf{O}_{\mathsf{P}}(1)$ for any $m \in [0, \infty)$ and uniformly in $c \in [c_+, \infty)$. To apply this theorem, we require Assumption 1.3.1(ia, iia, iii, iv, v) with $\eta = 1/4$ meaning that the initial estimators are tight. We then combine this fact with Assumption 1.3.1(ia, iia, iii, iv) to apply Lemma 1.C.1, which gives the expansion of the m-step FODR in terms of the m-step estimation errors for any $m \in [0, \infty)$

$$
\begin{aligned}
n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}(1_{(|u_i|>\sigma c)} - \gamma_c) - \frac{\mathsf{f}_u(c)}{\sigma}\zeta_c^{-\prime}\Sigma^{1/2}n^{1/2}(\widehat{\beta}_c^{(m)} - \beta) \\
- \frac{2c\mathsf{f}_u(c)}{\sigma}n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) + \mathsf{o}_{\mathsf{P}}(1),
\end{aligned}
$$

uniformly in $c \in [c_+, \infty)$. In a final step, we invoke Theorem 3.4 in Jiao (2022) using Assumption 1.3.1(ia, iia, iii, iv, v) with $\eta = 1/4$, which provides expansions of

the m-step estimation error of beta in terms of the initial estimation error for any $m \in [1, \infty)$. Following that theorem's proof, we can similarly derive an expansion of the m-step estimation error of sigma in terms of its initial estimation errors for any $m \in [1, \infty)$, which together yields

$$
\begin{aligned}
n^{1/2}(\widehat{\beta}_c^{(m)} - \beta) &= \varrho_{\beta\beta,c}^{(m)} n^{1/2}(\widehat{\beta}_c^{(0)} - \beta) + \varrho_{\beta\tilde{x}u,c}^{(m)} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} + \mathsf{o}_{\mathsf{P}}(1), \\
n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) &= \varrho_{\sigma\sigma,c}^{(m)} n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) + \frac{\sigma}{2} \varrho_{\sigma uu,c}^{(m)} n^{-1/2} \sum_{i=1}^{n} (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i| \leq \sigma c)} \\
&\quad + \frac{\mathsf{f}_u(c) \varrho_{\sigma\beta,c}^{(m)}}{\tau_2^c} (\frac{c^2 - \varsigma_c^2}{2} \zeta_c^- - \frac{2c}{\psi_c} \omega_c)' \Sigma^{1/2} n^{1/2} (\widehat{\beta}_c^{(0)} - \beta) \\
&\quad + \{ \frac{\varrho_{\sigma\tilde{x}u,c}^{(m)} (c^2 - \varsigma_c^2)}{2} \zeta_c^- - \frac{2c \varrho_{\sigma\tilde{x}u,c}^{(m)} + \varrho_{\sigma uu,c}^{(m)}}{\psi_c} \omega_c \}' \Sigma^{1/2} M_{\tilde{x}\tilde{x},n}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i 1_{(|u_i| \leq \sigma c)} \\
&\quad + \mathsf{o}_{\mathsf{P}}(1),
\end{aligned}
$$

uniformly in $c \in [c_+, \infty)$. All the coefficients $\varrho^{(m)}$ are presented in Theorem 1.C.2 ($R^{(m)}$ for $m \geq 1$). Substitute these two expansions of the estimation errors into the FODR expansion and rearrange the expression to complete the proof. ■

With the above lemma, we are ready to establish the Gaussian approximation results for the FODR $\widehat{\gamma}_c^{(m)}$ of Algorithm 1.R calculated at any iteration step $m \geq 1$.

Theorem 1.C.2. *Consider Algorithm 1.R. Suppose Assumption 1.3.1*(ia, iia, iii, iv) *holds. Let* $\mathbb{G}_n^{(m)}(c) = n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)$ *for* $c \in [c_+, \infty)$ *denote a sequence of processes of the FODR at iteration step* $m \in [1, \infty)$*. Then, as* $n \to \infty$

$$
\mathbb{G}_n^{(m)} \rightsquigarrow \mathsf{GP}(0, K^{(m)}),
$$

where $\mathsf{GP}(c; 0, K^{(m)})$ *for* $c \in [c_+, \infty)$ *refers to the Gaussian process with expected value* $\mathsf{E}\{\mathsf{GP}(t; 0, K^{(m)})\} = 0$ *and covariance* $K_{st}^{(m)} = \mathsf{Cov}\{\mathsf{GP}(s; 0, K^{(m)}), \mathsf{GP}(t; 0, K^{(m)})\}$

for any $c_+ \leq s \leq t < \infty$ *such that*

$$
\begin{aligned}
K_{st}^{(m)} &= \gamma_t(1-\gamma_s) - s\mathsf{f}_u(s)\varrho_{\sigma\sigma,s}^{(m)}(1-\gamma_t-\tau_2^t) - t\mathsf{f}_u(t)\varrho_{\sigma\sigma,t}^{(m)}(1-\gamma_s-\tau_2^s) \\
&\quad + \psi_s t\mathsf{f}_u(t)\varrho_{\sigma uu,t}^{(m)}(\varsigma_s^2-\varsigma_t^2) + st\mathsf{f}_u(s)\mathsf{f}_u(t)\{\varrho_{\sigma\sigma,s}^{(m)}\varrho_{\sigma\sigma,t}^{(m)}(\tau_4-1) + \varrho_{\sigma\sigma,s}^{(m)}\varrho_{\sigma uu,t}^{(m)}(\tau_4^t-\tau_2^t\varsigma_t^2) \\
&\quad + \varrho_{\sigma uu,s}^{(m)}(\varrho_{\sigma\sigma,t}^{(m)}+\varrho_{\sigma uu,t}^{(m)})(\tau_4^s-\tau_2^s\varsigma_s^2)\} + \tau_2^t\mathsf{f}_u(s)\mathsf{f}_u(t)[\{\varrho_{\beta\beta,s}^{(m)} + \frac{\tau_2^s\varrho_{\beta\tilde{x}u,s}^{(m)}}{\tau_2^t} \\
&\quad + \frac{s\mathsf{f}_u(s)\varrho_{\sigma\beta,s}^{(m)}(s^2-\varsigma_s^2)}{\tau_2^s} + \frac{s\tau_2^s\varrho_{\sigma\tilde{x}u,s}^{(m)}(s^2-\varsigma_s^2)}{\tau_2^t}\}\zeta_s^- - \frac{2s\varsigma_s^2}{\tau_2^t}\{\tau_2^t\frac{2s\mathsf{f}_u(s)\varrho_{\sigma\beta,s}^{(m)}}{(\tau_2^s)^2} + 2s\varrho_{\sigma\tilde{x}u,s}^{(m)} \\
&\quad + \varrho_{\sigma uu,s}^{(m)}\}\omega_s - 2s\varrho_{\sigma\sigma,s}^{(m)}\Omega]'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}[\{\varrho_{\beta\tilde{x}u,t}^{(m)} + t\varrho_{\sigma\tilde{x}u,t}^{(m)}(t^2-\varsigma_t^2)\}\zeta_t^- \\
&\quad - \frac{2t}{\psi_t}(2t\varrho_{\sigma\tilde{x}u,t}^{(m)} + \varrho_{\sigma uu,t}^{(m)})\omega_t] + \mathsf{f}_u(s)\mathsf{f}_u(t)[\{\varrho_{\beta\beta,s}^{(m)} + \tau_2^s\varrho_{\beta\tilde{x}u,s}^{(m)} + \frac{s\mathsf{f}_u(s)\varrho_{\sigma\beta,s}^{(m)}(s^2-\varsigma_s^2)}{\tau_2^s} \\
&\quad + s\tau_2^s\varrho_{\sigma\tilde{x}u,s}^{(m)}(s^2-\varsigma_s^2)\}\zeta_s^- - 2s\varsigma_s^2\{\frac{2s\mathsf{f}_u(s)\varrho_{\sigma\beta,s}^{(m)}}{(\tau_2^s)^2} + 2s\varrho_{\sigma\tilde{x}u,s}^{(m)} + \varrho_{\sigma uu,s}^{(m)}\}\omega_s - 2s\varrho_{\sigma\sigma,s}^{(m)}\Omega]' \\
&\quad \times \Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}[\{\varrho_{\beta\beta,t}^{(m)} + \frac{t\mathsf{f}_u(t)\varrho_{\sigma\beta,t}^{(m)}(t^2-\varsigma_t^2)}{\tau_2^t}\}\zeta_t^- - \frac{4t^2\mathsf{f}_u(t)\varrho_{\sigma\beta,t}^{(m)}}{\psi_t\tau_2^t}\omega_t - 2t\varrho_{\sigma\sigma,t}^{(m)}\Omega],
\end{aligned}
$$

where the expressions of the coefficients $\varrho_{\beta\beta,c}^{(m)}$, $\varrho_{\beta\tilde{x}u,c}^{(m)}$, $\varrho_{\sigma\sigma,c}^{(m)}$, $\varrho_{\sigma uu,c}^{(m)}$, $\varrho_{\sigma\beta,c}^{(m)}$, $\varrho_{\sigma\tilde{x}u,c}^{(m)}$ *are given in Lemma 1.C.2.*

Proof of Theorem 1.C.2. The proof here is analogous to that of Theorems 1.3.2 ($R^{(0)}$) and 1.C.1 ($R^{(1)}$). For any $m \in [1,\infty)$, we first establish the finite dimensional convergence for a sequence of the m-step FODR processes and proceed to prove weak convergence.

Algorithm 1.R chooses the full-sample 2SLS estimator as the initial estimator. By Assumption 1.3.1(ia, iia, iva, ivc), Lemma 1.B.1 shows that they are tight so that Assumption 1.3.1(v) with $\eta = 1/4$ holds. Together with Assumption 1.3.1(ia, iia, iii, iv), Lemma 1.C.2 gives the expansion of the FODR at step $m \in [1,\infty)$ in terms of the initial estimation errors, which holds uniformly in $c \in [c_+,\infty)$. We then insert the expansions of the initial estimation errors from Lemma 1.B.1 and rearrange the

expression to obtain for any iteration step $m \in [1, \infty)$

$$\mathbb{G}_n^{(m)}(c) = \begin{Bmatrix} 1 \\ -c\mathsf{f}_u(c)\varrho_{\sigma\sigma,c}^{(m)} \\ -c\mathsf{f}_u(c)\varrho_{\sigma uu,c}^{(m)} \end{Bmatrix}' n^{-1/2} \sum_{i=1}^{n} \begin{Bmatrix} 1_{(|u_i|>\sigma c)} - \gamma_c \\ \frac{u_i^2}{\sigma^2} - 1 \\ (\frac{u_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|u_i|\leq \sigma c)} \end{Bmatrix}$$
$$- \frac{\mathsf{f}_u(c)}{\sigma} \begin{bmatrix} \{\varrho_{\beta\beta,c}^{(m)} + \frac{c\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}(c^2-\varsigma_c^2)}{\tau_2^c}\}\zeta_c^- - \frac{4c^2\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}}{\psi_c\tau_2^c}\omega_c - 2c\varrho_{\sigma\sigma,c}^{(m)}\Omega \\ \{\varrho_{\beta\tilde{x}u,c}^{(m)} + c\varrho_{\sigma\tilde{x}u,c}^{(m)}(c^2-\varsigma_c^2)\}\zeta_c^- - \frac{2c(2c\varrho_{\sigma\tilde{x}u,c}^{(m)} + \varrho_{\sigma uu,c}^{(m)})}{\psi_c}\omega_c \end{bmatrix}'$$
$$\times \begin{pmatrix} \Sigma^{1/2} M_{\tilde{x}\tilde{x},n}^{-1} & 0_{d_x \times d_x} \\ 0_{d_x \times d_x} & \Sigma^{1/2} M_{\tilde{x}\tilde{x},n}^{-1} \end{pmatrix} \sum_{i=1}^{n} \begin{pmatrix} \tilde{x}_{in} u_i \\ \tilde{x}_{in} u_i 1_{(|u_i|\leq \sigma c)} \end{pmatrix} + o_{\mathsf{P}}(1).$$

To guarantee the uniformity argument over $c \in [c_+, \infty)$ in the above, we also require

$$\sup_{c_+ \leq c < \infty} \max\{|\varrho_{\beta\beta,c}^{(m)}|, |\varrho_{\beta\tilde{x}u,c}^{(m)}|, |\varrho_{\sigma\sigma,c}^{(m)}|, |\varrho_{\sigma uu,c}^{(m)}|, |\varrho_{\sigma\beta,c}^{(m)}|, |\varrho_{\sigma\tilde{x}u,c}^{(m)}|\} < \infty, \sup_{c_+ \leq c < \infty} |\mathsf{f}_u(c)\zeta_c^-| < \infty.$$

The first condition is implied by $\sup_{c_+ \leq c < \infty} \max\{|2c\mathsf{f}_u(c)/\psi_c|, |c\mathsf{f}_u(c)(c^2-\varsigma_c^2)/\tau_2^c|\} < 1$, which can be derived from Assumption 1.3.1(ia), see Theorem 3.6 in Johansen and Nielsen (2013). The second condition is directly shown by Assumption 1.3.1($iiib$). To derive the limiting distribution of the expansion, we need to apply $M_{\tilde{x}\tilde{x},n} \xrightarrow{\mathsf{P}} M_{\tilde{x}\tilde{x}} > 0$ by Assumption 1.3.1(iva), Lemma 1.A.3 by Assumption 1.3.1(ia, iva, ivc), Slutsky's theorem, and the fact that a linear transformation of a multivariate normal distribution is still normal. Thus, for any $m \in [1, \infty)$, $c \in [c_+, \infty)$, and as $n \to \infty$

$$\mathbb{G}_n^{(m)}(c) \xrightarrow{\mathsf{D}} \mathsf{N}(0, K_{cc}^{(m)}),$$

where

$$K_{cc}^{(m)} = \gamma_c(1-\gamma_c) - 2c\mathsf{f}_u(c)\varrho_{\sigma\sigma,c}^{(m)}(1-\gamma_c-\tau_2^c) + c^2\mathsf{f}_u^2(c)\{(\varrho_{\sigma\sigma,c}^{(m)})^2(\tau_4-1)$$
$$+ \varrho_{\sigma uu,c}^{(m)}(2\varrho_{\sigma\sigma,c}^{(m)} + \varrho_{\sigma uu,c}^{(m)})(\tau_4^c - \tau_2^c\varsigma_c^2)\} + \tau_2^c\mathsf{f}_u^2(c)[\{\varrho_{\beta\beta,c}^{(m)} + \varrho_{\beta\tilde{x}u,c}^{(m)} + \frac{c\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}(c^2-\varsigma_c^2)}{\tau_2^c}$$
$$+ c\varrho_{\sigma\tilde{x}u,c}^{(m)}(c^2-\varsigma_c^2)\}\zeta_c^- - \frac{2c}{\psi_c}\{\frac{2c\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}}{\tau_2^c} + 2c\varrho_{\sigma\tilde{x}u,c}^{(m)} + \varrho_{\sigma uu,c}^{(m)}\}\omega_c - 2c\varrho_{\sigma\sigma,c}^{(m)}\Omega]'$$
$$\times \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \Sigma^{1/2}[\{\varrho_{\beta\tilde{x}u,c}^{(m)} + c\varrho_{\sigma\tilde{x}u,c}^{(m)}(c^2-\varsigma_c^2)\}\zeta_c^- - \frac{2c}{\psi_c}(2c\varrho_{\sigma\tilde{x}u,c}^{(m)} + \varrho_{\sigma uu,c}^{(m)})\omega_c]$$
$$+ \mathsf{f}_u^2(c)[\{\varrho_{\beta\beta,c}^{(m)} + \tau_2^c\varrho_{\beta\tilde{x}u,c}^{(m)} + \frac{c\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}(c^2-\varsigma_c^2)}{\tau_2^c} + c\tau_2^c\varrho_{\sigma\tilde{x}u,c}^{(m)}(c^2-\varsigma_c^2)\}\zeta_c^-$$
$$- 2c\varsigma_c^2\{\frac{2c\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}}{(\tau_2^c)^2} + 2c\varrho_{\sigma\tilde{x}u,c}^{(m)} + \varrho_{\sigma uu,c}^{(m)}\}\omega_c - 2c\varrho_{\sigma\sigma,c}^{(m)}\Omega]'\Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}$$
$$\times [\{\varrho_{\beta\beta,c}^{(m)} + \frac{c\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}(c^2-\varsigma_c^2)}{\tau_2^c}\}\zeta_c^- - \frac{4c^2\mathsf{f}_u(c)\varrho_{\sigma\beta,c}^{(m)}}{\psi_c\tau_2^c}\omega_c - 2c\varrho_{\sigma\sigma,c}^{(m)}\Omega].$$

A similar exercise concludes the proof of the finite dimensional convergence of $\mathbb{G}_n^{(m)}$ by using the multivariate CLT.

Due to Assumption 1.3.1(ia, iia, iii, iv) and the fact that the initial full-sample 2SLS estimator is tight, Theorem 1.3.1 shows the tightness of $\mathbb{G}_n^{(m)}$ for $m \in [0, \infty)$. Thus, combining the finite dimensional convergence and tightness, we can demonstrate the weak convergence so that for any $m \in [1, \infty)$

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

To complete the whole proof, the covariance $K_{st}^{(m)}$ of the limiting Gaussian process between two cut-off values s and t needs to be calculated, where $c_+ \leq s \leq t < \infty$. Replace c by s and t, respectively, in the expansion of $\mathbb{G}_n^{(m)}(c)$ and apply similar reasoning as in Lemma 1.A.3 and Theorems 1.3.2 ($R^{(0)}$) and 1.C.1 ($R^{(1)}$) to compute the asymptotic covariance $K_{st}^{(m)}$. ■

Theorem 3.7 in Jiao (2022) and Lemma 1.B.3 show for any $m \in [1, \infty)$ that the split-half version of Algorithm 1.S has the same expansions of $n^{1/2}(\widehat{\beta}_c^{(m)} - \beta)$ and $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ as Algorithm 1.R. It immediately follows by Lemma 1.C.1 that the two algorithms also have the same expansions of the FODR $n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)$ for $m \in [1, \infty)$. Thus, to prove that both algorithms have the same Gaussian weak limit of their FODRs at any iteration step, it suffices to show equivalence of their initial step FODR expansions.

First, we investigate the initial iteration of the general split version of Algorithm 1.S where the subsamples $\mathcal{I}_1$ and $\mathcal{I}_2$ contain $n_1 = \text{int}[\alpha n]$ and $n_2 = n - n_1$ observations with $\alpha \in (0, 1)$ before focusing on the split-half version with $\alpha = 1/2$. For $i \in \mathcal{I}_j$, where $j = 1, 2$, we use the notation $z_{in_j} = n_j^{-1/2} z_i$, $\tilde{x}_{in_j} = n_j^{-1/2} \tilde{x}_i = \Pi' z_{in_j}$, and $M_{\tilde{x}\tilde{x},n_j}^{\mathcal{I}_j} = \sum_{i \in \mathcal{I}_j} \tilde{x}_{in_j} \tilde{x}_{in_j}'$.

Lemma 1.C.3. *Consider Algorithm 1.S. Suppose that the sample is split into a partition, $\mathcal{I}_1$ and $\mathcal{I}_2$, with number of observations $n_1 = \text{int}[\alpha n]$ and $n_2 = n - n_1$ where $\alpha \in (0, 1)$ represents the share of observations in subsample $\mathcal{I}_1$. Suppose Assumption 1.3.1(ia, iia, iii, iv) holds for each subsample. Then, as $n \to \infty$ and*

uniformly in $c \in [c_+, \infty)$, the expansion of the initial FODR is given by

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) = \sum_{j=1}^{2} (\frac{n_j}{n})^{1/2} \{ & n_j^{-1/2} \sum_{i \in \mathcal{I}_j} (1_{(|u_i| > \sigma c)} - \gamma_c) \\ & - (\frac{n_j}{n_{3-j}})^{1/2} c \mathsf{f}_u(c) n_{3-j}^{-1/2} \sum_{i \in \mathcal{I}_{3-j}} (\frac{u_i^2}{\sigma^2} - 1) \\ & - (\frac{n_j}{n_{3-j}})^{1/2} \frac{\mathsf{f}_u(c)}{\sigma} (\zeta_c^- - 2c\Omega)' \Sigma^{1/2} (M^{\mathcal{I}_{3-j}}_{\tilde{x}\tilde{x}, n_{3-j}})^{-1} \sum_{i \in \mathcal{I}_{3-j}} \tilde{x}_{in_{3-j}} u_i \} + \mathrm{o_P}(1). \end{aligned}$$

Proof of Lemma 1.C.3. We start with the definition of the FODR for Algorithm 1.S at the initial step and express it in terms of the two subsample counts

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) &= n^{-1/2} \sum_{i=1}^{n} \{ (1 - v_{i,c}^{(0)}) - \gamma_c \} \\ &= \sum_{j=1}^{2} (\frac{n_j}{n})^{1/2} n_j^{-1/2} \sum_{i \in \mathcal{I}_j} (1 - 1_{(|y_i - x_i' \widehat{\beta}_{3-j}| \leq \widehat{\sigma}_{3-j} c)} - \gamma_c). \end{aligned}$$

Apply similar reasoning as in Lemma 1.C.1 using the expansion of the empirical processes in Lemma 1.A.1 with $w_{in} = 1$ and $p = 0$. It then follows that

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) = \sum_{j=1}^{2} (\frac{n_j}{n})^{1/2} \{ & n_j^{-1/2} \sum_{i \in \mathcal{I}_j} (1_{(|u_i| > \sigma c)} - \gamma_c) \\ & - (\frac{n_j}{n_{3-j}})^{1/2} \frac{\mathsf{f}_u(c)}{\sigma} \zeta_c^{-\prime} \Sigma^{1/2} n_{3-j}^{1/2} (\widehat{\beta}_{3-j} - \beta) \\ & - (\frac{n_j}{n_{3-j}})^{1/2} \frac{2c\mathsf{f}_u(c)}{\sigma} n_{3-j}^{1/2} (\widehat{\sigma}_{3-j} - \sigma) \} + \mathrm{o_P}(1), \end{aligned}$$

uniformly in $c \in [c_+, \infty)$. To apply the empirical processes argument, we require Assumption 1.3.1(ia, iia, iii, iv) holding for each subsample $\mathcal{I}_j$ and the fact that the initial estimation errors $n_{3-j}^{1/2}(\widehat{\beta}_{3-j} - \beta)$, $n_{3-j}^{1/2}(\widehat{\sigma}_{3-j} - \sigma)$ are $\mathrm{O_P}(1)$ for $j = 1, 2$. The initial estimators of Algorithm 1.S are the usual 2SLS estimators but calculated on their respective subsamples, so Lemma 1.B.1 using Assumption 1.3.1(ia, iia, iva, ivc) gives their expansions

$$\begin{aligned} n_{3-j}^{1/2}(\widehat{\beta}_{3-j} - \beta) &= (M^{\mathcal{I}_{3-j}}_{\tilde{x}\tilde{x}, n_{3-j}})^{-1} \sum_{i \in \mathcal{I}_{3-j}} \tilde{x}_{in_{3-j}} u_i + \mathrm{O_P}(n_{3-j}^{-1/2}), \\ n_{3-j}^{1/2}(\widehat{\sigma}_{3-j} - \sigma) &= \frac{\sigma}{2} n_{3-j}^{-1/2} \sum_{i \in \mathcal{I}_{3-j}} (\frac{u_i^2}{\sigma^2} - 1) \\ &\quad - \Omega' \Sigma^{1/2} (M^{\mathcal{I}_{3-j}}_{\tilde{x}\tilde{x}, n_{3-j}})^{-1} \sum_{i \in \mathcal{I}_{3-j}} \tilde{x}_{in_{3-j}} u_i + \mathrm{O_P}(n_{3-j}^{-1/2}), \end{aligned}$$

for $j = 1, 2$. Finally, substitute the above expansions into $n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c)$ to complete the proof. To guarantee uniformity over $c \in [c_+, \infty)$, we also use the fact that $\sup_{c_+ \leq c < \infty} |\zeta_c^- \mathsf{f}_u(c)| < \infty$ and $\sup_{c_+ \leq c < \infty} |c\mathsf{f}_u(c)| < \infty$ by Assumption 1.3.1$(ia, iiib)$. ■

Equipped with Lemma 1.C.3, we demonstrate the Gaussian weak limit for the initial step FODR of Algorithm 1.S with any split of subsamples $\mathcal{I}_1$ and $\mathcal{I}_2$. The result is presented in Theorem 1.3.3 ($S^{(0)}$).

Proof of Theorem 1.3.3. Using Assumption 1.3.1(ia, iia, iii, iv), we can invoke Lemma 1.C.3, which gives the expansion of the FODR for Algorithm 1.S at the initial step

$$\begin{aligned}
n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) = {} & n^{-1/2} \sum_{i=1}^{n} (1_{(|u_i| > \sigma c)} - \gamma_c) \\
& - (\frac{n_2}{n})^{1/2} (\frac{n_2}{n_1})^{1/2} c\mathsf{f}_u(c) n_1^{-1/2} \sum_{i \in \mathcal{I}_1} (\frac{u_i^2}{\sigma^2} - 1) \\
& - (\frac{n_1}{n})^{1/2} (\frac{n_1}{n_2})^{1/2} c\mathsf{f}_u(c) n_2^{-1/2} \sum_{i \in \mathcal{I}_2} (\frac{u_i^2}{\sigma^2} - 1) \\
& - (\frac{n_2}{n})^{1/2} (\frac{n_2}{n_1})^{1/2} \frac{\mathsf{f}_u(c)}{\sigma} (\zeta_c^- - 2c\Omega)' \Sigma^{1/2} (M_{\tilde{x}\tilde{x}, n_1}^{\mathcal{I}_1})^{-1} n_1^{-1/2} \sum_{i \in \mathcal{I}_1} \tilde{x}_i u_i \\
& - (\frac{n_1}{n})^{1/2} (\frac{n_1}{n_2})^{1/2} \frac{\mathsf{f}_u(c)}{\sigma} (\zeta_c^- - 2c\Omega)' \Sigma^{1/2} (M_{\tilde{x}\tilde{x}, n_2}^{\mathcal{I}_2})^{-1} n_2^{-1/2} \sum_{i \in \mathcal{I}_2} \tilde{x}_i u_i + \mathsf{o_P}(1),
\end{aligned}$$

uniformly in $c \in [c_+, \infty)$. Assumption 1.3.1(ia, iva, ivc) holds for each sub-sample $\mathcal{I}_j$, $j = 1, 2$, thus $n_j^{-1/2} \sum_{i \in \mathcal{I}_j} (u_i^2/\sigma^2 - 1)$, $j = 1, 2$, have the same limiting behaviour, while $n_j^{-1/2} \sum_{i \in \mathcal{I}_j} \tilde{x}_i u_i$, $j = 1, 2$, share the same weak limit. All their asymptotics are studied in Lemma 1.A.3, and there are no covariances across the two subsamples. Furthermore, note that $M_{\tilde{x}\tilde{x}, n_j}^{\mathcal{I}_j}$, $j = 1, 2$, have the same probability limit $M_{\tilde{x}\tilde{x}}$, and we have four ratios $n_1/n \to \alpha, n_2/n \to 1-\alpha, n_1/n_2 \to \alpha/(1-\alpha)$, and $n_2/n_1 \to (1-\alpha)/\alpha$. Then, we apply the same reasoning as in Theorem 1.3.2 ($R^{(0)}$) to get the finite dimensional convergence so that for any $c \in [c_+, \infty)$ and as $n \to \infty$

$$\mathbb{G}_n^{(0)}(c) \xrightarrow{\mathsf{D}} \mathsf{N}(0, K_{cc}^{(0)}),$$

where the variance $K_{cc}^{(0)}$ is given below

$$K_{cc}^{(0)} = \gamma_c(1-\gamma_c) - 2c\mathsf{f}_u(c)(1-\gamma_c-\tau_2^c) + \frac{\alpha^3+(1-\alpha)^3}{\alpha(1-\alpha)}c^2\mathsf{f}_u^2(c)(\tau_4-1)$$
$$+ \frac{\alpha^3+(1-\alpha)^3}{\alpha(1-\alpha)}\mathsf{f}_u^2(c)(\zeta_c^- - 2c\Omega)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}(\zeta_c^- - 2c\Omega).$$

Since the initial estimators are the split-sample 2SLS estimators, they are bounded in probability. Combined with Assumption 1.3.1(ia, iia, iii, iv), we can apply Theorem 1.3.1 to show tightness of the initial FODR processes $\mathbb{G}_n^{(0)}$. Combining the finite dimensional convergence and tightness, we get the weak convergence result

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

where the covariance $K_{st}^{(0)}$ for $c_+ \leq s \leq t < \infty$ is given below as

$$K_{st}^{(0)} = \gamma_t(1-\gamma_s) - s\mathsf{f}_u(s)(1-\gamma_t-\tau_2^t) - t\mathsf{f}_u(t)(1-\gamma_s-\tau_2^s)$$
$$+ \frac{\alpha^3+(1-\alpha)^3}{\alpha(1-\alpha)}st\mathsf{f}_u(s)\mathsf{f}_u(t)(\tau_4-1)$$
$$+ \frac{\alpha^3+(1-\alpha)^3}{\alpha(1-\alpha)}\mathsf{f}_u(s)\mathsf{f}_u(t)(\zeta_s^- - 2s\Omega)'\Sigma^{1/2}M_{\tilde{x}\tilde{x}}^{-1}\Sigma^{1/2}(\zeta_t^- - 2t\Omega). \qquad \blacksquare$$

Choosing $\alpha = 1/2$ for Algorithm 1.S, i.e. the split-half version, at the beginning of Lemma 1.C.3, we can demonstrate that the expansion of $n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c)$ is identical to that of Algorithm 1.R. The next theorem shows the equality of the expansions of $n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)$ for $m \in [1, \infty)$ between Algorithm 1.R and the split-half version of Algorithm 1.S.

Theorem 1.C.3. *Consider the split-half version of Algorithm 1.S, where the sample is partitioned into two halves, $\mathcal{I}_1$ and $\mathcal{I}_2$, with number of observations $n_1 = \text{int}[n/2]$ and $n_2 = n - n_1$. Suppose Assumption 1.3.1(ia, iia, iii, iv) holds for each subsample. Then, as $n \to \infty$ and for any $m \in [0, \infty)$, the process of the FODR $n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)$ for $c \in [c_+, \infty)$ has the same asymptotic expansion as that of Algorithm 1.R and thus weakly converges to the same Gaussian processes reported in Theorems 1.3.2 ($R^{(0)}$), 1.C.1 ($R^{(1)}$), and 1.C.2 ($R^{(m)}$ for $m \geq 1$).*

Proof of Theorem 1.C.3. We first derive the expansion of the initial FODR for Algorithm 1.S in the special case where the sample is split in half, i.e. $\alpha = 0.5$ such

that $n_1 = \text{int}[n/2]$ and $n_2 = n - n_1$. Use Assumption 1.3.1(ia, iia, iii, iv) to invoke Lemma 1.C.3. The expansion can be simplified by noting that $n_j/n \to 1/2$ and $n_j/n_{3-j} \to 1$ as $n \to \infty$ for $j \in \{1, 2\}$. In addition, although $M^{\mathcal{I}_j}_{\tilde{x}\tilde{x},n_j} \neq M_{\tilde{x}\tilde{x},n}$ for $j = 1, 2$, they still have the same probability limit $M_{\tilde{x}\tilde{x}}$ by the LLN and Assumption 1.3.1(iva). Thus, we have the expansion

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) = n^{-1/2} \sum_{i=1}^{n} (1_{(|u_i|>\sigma c)} - \gamma_c) - c\mathsf{f}_u(c) n^{-1/2} \sum_{i=1}^{n} (\frac{u_i^2}{\sigma^2} - 1) \\ - \frac{\mathsf{f}_u(c)}{\sigma} (\zeta_c^- - 2c\Omega)' \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \sum_{i=1}^{n} \tilde{x}_{in} u_i + \mathsf{o}_{\mathsf{P}}(1), \end{aligned}$$

uniformly in $c \in [c_+, \infty)$. We then immediately find that the expansion of $\widehat{\gamma}_c^{(0)}$ for Algorithm 1.S is asymptotically equivalent to that for Algorithm 1.R, and thus the two algorithms have the same weak convergence for $\widehat{\gamma}_c^{(0)}$. After the initial iteration, Algorithms 1.R and 1.S proceed identically so that they have the same asymptotic expansions of the iterated estimators, see Theorem 3.7 in Jiao (2022) and Lemma 1.B.3, as well as the same expansion of the FODR $\widehat{\gamma}_c^{(m)}$ for $m \geq 1$. In summary, for any $m \in [0, \infty)$ and uniformly in $c \in [c_+, \infty)$, the process of the FODR $n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)$ in its split-half version weakly converges to the Gaussian processes shown in Theorems 1.3.2 ($R^{(0)}$), 1.C.1 ($R^{(1)}$), and 1.C.2 ($R^{(m)}$ for $m \geq 1$). ■

Next, we demonstrate that the FODR of Algorithm 1.I started from Algorithms 1.R or 1.S in its split-half version converges in probability to a fixed point upon infinite iteration ($R^{(*)}$ and split-half $S^{(*)}$). The fixed point result was presented in Theorem 1.3.4.

Proof of Theorem 1.3.4. By Assumption 1.3.1(ia, iia, iii, iv), Theorems 1.C.2 ($R^{(m)}$ for $m \geq 1$) and 1.C.3 (split-half $S^{(m)}$ for $m \geq 0$) provide the weak convergence results for the FODR processes $\mathbb{G}_n^{(m)}(c) = n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)$ for $c \in [c_+, \infty)$ at any iterated step $m \geq 1$. After letting $m \to \infty$ we can immediately establish the fixed point $\mathbb{G}_n^{(*)}$ such that as $n \to \infty$

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

with the asymptotic covariance for any $c_+ \leq s \leq t < \infty$

$$
\begin{aligned}
K_{st}^{(*)} &= \gamma_t(1-\gamma_s) + \psi_s t \mathsf{f}_u(t) \varrho_{\sigma uu,t}^{(*)}(\varsigma_s^2 - \varsigma_t^2) + st\mathsf{f}_u(s)\mathsf{f}_u(t)\varrho_{\sigma uu,s}^{(*)}\varrho_{\sigma uu,t}^{(*)}(\tau_4^s - \tau_2^s\varsigma_s^2) \\
&+ \tau_2^s \mathsf{f}_u(s)\mathsf{f}_u(t)[\{\varrho_{\beta\tilde{x}u,s}^{(*)} + s\varrho_{\sigma\tilde{x}u,s}^{(*)}(s^2 - \varsigma_s^2)\}\zeta_s^- - \frac{2s}{\psi_s}(2s\varrho_{\sigma\tilde{x}u,s}^{(*)} + \varrho_{\sigma uu,s}^{(*)})\omega_s]' \\
&\times \Sigma^{1/2} M_{\tilde{x}\tilde{x}}^{-1} \Sigma^{1/2}[\{\varrho_{\beta\tilde{x}u,t}^{(*)} + t\varrho_{\sigma\tilde{x}u,t}^{(*)}(t^2 - \varsigma_t^2)\}\zeta_t^- - \frac{2t}{\psi_t}(2t\varrho_{\sigma\tilde{x}u,t}^{(*)} + \varrho_{\sigma uu,t}^{(*)})\omega_t],
\end{aligned}
$$

where the coefficients have expressions

$$
\begin{aligned}
\varrho_{\beta\beta,c}^{(*)} &= 0, & \varrho_{\beta\tilde{x}u,c}^{(*)} &= \frac{1}{\psi_c - 2c\mathsf{f}_u(c)}, \\
\varrho_{\sigma\sigma,c}^{(*)} &= 0, & \varrho_{\sigma uu,c}^{(*)} &= \frac{1}{\tau_2^c - c\mathsf{f}_u(c)(c^2 - \varsigma_c^2)}, \\
\varrho_{\sigma\beta,c}^{(*)} &= 0, & \varrho_{\sigma\tilde{x}u,c}^{(*)} &= \frac{\mathsf{f}_u(c)}{\{\psi_c - 2c\mathsf{f}_u(c)\}\{\tau_2^c - c\mathsf{f}_u(c)(c^2 - \varsigma_c^2)\}}.
\end{aligned}
$$

The covariance expression shown in Theorem 1.3.4 is obtained by further simplifying and collecting the various $\varrho^{(*)}$ terms in the above expression for $K_{st}^{(*)}$. ■

We have now completed all the proofs for the Gaussian approximation results for the FODR. The final part of this appendix proves Theorem 1.3.5, which establishes a Poisson approximation to the number of falsely discovered outliers, which applies to all Algorithms $1.I$, $1.R$, and $1.S$ at any iteration step $m \in [0, \infty)$.

Proof of Theorem 1.3.5. Assumption 1.3.1(ia) implies $\mathsf{E}|u_i/\sigma|^l < \infty$ for some $l > 4$. Apply (1.3.1) and Chebyshev's inequality to get $\lambda/n = \mathsf{P}(|u_i| > \sigma c_n) \leq \mathsf{E}|u_i/\sigma|^l c_n^{-l}$. Thus, $c_n \leq (\mathsf{E}|u_i/\sigma|^l)^{1/l}\lambda^{-1/l}n^{1/l}$ 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, iia, iii, iv, v) with $\eta = 1/4$, Theorem 3.3 in Jiao (2022) shows the tightness of $\widehat{\beta}_{c_n}^{(m)}$, $(\widehat{\sigma}_{c_n}^{(m)})^2$. Assumption 1.3.1($iib$) shows $\max_{1\leq i\leq n}|r_i| = \mathrm{O}_{\mathsf{P}}(n^{\kappa}) = \mathrm{o}_{\mathsf{P}}(n^{1/4})$ for some $0 \leq \kappa < 1/4$. As $\tilde{x}_{in} = \Pi' z_{in}$, Assumption 1.3.1($ivb$) gives $\max_{1\leq i\leq n}|\tilde{x}_{in}| = \mathrm{O}_{\mathsf{P}}(n^{\kappa-1/2}) = \mathrm{o}_{\mathsf{P}}(n^{-1/4})$. Thus, for all $\epsilon > 0$ there exists a large constant A_0 so that the set

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

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

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

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

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

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

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

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

A similar argument shows

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

Thus, we get the lower and upper bounds for the indicators uniformly in m and i such that

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

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

$$n\mathsf{P}(|u_i/\sigma| > c_n + n^{-1/4} A_1) \to \lambda, \qquad n\mathsf{P}(|u_i/\sigma| > c_n - n^{-1/4} A_1) \to \lambda. \tag{1.C.2}$$

Since $n\mathsf{P}(|u_i/\sigma| > c_n) = \lambda$ by (1.3.1), it suffices to show that

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

Note $|u_i/\sigma| \sim \mathsf{g}_u, \mathsf{G}_u$ and $u_i/\sigma \sim \mathsf{f}_u, \mathsf{F}_u$ such that $\mathsf{g}_u = 2\mathsf{f}_u 1_{[0,\infty)}$, $\mathsf{G}_u = (2\mathsf{F}_u - 1)1_{[0,\infty)}$. By this and (1.3.1), $2\{1 - \mathsf{F}_u(c_n)\} = \lambda/n$. Apply the mean value theorem to $\mathcal{P}_n$ and use the above identity to obtain

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

where $|\tilde{c} - c_n| \leq n^{-1/4} A_1$. Then, we rearrange the whole term to get

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

Since $c_n - n^{-1/4}A_1 \leq \tilde{c}$ and f_u has a 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 that the second and third ratio are bounded. Then, use $n^{-1/4}c_n = \mathrm{o}(1)$ to get $\mathcal{P}_n = \mathrm{o}(1)$.

4. *Poisson approximation.* On the set $\mathcal{A}_n$, apply (1.C.1) to obtain

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

Using (1.C.2), the Poisson central limit theorem shows that the lower and upper bounds converge to the Poisson distribution with the parameter λ. Thus, as $n \to \infty$ then $n\widehat{\gamma}_{c_n}^{(m)} \xrightarrow{\mathsf{D}} \text{Poisson}(\lambda)$ for $0 \leq m < \infty$. ■

1.D Additional Simulations

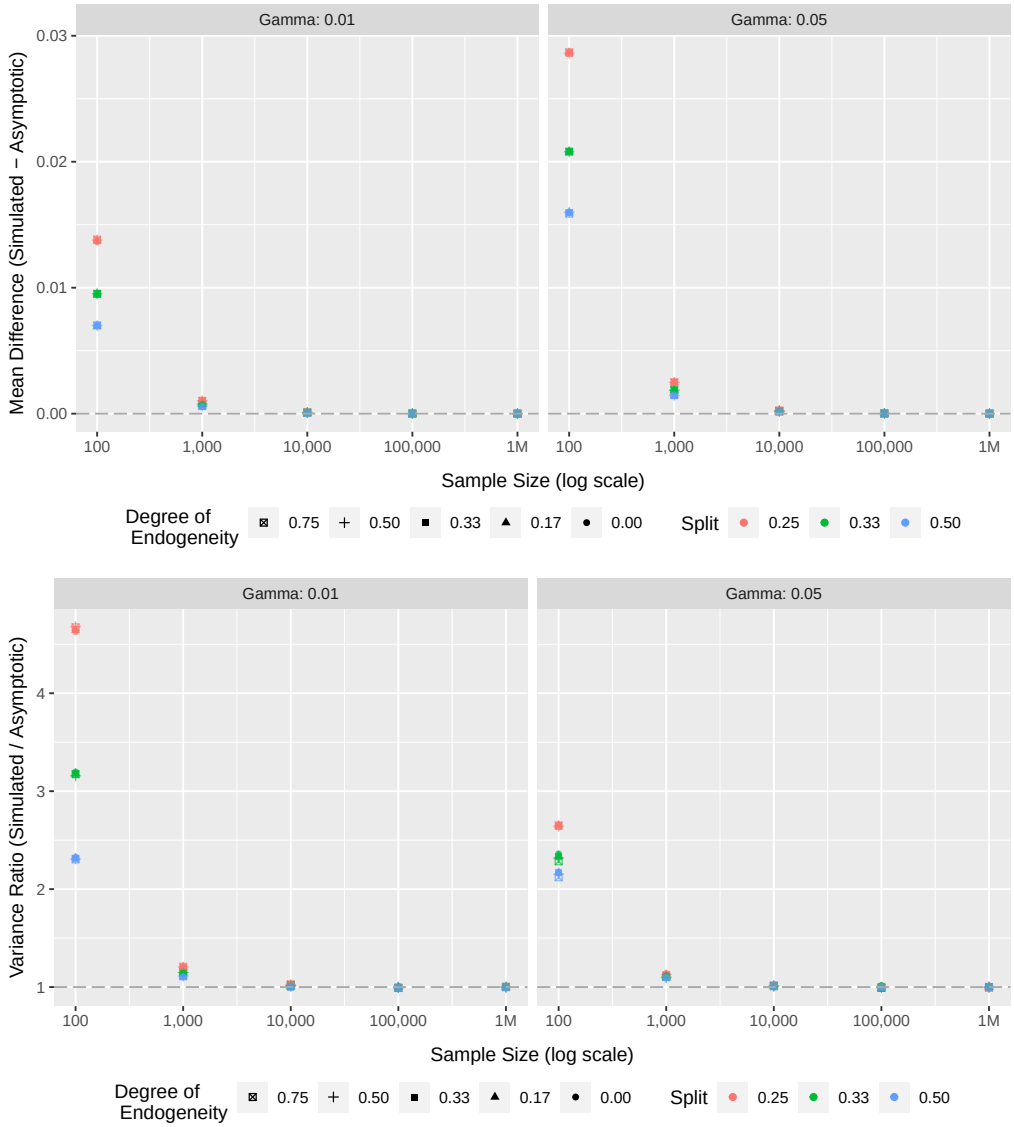
We present two additional simulation results here, which lend further support to our asymptotic theory. The first set of Monte Carlo experiments is to assess the finite sample performance of the Gaussian approximation to the FODR of Saturated 2SLS Algorithm 1.S with unequal splits α when $m = 0$. Second, we evaluate the convergent iterations of the algorithms.

1.D.1 Performance of Saturated 2SLS with Unequal Splits

We now focus on the performance of Saturated 2SLS Algorithm 1.S when $m = 0$. For this case, we derived the asymptotic distribution of the FODR in Theorem 1.3.3 ($S^{(0)}$) for general split shares $\alpha \in (0, 1)$. The two panels in Figure 1.D.1 are analogous to the figures in Section 1.4.1 except that they show the performance for three different splits, $\alpha \in \{1/4, 1/3, 1/2\}$.

We find that both the deviation from the theoretical mean and from the theoretical variance is larger for more unequal splits, especially in small samples. Splitting a sample to obtain the initial subsample estimates $\widehat{\beta}_j$ and $\widehat{\sigma}_j$ already means that the estimates are based on a smaller subsample and are therefore further away from the asymptotic approximation. Splitting the sample unequally exacerbates this problem for one of the subsample estimates. Therefore, the *robust2sls* R package (Kurle 2022b) raises a warning for very unequal splits and warns the user that this should

Figure 1.D.1: Evaluation of the FODR: Saturated 2SLS, $m = 0$



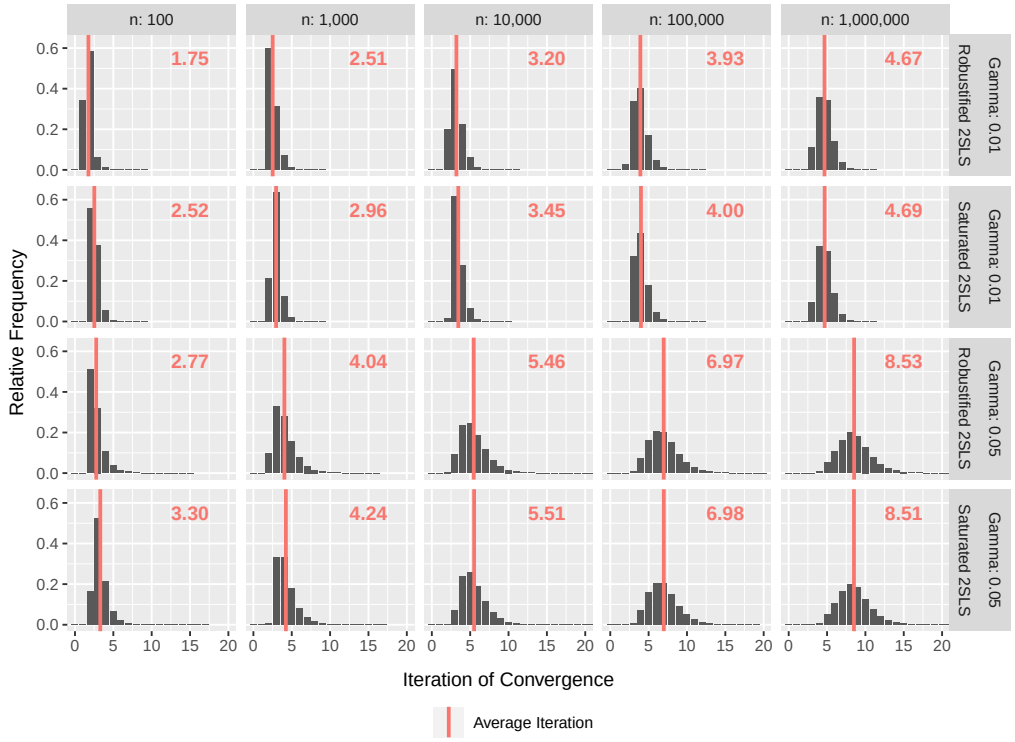
Comparison of the theoretical asymptotic mean and variance of the FODR with the simulated mean and variance of the FODR. All simulations refer to the initial step ($m = 0$) of Saturated 2SLS. Several settings are varied: the split α, the degree of endogeneity Ω_2 in the DGP, the cut-off c (or equally γ_c), and the sample size n. All Monte Carlo simulations used $W = 50{,}000$ replications.

only be done for large samples. For tight significance levels, such as $\gamma_c = 0.01$, a sample size of 1000 might already be enough to be close to the asymptotic approximation but a larger sample size tends to be required for looser significance levels.

1.D.2 Evaluation of the Convergent Iterations

The algorithms were supposed to be iterated until $\|\widehat{\beta}_c^{(m)} - \widehat{\beta}_c^{(m-1)}\|_2 = 0$, which is only fulfilled when the updated estimate and its previous estimate coincide exactly. In practice, this will be the case when the algorithm has converged and the same subsample was used in iteration m and $m-1$. To ensure that the algorithm stops, we add an additional stopping criterion of reaching a maximum of 50 iterations.

Figure 1.D.2: Evaluation of the Convergent Iterations



Histograms of the iteration that was the convergent iteration. The red line and number represent the average number of iterations until the algorithm converged. The histograms are cut off for values larger than 20. The plots represent different algorithms, significance levels γ_c, and sample sizes n. The degree of endogeneity in the DGP is $\Omega_2 = 0.75$. Each plot is based on $W = 50{,}000$ replications.

There does not seem to be a lot of difference between degrees of endogeneity, so that Figure 1.D.2 only shows the results for the highest degree of endogeneity, $\Omega_2 = 0.75$.

We find several convergence patterns. First, the iteration at which the algorithm converges is broadly the same for Robustified 2SLS and for Saturated 2SLS. It is slightly higher for Saturated 2SLS but the difference vanishes as the sample size increases. Second, the number of iterations required for convergence is higher as the

significance level increases. Third, convergence is also reached later as the sample size increases.

The histograms in Figure 1.D.2 are cut off for iterations higher than 20, which only reflect a very small share of iterations in each Monte Carlo setting. Out of the 20 settings, only four settings recorded iterations larger than 20. All of them were for $\gamma_c = 0.05$ but they represented a tiny share of all 50,000 iterations. The highest share of convergent iterations above 20 was recorded for split-half Saturated 2SLS, $\gamma_c = 0.05$, $n = 1{,}000{,}000$, and only represented 0.024% of iterations. In fact, none of the $20 \times 50{,}000 = 1{,}000{,}000$ replications presented in Figure 1.D.2 reached the limit of 50 iterations, so the replications all converged before iteration $m = 50$ for this degree of endogeneity $\Omega_2 = 0.75$.

Considering results for other degrees of endogeneity, we also record very few high iterations. Overall, only 0.0011% of iterations were stopped at $m = 50$ rather than converging before. The maximum share of iterations that reached the maximum number of 50 iterations within a single Monte Carlo setting was 0.016%. Hence, the effect of iterations that have not fully converged is negligible and all results can be interpreted as results for the convergence of the algorithms.

1.E Robustness of Algorithms

We do not formally analyse the robustness properties of Algorithms 1.R and 1.S. Future work could assess their robustness properties in terms of concepts that were developed in the robust statistics literature, such as the estimators' influence function (Hampel 1968, 1974; Hampel et al. 1986) or breakdown point (Hampel 1968; Donoho and Huber 1983). Due to the presence of indicator functions in the algorithms we analyse, this is not a trivial exercise. Instead, we explain the intuition why Algorithm 1.R is expected to be less robust than Algorithm 1.S in certain settings and provide preliminary simulation results.

1.E.1 Intuition for Differences in Robustness

First, note that Algorithm 1.R is likely to be not robust since it is based on the classical full-sample 2SLS estimator, which is not itself robust. Zhelonkin, Genton and Ronchetti (2012) show that the classical 2SLS estimator has an unbounded influence function in the dependent variable of the structural equation (y), the exogenous and endogenous regressors (x_1, x_2), and the excluded instruments (z_2). This means that an infinitesimal deviation from the true distribution of the data can in the worst case induce an infinite change in the estimate of the location parameter β. Hence, even the presence of a single outlier may cause poor performance.

Imagine that an outlier causes the initial 2SLS estimate $\widehat{\beta}^{(0)}$ to be far from the true β, such that the calculated residuals are far from the true errors u and the classification of outliers is poor — not detecting the true outliers and potentially classifying "good" observations as outliers. Subsequent iterations of Algorithm 1.R may not improve upon the initial classification and the updated β-estimates might remain far from their true value.

Algorithm 1.S is more robust than Algorithm 1.R when the outliers are fully contained in one of the subsamples into which Algorithm 1.S partitions the sample. Remember that Algorithm 1.S first partitions the sample into two subsamples, then estimates β and σ separately on each subsample, and finally classifies outliers in one subsample based on the residuals using the estimates from the other subsample. The subsample estimates are each subject to the same lack of robustness as the full-sample estimates. However, intuitively, the updated ($m \geq 1$) estimator of Algorithm 1.S is robust when the outlier contamination is contained in only one of the two subsamples.

Suppose that there is at least one outlier in the first half of the sample but none in the second half. The parameter estimates $(\widehat{\beta}_1, \widehat{\sigma}_1^2)$ in the first half may be severely affected by the outlier(s) and hence the classification of observations as outliers in the second half may be very wrong. Since the second half only consists of "good" observations, some (or in the worst case all) might be falsely classified as outliers. This is inefficient but does not cause true outliers to be classified as "good"

observations that remain in the sample. Turning to the parameter estimates of the second half, $(\widehat{\beta}_2, \widehat{\sigma}_2^2)$ are consistent for the true parameters β, σ^2 because the second half contains no outliers. Therefore, the residuals of the first half, calculated using the estimates of the second half, are consistent for the true error and asymptotically, we classify the true outliers as outliers. Combining the classification of the first and second half, we obtain a subsample of non-outlying observations for the next iteration $m = 1$, which only contains "good" observations.

Algorithm 1.S is a stylised version of IIS (Hendry, Johansen and Santos 2008), the latter likely being robust in even more settings. The main shortcoming of Algorithm 1.S is that its search for outliers is very simple by partitioning the sample into only two parts. If both partitions contain outliers, the iterated estimates are also not robust. IIS uses multiple block searches, which might allow the algorithm to detect outliers even when they are spread out across the sample.

1.E.2 Monte Carlo Illustration

The following Monte Carlo simulations confirm the expected robustness properties of the two Algorithms. The structural equation of our DGP is given as

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad \text{for } i = 1, \ldots, n, \tag{1.E.1}$$

where the true parameters are $\beta_0 = \beta_1 = 0$. The endogenous regressor x_i is defined by the first stage projection

$$x_i = \pi z_i + r_i, \quad \text{for } i = 1, \ldots, n, \tag{1.E.2}$$

with $\pi = 1$ and x_i positively correlated with the structural error u_i. The instrument z_i is uniformly distributed, independently of the normally distributed errors:

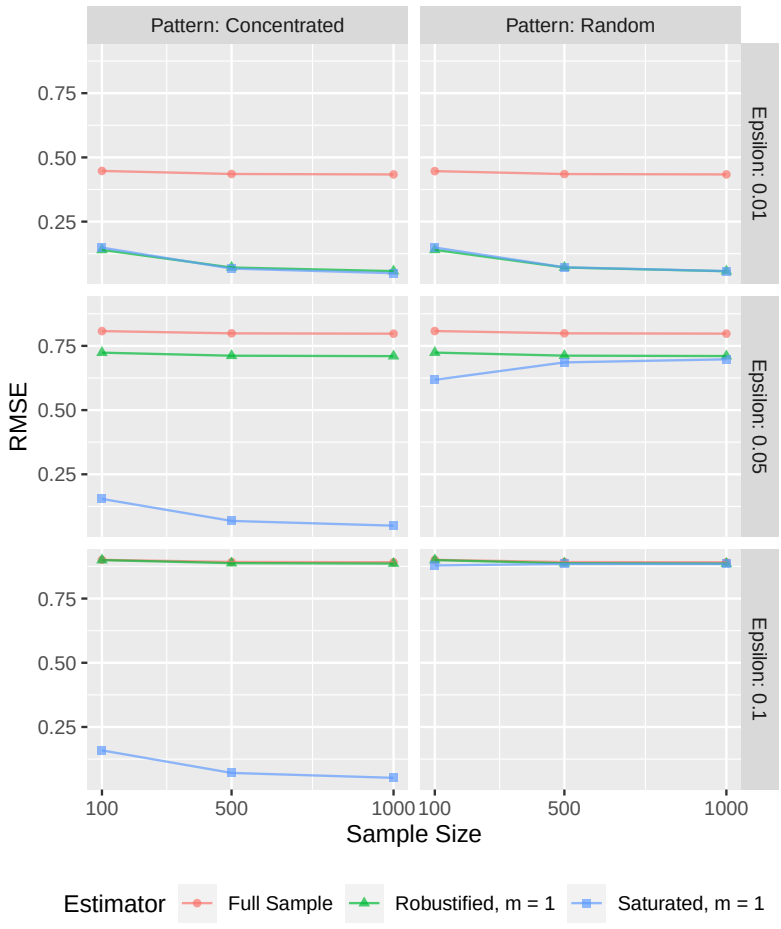
$$\begin{aligned} z_i &\overset{\mathsf{D}}{=} \mathsf{U}[-2, 2], \\ \begin{pmatrix} u_i \\ r_i \end{pmatrix} &\overset{\mathsf{D}}{=} \mathsf{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right\}. \end{aligned}$$

To ensure that the outliers exert a strong influence on the classical 2SLS estimator, we create the outliers as bad leverage points which are not only outliers in the

y-direction but also in z and x. The outliers are defined by $(z, y) = (10, 10)$ and $r \overset{\mathsf{D}}{=} \mathsf{N}(0, 1)$.

We consider two contamination patterns and three contamination intensities (ϵ). In the "concentrated" pattern, we contaminate only observations in the first half of the sample. In the "random" pattern, we randomly select indices throughout the whole sample and contaminate those observations. It is therefore likely that outliers are present in more than one half of the sample. As for the contamination intensities, we contaminate 1%, 5%, or 10% of the sample. The sample size varies across $n \in \{100, 500, 1000\}$.

Figure 1.E.1: Comparison of RMSE



Comparison of the root mean square error (RMSE) of the parameter estimates of β for (i) different estimators (full-sample 2SLS, step $m = 1$ Robustified 2SLS, step $m = 1$ split-half Saturated 2SLS), (ii) different contamination patterns (concentrated in one half, randomly distributed across the sample), (iii) different shares of contamination (1%, 5%, 10%), and (iv) different sample sizes (100, 500, 1000). Each simulation is based on $W = 10{,}000$ replications.

To evaluate the performance of the different algorithms, we first compare their root mean square error (RMSE): $RMSE = \sqrt{\frac{1}{W}\sum_{w=1}^{W}(\widehat{\beta}_{0,w}^2 + \widehat{\beta}_{1,w}^2)}$, where $w = 1, \ldots, 10{,}000$ indicates the replication of the Monte Carlo simulation.

As a benchmark, we report the RMSE of the classical full-sample 2SLS estimator. We also calculate the RMSE of the step $m = 1$ estimates of Algorithm $1.R$ (Robustified 2SLS) and Algorithm $1.S$ (Saturated 2SLS), which are the updated estimates after having removed outliers from the sample. The detection significance level was chosen as $\gamma_c = 0.05$, using a cut-off value of $c = 1.96$ assuming normality. The results are presented in Figure 1.E.1.

We find that for all contamination settings, the RMSE of the full-sample 2SLS estimates is the highest and it increases in the share of outliers. Robustified 2SLS performs better but loses its advantage over the full-sample estimator as the share of outliers increases. There is almost no difference between the contamination patterns. Focusing on the "concentrated" pattern where the outliers only occur in one half of the sample, however, we see that Saturated 2SLS performs much better than the other two estimators, even when the share of outliers is high. When the sample is contaminated randomly, Saturated 2SLS performs similarly to Robustified 2SLS and also yields no improvement over the full-sample estimator when the contamination becomes too large.

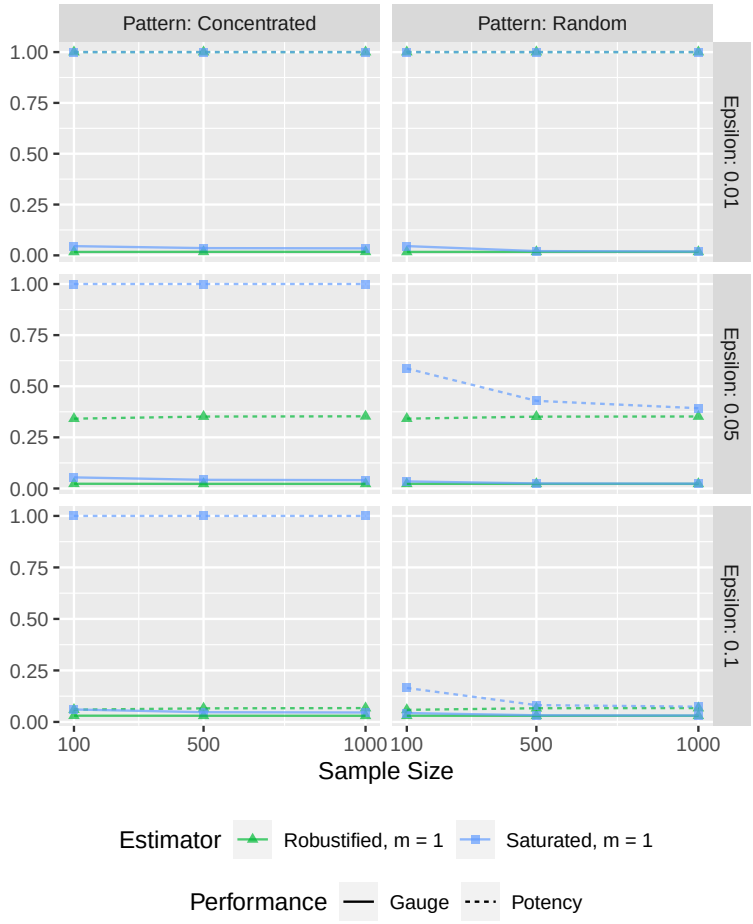
The results confirm our intuition that Saturated 2SLS performs better than Robustified 2SLS when the contamination is fully contained in one of the sample splits. Random contamination means that both sample splits are likely to contain at least some outliers, making Saturated 2SLS no better than Robustified 2SLS. In this case, we would expect IIS to likely outperform the two algorithms.

In addition to the RMSE, we also show the gauge and potency, concepts introduced in Castle, Doornik and Hendry (2011) to evaluate the performance of model or outlier detection. The gauge measures the share of falsely detected outliers. It coincides with what we call the FODR (see Section 1.2.3). While we analyse its asymptotic properties when there are no outliers in the sample, here we apply the concept to a contaminated setting. Potency refers to the share of correctly detected

outliers. To fix ideas, suppose the sample size is $n = 100$ and five observations are true outliers. If the algorithm detected a total of six outliers, four of the true outliers and two which are not true outliers, we would obtain a potency of $4/5 = 0.8$ and a gauge of $2/95 \approx 0.02$.

Figure 1.E.2 shows the gauge (solid line) and potency (dashed line) of the initial outlier classification $v_{i,c}^{(0)}$ (see equation (1.2.13) for Algorithm 1.R and equation (1.2.21) for Algorithm 1.S). The updated step $m = 1$ estimates and their RMSE shown in Figure 1.E.1 are based on this classification. In all contamination settings and for both algorithms, the gauge is close to or even below 0.05. The potency is very high — close to or equal to 1 — when only a small share of observations is contaminated ($\epsilon = 0.01$). It remains high for Algorithm 1.S when the outliers are concentrated in one half of the sample but it is low for Algorithm 1.R, so Saturated 2SLS correctly detects the outliers while Robustified 2SLS does not. This explains its superior performance in terms of the RMSE shown in Figure 1.E.1. When the outliers are randomly allocated in the sample and $\epsilon \geq 0.05$, the potency is low and falling in ϵ for both algorithms.

Figure 1.E.2: Comparison of Gauge and Potency



Comparison of the gauge and potency for (i) different estimators (step $m = 1$ Robustified 2SLS, step $m = 1$ split-half Saturated 2SLS), (ii) different contamination patterns (concentrated in one half, randomly distributed across the sample), (iii) different shares of contamination (1%, 5%, 10%), and (iv) different sample sizes (100, 500, 1000). Each simulation is based on $W = 10{,}000$ replications.

CHAPTER 2

Testing for Outliers in Robust Two-Stage Least Squares Regressions*

Abstract Recent literature has shown that 2SLS estimates are highly sensitive to outliers. This chapter proposes two types of test statistics for the presence of outliers in 2SLS regression models. The first type of test takes the outlier frequency into account while the second type also accounts for outlier magnitude. We derive the asymptotic distribution of the test statistics assuming a normal reference distribution for outlier classification. Monte Carlo simulations compare the performance of the different tests. The results show that the size can be well-controlled and that the power of the tests tends to increase both in the outlier frequency and outlier magnitude. An empirical illustration shows the usage of the tests, which are available from the R package *robust2sls*.

∗. We are grateful to participants at Dynamic Econometrics 2023 and IAAE 2023 for helpful comments and discussions. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (Richards 2015).

2.1 Introduction

Recent literature has shown that two-stage least squares (2SLS) estimates are highly sensitive to outliers. Young (2022) analyses the results from over 1300 2SLS regressions and finds that often the statistical significance and even the sign of the estimates change when removing just one or two observations from the sample. Similarly, Broderick, Giordano and Meager (2023) find that some empirical results are strongly driven by a small share of observations.

To show robustness of empirical research results, it is therefore important to assess whether the sample contains outliers that might drive the results. Applied researchers often use simple outlier-robust heuristics to obtain robust estimates and to assess the sensitivity of their results.

This chapter contributes to the issue of outlier detection by proposing statistical tests for the presence of outliers in 2SLS regression models. We develop the tests for the classical textbook 2SLS regression model with homoskedastic errors because it is a common starting point for empirical researchers in order to address endogeneity. Both the order and the rank condition are assumed to be fulfilled, and we abstract from any weak or many instrument issues. We derive the asymptotic distribution of the test statistics under assumptions that are fulfilled in most cross-sectional settings and for the most common choice of normally distributed errors as reference distribution for outlier classification. Extensive Monte Carlo simulations for the finite sample performance show that the size can be controlled for all tests and that the tests achieve high power.

The tests build on popular outlier robust heuristics based on the size of regression residuals, which many applied researchers already use (see for example Albouy 2012; Acemoglu et al. 2019). These procedures first run 2SLS on the full sample and compare the standardised residuals against a chosen cut-off value from a reference distribution. Residuals beyond that cut-off value are unlikely to have occurred under the reference distribution and are therefore classified as outliers and removed from the sample. Subsequently, the model is re-estimated by 2SLS on the remaining

subsample of non-outlying observations. We improve and formalise the heuristics into three different algorithms. The algorithms differ in the choice of initial estimators that yield the standardised residuals, based on which observations are then classified as outliers.

The algorithms are expected to classify a certain share of observations as outliers even when the sample contains no outliers. The reason is that even the "good" reference distribution can produce unusually large errors despite being unlikely. For example, the normal reference distribution produces absolute values larger than 1.96 in five percent of the draws. These values are expected to be incorrectly classified as outliers because they come from the reference distribution and are not true outliers.

The underlying idea is then to test whether the actual share of detected outliers is significantly larger than the expected share under the null hypothesis of no outliers for a given reference distribution, which is the normal distribution in this chapter. To judge whether the difference is statistically significant, we derive the asymptotic distribution of the test statistics assuming a joint normal distribution of the errors and certain moment conditions that are fulfilled in many cross-sectional settings.

We suggest two different types of test statistics for the presence of outliers in 2SLS regression models, called "simple" and "scaling" tests in reference to Jiao and Pretis (2022). Simple tests are based on a single cut-off value and compare the actual frequency of detected outliers to the expected frequency when there is no outlier contamination. Scaling tests not only take the relative frequency but also the magnitude of outliers into account by running a scale (range) of cut-off values. Using multiple cut-off values avoids the sometimes arbitrary choice of a single cut-off value, potentially leading to higher power for finding contaminated samples.

The asymptotic distribution of the test statistics is derived assuming a joint normal distribution of the structural and the first stage errors. In practice, researchers need to choose a reference distribution against which observations are judged to be outliers or not. Without such a reference distribution, the concept of an outlier, i.e. of an "unusual" draw, is ill-defined. We focus on the normal distribution because this seems to be the most common (implicit or explicit) choice of empirical

researchers. The theory developed in Chapter 1 allows for other choices of reference distribution and future research could expand the tests to other distributional assumptions.

The performance of the tests is assessed for two different contamination scenarios. First, we consider the ϵ-contamination framework introduced by Huber (1964). In this case, the structural error is drawn from a normal reference distribution with probability $1 - \epsilon$. With probability ϵ, however, the structural error is drawn from a discrete contaminating distribution that takes on values that are unlikely under a normal distribution. The second contamination scenario considers leverage points (data points that are not only outliers in the y- but also the x-direction) and is adapted from the Least Trimmed Squares (LTS) contamination introduced in Berenguer-Rico and Nielsen (2023a) and Berenguer-Rico, Johansen and Nielsen (2023).

The Monte Carlo simulations show that the size can be well-controlled for all tests and algorithm settings. The power of the tests increases with the sample size, outlier magnitude, and outlier frequency. If the outlier frequency becomes too high, the simple tests start to lose power again but the scaling tests keep high power.

We illustrate the usage of the different tests with the seminal work of Angrist and Krueger (1991). Our proposed tests for the presence of outliers reject the null hypothesis of no outliers, supported by both the simple and the scaling tests. The robust parameter estimates of the return to education tend to be smaller than the original estimates but are still statistically significant.

The tests that we propose in this chapter and the outlier detection algorithms are readily available in the companion R package *robust2sls* (Kurle 2022b), making it easy for applied researchers to use them. The researcher can specify the cut-off value, one of three algorithms for outlier detection, and the number of iterations using a single command. Subsequently, they can use the commands for the simple and the scaling tests to test the null hypothesis of no contamination.

This chapter builds on Jiao and Pretis (2022) by extending their proposed tests for the presence of outliers from the ordinary least squares (OLS) regression setting

to the 2SLS regression setting. Our work is also related to the following literature. Jiao (2022) and Kaji (2018) provide Hausman-type tests to test whether the structural parameters estimated by the ordinary and the robust 2SLS regression are significantly different. So rather than testing for the presence of outliers directly, they provide a test for whether the outliers have a significant effect on the parameter estimates. Jiao, Pretis and Schwarz (2024) provide such a test on the coefficient estimates for the OLS case. Testing for the presence of outliers can also be useful because outlier removal affects subsequent misspecification tests, such as tests for normality (Berenguer-Rico and Nielsen 2023b) or homoskedasticity (Berenguer-Rico and Wilms 2021).

The chapter is structured as follows: Section 2.2 introduces the notation that we use for the classical textbook 2SLS regression model. Section 2.3 explains the underlying testing idea and provides an overview of the different tests. Section 2.4 presents the formalised outlier detection algorithms. Section 2.5 states the required assumptions for the asymptotic theory and derives the asymptotic distributions of the test statistics. Section 2.6 evaluates extensive Monte Carlo simulations to assess the size and power performance of the tests in finite samples and compares them. Section 2.7 discusses the practical choices of the testing procedures that researchers need to make, taking both theoretical aspects and the simulation results into account. Section 2.8 illustrates the usage of the tests by revisiting the seminal work of Angrist and Krueger (1991) and Section 2.9 concludes.

2.2 Classical 2SLS Regression Model and Notation

We develop the statistical tests for the classical textbook 2SLS regression model because it is a common starting point for empirical researchers in order to address endogeneity. We assume that both the order and the rank condition are fulfilled, abstracting from any weak or many instrument issues. Furthermore, both the structural errors and the first stage errors are assumed to be homoskedastic. For the reference distribution, we assume joint normality of the errors.

The notation follows the exposition in Chapter 1. Suppose we have independently and identically distributed cross-sectional data $\{(y_i, x_i, z_i)\}_{i=1}^n$, where $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^{d_x}$, and $z_i \in \mathbb{R}^{d_z}$. The structural equation that we want to estimate is

$$y_i = x_i'\beta + u_i, \quad i = 1, 2, \ldots, n. \tag{2.2.1}$$

We partition the set of regressors into a set of exogenous regressors $x_{1i} \in \mathbb{R}^{d_{x_1}}$ for which $\mathsf{E}x_{1i}u_i = 0_{d_{x_1}}$, and a set of endogenous regressors $x_{2i} \in \mathbb{R}^{d_{x_2}}$ for which $\mathsf{E}x_{2i}u_i \neq 0_{d_{x_2}}$. Suppose we have a set of instrumental variables $z_i \in \mathbb{R}^{d_z}$ available, which consists of the exogenous regressors of the structural equation and a set of excluded instruments $z_{2i} \in \mathbb{R}^{d_{z_2}}$, such that we have $z_i = (x_{1i}', z_{2i}')'$. The first stage equation is given by

$$x_i = \Pi' z_i + r_i \iff \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} I_{d_{x_1}} & 0_{d_{x_1} \times d_{z_2}} \\ \Pi_1 & \Pi_2 \end{pmatrix} \begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix} + \begin{pmatrix} 0_{d_{x_1}} \\ r_{2i} \end{pmatrix}, \quad i = 1, 2, \ldots, n, \tag{2.2.2}$$

where the exogenous regressors can be fitted perfectly and their corresponding errors are therefore zero.

We assume that the order condition is fulfilled, i.e. that $d_{z_2} \geq d_{x_2}$. Moreover, the instruments are assumed to be valid, $\mathsf{E}z_i u_i = 0_{d_z}$, and informative, which can be written as the rank condition $\operatorname{rank}\Pi = d_x$ and $\operatorname{rank}\mathsf{E}z_i z_i' = d_z$. The instruments are assumed to be orthogonal to the first stage errors, i.e. $\mathsf{E}z_i r_i' = 0_{d_z \times d_x}$.

In this chapter, we focus on errors with a joint normal distribution as the reference distribution — an assumption that is often made by empirical researchers when checking for outliers (see for example Acemoglu et al. 2019). The i.i.d. structural error u_i is assumed to be mean zero with variance σ^2 and we use $\phi_u(y)$ to denote the standard normal density of the scaled error u_i/σ with distribution function $\Phi_u(y) = \mathsf{P}(u_i/\sigma \leq y)$. The first stage error is also assumed to be mean zero with dispersion matrix $\Sigma \in \mathbb{R}^{d_x \times d_x}$. The standard normal density of the scaled first stage error $\Sigma^{-1/2} r_i$ is denoted by $\phi_r(x)$ and its distribution function by $\Phi_r(x)$. The joint distribution of the scaled structural and first stage errors $(u_i/\sigma, \Sigma^{-1/2} r_i)$ is denoted by $\phi_{u,r}(y, x)$ and $\Phi_{u,r}(y, x)$ accordingly. We use $\Omega = \mathsf{Cov}(u_i/\sigma, \Sigma^{-1/2} r_i) \in \mathbb{R}^{d_x}$ to

denote the covariance between the scaled structural error and the scaled first stage errors, which is generally non-zero due to the presence of endogeneity.

We can hence write the joint distribution of the scaled errors as follows

$$\begin{pmatrix} u_i/\sigma \\ 0_{d_{x_1}} \\ \Sigma_2^{-1/2} r_{2i} \end{pmatrix} \overset{\mathsf{D}}{=} \mathsf{N}\left\{ \begin{pmatrix} 0 \\ 0_{d_{x_1}} \\ 0_{d_{x_2}} \end{pmatrix}, \begin{pmatrix} 1 & 0'_{d_{x_1}} & \Omega_2' \\ 0_{d_{x_1}} & 0_{d_{x_1}\times d_{x_1}} & 0_{d_{x_1}\times d_{x_2}} \\ \Omega_2 & 0_{d_{x_2}\times d_{x_1}} & I_{d_{x_2}} \end{pmatrix} \right\}, \tag{2.2.3}$$

where r_{2i} is always exactly zero and therefore also has a variance of zero. Furthermore, we use $\psi_c = \mathsf{P}(|u_i| \leq \sigma c)$ to denote the probability of the scaled structural error lying inside the interval $[-c, c]$.

Under the normal distribution, the asymptotic theory of the gauge developed in Chapter 1 simplifies as shown in Appendix 2.A, such that only the following moments appearing in the formulas of the test statistics need to be introduced. We define the k-th moment and truncated moment of the scaled structural error as

$$\tau_k = \int_{-\infty}^{\infty} y^k \phi_u(y) dy, \qquad \tau_k^c = \int_{-c}^{c} y^k \phi_u(y) dy. \tag{2.2.4}$$

The variance of the scaled structural error conditional on being smaller than the cut-off value c is defined as

$$\varsigma_c^2 = \frac{\tau_2^c}{\psi_c} = \frac{\int_{-c}^{c} y^2 \phi_u(y) dy}{\mathsf{P}(|u_i| \leq \sigma c)}, \tag{2.2.5}$$

which also acts as the bias correction factor for the variance estimator in the algorithms presented in Section 2.4. The conditional expected value of the scaled first stage error given that the scaled structural error is equal to the cut-off value c is defined as

$$\xi_y = \mathsf{E}(\Sigma^{-1/2} r_i | u_i/\sigma = y) = \int_{\mathbb{R}^{d_x}} x \phi_{r|u}(x|y)(dx). \tag{2.2.6}$$

For convenience, we also introduce the notation $\zeta_y^+ = \xi_y + \xi_{-y}$ and $\zeta_y^- = \xi_y - \xi_{-y}$. Finally, we introduce the truncated covariance between the scaled structural and first stage error

$$\omega_c = \mathsf{E}(\Sigma^{-1/2} r_i \frac{u_i}{\sigma} 1_{(|u_i| \leq \sigma c)}) = \iint_{y\in\mathbb{R}, x\in\mathbb{R}^{d_x}} xy 1_{(|y|\leq c)} \phi_{u,r}(y, x) dy(dx), \tag{2.2.7}$$

where (dx) describes the exterior product $dx_1 dx_2 \cdots dx_k$ of a k-dimensional multivariate integral.

Having defined the notation, the next section describes the testing idea, the underlying concepts, and the test statistics.

2.3 Test Statistics

2.3.1 Outlier Share and Outlier Count

The tests are based on algorithms that classify observations with large standardised residuals as outliers. The algorithms can be iterated and are introduced in detail in Section 2.4. Each of the algorithms yields a classification of observations into outliers and non-outliers. This classification is represented by an indicator function $v_{i,c}^{(m)}$ which takes on the value 0 if the observation is classified as an outlier and 1 of it is not. Subscript i represents the observational index, superscript m the iteration step of the algorithm, and c denotes the cut-off value, which is a tuning parameter chosen by the researcher. Using $\widehat{u}_i^{(m)}$ to denote the residual of observation i and $\widehat{\sigma}_c^{(m)}$ the scale estimator after m iterations of the algorithm, we can formally define the classification as

$$v_{i,c}^{(m)} = 1_{(|\widehat{u}_i^{(m)}/\widehat{\sigma}_c^{(m)}| \leq c)}. \tag{2.3.1}$$

The cut-off value c is usually chosen with respect to a reference distributions such that absolute values larger than c are unlikely. For example, researchers often choose a cut-off value $c = 1.96$ and assume (implicitly or explicitly) normality of the errors, which yields a probability of 0.05 to draw a true absolute error beyond c simply due to sampling variation.

The idea of the tests suggested in this chapter is to test whether we detected significantly more outliers in the sample than we would expect purely due to random chance when there is actually no contamination in the errors. The test statistics are either based on the share of detected outliers or the number of detected outliers in the sample.

We use $\widehat{\gamma}_c^{(m)}$ to denote the share of detected outliers in the sample

$$\widehat{\gamma}_c^{(m)} = \frac{1}{n} \sum_{i=1}^{n} (1 - v_{i,c}^{(m)}). \tag{2.3.2}$$

In Chapter 1, $\widehat{\gamma}_c^{(m)}$ was introduced as the false outlier detection rate (FODR) because we analysed the asymptotic properties of $\widehat{\gamma}_c^{(m)}$ under the null hypothesis of no outliers, such that any detected outlier represented a false detection. Now that we are testing for outliers and we might be under the null or the alternative hypothesis, we use $\widehat{\gamma}_c^{(m)}$ to simply denote the share of detected outliers, or outlier share in short. In Section 2.5.2, we will use the results from Chapter 1 to derive the distribution of the test statistics under the null hypothesis of no outliers, in which case the outlier share becomes the FODR.

We use γ_c to represent the probability limit of $\widehat{\gamma}_c^{(m)}$. Under certain assumptions and under the null hypothesis of no outliers, Theorem 1.3.1 in Chapter 1 shows that $\widehat{\gamma}_c^{(m)} \xrightarrow{\mathsf{P}} \mathsf{P}(|u_i| > \sigma c)$, which is fully determined by the choice of reference distribution (the normal distribution here) and the choice of cut-off value c. It can therefore be computed numerically to be used as the hypothesised value γ_c^0 in hypothesis testing.

Alternatively, we also suggest a test based on the outlier count instead of the outlier share. We use $\widehat{\lambda}_{c_n}^{(m)}$ to denote the number of detected outliers in the sample

$$\widehat{\lambda}_{c_n}^{(m)} = n\widehat{\gamma}_{c_n}^{(m)}, \tag{2.3.3}$$

where the cut-off value c_n is now indexed by the sample size n because it must increase with the sample size in such a way that it yields a fixed expected number of outliers. When there are no outliers, Theorem 1.3.5 in Chapter 1 shows that $\widehat{\lambda}_{c_n}^{(m)}$ converges to a Poisson distribution with expected value $\lambda = n\mathsf{P}(|u_i| > \sigma c_n)$. As with the outlier share, assuming a specific reference distribution yields a specific value for the expected outlier count, which can be used in hypothesis testing and which we denote by λ^0.

Our null hypothesis is that there are no outliers in the sample, which translates into the parametric null hypothesis $H_0 : \gamma_c = \gamma_c^0$ for the outlier share and $H_0 : \lambda = \lambda^0$ for the outlier count. Both a two-sided alternative hypothesis, $H_1^{two} : \gamma_c \neq \gamma_c^0$ or $H_1^{two} : \lambda \neq \lambda^0$, and a one-sided alternative hypothesis, $H_1^{one} : \gamma_c > \gamma_c^0$ or $H_1^{two} : \lambda > \lambda^0$, are possible. For the one-sided alternative, we reject the null hypothesis of no outliers when the detected share or number of detected outliers is significantly

larger than the hypothesised share or number, respectively. This is closest to the interpretation of the test as a test for the presence of outliers, which we usually think of as unusually large values. Alternatively, one could also reject the null hypothesis against the two-sided alternative. In this case, the null hypothesis is rejected either when the number or share of detected outliers is significantly larger or significantly smaller than hypothesised. This interprets the tests more as a misspecification test for the correct choice of reference distribution because we also reject when we detect significantly fewer outliers than we hypothesised under the null hypothesis of no outliers.

It is important to note that the distributions of the test statistics are derived both assuming no contamination and the correct specification of the reference distribution. A rejection of the null hypothesis (both against the one- and the two-sided alternative) could therefore also be interpreted as an incorrect choice of reference distribution rather than the presence or absence of outliers.

2.3.2 Stylised Illustration

Figure 2.3.1 illustrates the testing idea using artificial data. Suppose a researcher has a sample of $n = 100$ observations, runs their 2SLS regression on the full sample and then calculates the standardised residuals for each observation. The scatter plots in the left panels show such potential standardised residuals. The top panel displays random draws from the standard normal distribution, which represents the non-contaminated case without outliers in green. In the panel below, five randomly selected values are replaced with outliers of magnitude 2.3, which are highlighted by the red vertical lines. The contaminated case is shown in red.

Next, the researcher chooses a cut-off value c. Observations with standardised residuals beyond the chosen cut-off value c are classified as outliers. Even in the non-contaminated (green) case, some values lie beyond the cut-off value simply due to random sampling chance and in the limit, as $n \to \infty$, the share of such false detections converges to $\mathsf{P}(|u_i| > \sigma c)$. The grey lines in the left panels indicate three potential cut-off values, translating into their correspond-

Figure 2.3.1: Illustration of Testing Idea

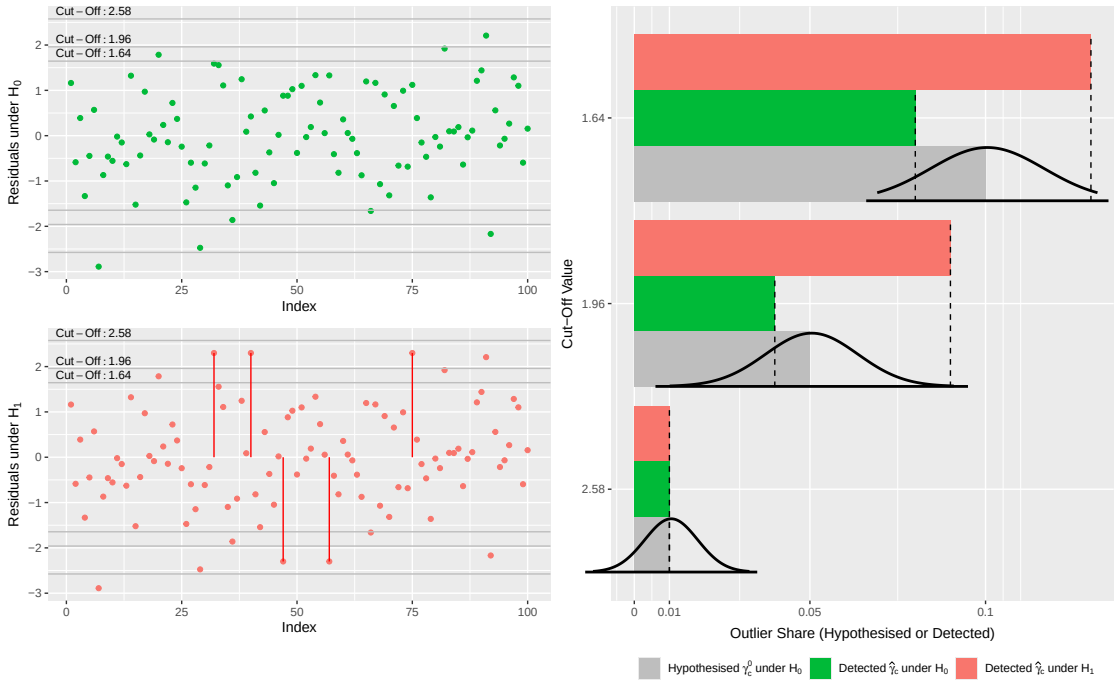


Illustration based on artificial data. The top left panel shows $n = 100$ random error draws from the standard normal distribution. The bottom left panel shows the same data except that five random values have been replaced with outliers of magnitude 2.3, indicated by vertical lines. The horizontal lines in the two left panels represent the cut-off values $c \in \{1.64, 1.96, 2.58\}$ corresponding to $\mathsf{P}(|u_i| > \sigma c) \in \{0.1, 0.05, 0.01\}$ under normality and under the null hypothesis of no outliers. The right panel shows the share of detected outliers for the different tuning parameters c resulting from the contaminated (red) and the non-contaminated (green) scatter plots on the left and compares it to the hypothesised share of detected outliers under the null hypothesis of no contamination (grey), γ_c^0. The normal distributions (black) indicate the uncertainty around γ_c^0, which could be used in a formal hypothesis test to determine whether the difference between the detected and the hypothesised share is statistically significant.

ing false outlier detection rates under normality and when there are no outliers: $(c, \mathsf{P}(|u_i| > \sigma c)) \in \{(1.64, 0.1), (1.96, 0.05), (2.58, 0.01)\}$.

The bar chart in the panel on the right compares the share of detected outliers for both the contaminated case with outliers (red) and the non-contaminated case without outliers (green) to their hypothesised value when there are no outliers (grey). The normal distributions indicate the asymptotic distribution of the outlier share around its probability limit under the null hypothesis of no outliers, which can be used to test whether the detected share of outliers is significantly different from the hypothesised value.

In this example, we detect slightly fewer outliers than expected for all shown

cut-off values in the non-contaminated scenario, which is simply due to random sampling chance. Equally, we might have detected more outliers than expected even when the sample is not actually contaminated. Furthermore, we can see that in the contaminated case, we have evidence for the presence of outliers for the cut-off values 1.64 and 1.96 but not 2.58. For $c = 1.96$ ($c = 1.64$), we expect 5% (10%) of observations to be declared as outliers even when the sample is not contaminated. In reality, however, we classify 9% (13%) of observations as outliers, which is unlikely if the sample is really not contaminated. This is shown by the dashed lines, which show that such realisations are in the tails of the normal distribution. For $c = 2.58$, we find no evidence of outliers. Remember that the outlier magnitude was 2.3 and therefore smaller than this cut-off value.

The fact that we have evidence for the presence of outliers using some cut-off values but not others reveals a potential weakness of using a single cut-off value c. Given a chosen cut-off value c, the power to reject the null hypothesis of no contamination will depend only on the frequency of outliers beyond c. Therefore, Jiao and Pretis (2022) propose using multiple cut-off values. Each cut-off value yields a certain difference between the detected and the hypothesised outlier share, which then can all be combined into a single test statistic. For example, one could take the maximum of the deviations between $\widehat{\gamma}_c^{(m)}$ and γ_c^0 across different cut-off values in the right panel of Figure 2.3.1. Using multiple cut-off values implicitly also takes outlier magnitude into account.

The next section provides an overview over the different test statistics.

2.3.3 Overview Over Tests

This section formally introduces the proposed tests. They are based on either the outlier share or the outlier count, which were introduced in Section 2.3.1. Following Jiao and Pretis (2022), we call the tests that are based on a single cut-off value "simple" tests and those that are based on multiple cut-off values "scaling" tests.

Simple Tests

Simple tests build on the current practice of choosing a single cut-off value c to

decide whether an observation should be classified as an outlier or not. Given a certain value c, we test whether more observations were classified as outliers than expected under the null hypothesis of no outliers. Hence, the frequency of outliers judged with respect to the cut-off value c determines the power of the simple tests. There are two simple tests, which are called proportion and count test.

2.3.3.1 Proportion Test

The proportion test is based on the outlier share $\widehat{\gamma}_c^{(m)}$. The idea is to compare the detected share of outliers to the expected share of outliers if the sample is not contaminated.

When there are no outliers and under certain assumptions, Theorem 1.3.1 in Chapter 1 shows that $\widehat{\gamma}_c^{(m)}$ converges in probability to $\gamma_c = \mathsf{P}(|u_i| > \sigma c)$, which is fully determined by c and the reference distribution. Under our null hypothesis of no outliers and assuming normality for the reference distribution, this probability can be numerically calculated and used as the hypothesised value γ_c^0 in the parametric null hypothesis $H_0 : \gamma_c = \gamma_c^0$.

To conduct the proportion test, proceed as follows. First, choose a specific cut-off value c and number of iterations m for the outlier detection algorithm, to be defined in Section 2.4. Next, run the algorithm to obtain the detected share of outliers, $\widehat{\gamma}_c^{(m)}$. The test statistic of the proportion test is then given as

$$T_{prop,c}^{(m)} = \frac{n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c^0)}{\sqrt{K_{cc}^{(m)}}},$$

where $K_{cc}^{(m)}$ is an asymptotic variance which is used to normalise the test statistic and which will be specified in Theorem 2.5.1.

2.3.3.2 Count Test

The count test is based on the outlier count $\widehat{\lambda}_{c_n}^{(m)}$, which we compare to the number of outliers that we would expect when the sample is not contaminated.

When there are no outliers, $\widehat{\lambda}_{c_n}^{(m)}$ asymptotically follows a Poisson distribution with mean $\lambda = n\mathsf{P}(|u_i| > \sigma c_n)$. The null hypothesis of no outliers can be written as $H_0 : \lambda = \lambda^0$, where the hypothesised value λ^0 can be numerically computed as

$n\mathsf{P}(|u_i| > \sigma c_n)$ using the normal reference distribution.

The procedure for the count test is similar to that for the proportion test. First, choose a specific target for the number of falsely detected outliers under no contamination, λ, which together with the sample size n also determines the cut-off value c_n. Specifically, using the standard normal distribution as the reference distribution for outlier detection, the cut-off is given as $c_n = \Phi^{-1}\{1 - \lambda/(2n)\}$, where Φ^{-1} represents the quantile function of the standard normal distribution. This ensures that when there is no contamination, the expected number of detected outliers λ stays constant as the sample size $n \to \infty$. Then, choose a number of iterations m and run the outlier detection algorithm to obtain the detected number of outliers, $\widehat{\lambda}_{c_n}^{(m)}$. The test statistic of the count test is given as

$$T_{count,\lambda}^{(m)} = \widehat{\lambda}_{c_n}^{(m)} = n\widehat{\gamma}_{c_n}^{(m)}.$$

Scaling Tests

Instead of using a single cut-off value c, scaling tests rely on a range (scale) of K different cut-off values $\mathbb{S}_c = \{c_1, \ldots, c_K\}$. This provides two advantages over the simple tests. Simple tests only compare the actual to the hypothesised frequency of detected outliers at a single cut-off value, while scaling tests additionally take outlier magnitude into account by considering multiple cut-off values. Second, choosing a range of cut-off values avoids the somewhat arbitrary choice of a single cut-off value defining what constitutes an outlier.

The first scaling test is a Bonferroni-type correction for multiple hypothesis testing after computing several simple proportion or count tests across a range of cut-off values. Instead of running K separate simple tests, the sum and the supremum test combine the results from the K different cut-off values into a single metric by taking either the sum or the supremum over the differences between the outlier shares and their hypothesised values.

2.3.3.3 Multiple Proportion and Count Test

We suggest the Simes (1986) procedure for multiple hypothesis testing because it can improve upon the Bonferroni correction when the test statistics are correlated, as

is the case in our setting. For example, having many large errors beyond a certain cut-off value c means that we are also more likely to have many errors beyond a smaller cut-off $c' < c$.

To conduct the multiple tests, first choose a set of K different cut-off values $\mathbb{S}_c = \{c_1, \ldots, c_K\}$ with a corresponding set of hypothesised shares of falsely detected outliers $\mathbb{S}_\gamma = \{\gamma^0_{c_1}, \ldots, \gamma^0_{c_K}\}$ for the proportion test or a set of hypothesised numbers of falsely detected outliers $\mathbb{S}_\lambda = \{\lambda^0_1, \ldots, \lambda^0_K\}$ for the count test. Choose a specific iteration step m for one of the outlier detection algorithms defined in Section 2.4 and obtain the outlier share $\widehat{\gamma}^{(m)}_{c_k}$ or outlier count $\widehat{\lambda}^{(m)}_{c_k,n}$, respectively, for each cut-off value $c_k \in \mathbb{S}_c$.

Section 2.5.2.3 explains how to use the outlier shares and outlier counts to test the global null hypothesis of no contamination, which is $H_0 : \gamma_{c_k} = \gamma^0_{c_k} \, \forall \, k \in \{1, \ldots, K\}$ for the multiple proportion test and $H_0 : \lambda_k = \lambda^0_k \, \forall \, k \in \{1, \ldots, K\}$ for the multiple count test.

2.3.3.4 Sum Test

The sum test follows the same intuition as the proportion test. It compares the detected to the hypothesised share of outliers when there is no contamination but across several cut-off values $c_k \in \mathbb{S}_c = \{c_1, \ldots, c_K\}$. Since the outlier shares $\widehat{\gamma}^{(m)}_{c_k}$ at different cut-off values are correlated, we need to take their variance-covariance structure into account.

The sum test can be conducted as follows. Choose an ordered set of K different cut-off values $\mathbb{S}_c = \{c_1, \ldots, c_K\}$ in increasing order $c_1 < \cdots < c_K$ with a corresponding ordered set of hypothesised shares of falsely detected outliers when there is no contamination $\mathbb{S}_\gamma = \{\gamma^0_{c_1}, \ldots, \gamma^0_{c_K}\}$ in decreasing order $\gamma^0_{c_1} > \cdots > \gamma^0_{c_K}$. Choose a specific iteration step m and apply one of the outlier detection algorithms defined in Section 2.4 for each of the cut-off values $c_k \in \mathbb{S}_c$. Record the scaled difference for each setting and sum them up $\sum_{k=1}^{K} n^{1/2}(\widehat{\gamma}^{(m)}_{c_k} - \gamma^0_{c_k})$. Using the asymptotic variance $K^{(m)}_{\mathbb{S}_c}$ of this sum of deviations as a scaling factor, we obtain the sum test statistic

for the global null hypothesis of $H_0 : \gamma_{c_k} = \gamma^0_{c_k} \, \forall \, k \in \{1, \ldots, K\}$ as

$$T^{(m)}_{sum,\mathbb{S}_c} = \frac{\sum_{k=1}^{K} n^{1/2}(\widehat{\gamma}^{(m)}_{c_k} - \gamma^0_{c_k})}{\sqrt{K^{(m)}_{\mathbb{S}_c}}}.$$

The scaling factor $K^{(m)}_{\mathbb{S}_c}$ and the asymptotic distribution of the sum test statistic will be derived in Theorem 2.5.3.

2.3.3.5 Supremum Test

The previous sum test simply calculates the sum of the deviations between the outlier share and its hypothesised value at different cut-off values. This means that positive and negative deviations at different cut-off values partially cancel each other. The supremum test instead takes the largest absolute deviation across the K different cut-off values. This means that the test is only one-sided but actually considers positive and negative deviations equally. As for the sum test, the supremum test must take the variance-covariance structure across different cut-off values into account.

The supremum test is conducted in the same way as the sum test. However, instead of taking the sum of the deviations, we now take the supremum over the absolute deviations across the cut-off values. Since in practice, we have a finite number K of cut-off values, this is equivalent to taking the maximum.

To test the global null hypothesis of $H_0 : \gamma_{c_k} = \gamma^0_{c_k} \, \forall \, k \in \{1, \ldots, K\}$, we use the supremum test statistic

$$T^{(m)}_{sup,\mathbb{S}_c} = \sup_{1 \leq k \leq K} |n^{1/2}(\widehat{\gamma}^{(m)}_{c_k} - \gamma^0_{c_k})|,$$

whose asymptotic distribution is derived in Theorem 2.5.4.

Table 2.3.1 summarises the different tests in terms of their classification into a simple or scaling test depending on the number of tuning parameters involved, their null hypothesis, and their test statistic.

The next section provides more detail on the tests, first defining the algorithms which yield the outlier share $\widehat{\gamma}^{(m)}_c$ and outlier count $\widehat{\lambda}^{(m)}_{c_n}$, describing the required assumptions for the asymptotic theory, and deriving the asymptotic distributions of the test statistics.

Table 2.3.1: Overview Over Tests

Type	Tuning Par.	H_0	Test	Test Statistic
Simple	single c	$\gamma_c = \gamma_c^0$	Proportion Test	$\frac{n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c^0)}{\sqrt{K_{cc}^{(m)}}}$
		$\lambda = \lambda^0$	Count Test	$\widehat{\lambda}_{c_n}^{(m)} = n\widehat{\gamma}_{c_n}^{(m)}$
Scaling	multiple $\mathbb{S}_c = \{c_1, ..., c_K\}$	$\gamma_{c_k} = \gamma_{c_k}^0$ $\forall\, k = 1, \ldots, K$	Multiple Proportion Test	see Proportion Test
		$\lambda_k = \lambda_k^0$ $\forall\, k = 1, \ldots, K$	Multiple Count Test	see Count Test
		$\gamma_{c_k} = \gamma_{c_k}^0$ $\forall\, k = 1, \ldots, K$	Sum Test	$\frac{\sum_{k=1}^K n^{1/2}(\widehat{\gamma}_{c_k}^{(m)} - \gamma_{c_k}^0)}{\sqrt{K_{\mathbb{S}_c}^{(m)}}}$
			Supremum Test	$\sup_{1 \le k \le K} \lvert n^{1/2}(\widehat{\gamma}_{c_k}^{(m)} - \gamma_{c_k}^0) \rvert$

2.4 Outlier Detection Algorithms

The tests are based on a class of popular heuristics to find outliers in a sample, which we improve and formalise into three different algorithms. The algorithms provide a classification of observations into outliers and non-outliers, which forms the basis of all tests through $\widehat{\gamma}_c^{(m)}$ and $\widehat{\lambda}^{(m)}$.

All algorithms presented in this section are based on the same approach: they use some initial estimator to obtain standardised residuals for each observation in the sample and then classify observations with residuals beyond a chosen cut-off value as outliers. However, the algorithms differ in (i) the choice of initial estimator and (ii) whether the procedure is iterated or stopped after the first classification.

The description of the algorithms follows Chapter 1 closely but we keep their numbering separate because one of them is restricted to a special case here. The algorithms are formalisations and potential improvements over the outlier detection heuristic that applied researchers often use. In the heuristic approach, the researcher chooses a cut-off value c with a reference distribution for the structural error in mind. Values beyond the cut-off are unlikely and therefore considered outliers. After obtaining some initial estimates of β and σ^2, usually using the full-sample 2SLS estimates, the researcher computes the standardised residuals of each observation, classifies those with absolute standardised residuals larger than the cut-off value as

outliers, and removes them from the sample. Subsequently, the 2SLS regression is re-estimated on the remaining subsample, which yields the robust parameter estimates.

Following Jiao (2022), we embed three improvements in the below algorithms. First, we use a bias correction factor for the updated estimator of the variance. This correction is necessary when the sample is not contaminated because the algorithms have a positive probability of falsely removing observations with large residuals from the sample, which causes the ordinary variance estimator to be downward biased. Second, we suggest an alternative, more robust initial estimator than the full-sample estimator, because the latter is not robust and therefore might not be a good starting point for the algorithm. Third, we propose to iterate the outlier detection procedure after obtaining the updated estimates based on the subsample of non-outlying observations. This might achieve higher robustness because if the procedure is iterated until it reaches a fixed point, it has the same first order asymptotics as the robust Huber-skip estimator in the 2SLS regression setting, as shown in Jiao (2022).

Algorithm 2.I is a general representation of such an iterated procedure, which incorporates the bias correction factor ς_c^{-2} for the scale estimator in equation (2.4.4). The algorithm is general because it does not specify the initial estimators $\widehat{\beta}_c^{(0)}$ and $\widehat{\sigma}_c^{(0)}$ in step 1. The asymptotic analysis in Section 2.5.2 requires these initial estimators to be consistent and bounded in probability.

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

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

2. Compute the indicator variables for selecting non-outlying observations

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \leq \widehat{\sigma}_c^{(m)} c)}. \tag{2.4.1}$$

3. Compute the updated least squares estimator for Π

$$\widehat{\Pi}_c^{(m+1)} = (\sum_{i=1}^{n} z_i z_i' v_{i,c}^{(m)})^{-1} (\sum_{i=1}^{n} z_i x_i' v_{i,c}^{(m)}). \tag{2.4.2}$$

4. Compute the updated 2SLS estimators for β and σ^2

$$\widehat{\beta}_c^{(m+1)} = (\widehat{\Pi}_c^{(m+1)\prime} \sum_{i=1}^{n} z_i z_i' v_{i,c}^{(m)} \widehat{\Pi}_c^{(m+1)})^{-1} (\widehat{\Pi}_c^{(m+1)\prime} \sum_{i=1}^{n} z_i y_i v_{i,c}^{(m)}), \tag{2.4.3}$$

$$(\widehat{\sigma}_c^{(m+1)})^2 = \varsigma_c^{-2} (\sum_{i=1}^{n} v_{i,c}^{(m)})^{-1} \{\sum_{i=1}^{n} (y_i - x_i' \widehat{\beta}_c^{(m+1)})^2 v_{i,c}^{(m)}\}. \tag{2.4.4}$$

5. Let $m = m + 1$ and repeat steps 2 - 5 until the desired number of iterations is reached.

The algorithm delivers the outlier classification indicators $v_{i,c}^{(m)}$ on which the outlier share $\widehat{\gamma}_c^{(m)}$ and the outlier count $\widehat{\lambda}_{c_n}^{(m)}$ that were introduced in Section 2.3.1 are based.

The most common heuristic practice of using the full-sample 2SLS estimators as the initial estimators is formalised in Algorithm 2.R below. Analogous to the least squares literature, this procedure is also called Robustified 2SLS (Johansen and Nielsen 2009).

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

1.1. Compute the least squares estimator for Π based on the whole sample

$$\widehat{\Pi}_c^{(0)} = (\sum_{i=1}^{n} z_i z_i')^{-1} (\sum_{i=1}^{n} z_i x_i'). \tag{2.4.5}$$

1.2. Compute the 2SLS estimators for β and σ^2 based on the whole sample

$$\widehat{\beta}_c^{(0)} = (\widehat{\Pi}_c^{(0)\prime} \sum_{i=1}^{n} z_i z_i' \widehat{\Pi}_c^{(0)})^{-1} (\widehat{\Pi}_c^{(0)\prime} \sum_{i=1}^{n} z_i y_i), \qquad (\widehat{\sigma}_c^{(0)})^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' \widehat{\beta}_c^{(0)})^2. \tag{2.4.6}$$

1.3. Select $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ given in (2.4.6) *as initial estimators and let $m = 0$.*

2. Follow the steps 2 - 5 in Algorithm 2.I.

Unfortunately, the full-sample 2SLS estimator is itself not robust, such that it might not be a good starting point for the outlier detection procedure. As an alternative, we suggest the Saturated 2SLS estimator, which is a stylised but analytically tractable approximation to impulse indicator saturation (IIS). IIS was originally introduced in the OLS regression setting in Hendry (1999) to analyse U.S. food expenditure data that was contaminated with many outliers. IIS was later formalised

in Hendry, Johansen and Santos (2008). A simplified version of IIS, called Saturated 2SLS, is described in Algorithm 2.S and works as follows. First, split the sample into two halves and obtain 2SLS estimates for each half. Then, use the estimates from one half to compute standardised residuals in the other half. As in the general outlier detection procedure, observations are again classified as outliers if their absolute standardised residuals are larger than the cut-off value c and removed from the sample. Subsequently, take the union of the remaining observations of each half and re-estimate the 2SLS regression, which yields the robust parameter estimates.

Algorithm 2.*S* (Saturated 2SLS, Split-Half). *Choose a cut-off value* $c > 0$.

1.1. Create two halves $\mathcal{I}_1$ *and* $\mathcal{I}_2$ *with* $n_1 = int[n/2]$ *and* $n_2 = n - n_1$ *observations.*

1.2. Compute the least squares estimators for Π *for each half* $\mathcal{I}_j$, $j = 1, 2$

$$\widehat{\Pi}_j = (\sum_{i \in \mathcal{I}_j} z_i z_i')^{-1} (\sum_{i \in \mathcal{I}_j} z_i x_i'). \tag{2.4.7}$$

1.3. Compute the 2SLS estimators for β *and* σ^2 *for each half* $\mathcal{I}_j$, $j = 1, 2$

$$\widehat{\beta}_j = (\widehat{\Pi}_j' \sum_{i \in \mathcal{I}_j} z_i z_i' \widehat{\Pi}_j)^{-1} (\widehat{\Pi}_j' \sum_{i \in \mathcal{I}_j} z_i y_i), \qquad \widehat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} (y_i - x_i' \widehat{\beta}_j)^2. \tag{2.4.8}$$

1.4. Compute the indicator variables for selecting non-outlying observations

$$v_{i,c}^{(0)} = 1_{(i \in \mathcal{I}_1)} 1_{(|y_i - x_i' \widehat{\beta}_2| \leq \widehat{\sigma}_2 c)} + 1_{(i \in \mathcal{I}_2)} 1_{(|y_i - x_i' \widehat{\beta}_1| \leq \widehat{\sigma}_1 c)}. \tag{2.4.9}$$

1.5. Compute $\widehat{\Pi}_c^{(1)}$, $\widehat{\beta}_c^{(1)}$, $(\widehat{\sigma}_c^{(1)})^2$ *using* (2.4.2), (2.4.3), (2.4.4) *with* $m = 0$ *and let* $m = 1$.

2. Follow the steps 2 - 5 in Algorithm 2.I.

Note that Algorithm 2.S presents the *split-half* version of Saturated 2SLS in which the sample is split into two halves, whereas Algorithm 1.S in Chapter 1 presents a more general version in which the sample can also be split into unequal parts. While unequal splits might increase robustness, we believe that the split-half version is most useful in practice because unequal splits require a lot of data for the smaller subsample to still provide good parameter estimates. Instead of using unequal splits in the simple Saturated 2SLS algorithm, we suggest using the

IIS algorithm below, which searches for outliers across many small subsets of the sample.

We describe IIS in more detail but not as a formalised algorithm due to its complexity, which also prevents us from providing asymptotic theory for IIS. IIS is a block-wise multi-path search algorithm that consecutively adds blocks of dummy variables (impulse indicators), which can perfectly fit an observation, to the set of regressors and tests for their statistical significance. Suppose that the sample size is 100, in which case four blocks of 25 impulse indicators each may be used as the starting points for the multi-path search. After adding a block of 25 indicators to the set of regressors, the algorithm explores multiple paths, dropping insignificant impulse indicators. Once all blocks have been searched, the algorithm takes the union of all previously statistically significant impulse indicators, adds them jointly to the regression and tests again for their significance. The observations corresponding to the retained indicators can then be considered outliers, i.e. their initial classification is set to $v_{i,c}^{(0)} = 0$ in equation (2.4.1). To iterate the outlier detection procedure, simply follow steps 2 to 5 in the general Algorithm 2.I.

IIS might have better power than the other algorithms due to its multi-path search. It might detect outlier patterns that are invisible to the other algorithms, for example if the outliers are concentrated in a small part of the sample. We do not provide any asymptotic theory for IIS but the asymptotic approximations derived in Section 2.5.2 might still work well if IIS is close enough to Algorithm 2.S.

The R package *ivgets* (Kurle 2022a) implements — among other functionalities — IIS for 2SLS regression models, building on the multi-block search algorithm of the R package *gets*. See Pretis, Reade and Sucarrat (2018) for a more detailed description of the search algorithm. The simulation study in Section 2.6 also assesses the performance of IIS in a subset of the Monte Carlo experiments. The tests for the presence of outliers based on IIS are implemented using the asymptotic approximation of Algorithm 2.S.

2.5 Asymptotic Theory

This section first states the required assumptions for deriving the asymptotic theory. We then provide the asymptotic distributions of the different test statistics and explain the testing procedures.

2.5.1 Assumptions

The derivations of the large sample properties of the test statistics are based on the asymptotic theory developed in Chapter 1. In the present chapter, we make the further assumption of joint normality of the errors as in equation (2.2.3), which means that many of the more general assumptions in Chapter 1 are automatically fulfilled, see Appendix 2.B. What remains are assumptions on the existence of certain moments of the instruments and on the initial estimators.

Assumption 2.5.1. *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 *are* $\mathcal{F}_{i-1}$ *measurable while* r_i*,* u_i *are independent of* $\mathcal{F}_{i-1}$*. Assume* $d_x \leq d_z$ *and* $\operatorname{rank}\Pi = d_x$*. Choose* $\kappa \in [0, 1/4)$*. Furthermore, assume*

(i) *the errors* $(u_i/\sigma, \Sigma^{-1/2} r_i)$ *are jointly normally distributed as in equation* (2.2.3)*;*

(ii) *the instruments* z_i *satisfy*

- (a) $M_{zz,n} = n^{-1} \sum_{i=1}^{n} z_i z_i' \xrightarrow{\mathsf{P}} M_{zz} > 0$*;*
- (b) $n^{-\kappa} \max_{1 \leq i \leq n} |z_i| = \mathrm{O}_{\mathsf{P}}(1)$*;*
- (c) $\mathsf{E}|z_i|^9 = \mathrm{O}(1)$*;*

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

- (a) $n^{1/2}(\widetilde{\beta} - \beta) = \mathrm{O}_{\mathsf{P}}(1)$*;*
- (b) $n^{1/2}(\widetilde{\sigma}^2 - \sigma^2) = \mathrm{O}_{\mathsf{P}}(1)$.

Assumption 2.5.1(i) specifies that the structural error and the first stage errors are jointly normally distributed, which is often assumed in practice. For example, Acemoglu et al. (2019) trim observations with absolute standardised residuals beyond 1.96, either based only on the second stage or on both stages.

Assumption 2.5.1(ii) imposes conditions on moments of the instruments. As-

sumption 2.5.1(*iia*) requires the expected value $\mathsf{E}z_i z_i' \equiv M_{zz}$ to exist, such that the Law of Large Numbers (LLN) can be applied to the sample average. M_{zz} is also required to be positive definite. This implies that it is full rank d_z, such that together with the assumption of $\operatorname{rank}\Pi = d_x$, the rank condition for identification holds. Assumption 2.5.1(*iib*) requires the maximum absolute value of z_i to be stochastically bounded when it is shrunk by a factor n^{κ}. Johansen and Nielsen (2016a) show that this condition holds for stationary instruments. Finally, Assumption 2.5.1(*iic*) assumes the existence of at least the 9-th moment of the absolute value of the instruments.

The last Assumption 2.5.1(*iii*) requires tightness of the initial estimators, on which the initial classification of outliers is based. This is automatically fulfilled for Algorithms 2.R and 2.S, which form the basis for the proportion, sum, and supremum test. The count test allows for different initial estimators as long as they fulfil Assumption 2.5.1(*iii*) and therefore includes but is not limited to Algorithms 2.R and 2.S.

2.5.2 Asymptotic Distributions of the Test Statistics

This section derives the asymptotic distributions of the test statistics and describes how the different hypothesis tests are conducted. The asymptotic theory is derived under the null hypothesis of no outliers, in which case the share of detected outliers coincides with the FODR. Hence, the results from Chapter 1 apply and we will use outlier share and FODR synonymously. Throughout this section, we use c_+ to denote any small finite, positive real number and $*$ to denote the fixed point of the iterated outlier detection algorithms.

2.5.2.1 Proportion Test

The derivation of the asymptotic distribution of the proportion test builds on the asymptotic analysis of the FODR for a given cut-off value $c \in [c_+, \infty)$ and iteration step $m \in [0, \infty)$, which is shown in Theorems 1.C.2 and 1.C.3 in Chapter 1. The following theorem states the asymptotic distribution of the proportion test statistic, which holds for Algorithms 2.R and 2.S.

Theorem 2.5.1 (Proportion Test). *Consider Algorithm 2.R or 2.S. Suppose Assumption 2.5.1(i, ii) holds. Choose a specific cut-off value $c \in [c_+, \infty)$, which determines $\gamma_c^0 = \mathsf{P}(|u_i| > \sigma c)$ under no contamination. Choose a specific iteration step $m \in [0, \infty)$, which includes the fixed point $*$ when $m \to \infty$. Then, under the null hypothesis $H_0 : \gamma_c = \gamma_c^0$ and as $n \to \infty$,*

$$T_{prop,c}^{(m)} = \frac{n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c^0)}{\sqrt{K_{cc}^{(m)}}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1),$$

where the asymptotic variance $K_{cc}^{(m)}$ is given as

$$\begin{aligned} K_{cc}^{(m)} &= \gamma_c^0 \psi_c - 2c\phi_u(c)\varrho_{\sigma\sigma,c}^{(m)}(\psi_c - \tau_2^c) \\ &\quad + c^2\phi_u^2(c)\{2(\varrho_{\sigma\sigma,c}^{(m)})^2 + \varrho_{\sigma uu,c}^{(m)}(2\varrho_{\sigma\sigma,c}^{(m)} + \varrho_{\sigma uu,c}^{(m)})(\tau_4^c - \tau_2^c\varsigma_c^2)\} \end{aligned} \tag{2.5.1}$$

and the coefficients have expressions

$$\varrho_{\sigma\sigma,c}^{(m)} = \{\frac{c\phi_u(c)(c^2 - \varsigma_c^2)}{\tau_2^c}\}^m, \qquad \varrho_{\sigma uu,c}^{(m)} = \frac{(\tau_2^c)^m - \{c\phi_u(c)(c^2 - \varsigma_c^2)\}^m}{(\tau_2^c)^m\{\tau_2^c - c\phi_u(c)(c^2 - \varsigma_c^2)\}}.$$

Based on this limiting distribution, we can compare the test statistic to the critical values of the standard normal distribution. We use $\Phi(\cdot)$ to denote the cumulative distribution function (CDF) of the standard normal distribution and $\Phi^{-1}(\cdot)$ the corresponding quantile function. Using α to represent the nominal significance level, we have the following decision rules:

Reject H_0 in favour of $H_1^{one} : \gamma_c > \gamma_c^0$ if $T_{prop,c}^{(m)} > \Phi^{-1}(1-\alpha)$.
Reject H_0 in favour of $H_1^{two} : \gamma_c \neq \gamma_c^0$ if $|T_{prop,c}^{(m)}| > \Phi^{-1}(1-\alpha/2)$.

Note that because of our assumption of normality, the asymptotic variance $K_{cc}^{(m)}$ of the FODR is completely determined by the assumed normal reference distribution of the structural error, ϕ_u. Therefore, no estimation of the asymptotic variance is required. Moreover, the asymptotic variance in equation (2.5.1) represents the general iteration step $m \in [0, \infty)$. The formula simplifies in the special cases when the proportion test is based on the outlier share obtained from the initial step ($m = 0$) or the fixed point ($m = *$) of the outlier detection algorithm, see the following remark.

Remark 2.5.1. *For $m = 0$ and $m = *$, the equation for the asymptotic variance of the proportion test statistic in equation* (2.5.1) *simplifies to*

$$K_{cc}^{(0)} = \gamma_c^0 \psi_c - 2c^2 \phi_u^2(c), \qquad K_{cc}^{(*)} = \gamma_c^0 \psi_c + \frac{c^2 \phi_u^2(c)(\tau_4^c - \tau_2^c \varsigma_c^2)}{\{\tau_2^c - c(c^2 - \varsigma_c^2)\phi_u(c)\}^2}.$$

2.5.2.2 Count Test

The count test follows from the Poisson exceedance theory for the number of falsely detected outliers derived in Theorem 1.3.5 of Chapter 1. In the asymptotic analysis, the cut-off value c_n must increase with the sample size in order to keep the expected number of outliers, $\lambda = n\mathsf{P}(|u_i| > \sigma c_n)$, fixed. In other words, the asymptotic approximation requires $c_n \to \infty$ as $n \to \infty$.

The asymptotic distribution of the test statistic is derived for the general Algorithm 2.*I*. This means that the algorithm may start with any tight initial estimators of (β, σ^2), including the full-sample or split-half estimators specified in Algorithms 2.*R* or 2.*S*.

Theorem 2.5.2 (Count Test). *Consider Algorithm 2.I. Suppose Assumption 2.5.1 holds. Note that Assumption 2.5.1(iii) is fulfilled for Algorithms 2.R and 2.S. Choose a specific target for the number of falsely detected outliers $\lambda^0 = n\mathsf{P}(|u_i| > \sigma c_n)$, which together with the sample size n also determines the cut-off value $c_n \in [c_+, \infty)$. Choose a specific iteration step $m \in [0, \infty)$, which includes the fixed point $*$ when $m \to \infty$. Then, under the null hypothesis $H_0 : \lambda = \lambda^0$ and as $n \to \infty$,*

$$T_{count,\lambda}^{(m)} = \widehat{\lambda}_{c_n}^{(m)} = n\widehat{\gamma}_{c_n}^{(m)} \xrightarrow{\mathsf{D}} \text{Poisson}(\lambda^0).$$

We can now test whether the count test statistic is an unlikely realisation under a Poisson distribution with mean λ^0. Using $\mathsf{F}(\cdot|\lambda^0)$ to denote the CDF of a Poisson distribution with expected value λ^0, and using α to denote the nominal significance level, we have the following decision rules:

Reject H_0 in favour of $H_1^{one} : \lambda > \lambda^0$ if $1 - \mathsf{F}(n\widehat{\gamma}_{c_n}^{(m)} - 1|\lambda^0) < \alpha$.
Reject H_0 in favour of $H_1^{two} : \lambda \neq \lambda^0$ if $p_{\text{two-sided}} < \alpha$,

where $p_{\text{two-sided}}$ refers to the two-sided p-value. There are several ways to calculate two-sided p-values for discrete distributions. For example, the "central" (also called

"twice the smaller tail" in Hirji (2006)) approach takes "2 times the minimum of the one-sided p-values bounded above by 1 (Fay 2010)." That is, it can be calculated as $p^{\text{central}}_{\text{two-sided}} = \min[1, 2 \times \min\{\mathsf{F}(n\widehat{\gamma}^{(m)}_{c_n}|\lambda^0), 1 - \mathsf{F}(n\widehat{\gamma}^{(m)}_{c_n} - 1|\lambda^0)\}]$.

The discrete nature of the Poisson distribution causes the size in general to differ from the nominal significance level. This is discussed in more detail in Section 2.6.1 and the researcher is free to choose a different decision rule than the one described above, e.g. a randomised decision rule to obtain the desired exact size in expectation (on average).

2.5.2.3 Multiple Proportion and Count Test

Since the simple tests are applied multiple times using a Bonferroni-type correction, they require the same assumptions as the underlying individual simple tests, i.e. the assumptions stated in Theorem 2.5.1 for the multiple proportion test and those in Theorem 2.5.2 for the multiple count test, respectively.

After having obtained the K proportion or count test statistics as described in Section 2.3.3.3, do the following. For each of the cut-off values, test the individual null hypothesis $H_0 : \gamma_{c_k} = \gamma^0_{c_k}$ or $H_0 : \lambda_k = \lambda^0_k$ and record the appropriate p-value.

Then, test the global null hypothesis using the Simes (1986) procedure to account for multiple hypothesis testing as follows. First, choose a global nominal significance level α. Second, order the p-values of the K individual tests in increasing order, $p_{(1)} \leq \cdots \leq p_{(K)}$. Third, compare the p-values to adjusted significance levels of $\alpha \times k/K$ and reject the global null hypothesis of no outliers when there exists a $k \in \{1, \ldots, K\}$ such that the following condition holds

$$p_{(k)} \leq \alpha \frac{k}{K}.$$

Instead of comparing all p-values to the same adjusted significance level of α/K, we compare them to a multiple of that value according to their order statistics. This type of adjustment leads to weakly more rejections of the global null hypothesis than comparing all p-values against the smallest adjusted significance level of α/K.

2.5.2.4 Sum Test

The asymptotic distribution for the sum test statistic is derived for both Algorithm 2.R and 2.S. The derivations build on Theorems 1.C.2 and 1.C.3 in Chapter 1. The sum of deviations across cut-off values between the different outlier shares and their hypothesised values is normalised by the asymptotic variance $K_{\mathbb{S}_c}^{(m)}$.

Theorem 2.5.3 (Sum Test). *Consider Algorithm 2.R or 2.S. Suppose Assumption 2.5.1(i, ii) holds. Choose an ordered set of K different cut-off values $\mathbb{S}_c = \{c_1, \ldots, c_K\}$ in increasing order $c_1 < \cdots < c_K$ with a corresponding ordered set of hypothesised shares of falsely detected outliers $\mathbb{S}_\gamma = \{\gamma_{c_1}^0, \ldots, \gamma_{c_K}^0\}$ in decreasing order $\gamma_{c_1}^0 > \cdots > \gamma_{c_K}^0$. Choose a specific iteration step $m \in [0, \infty)$, which includes the fixed point $*$ when $m \to \infty$. Then, under the global null hypothesis $H_0 : \gamma_{c_k} = \gamma_{c_k}^0 \; \forall \, k \in \{1, \ldots, K\}$ and as $n \to \infty$,*

$$T_{sum,\mathbb{S}_c}^{(m)} = \frac{\sum_{k=1}^K n^{1/2}(\widehat{\gamma}_{c_k}^{(m)} - \gamma_{c_k}^0)}{\sqrt{K_{\mathbb{S}_c}^{(m)}}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1),$$

where the asymptotic variance $K_{\mathbb{S}_c}^{(m)}$ is given as

$$\begin{aligned} K_{\mathbb{S}_c}^{(m)} = {} & \sum_{k=1}^K \psi_{c_k}\gamma_{c_k}^0 + \sum_{k=1}^K \{c_k\phi_u(c_k)\varrho_{\sigma uu,c_k}^{(m)}\}^2(\tau_4^{c_k} - \tau_2^{c_k}\varsigma_{c_k}^2) + 2\sum_{k=1}^K \{c_k\phi_u(c_k)\varrho_{\sigma\sigma,c_k}^{(m)}\}^2 \\ & + 2\{\sum_{k=1}^K c_k\phi_u(c_k)\varrho_{\sigma\sigma,c_k}^{(m)}\}\{\sum_{k=1}^K c_k\phi_u(c_k)\varrho_{\sigma uu,c_k}^{(m)}(\tau_4^{c_k} - \tau_2^{c_k}\varsigma_{c_k}^2) + \tau_2^{c_k} - \psi_{c_k}\} \quad (2.5.2) \\ & + 2\sum_{1\le k<l\le K} c_k c_l \phi_u(c_k)\phi_u(c_l)\{2\varrho_{\sigma\sigma,c_k}^{(m)}\varrho_{\sigma\sigma,c_l}^{(m)} + \varrho_{\sigma uu,c_k}^{(m)}\varrho_{\sigma uu,c_l}^{(m)}(\tau_4^{c_k} - \tau_2^{c_k}\varsigma_{c_k}^2)\} \\ & + 2\sum_{1\le k<l\le K} \psi_{c_k}\{\gamma_{c_l}^0 + c_l\phi_u(c_l)\varrho_{\sigma uu,c_l}^{(m)}(\varsigma_{c_k}^2 - \varsigma_{c_l}^2)\} \end{aligned}$$

and the coefficients $\varrho_{\sigma\sigma,c}^{(m)}$ and $\varrho_{\sigma uu,c}^{(m)}$ have the same expressions as in Theorem 2.5.1.

Using $\Phi(\cdot)$ to denote the CDF of the standard normal distribution with $\Phi^{-1}(\cdot)$ being the corresponding quantile function, and α to denote the nominal significance level, we have the following decision rules:

Reject H_0 in favour of $H_1^{one} : \exists \, k \in \{1, \ldots, K\} : \gamma_{c_k} > \gamma_{c_k}^0$ if $T_{sum,\mathbb{S}_c}^{(m)} > \Phi^{-1}(1-\alpha)$.
Reject H_0 in favour of $H_1^{two} : \exists \, k \in \{1, \ldots, K\} : \gamma_{c_k} \neq \gamma_{c_k}^0$ if $|T_{sum,\mathbb{S}_c}^{(m)}| > \Phi^{-1}(1-\alpha/2)$.

For the special cases when the tests are based on the initial iteration step $m = 0$ or the fixed point $m = *$ of the outlier detection algorithm, the formula of the asymptotic variance of the sum test simplifies as shown in the following remark.

Remark 2.5.2. *For $m = 0$ and $m = *$, the equation for the asymptotic variance of the sum test statistic in equation* (2.5.2) *simplifies to*

$$K_{\mathbb{S}_c}^{(0)} = \sum_{k=1}^{K} \psi_{c_k} \gamma_{c_k}^0 + 2 \sum_{k=1}^{K} \{c_k \phi_u(c_k)\}^2 + 2\{\sum_{k=1}^{K} c_k \phi_u(c_k)\}\{\sum_{k=1}^{K} (\tau_2^{c_k} - \psi_{c_k})\}$$
$$+ 2 \sum_{1 \le k < l \le K} \psi_{c_k} \gamma_{c_l}^0 + 4 \sum_{1 \le k < l \le K} c_k c_l \phi_u(c_k) \phi_u(c_l)$$

and

$$K_{\mathbb{S}_c}^{(*)} = \sum_{k=1}^{K} \psi_{c_k} \gamma_{c_k}^0 + \sum_{k=1}^{K} \{\frac{c_k \phi_u(c_k)}{\tau_2^{c_k} - c_k \phi_u(c_k)(c_k^2 - \varsigma_{c_k}^2)}\}^2 (\tau_4^{c_k} - \tau_2^{c_k} \varsigma_{c_k}^2)$$
$$+ 2 \sum_{1 \le k < l \le K} \frac{c_k c_l \phi_u(c_k) \phi_u(c_l)(\tau_4^{c_k} - \tau_2^{c_k} \varsigma_{c_k}^2)}{\{\tau_2^{c_k} - c_k \phi_u(c_k)(c_k^2 - \varsigma_{c_k}^2)\}\{\tau_2^{c_l} - c_l \phi_u(c_l)(c_l^2 - \varsigma_{c_l}^2)\}}$$
$$+ 2 \sum_{1 \le k < l \le K} \psi_{c_k} \{\gamma_{c_l}^0 + \frac{c_l \phi_u(c_l)(\varsigma_{c_k}^2 - \varsigma_{c_l}^2)}{\tau_2^{c_l} - c_l \phi_u(c_l)(c_l^2 - \varsigma_{c_l}^2)}\}.$$

2.5.2.5 Supremum Test

As for the sum test, the asymptotic distribution of the supremum test is derived using Theorems 1.C.2 and 1.C.3 in Chapter 1. The asymptotic theory of the test statistic applies to Algorithms 2.R and 2.S.

Theorem 2.5.4 (Supremum Test). *Consider Algorithm 2.R or 2.S. Suppose Assumption 2.5.1(i, ii) holds. Choose an ordered set of K different cut-off values $\mathbb{S}_c = \{c_1, \ldots, c_K\}$ in increasing order $c_1 < \cdots < c_K$ with a corresponding ordered set of hypothesised shares of falsely detected outliers $\mathbb{S}_\gamma = \{\gamma_{c_1}^0, \ldots, \gamma_{c_K}^0\}$ in decreasing order $\gamma_{c_1}^0 > \cdots > \gamma_{c_K}^0$. Choose a specific iteration step $m \in [0, \infty)$, which includes the fixed point $*$ when $m \to \infty$. Then, under the global null hypothesis $H_0 : \gamma_{c_k} = \gamma_{c_k}^0 \; \forall \, k \in \{1, \ldots, K\}$ and as $n \to \infty$,*

$$T_{sup,\mathbb{S}_c}^{(m)} = \sup_{1 \le k \le K} |n^{1/2}(\widehat{\gamma}_{c_k}^{(m)} - \gamma_{c_k}^0)| \xrightarrow{\mathsf{D}} \sup_{1 \le k \le K} |\mathsf{GP}(c_k; 0, K^{(m)})|,$$

where $\mathsf{GP}(c; 0, K^{(m)})$ for $c \in [c_+, \infty)$ refers to the Gaussian process with expected value $\mathsf{E}\{\mathsf{GP}(c; 0, K^{(m)})\} = 0$ and covariance $K_{st}^{(m)} = \mathsf{Cov}\{\mathsf{GP}(s; 0, K^{(m)}), \mathsf{GP}(t; 0, K^{(m)})\}$

for any $c_+ \leq s \leq t < \infty$ *with*

$$K_{st}^{(m)} = \psi_s\gamma_t^0 - s\phi_u(s)\varrho_{\sigma\sigma,s}^{(m)}(\psi_t - \tau_2^t) - t\phi_u(t)\varrho_{\sigma\sigma,t}^{(m)}(\psi_s - \tau_2^s) + \psi_s t\phi_u(t)\varrho_{\sigma uu,t}^{(m)}(\varsigma_s^2 - \varsigma_t^2) \tag{2.5.3}$$

$$+ st\phi_u(s)\phi_u(t)[\varrho_{\sigma\sigma,s}^{(m)}\{2\varrho_{\sigma\sigma,t}^{(m)} + \varrho_{\sigma uu,t}^{(m)}(\tau_4^t - \tau_2^t\varsigma_t^2)\} + \varrho_{\sigma uu,s}^{(m)}(\varrho_{\sigma\sigma,t}^{(m)} + \varrho_{\sigma uu,t}^{(m)})(\tau_4^s - \tau_2^s\varsigma_s^2)]$$

and the coefficients $\varrho_{\sigma\sigma,c}^{(m)}$ *and* $\varrho_{\sigma uu,c}^{(m)}$ *have the same expressions as in Theorem 2.5.1.*

The supremum norm over a multivariate normal process has no closed form for the cumulative distribution function (CDF) and must be simulated. For that purpose, draw S random vectors $x_s = (x_{1,s}, \ldots, x_{K,s})' \in \mathbb{R}^K$ from a multivariate normal distribution with variance-covariance structure specified by $K_{st}^{(m)}$. The simulated CDF for the supremum norm over the multivariate normal process can be obtained as $\widehat{\mathsf{F}}_{\sup|X|}(x|\mathbb{S}_c) = \sum_{s=1}^{S} 1(\sup_{1\leq k\leq K}|x_{k,s}| \leq x)/S$, where $1(\cdot)$ is an indicator function that evaluates to 1 if the condition inside the parentheses is true and 0 otherwise. Using α to denote the nominal significance level, we have the following decision rule:

Reject H_0 in favour of $H_1 : \exists\, k \in \{1, \ldots, K\} : \gamma_{c_k} \neq \gamma_{c_k}^0$ if $T_{sup,\mathbb{S}_c}^{(m)} > \widehat{\mathsf{F}}_{\sup|X|}^{-1}(1-\alpha|\mathbb{S}_c)$.

The formulas for the variance-covariance structure of the Gaussian process simplify for the initial iteration step $m = 0$ and the fixed point $m = *$ of the outlier detection algorithm. The formulas are given in the following remark.

Remark 2.5.3. *For* $m = 0$ *and* $m = *$*, the variance-covariance structure of the Gaussian process in the supremum test statistic in equation* (2.5.3) *simplifies to*

$$K_{st}^{(0)} = \psi_s\gamma_t^0 - s\phi_u(s)(\psi_t - \tau_2^t) - t\phi_u(t)(\psi_s - \tau_2^s) + 2st\phi_u(s)\phi_u(t)$$

and

$$K_{st}^{(*)} = \psi_s\gamma_t^0 + \frac{\psi_s t\phi_u(t)(\varsigma_s^2 - \varsigma_t^2)}{\tau_2^t - t\phi_u(t)(t^2 - \varsigma_t^2)} + \frac{st\phi_u(s)\phi_u(t)(\tau_4^s - \tau_2^s\varsigma_s^2)}{\{\tau_2^s - s\phi_u(s)(s^2 - \varsigma_s^2)\}\{\tau_2^t - t\phi_u(t)(t^2 - \varsigma_t^2)\}}.$$

Now that we have defined the different tests, we use Monte Carlo experiments to assess their finite sample performance in terms of size and power.

2.6 Simulation Study

This section assesses the finite sample performance of the different types of statistical tests using extensive Monte Carlo simulations.[1] The size of the tests is close to the nominal significance level for large sample sizes but varies considerably by the choice of algorithm and iteration step in small samples. The power of the tests increases with the sample size, outlier magnitude, and outlier frequency. If the outlier frequency becomes too high, the one-sided simple tests start to lose power again but the scaling tests keep high power. Even though the algorithms only use the residuals of the structural equation to identify outliers, the tests also have some power in the presence of leverage points.

In the simulations, we vary the following settings: the outlier detection Algorithms 2.R, 2.S, and IIS; the iteration step $m \in \{0, *\}$; the tuning parameter c for the proportion test, c_n for the count test, and the set of tuning parameters $\mathbb{S}_c$ for the scaling tests. We also vary the sample size n to examine its impact on the performance of the tests. When using Algorithms 2.R or 2.S, we use $M = 10{,}000$ replications per Monte Carlo setting. For Algorithm IIS, we restrict the settings that we assess and their number of replications for computational reasons. The exact settings and number of replications are described below.

2.6.1 Size

To assess the size of the tests, we use the same data generating process (DGP) as in Chapter 1. The simulations show that the empirical size of all tests is close to the nominal size when using Algorithm 2.R but that Algorithm 2.S yields oversized tests unless the algorithm is iterated and/or the sample size is large.

1. The simulations were performed using R Statistical Software (v4.1.2; R Core Team 2022) in RStudio (Posit Team 2023). We used the tidyverse R packages (Wickham et al. 2019) to evaluate and visualise the simulation results.

2.6.1.1 DGP

The structural equation consists of an intercept and one endogenous regressor, which can be written as

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \qquad i = 1, 2, \ldots, n,$$

where $x_{1i} = 1$ for all i and the structural parameters are chosen as $\beta = (\beta_1, \beta_2)' = (2, 4)'$. We have one excluded and informative instrument, such that the first stage projection is

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix} + \begin{pmatrix} 0 \\ r_{2i} \end{pmatrix}, \qquad i = 1, 2, \ldots, n,$$

where the first stage coefficient matrix Π is the identity matrix I_2 and therefore has full rank. For the reference distribution, we assume joint normality of the errors $\phi_{u,r}$, variance $\sigma^2 = 1$ of the structural error and variance $\Sigma_2 = 1$ of the first stage error, that is

$$\begin{pmatrix} u_i \\ r_{2i} \end{pmatrix} \overset{\mathsf{D}}{=} \mathsf{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \Omega_2 \\ \Omega_2 & 1 \end{pmatrix} \right\}.$$

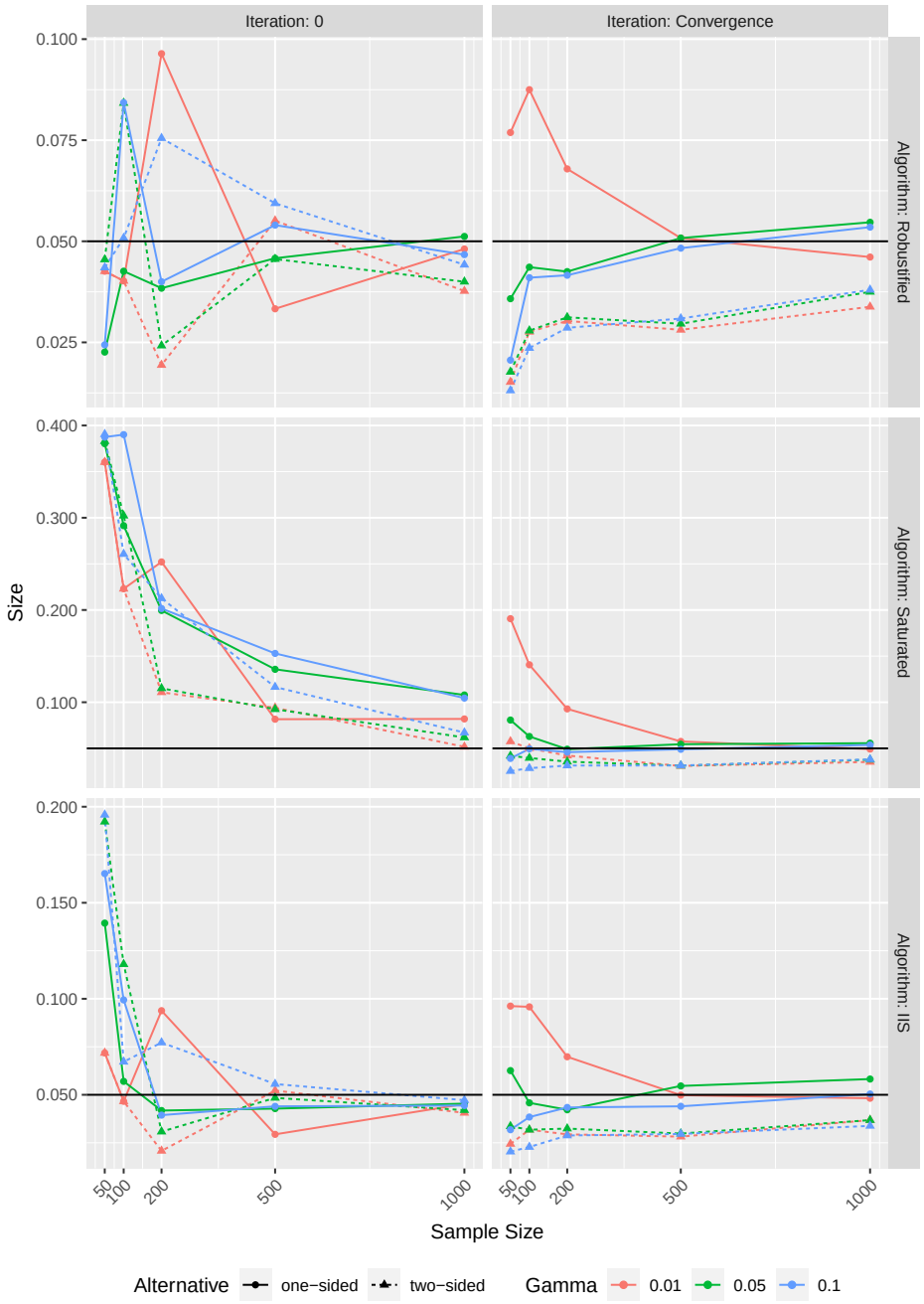
Under joint normality, the asymptotic distributions of the tests statistics do not depend on the degree of endogeneity, measured by the covariance between the first stage and structural error, Ω_2. The simulations in Chapter 1 furthermore indicate that the finite sample performance of the asymptotic approximation to the FODR is also virtually the same across a range of Ω_2 values from 0 to 0.75. We therefore focus on the case with the highest endogeneity and choose $\Omega_2 = 0.75$. The above error distribution is the reference distribution and hence used in the non-contaminated setting to assess the size of the statistical tests. The sample size takes on the values $n \in \{50, 100, 200, 500, 1000\}$.

2.6.1.2 Proportion Test

Figure 2.6.1 shows the size of the one- and two-sided proportion test with tuning parameters $c \in \{2.58, 1.96, 1.64\}$, corresponding to $\gamma_c^0 \in \{0.01, 0.05, 0.1\}$. We can see that the size of the proportion test is closer to the nominal significance level α when the procedure is iterated. For Algorithm 2.R, the size is close to α for

moderate sample sizes of $n \geq 500$. The size of Algorithms 2.S and IIS exceeds α for small n but iterating until convergence allows for good size control for moderate sample sizes of $n \geq 200$.

Figure 2.6.1: Size of Proportion Test, $\alpha = 0.05$



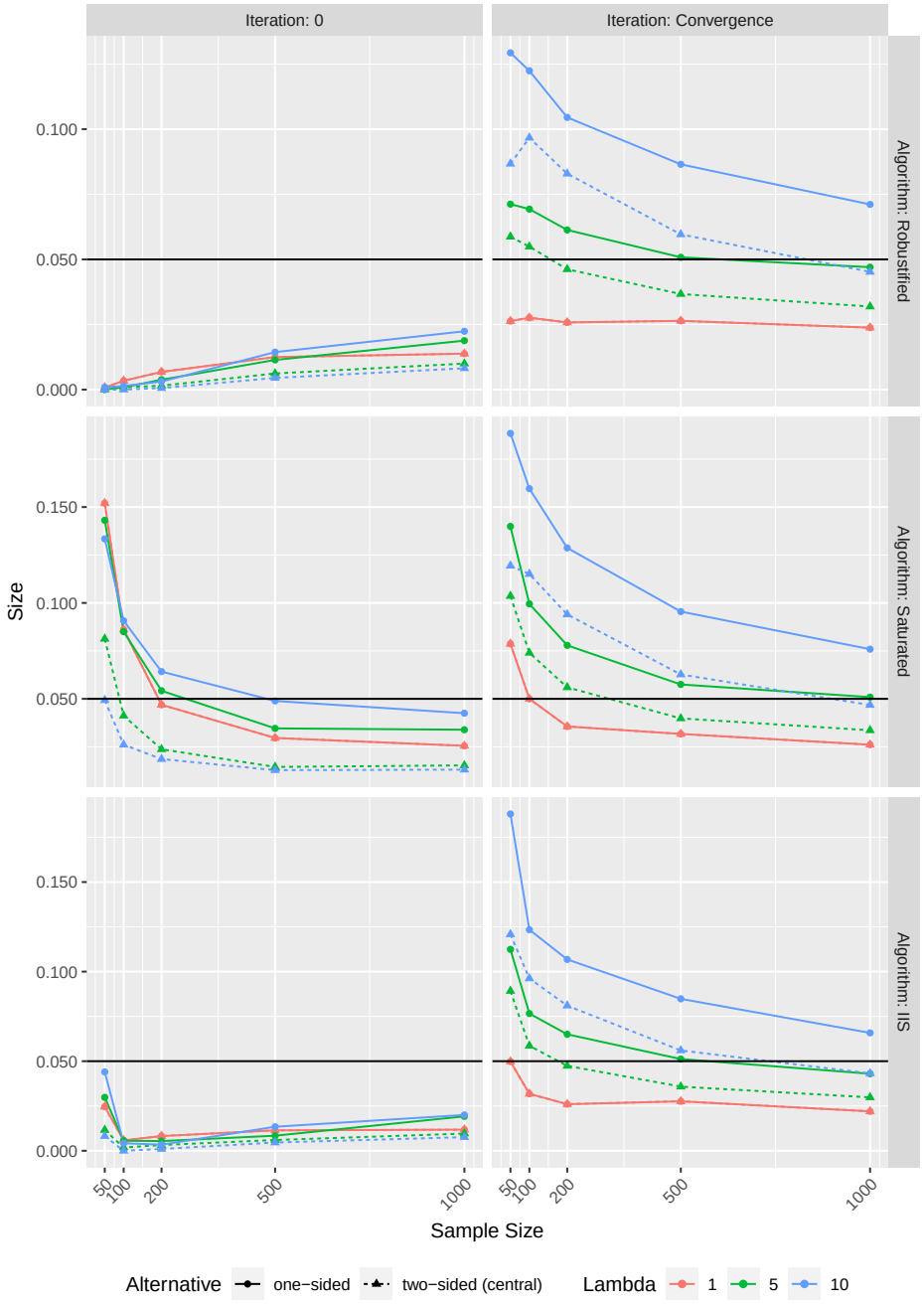
Comparison of the simulated size to the nominal significance level $\alpha = 0.05$ (black). Several settings are varied: the sample size n, the iteration step m, the algorithm, the tuning parameter c to yield γ^0, and the alternative hypothesis. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 5000$ replications.

2.6.1.3 Count Test

Figure 2.6.2 displays the size of the one- and two-sided count test with tuning parameter c_n chosen to result in $\lambda^0 \in \{1, 5, 10\}$. The count test tends to be undersized for iteration $m = 0$ for moderate to large n, which is due to the discreteness of the Poisson distribution, as explained above. As was the case for the proportion test, Figure 2.6.2 shows that iterating the algorithms leads to better size control for a sample size of $n = 500$ and $n = 1000$ but asymptotically, the test is still expected to be undersized. Targeting a higher expected number of outliers leads to higher size, including oversized performance for small sample sizes. Note that one would typically not choose a target level of 10 expected falsely classified outliers when the sample size is only 50 or 100, such that the oversized performance is not too concerning in these cases. The asymptotic theory, however, requires the expected value of the Poisson distribution, λ, to be fixed across the sample size n by letting the cut-off value c_n grow with n and we therefore still have the same hypothesised λ^0 across the whole range of n.

For the count test, the size may not just deviate from the nominal significance level α due to finite sample issues but also because the count test follows a discrete distribution. That is, even as $n \to \infty$ or if drawing from an exact Poisson distribution, the size will generally differ from α. To illustrate this, suppose that our count test statistic were to exactly follow a Poisson distribution, $X \overset{D}{=} \text{Poisson}(\lambda)$ with $\lambda = 1$. Suppose we do a one-sided test using a significance level of $\alpha = 0.05$, i.e. we reject if the p-value is less than 0.05. The issue is that the $0.95-$quantile is not uniquely defined and $\mathsf{P}(X \geq 3) \approx 0.08$ and $\mathsf{P}(X \geq 4) \approx 0.019$, which means that we only reject the null hypothesis when the count test statistic is 4 or larger, which only occurs in 1.9% of the cases, not in 5%. The chosen decision rule in the simulations is therefore conservative and applied researchers are free to choose other decision rules that have been proposed in the literature.

Figure 2.6.2: Size of Count Test, $\alpha = 0.05$



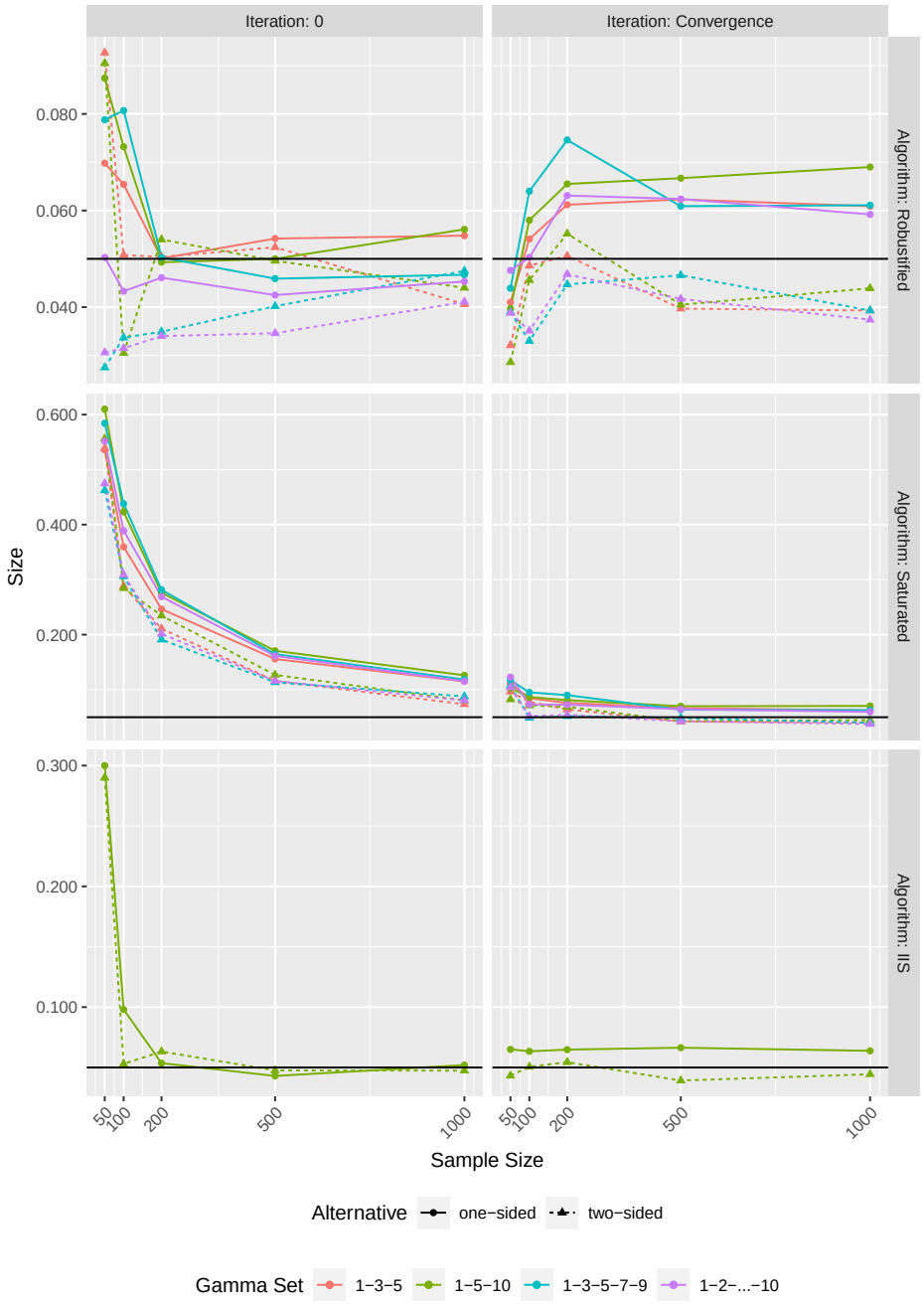
Comparison of the simulated size to the nominal significance level $\alpha = 0.05$ (black). Several settings are varied: the sample size n, the iteration step m, the algorithm, the tuning parameter c_n to yield λ^0, and the alternative hypothesis. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 5000$ replications.

2.6.1.4 Multiple Proportion and Count Test

The multiple proportion and count tests in Figures 2.6.3 and 2.6.4 follow the same patterns as their single cut-off counterparts. On average, however, the size distortions are exacerbated for small sample sizes. For example, for small sample sizes n

and for Algorithms 2.S and IIS, the size of the multiple proportion test is now even higher than for the single proportion test. Again, iterating the algorithms tends to yield sizes closer to the nominal significance level.

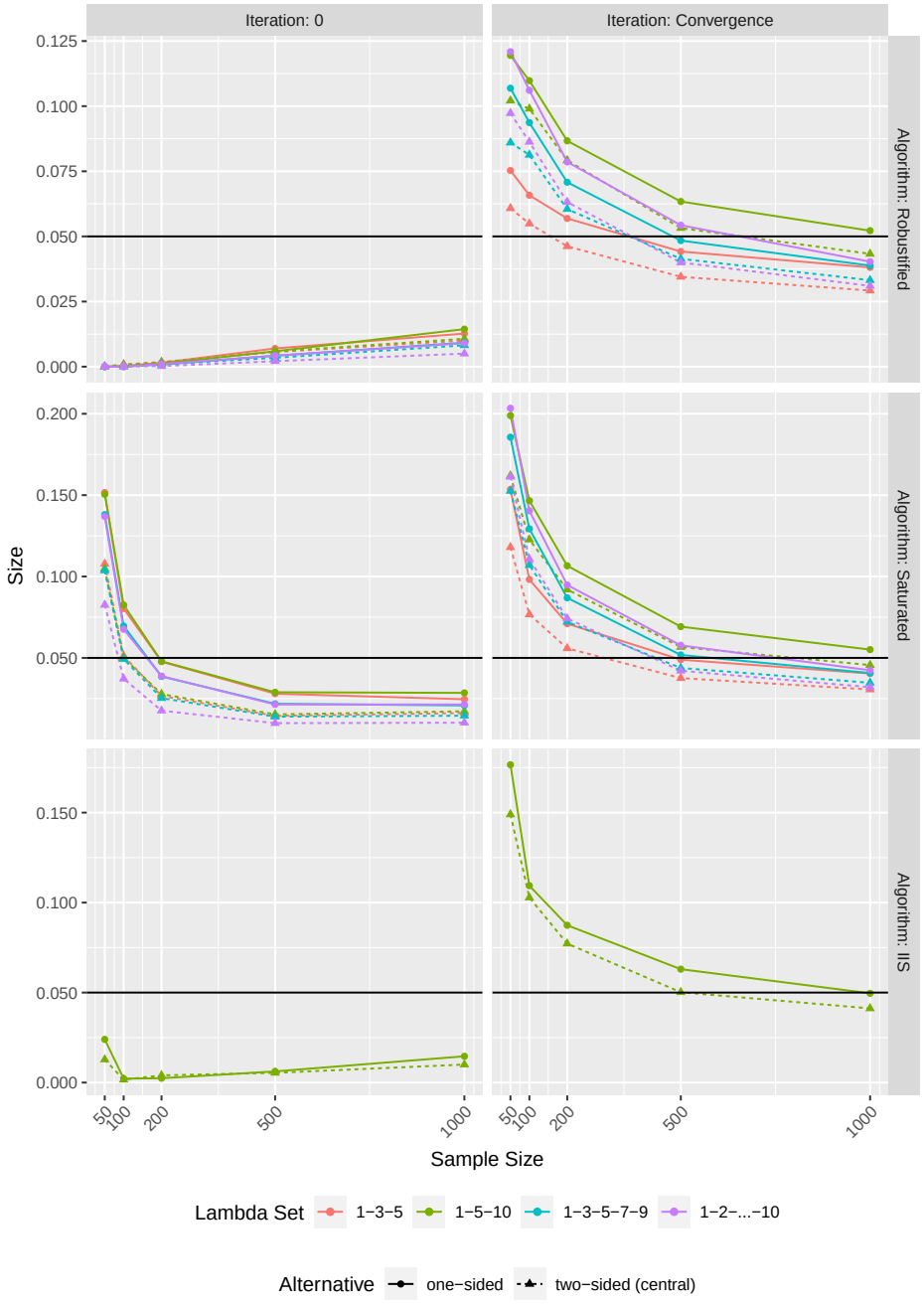
Figure 2.6.3: Size of Multiple Proportion Test, $\alpha = 0.05$



Comparison of the simulated size to the nominal significance level $\alpha = 0.05$ (black). Several settings are varied: the sample size n, the iteration step m, the algorithm, the set of tuning parameters $\mathbb{S}_c$ to yield $\mathbb{S}_\gamma$, and the alternative hypothesis. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 5000$ replications.

For Algorithms 2.R and 2.S, four types of multiple proportion (count) tests

Figure 2.6.4: Size of Multiple Count Test, $\alpha = 0.05$



Comparison of the simulated size to the nominal significance level $\alpha = 0.05$ (black). Several settings are varied: the sample size n, the iteration step m, the algorithm, the set of tuning parameters $\mathbb{S}_c$ to yield $\mathbb{S}_\lambda$, and the alternative hypothesis. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 5000$ replications.

with varying sets of tuning parameters $\mathbb{S}_c$ to target $\mathbb{S}_\gamma$ ($\mathbb{S}_\lambda$) are displayed. For Algorithm IIS, we only show one test with 3 different λ^0 values in order to limit the computational costs.

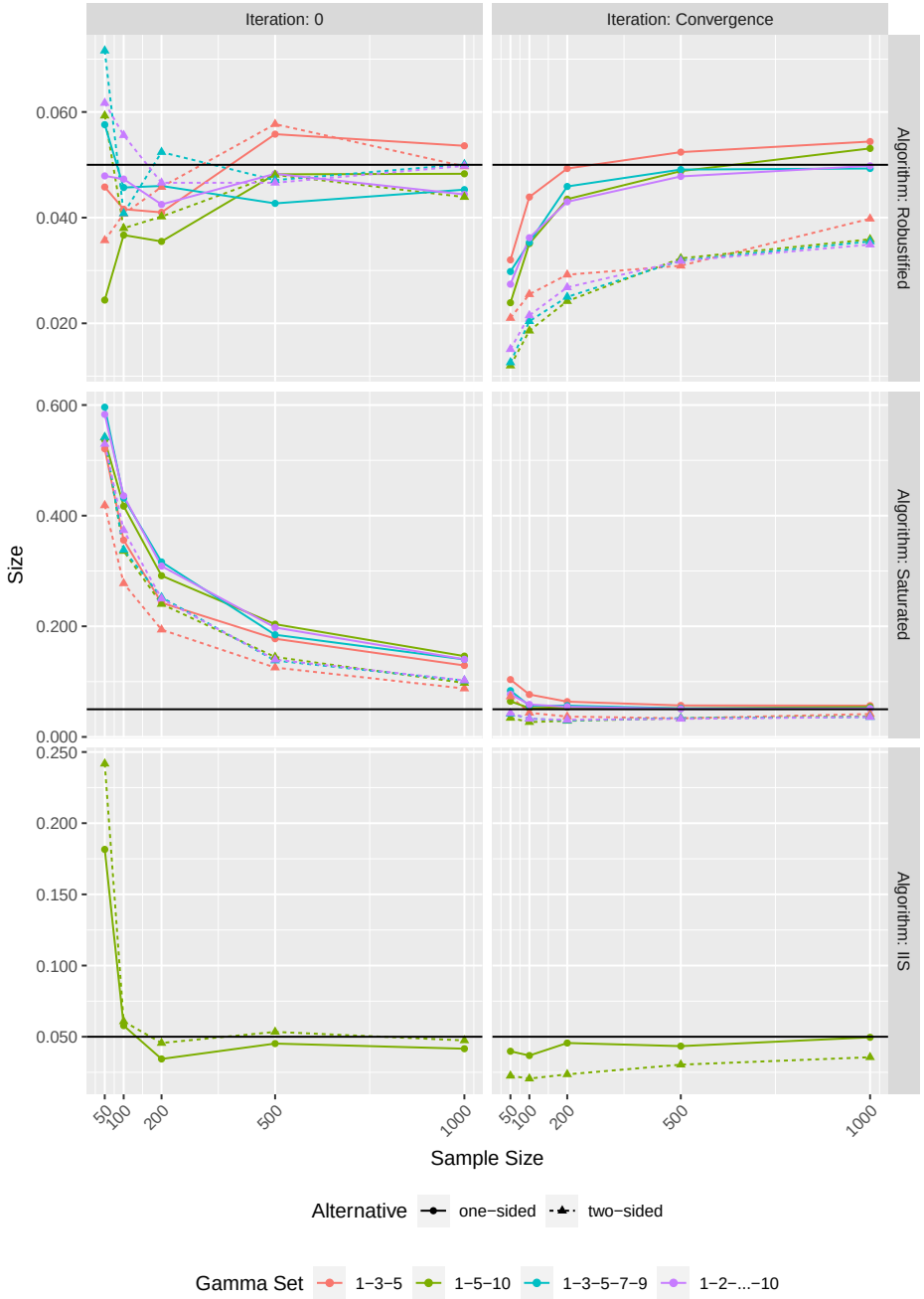
For the multiple proportion test, the size performance is very similar across

the different types of tests. For the multiple count test, the tests that include large values of λ^0 tend to have higher sizes. For example, $\mathbb{S}_\lambda = \{1, 3, 5, 7, 9\}$ and $\mathbb{S}_\lambda = \{1, 2, \ldots, 10\}$ have higher size than the test with $\mathbb{S}_\lambda = \{1, 3, 5\}$. This is probably driven by the fact that the underlying individual count tests with a higher λ^0 value tend to have higher size, as shown in Figure 2.6.2, especially for small sample sizes. As the sample size grows, this difference in size between the different sets of $\mathbb{S}_\lambda$ values becomes smaller.

2.6.1.5 Sum and Supremum Test

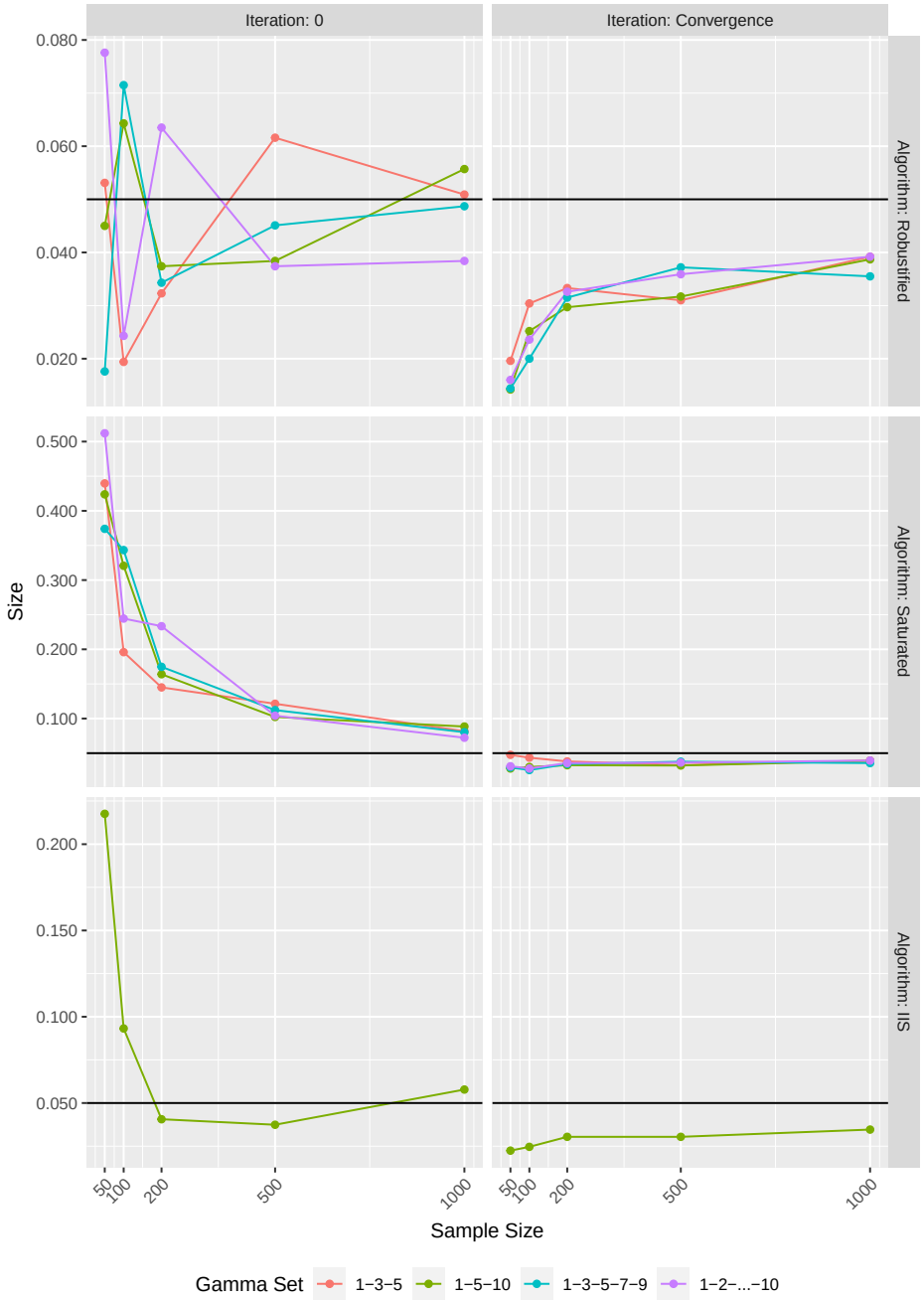
The sum and supremum tests shown in Figures 2.6.5 and 2.6.6 cover the same sets of tuning parameters $\mathbb{S}_c$ as the multiple proportion test in Figure 2.6.3. The performance between these three types of scaling tests is similar. The supremum test is one-sided only and performs similarly to the two-sided sum test, which is likely because they both reject for deviations from the null hypothesis in both directions — by taking the maximum over absolute deviations, the supremum test weighs positive and negative deviations from the null hypothesis equally. The sum and the supremum test tend to have slightly smaller size distortions than the multiple proportion test but are more conservative for the iterated procedures, i.e. they tend to be undersized.

Figure 2.6.5: Size of Sum Test, $\alpha = 0.05$



Comparison of the simulated size to the nominal significance level $\alpha = 0.05$ (black). Several settings are varied: the sample size n, the iteration step m, the algorithm, the set of tuning parameters $\mathbb{S}_c$ to yield $\mathbb{S}_\gamma$, and the alternative hypothesis. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 5000$ replications.

Figure 2.6.6: Size of Supremum Test, $\alpha = 0.05$



Comparison of the simulated size to the nominal significance level $\alpha = 0.05$ (black). Several settings are varied: the sample size n, the iteration step m, the algorithm, and the set of tuning parameters $\mathbb{S}_c$ to yield $\mathbb{S}_\gamma$. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 5000$ replications.

2.6.2 Power Epsilon-Contamination

In this section, we contaminate the structural error with outliers, corresponding to the approach of identifying outliers in the algorithms based on the standardised residuals of the structural equation. For all tests except the count test, size-adjusted power increases with the sample size and outlier magnitude. Since the count test targets an expected number of falsely detected outliers, the cut-off value increases more and more as the sample size grows, such that it loses power against smaller outliers. For the simple tests, power increases initially as the outlier frequency increases but starts to fall again if the contamination becomes too high. The scaling tests keep their power even in highly contaminated settings. Except in a few cases, iterating the algorithms tends to yield higher size-adjusted power by controlling size better while only slightly losing raw power.

2.6.2.1 DGP

We use Huber's (1964) ϵ-contamination framework, where the error is drawn from the good reference distribution f with probability $1-\epsilon$ and from the contaminating distribution f^{cont} with probability ϵ. We only contaminate the structural error u_i and r_{2i} still follows $\mathsf{N}(0,1)$ in the contaminated scenario. With probability $1-\epsilon$, u_i is drawn from a standard normal distribution as in Section 2.6.1.1. With probability ϵ, however, u_i is drawn from a discrete distribution $\mathsf{f}_u^{\text{cont}}(y)$, where

$$\mathsf{f}_u^{\text{cont}}(y) = \begin{cases} 1/2 & \text{for } y = \delta \\ 1/2 & \text{for } y = -\delta \\ 0 & \text{otherwise.} \end{cases}$$

The parameter δ represents the outlier magnitude. The different scenarios translate into different true values of γ_c^1, the probability of drawing an error beyond the cut-off value under the contamination alternative. The exact probabilities γ_c^1 for these two cases are given as

$$\gamma_c^1 = \mathsf{P}(|u_i| > \sigma c) = \begin{cases} \epsilon + (1-\epsilon)\gamma_c^0 & > \gamma_c^0 \text{ if } \delta > c, \\ (1-\epsilon)\gamma_c^0 & < \gamma_c^0 \text{ if } \delta \leq c. \end{cases} \tag{2.6.1}$$

Note that $\epsilon \in (0,1)$ is the probability of drawing from the contaminated distribution and that $\gamma_c^0 \in (0,1)$, determined by the tuning parameter c, represents the probability of drawing an error beyond the cut-off value c under the normal reference distribution.

In the first case when $\delta > c$, the error is draw from the contaminated distribution with probability ϵ, which yields a value that is larger than c in absolute value with certainty. With probability $1-\epsilon$, the error is drawn from the reference distribution, which produces absolute errors beyond the cut-off value c with probability γ_c^0. In the second case when $\delta \leq c$, only the reference distribution can produce absolute errors beyond the cut-off value c, which occurs with probability $(1-\epsilon)\gamma_c^0$.

To summarise, if the outlier magnitude δ is larger than the chosen cut-off value c, this yields a γ_c^1 that is larger than γ_c^0. If however the outlier magnitude δ is smaller than the chosen cut-off value c, this yields a γ_c^1 that is smaller than γ_c^0 and therefore is unlikely to be detected when testing against the one-sided alternative only.

Table 2.6.1: Contamination Scenarios and Values of γ_c^1

			ϵ				
γ_c^0	c	δ	0.001	0.005	0.01	0.02	0.03
0.01	2.58	2	0.00999	0.00995	0.0099	0.0098	0.0097
		3	0.01099	0.01495	0.0199	0.0298	0.0397
		4	0.01099	0.01495	0.0199	0.0298	0.0397
0.05	1.96	2	0.05095	0.05475	0.0595	0.069	0.0785
		3	0.05095	0.05475	0.0595	0.069	0.0785
		4	0.05095	0.05475	0.0595	0.069	0.0785
			ϵ				
γ_c^0	c	δ	0.05	0.075	0.1	0.125	0.15
0.01	2.58	2	0.0095	0.00925	0.009	0.00875	0.0085
		3	0.0595	0.08425	0.109	0.13375	0.1585
		4	0.0595	0.08425	0.109	0.13375	0.1585
0.05	1.96	2	0.0975	0.12125	0.145	0.16875	0.1925
		3	0.0975	0.12125	0.145	0.16875	0.1925
		4	0.0975	0.12125	0.145	0.16875	0.1925

Note: Values of gamma under the Alternative, γ_c^1, for different combinations of cut-off value c, outlier magnitude δ, and outlier frequency ϵ. Highlighted in grey are scenarios where γ_c^1 under contamination is smaller than the hypothesised outlier share γ_c^0.

Table 2.6.1 illustrates how the different deviations from the null hypothesis translate into different γ_c^1 for different choices of the tuning parameter c. It is worth pointing out the special case where the probability of a true outlier, ϵ, is equal to the hypothesised value of detected outliers when there is in fact no contamination, i.e. γ_c^0. For example, suppose we choose $c = 1.96$, which means that we expect to (falsely) classify 5% of the sample as outliers even when there is no contamination, i.e. $\gamma_c^0 = 0.05$. Further suppose that the sample is contaminated and that the probability of drawing an error from the contaminating distribution is also $\epsilon = 5\%$. In that case, we might expect intuitively that we have no power. However, Table 2.6.1 shows that $\gamma_c^1 = 0.0975$ because even the reference distribution produces "large" values with probability γ_c^0. Therefore, because $\gamma_c^1 > \gamma_c^0$, we still have a chance to reject the null hypothesis of no outliers even when the true outlier frequency equals the hypothesised frequency when there is no contamination.

To summarise, we deviate from the null hypothesis of no outliers in two dimensions: the probability of drawing an error from the contaminating distribution ϵ and the outlier magnitude of the contaminating distribution δ. In the simulations, they take on the values $\epsilon \in \{0.001, 0.005, 0.01, 0.02, 0.03, 0.05, 0.075, 0.1, 0.125, 0.15\}$ and $\delta \in \{2, 3, 4\}$.

Note that due to the stochastic nature of the contamination scheme, the signs of the outliers need not be balanced and not every Monte Carlo replication has exactly $\epsilon \times n$ outliers — some replications will have more, others will have fewer. Furthermore, the location of the outliers is chosen randomly as a uniform distribution across all observations. The sample size is varied across $n \in \{100, 500, 1000\}$.

In order to compare the power across tests, we calculate size-adjusted power with a target significance level of $\alpha = 0.05$. For the proportion, count, sum, and supremum test, this means finding the critical values such that the size simulations would yield an exact false rejection rate of 5%. These adjusted critical values are then also used to calculate power. The multiple testing procedures on the proportion and count test are adjusted differently. The global significance level α is chosen such that the size of the global null hypothesis of no outliers is as close to 5% as possible.

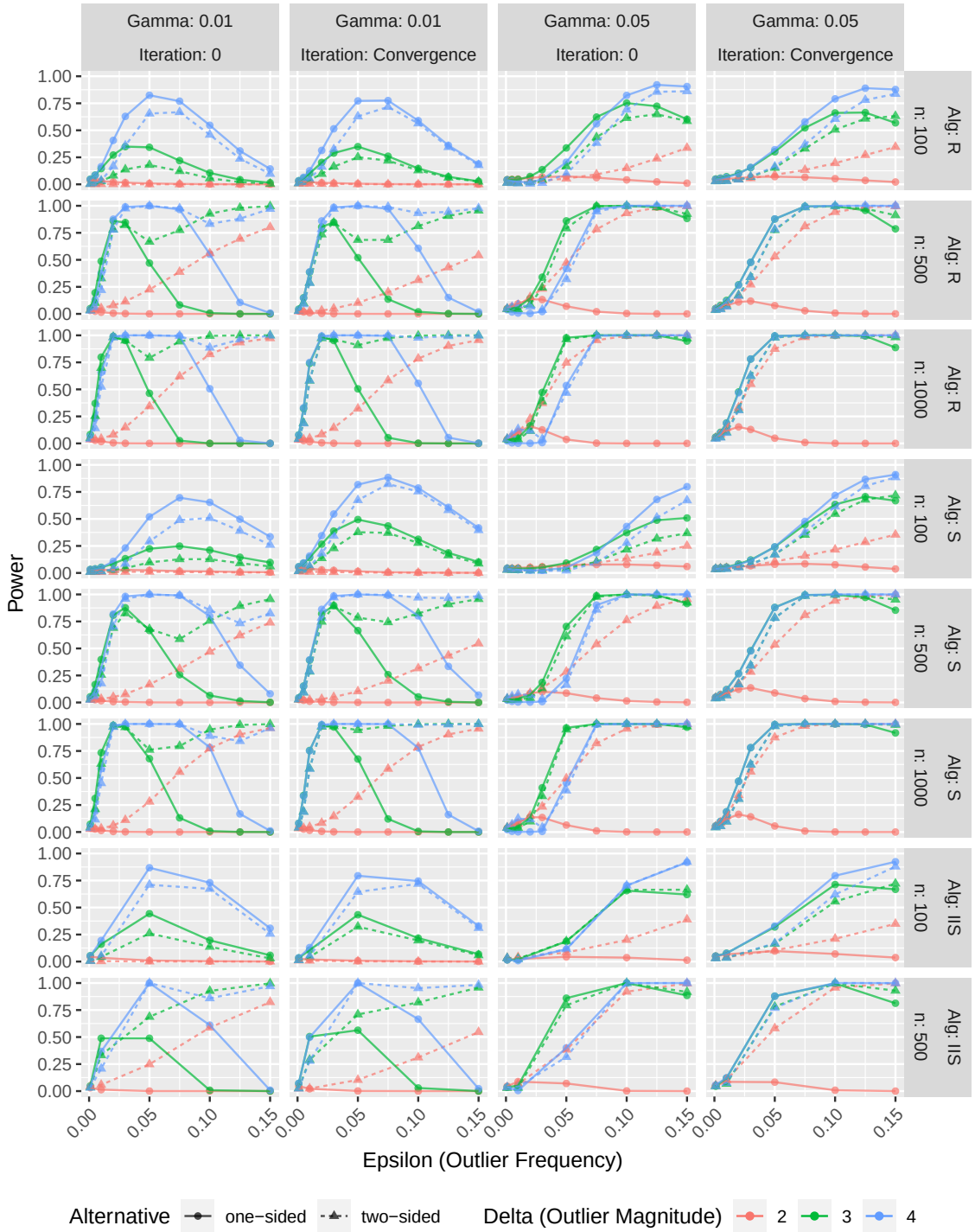
This involves solving for such an adjusted α using numerical optimisation, which sometimes does not yield an exact size of 5% but is still closer to our target than using an unadjusted nominal significance level of 0.05.

2.6.2.2 Proportion Test

Figure 2.6.7 plots the power of the proportion test against the outlier frequency ϵ for a significance level of $\alpha = 0.05$. The first two columns compare the initial step and fixed point results for the tuning parameter $c = 2.58$, while columns 3 and 4 show the corresponding results for $c = 1.96$. The rows represent different sample sizes and initial estimators. Each group of two or three rows represents a different algorithm and within each algorithm, the rows are ordered in increasing sample size. Different colours represent different outlier magnitudes δ and the marker and line type indicate whether a one- or two-sided test was conducted.

Several patterns are visible from the simulations. First, power tends to increase with the sample size. Second, power initially increases as the outlier frequency ϵ increases but starts to fall for the one-sided test if it becomes too high. The intuition is that as outliers become too frequent, the initial estimate of the variance, which is used to classify observations as outliers in step $m = 0$, becomes inflated. An overestimated variance causes the standardised residuals to be smaller and therefore less likely to be beyond the cut-off value c. As a consequence, fewer observations are classified as outliers and therefore $\widehat{\gamma}_c^{(m)}$ is likely smaller than γ_c^0. While the one-sided test has no power in this case, the two-sided test tends to either keep or start gaining power for larger values of ϵ. Third, power tends to increase in accordance with the outlier magnitude δ. As an exception, for $\gamma_c^0 = 0.05$ and iteration step $m = 0$, the power is higher for the smaller outlier magnitude $\delta = 3$ than $\delta = 4$. This could be for similar reasons as the drop in power if ϵ becomes too large: a larger outlier magnitude is more likely to cause an inflated variance estimate and therefore fewer detected outliers than γ_c^0. However, the difference in power vanishes when the outlier detection algorithm is iterated. The size-adjusted power performances between the initial step and the fixed point seem similar or increasing in the iteration step.

Figure 2.6.7: Power of Proportion Test, $\alpha = 0.05$



Simulated power (size-adjusted) for different outlier frequencies ϵ and outlier magnitudes δ. Several settings are varied: the sample size n, the iteration step m, the algorithm, and the tuning parameter c to yield γ^0. Significance level $\alpha = 0.05$. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 2500$.

For Algorithm 2.S and for the smaller sample sizes 100 and 500, the raw power (not shown) is higher for the initial step than for the fixed point, which is due to the fact that the proportion test is highly oversized in these cases. The size-

adjusted power shown in Figure 2.6.7 corrects for this and in order to control size, it is recommended to use the test based on the fixed point or to use Algorithm 2.R instead.

2.6.2.3 Count Test

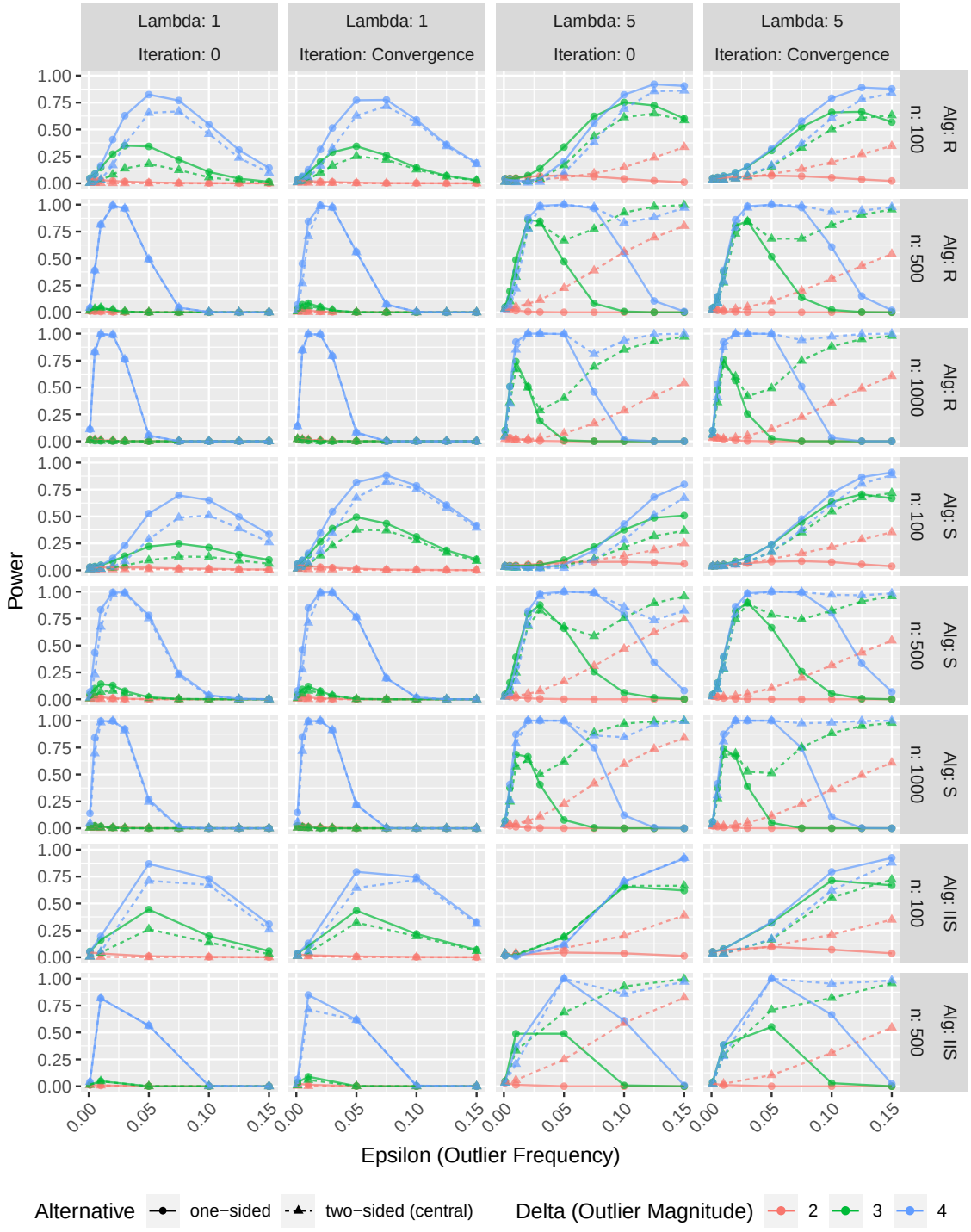
The count test has similar power to that of the proportion test when $n = 100$, i.e. when they use the same cut-off values (1.96 for $\gamma_c^0 = 0.05, \lambda_{c_{100}}^0 = 5$ and 2.58 for $\gamma_c^0 = 0.01, \lambda_{c_{100}}^0 = 1$), as shown in Figure 2.6.8. For larger sample sizes and the smaller outlier magnitudes $\delta = 2$ and 3, the power of the one-sided test is quickly falling to zero. The reason is that targeting only one or five expected false detections as the sample size grows means that the cut-off must grow. For $\lambda_{c_{500}}^0 = 1$, this means a cut-off of $c_{100} \approx 3.1$ and for $\lambda_{c_{1000}}^0 = 1$, the cut-off needs to be set at $c_{1000} \approx 3.3$. Therefore, actual outliers of magnitude 2 and 3 are not detected any more. The conservative choice of only allowing for a small expected number of falsely detected outliers under the null comes at the expense of limited power against outliers that are smaller than the cut-off value. Iterating the procedure does not seem to increase size-adjusted power much, such that iteration step $m = 0$ might be preferable because it has lower size. Moreover, as with the proportion test, the power performance of Algorithms 2.R and 2.S are comparable.[2]

2.6.2.4 Multiple Proportion and Count Test

Figures 2.6.9 and 2.6.10 display the power of the multiple proportion and count test using the Simes (1986) procedure to correct for multiple hypothesis testing. Testing across a range of tuning parameters increases power greatly. Most prominently, the tests have higher power against both very large and very frequent outliers.

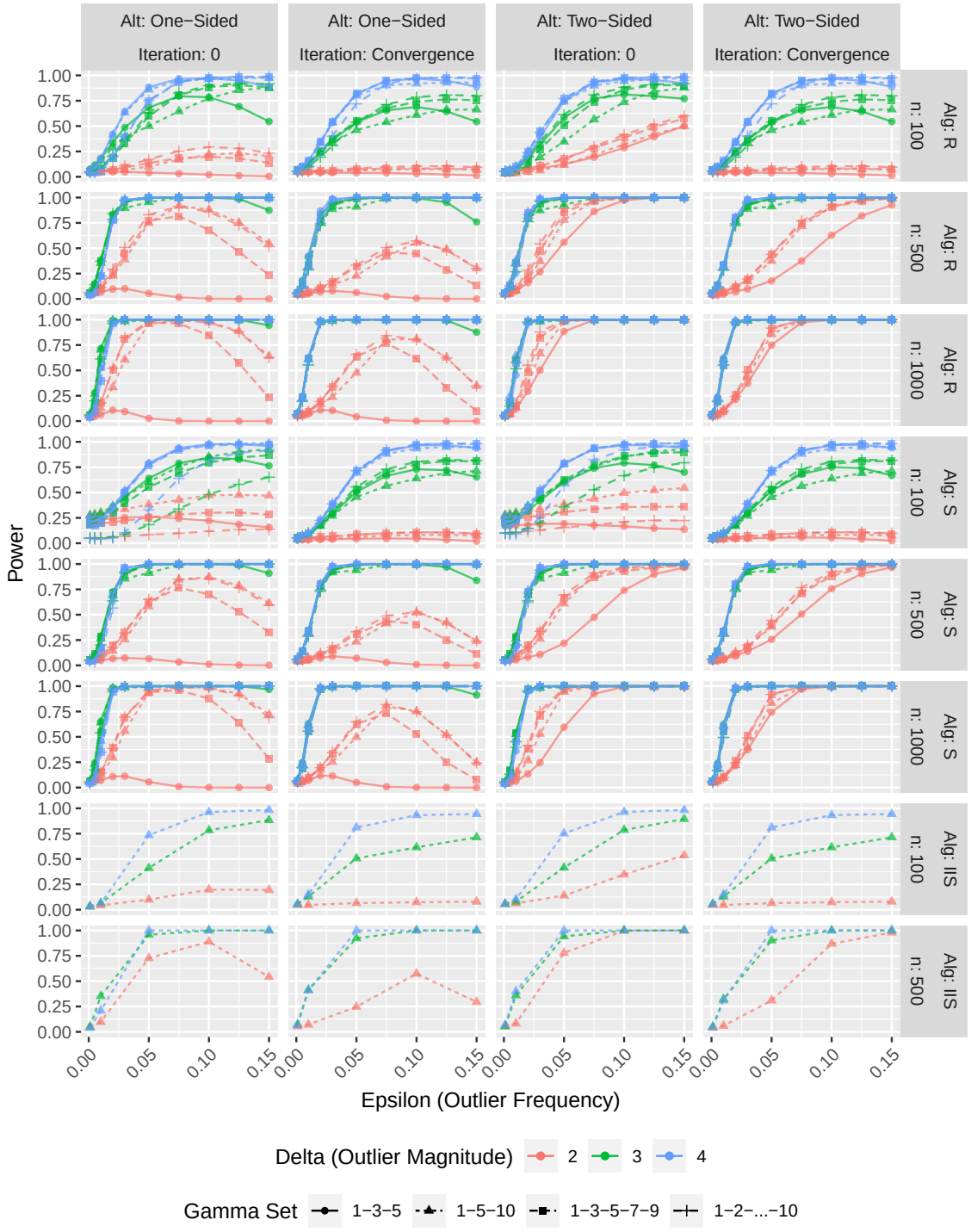
2. The one- and two-sided tests have the same *raw* power for $\lambda_{c_n}^0 = 1$ and therefore also have similar size-adjusted power. The reason for the identity of raw power is as follows. For negative deviations from the expected value $\lambda_{c_n}^0 = 1$, only the value 0 is possible. However, such a value is highly likely when the expected value is 1, namely $\mathsf{P}(X = 0) \approx 0.37$ and therefore much larger than the significance level of $\alpha = 0.05$. Hence, if we reject the null hypothesis, it must be for positive deviations, i.e. for the one-sided alternative $\lambda > \lambda^0$. For this one-sided test, we reject when 4 or more outliers have been detected. The associated p-value is $\mathsf{P}(X \geq 4) \approx 0.019$, which also yields a two-sided p-value below the significance level of 5% when following the "twice the smaller tail" method, $p_{\text{two-sided}}^{\text{central}} \approx 0.038$. As a consequence, the one- and two-sided tests yield the same decision in this case.

Figure 2.6.8: Power of Count Test, $\alpha = 0.05$



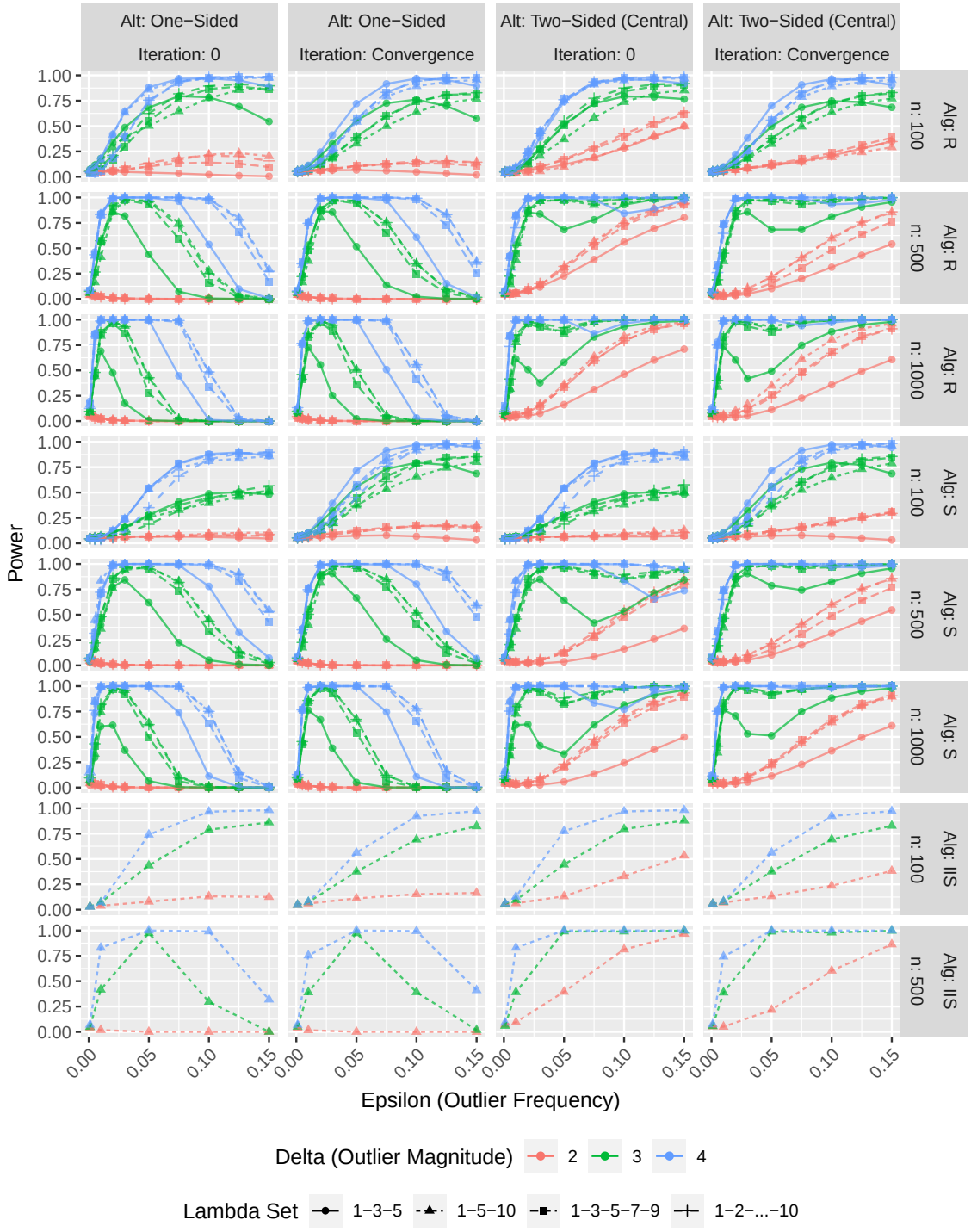
Simulated power (size-adjusted) for different outlier frequencies ϵ and outlier magnitudes δ. Several settings are varied: the sample size n, the iteration step m, the algorithm, the tuning parameter c_n to yield λ^0, and the alternative hypothesis. Significance level $\alpha = 0.05$. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 2500$.

Figure 2.6.9: Power of Multiple Proportion Test, $\alpha = 0.05$



Simulated power (size-adjusted) for different outlier frequencies ϵ and outlier magnitudes δ. Several settings are varied: the sample size n, the iteration step m, the algorithm, and the set of tuning parameters $\mathbb{S}_c$ to yield $\mathbb{S}_\gamma$. Significance level $\alpha = 0.05$. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 2500$.

Figure 2.6.10: Power of Multiple Count Test, $\alpha = 0.05$



Simulated power (size-adjusted) for different outlier frequencies ϵ and outlier magnitudes δ. Several settings are varied: the sample size n, the iteration step m, the algorithm, and the set of tuning parameters $\mathbb{S}_c$ to yield $\mathbb{S}_\lambda$. Significance level $\alpha = 0.05$. Monte Carlo simulations for Algorithms $2.R$ and $2.S$ used $M = 10{,}000$ replications, Algorithm IIS used $M = 2500$.

2.6.2.5 Sum Test

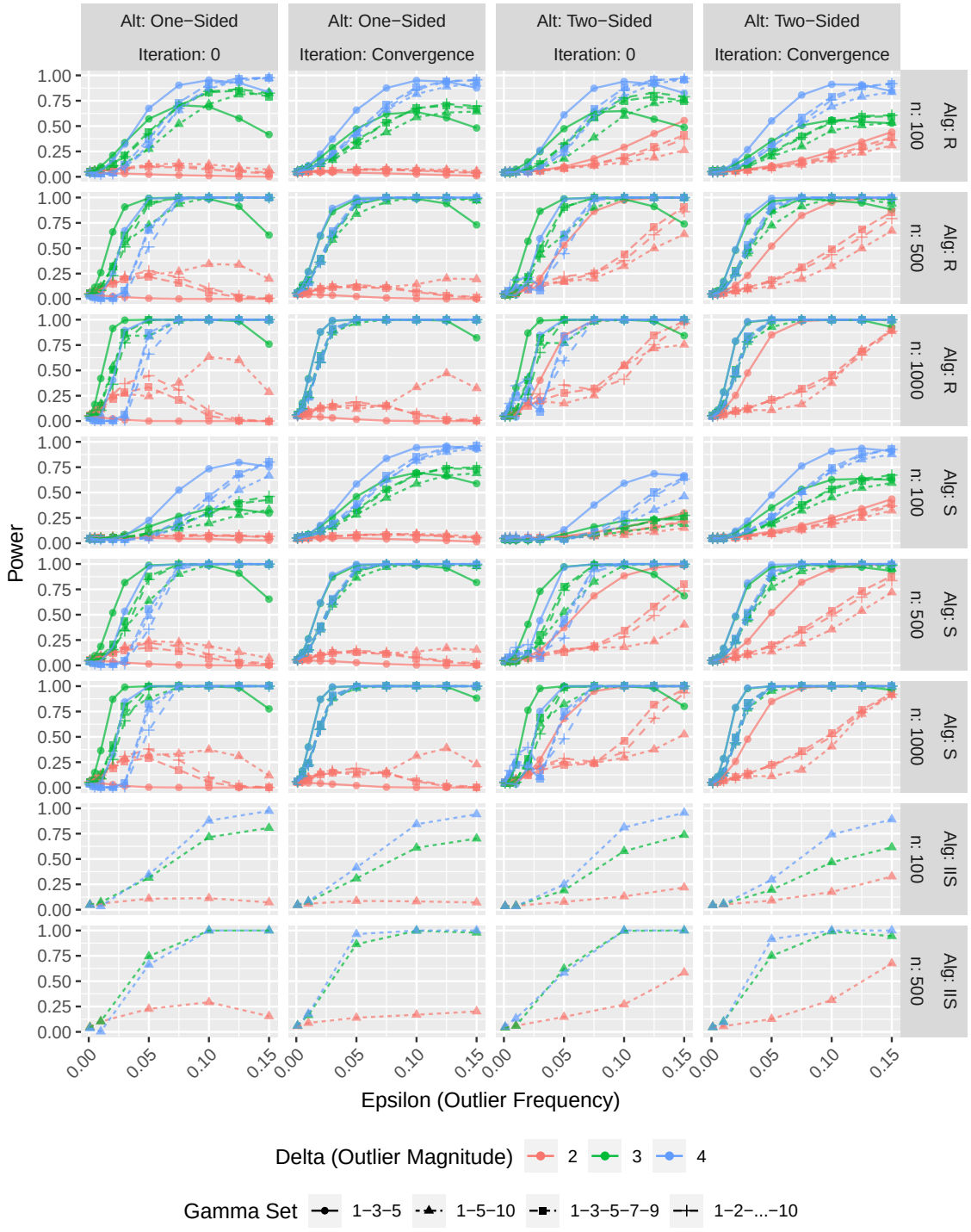
Figure 2.6.11 plots the power of the sum test against the outlier frequency ϵ. The first two columns show the results for the test against the one-sided alternative and the last two columns for the two-sided test. Within the pair of columns, the iteration step is varied, showing the performance of the sum test based on the initial step and the fixed point. The rows vary the algorithm and within each algorithm, two or three different sample sizes are shown in increasing order. The colour of the lines represents the outlier magnitude δ and the marker and line type represent different sets of tuning parameters $\mathbb{S}_c$.

As the sample size increases, power tends to increase for all settings. In contrast to the simple tests, the one-sided sum test now also has — at least some — power when the outlier magnitude is $\delta = 2$. Using a wide range of cut-off values allows us to detect outliers at different magnitudes. For example, the test with $\mathbb{S}_\gamma = \{0.05, 0.03, 0.01\}$ uses the cut-off values $\mathbb{S}_c = \{1.96, 2.17, 2.58\}$, which means that only one of the cut-off values is barely below the outlier magnitude. In contrast, the test with $\mathbb{S}_\gamma = \{0.1, 0.05, 0.01\}$ uses the cut-off values $\mathbb{S}_c = \{1.64, 1.96, 2.58\}$ and therefore has better chances to detect outliers of magnitude $\delta = 2$. Indeed, the sum test with the wider range of tuning parameters has power where the other test does not, especially for larger sample sizes. The two-sided test still has higher power for $\delta = 2$, which is likely due to an inflated variance estimate as with the proportion test. The iterated version of the sum test seems to have higher power than the non-iterated version, especially improving power against the larger outlier magnitudes $\delta = 3$ and 4.

2.6.2.6 Supremum Test

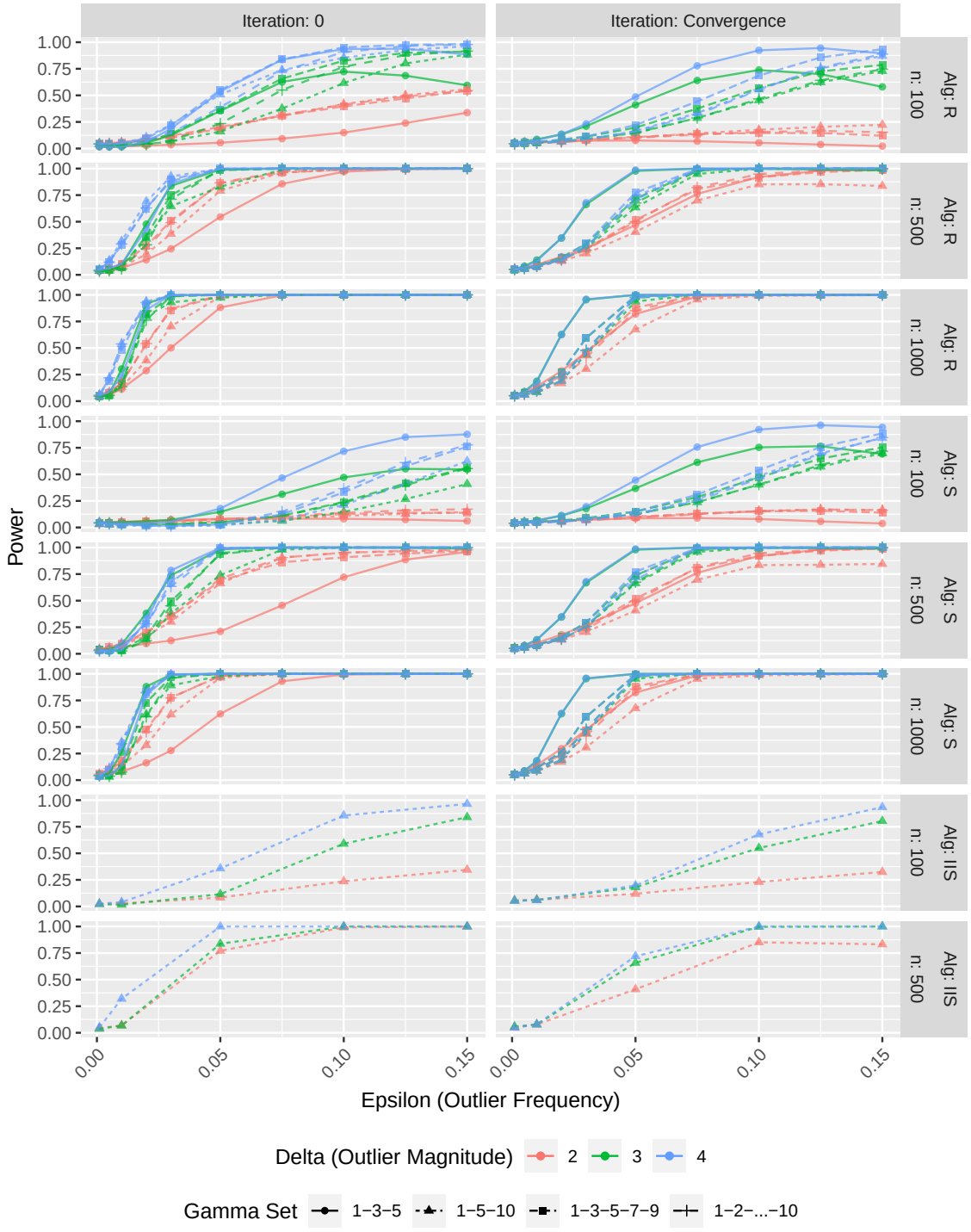
Figure 2.6.12 shows the power of the supremum test against the outlier frequency ϵ. The first column shows the power of the test based on the initial step, while the second column is based on the fixed point. The rows show the results for different algorithms in groups of two and three, and the sample size increases within each group. Different colours represent different outlier magnitudes δ and the marker and

Figure 2.6.11: Power of Sum Test, $\alpha = 0.05$



Simulated power (size-adjusted) for different outlier frequencies ϵ and outlier magnitudes δ. Several settings are varied: the sample size n, the iteration step m, the algorithm, and the set of tuning parameters $\mathbb{S}_c$ to yield $\mathbb{S}_\gamma$. Significance level $\alpha = 0.05$. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 2500$.

Figure 2.6.12: Power of Supremum Test, $\alpha = 0.05$



Simulated power (size-adjusted) for different outlier frequencies ϵ and outlier magnitudes δ. Several settings are varied: the sample size n, the iteration step m, the algorithm, and the set of tuning parameters $\mathbb{S}_c$ to yield $\mathbb{S}_\gamma$. Significance level $\alpha = 0.05$. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 2500$.

line type represent different sets of cut-off values $\mathbb{S}_c$ resulting in different sets of $\mathbb{S}_\gamma$.

The power of the supremum test increases both in the outlier magnitude δ and the outlier frequency ϵ. Unlike the other tests, we do not occasionally find lower power against larger outliers. The supremum test also seems more resilient against high outlier frequency, probably because it is based on absolute deviations. Therefore, even if the outlier frequency becomes large, it mostly keeps its power even if that could result in an inflated variance estimate and hence fewer detected outliers. For $\delta = 3$ and high $\epsilon > 0.1$ we see a slight drop in power for one of the sets of tuning parameters but this drop is smaller than for the sum test or the simple tests. Similarly, the supremum test also has power against the small outlier magnitude of $\delta = 2$.

Iterating the outlier detection algorithms leads to slightly smaller size-adjusted power but for Algorithm 2.S with small sample sizes, we would still recommend to use the iterated version because the test is otherwise oversized. Algorithm 2.R has size close to the nominal significance level for iteration 0 but is undersized for the iterated version, so in that case iteration step 0 might be preferable for better size control while at the same time yielding higher raw power.

2.6.3 Power LTS-Contamination

The ϵ-contamination model in Section 2.6 only considers outliers in the structural error u and hence in the y-direction. This is also how the algorithms classify observations as outliers. There exist other types of outliers, however, and we therefore present some simulation results in the presence of leverage points, which are outliers that also have unusual regressor values.

As in Section 2.6.2, power tends to increase as the sample size increases and as the outlier frequency increases unless it becomes too large. Iterating the outlier detection algorithms yields more stable power, especially when the null hypothesis coincides with the actual share of outliers in the data ($\gamma_c^0 = \epsilon$), e.g., when we expect 5% of (falsely) detected outliers under the null while the data truly contains 5% outliers.

2.6.3.1 DGP

We implement leverage points using the LTS-contamination scheme that was introduced in Berenguer-Rico and Nielsen (2023a) and Berenguer-Rico, Johansen and Nielsen (2023).

In the LTS-contamination, the set $\mathbb{H}$ of exactly $h = n \times (1 - \epsilon)$ errors is drawn from the "good" distribution, which is the same joint normal distribution as in Section 2.6.1.1

$$\begin{pmatrix} u_i \\ r_{2i} \end{pmatrix} \overset{\mathsf{D}}{=} \mathsf{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix} \right\}.$$

The $n-h$ outliers in set $\mathbb{H}^{\complement}$ are constructed such that the endogenous regressor x_2 is unusually large compared to the "good" data and the structural error u is extremely negative, thereby exerting a strong influence on the regression line. Formally, we have

$$u_i = \min_{j \in \mathbb{H}} u_j - \mathsf{U}[0,1] - 30,$$

$$x_{2i} = \max_{j \in \mathbb{H}} x_{2j} + \mathsf{U}[0,1],$$

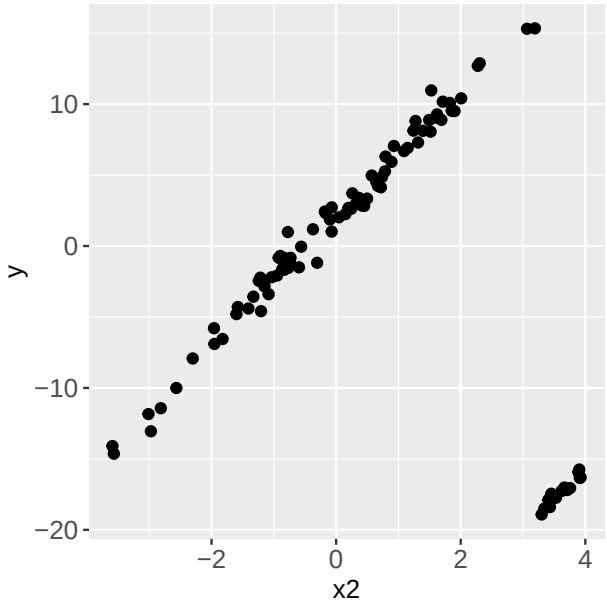
where $\mathsf{U}[0,1]$ represents the uniform distribution between 0 and 1. This implicitly also adjusts r_{2i}, possibly making it unusually large. The contamination scenario is modified from DGP 5 in Berenguer-Rico and Nielsen (2023a) and illustrated in Figure 2.6.13.

For computational reasons and for brevity, we restrict the analysis to a subset of the tests for outliers presented in Section 2.3.3 using fewer values for the outlier frequency and fewer cut-off values. In the Monte Carlo experiments, we vary the sample size $n \in \{100, 500, 1000\}$, the outlier frequency $\epsilon \in \{0.01, 0.05, 0.1, 0.15\}$, the iteration step $m \in \{0, *\}$, and the outlier detection algorithm. As in Section 2.6.2, we calculate size-adjusted power with a target significance level of $\alpha = 0.05$.

2.6.3.2 Proportion and Count Test

Figure 2.6.14 presents the results, some of which mirror those in Section 2.6.2. Power tends to increase as the sample size increases and as the outlier frequency increases unless it becomes too large. As before, the count test sometimes loses power as the

Figure 2.6.13: Illustration of LTS-Contamination

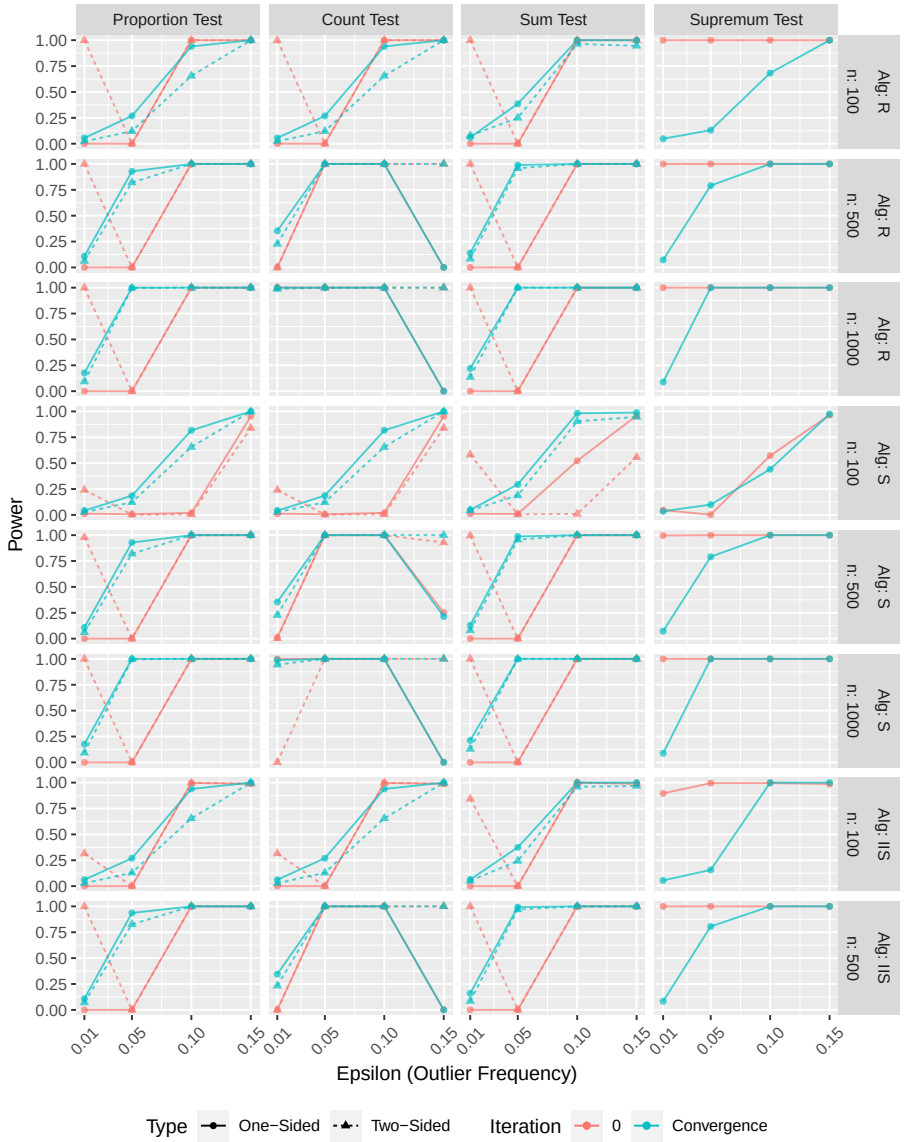


sample size increases because it keeps the target number of detected outliers fixed, such that the cut-off value c_n becomes larger and larger, making it less likely to detect outliers.

We find an interesting pattern for the non-iterated two-sided proportion test. Power is very high (often close to 1) for $\epsilon = 0.01$ but zero for $\epsilon = 0.05$, and resurges to a higher level for $\epsilon = 0.1, 0.15$. The reason is simple. Since $c = 1.96$, we expect to declare 5% of the sample as outliers under the null hypothesis of no outliers. In this specific LTS setting, it is easy to identify the outliers, of which there are 5% when $\epsilon = 0.05$ and we therefore have no power in this case. For higher values of ϵ, we detect significantly more than 5% and therefore have high power. Power is also high when $\epsilon = 0.01$ because we detect much fewer than 5% outliers, which also yields power for the two-sided but not the one-sided test. A similar pattern emerges for the sum and the count test, albeit the interaction is more complicated for the count test as n increases because the cut-off value changes.

Importantly, we find that iterating the algorithm yields more stable results. Intuitively, we find the true 5% of outliers in step $m = 0$ but then iterating the procedure causes us to detect some more "outliers" among the "good" observations, thereby yielding more than 5% detected outliers in total and therefore an improvement in

Figure 2.6.14: Power of the Tests, $\alpha = 0.05$



Simulated power (size-adjusted) for different outlier frequencies ϵ, sample sizes n, iteration steps m, and outlier detection algorithms. The proportion test uses $c = 1.96$ to yield $\gamma^0 = 0.05$, the count test uses c_n to yield $\lambda^0 = 5$, and the sum and supremum tests use $\mathbb{S}_c = \{2.58, 1.96, 1.64\}$ to yield $\mathbb{S}_\gamma = \{0.01, 0.05, 0.1\}$. Significance level $\alpha = 0.05$. Monte Carlo simulations for Algorithms 2.R and 2.S used $M = 10{,}000$ replications, Algorithm IIS used $M = 2500$.

power for the one-sided test. This is similar to the previous setting, where we had some power against $\epsilon = 0.05$ even when γ_c^0 was also 0.05, i.e. even when both the null hypothesis and the true contamination implied 5% outliers in the sample.

2.6.3.3 Sum and Supremum Test

The sum test follows the same pattern as the proportion test, providing more stable power when using the iterated algorithms. For the supremum test, iterating the outlier detection algorithms yields lower power, a phenomenon we already observed in the other contamination scenario. The most likely reason is simply that the asymptotic variance of the test increases strongly with the iteration step. Comparing power across the two contamination scenarios, the power of the convergent iteration is similar between the two but the non-iterated test has much more power in the LTS setting.

Overall, we have seen that the tests also have power in a setting with leverage points albeit with a more complicated pattern. This is despite the fact that the algorithms only detect outliers based on the structural error u, not other unusual values in x, z, or r. Future work could assess more sophisticated algorithms which also declare observations as outliers when their first stage error r is large or when the regressor/instrument is unusual. Such an analysis would require the derivation of additional empirical processes.

2.7 Discussion

In this section, we discuss what to consider when choosing which test, alternative hypothesis, algorithm, and iteration step to use, taking both theoretical aspects and the simulation results into account. We also briefly discuss the issue of pretest bias in case applied researchers want to use the tests for the presence of outliers to select between a robust and non-robust estimator.

The choice between simple or scaling test may depend on the computational and time cost of running the outlier detection algorithm only once versus several times. In general, it seems preferable to use a range of cut-off values to find outliers of different sizes and to avoid the somewhat arbitrary choice of what constitutes an outlier based on a single value. For very large data sets where the outlier detection algorithms take long to run, the researcher might prefer to choose a single cut-off

value and rely on one of the simple tests instead.

Among the simple tests, the proportion test controls the expected share of false detections when there is no contamination while the count test controls the expected number of false detections. Broderick, Giordano and Meager (2023) find that sensitivity of empirical results to outliers does not vanish as the sample size increases, such that targeting the share of false detections might be more suitable. For example, depending on what causes the outliers, it could be that increasing the sample size also increases the number of outliers in the sample. Alternatively, the count test fixes the expected number of false detections and its asymptotic approximation requires the cut-off value to grow with the sample size. Choosing a small number of detected outliers under the null controls the false detection rate very tightly but such a conservative choice comes at the cost of losing power against medium-sized outliers, as was evident in both power simulations presented in Sections 2.6.2 and 2.6.3.

Scaling tests avoid the often arbitrary choice of a single cut-off value by running the outlier detection algorithm across a range of cut-off values. This can improve the power of the tests to find evidence for outliers in the sample, as seen in the power simulations. Instead of only taking outlier frequency with respect to a single cut-off value into account, the scaling tests also check for the presence of outliers of different magnitudes. Among the scaling tests, there is no clear recommendation which one to choose based on the power simulations. The sum test has the disadvantage that positive and negative deviations between detected and hypothesised outliers across cut-off values can cancel. The supremum test does not have this drawback because it relies on the largest absolute deviation.

The researcher has the choice between testing against a one- or two-sided alternative. The one-sided alternative is closer to the intuition of rejecting the null hypothesis of no outliers when more outliers were detected than hypothesised. However, the distribution of the tests statistics is also derived under the auxiliary assumption of having specified the correct reference distribution. Testing against the two-sided alternative might yield higher power to reject when this auxiliary assumption is vi-

olated, in which case the two-sided alternative might be preferable. The supremum test already weighs positive and negative deviations equally by taking their absolute value. To increase confidence in the choice of reference distribution, one can use a Q-Q plot or a density plot of the (truncated) residuals and compare it to the true (truncated) density of the assumed reference distribution. See the empirical illustration in Section 2.8 for an example. The simulations in Section 2.6.2 have also shown that the two-sided tests tend to have higher power when the sample is highly contaminated such that the variance is considerably overestimated, leading to few(er) outliers being detected.

In current practice, most researchers use Algorithm 2.R and start with the full-sample 2SLS estimates to find outliers in the sample. Both power simulations show good performance of Algorithm 2.R, often yielding at least slightly higher power than Algorithm 2.S. However, there is a concern that starting with a non-robust estimator could prevent the correct detection of outliers if the initial estimates are highly affected by the outliers. To address this concern, one should either use Algorithm 2.S or IIS unless the sample size is small. Algorithm 2.S is highly oversized for small n because the asymptotic approximation for the FODR of Algorithm 2.S requires relatively large sample sizes of $n \geq 10{,}000$ to work well, as shown in Chapter 1. For small sample sizes, it underestimates the true variability of the FODR. This means that "extreme" values of the FODR are more likely than the asymptotic theory predicts, which translates into a rejection frequency that is higher than the target significance level α. Intuitively, splitting the sample into two halves means that for small sample sizes, there is a significant loss of information when estimating the model separately for each half. For example, with a total sample size of $n = 50$, each subsample only contains 25 observations. This may lead to very different parameter estimates $(\widehat{\beta}_1, \widehat{\sigma}_1)$ and $(\widehat{\beta}_2, \widehat{\sigma}_2)$ between the two subsamples. Since the parameter estimates of one subsample are used to compute the residuals for the other subsample, this likely leads to large residuals in both halves and therefore to many incorrect classifications as outliers. This issue vanishes as the sample size grows because both subsample estimators are consistent for the true parameter

values (β, σ). In fact, Chapter 1 shows that the asymptotic expansion of the FODR is identical for Algorithms 2.R and the split-half version of Algorithm 2.S, such that the size performance should become close to that of Algorithm 2.R as the sample size grows.

The researcher also has to decide which iteration step to use. Iterating the algorithms might achieve higher robustness. If removing outliers from the sample improves the parameter estimates, then re-calculating the standardised residuals and updating the classification of outliers should also be an improvement. Jiao (2022) shows that the fixed point estimators have the same first order asymptotics as the robust Huber-skip estimator in the 2SLS regression setting. Our simulations provide two further reasons for iterating the algorithms. First, while the size performance of the tests based on Algorithm 2.R is generally good, Algorithms 2.S and IIS can result in highly oversized tests for small sample sizes. But iterating the algorithms tends to improve the size of the tests considerably. Therefore, iterating to the fixed point helps control the false rejection rate. Second, iterating helps achieve power when the true share of outliers in the sample happens to coincide with the hypothesised share of detected outliers under the null hypothesis of no outliers, i.e. when $\epsilon = \gamma_c^0$.

Finally, the researcher has to decide what to do if they find evidence of outliers in their sample. Barnett (1978) provides a useful discussion of this question, distinguishing between four ways of handling outliers: (1) accommodating the outlier by drawing inference despite its presence, e.g. using robust estimators; (2) incorporating the outlier by changing the assumed reference distribution; (3) adapting the model such that the data point is not an outlier any more, e.g. by adding new covariates; and (4) rejecting the data point. The discussion shows that often, finding outliers is just the starting point for further data analysis.

When using the tests for outliers to select an appropriate estimator (robust versus full-sample 2SLS) rather than as a misspecification test, one has to be careful about pretest biases. The problem of pretest bias has originally been shown in the context of model selection (Bancroft 1944; Wallace 1977; Danilov and Magnus 2004). Outlier detection itself can be seen as a problem of model selection by in-

cluding an indicator variable that perfectly fits the observation and may therefore also be subject to pretest biases, such as the well-known impossibility result of non-uniform convergence after model selection (Leeb and Pötscher 2005, 2008). There is no pretest bias when the variable that is selected over is uncorrelated with the other included variables (not selected over), which in our setting would translate into observations with large residuals not also having unusual values of the regressors. Hendry and Johansen (2015) recommend orthogonalising the selection variables (e.g. indicators when searching for outliers) with respect to the other included variables. If the selection variables are truly irrelevant, the asymptotic distribution of the included variables is not impacted by model selection; if they are relevant, model selection can still yield improvements over not searching for the correct model specification/outliers. Finally and closest to our setting of choosing between the robust or full-sample 2SLS estimator, Chmelarova and Hill (2010) analyse the pretest bias when using a test for endogeneity of regressors before deciding whether to use the OLS or an instrumental variables (IV) estimator. They show that when there is no endogeneity, the mean squared error (MSE) of the endogenous variable coefficient estimator is not much higher when using a pretest estimator instead of OLS but that the MSE is substantially lower for moderate to high degrees of endogeneity, so it might still be beneficial to use a pretest estimator. A systematic assessment of the importance of pretest bias when using our tests for outliers is left for future research but applied researchers should be aware of the potential drawbacks.

2.8 Empirical Illustration

Using the seminal work of Angrist and Krueger (1991), we show that our proposed tests for the presence of outliers reject the null hypothesis of no outliers. Both the simple proportion tests at varying cut-off values and the scaling supremum and sum test yield that conclusion. The robust parameter estimates of the return to education tend to be smaller than the original estimates but are still statistically significant.

Angrist and Krueger (1991) use quarter of birth (QOB) as instruments for education when estimating the effect of education on wages. Educational attainment is likely endogenous because it is also related to an individual's characteristics such as motivation and intelligence, which are themselves direct determinants of their wage. To account for that endogeneity, the authors suggest using QOB as an excluded and valid instrument for the number of completed years of schooling. The instrument is plausibly valid and excluded from the structural equation because QOB is unlikely to be related to other characteristics that determine a person's wage nor a direct determinant thereof. The authors also argue and provide evidence that the instrument is relevant because QOB interacts with compulsory schooling laws to force some people to stay in high school for a longer time than they otherwise would. Compulsory schooling laws in the U.S. usually require students to attend school until they are 16 or 17 years old. At the same time, students are required to be six years old by 1 January when they enter school, which means that students who are born earlier in the year start school at an older age. These students reach the age at which they can leave school (16 or 17) earlier in their school career. Therefore, QOB in combination with compulsory schooling laws cause some students born later in the year to attend school for longer.

Equations (2.8.1) and (2.8.2) reproduce the first stage and structural equation, respectively, using the notation of Angrist and Krueger (1991). E_i represents the years of completed education for individual i, X_i is a vector of controls, Y_{ic} a dummy for birth year c, and Q_{ij} is a dummy for QOB j. The instruments are a set of interactions between birth year and QOB, $Y_{ic}Q_{ij}$. The dependent variable of the structural equation is the log of weekly wages of person i, $\ln W_i$.

$$E_i = X_i\pi + \sum_c Y_{ic}\delta_c + \sum_c \sum_j Y_{ic}Q_{ij}\theta_{jc} + \epsilon_i, \tag{2.8.1}$$

$$\ln W_i = X_i\beta + \sum_c Y_{ic}\xi_c + \rho E_i + \mu_i. \tag{2.8.2}$$

Our illustration focuses on the regression specification in column (6) of Table V in Angrist and Krueger (1991). Table V presents the results for men aged 40 to 49, "whose wages are hardly related to age [...] to avoid the effects of life-cycle changes

in earnings that are correlated with quarter of birth (Angrist and Krueger 1991, p. 995)." The specification in column (6) includes several controls, such as race and marital status.

The outlier detection algorithms and the proposed tests for the presence of outliers are readily available from the R package *robust2sls* (Kurle 2022b). The package also implements the IIS algorithm as a starting point for outlier detection based on the R package *ivgets* (Kurle 2022a), as well as the corrected inference and tests for β proposed in Jiao (2022).

We apply Algorithms 2.R and 2.S with tuning parameters $c \in \{1.64, 1.96, 2.58\}$ to the original sample of 329,509 men. We iterate the algorithms until the squared difference between two consecutive β-estimates is smaller than 0.0001, which we interpret as convergence. The results are presented in Table 2.8.1.

For all different cut-off values and algorithms, we detect a larger share of outliers than would be expected under the null hypothesis of no outliers and using normality as the reference distribution. For example, for $c = 1.96$, we would expect 5% of observations to be declared as outliers when there is no contamination. Both Algorithms 2.R and 2.S detected significantly more outliers than expected, namely 13.5% of the sample. The simple proportion tests $T^{(*)}_{prop,c}$ all individually reject the

Table 2.8.1: Comparison of Full-Sample and Robust Estimation Results

Alg.	c	γ^0_c	$\widehat{\gamma}^{(*)}_c$	n	FP	$T^{(*)}_{prop,c}$	$\widehat{\beta}^{(*)}_c$	$se(\widehat{\beta}^{(*)}_c)$	$t_{\beta=0}$
full	–	–	–	329,509	–	–	0.081	0.016	4.915
R	2.58	0.01	0.058	310,443	5	233.520	0.070	0.011	6.184
S	2.58	0.01	0.055	311,314	3	221.370	0.071	0.012	6.183
R	1.96	0.05	0.135	285,090	10	141.270	0.050	0.012	4.116
S	1.96	0.05	0.135	285,096	10	141.240	0.050	0.012	4.107
R	1.64	0.1	0.190	266,903	7	87.830	0.045	0.014	3.309
S	1.64	0.1	0.190	266,814	7	88.100	0.045	0.014	3.270

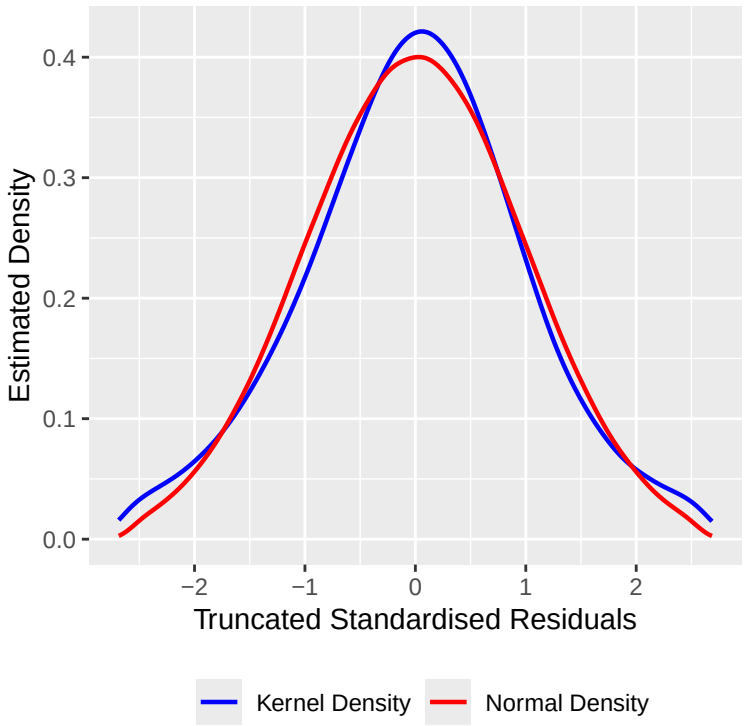
Note: Comparison of full-sample and robust estimation results across Algorithms 2.R and 2.S and different tuning parameters c. The column $\widehat{\gamma}^{(*)}_c$ reports the detected share of outliers, n represents the remaining sample size, FP the iteration at which the algorithm converged, $T^{(*)}_{prop,c}$ the value of the proportion test statistic, $\widehat{\beta}^{(*)}_c$ the coefficient measuring the return to education, $se(\widehat{\beta}^{(*)}_c)$ its standard error using the bias-corrected variance estimate from Jiao (2022) and $t_{\beta=0}$ the t-statistic for testing whether the return to education is zero.

null hypothesis of no outliers at any conventional significance level. Conducting the scaling sum and supremum tests for Algorithm 2.R across the set of tuning parameters $\mathbb{S}_c = \{1.64, 1.96, 2.58\}$ yields p-values below 0.0001, also rejecting the null hypothesis.

Examining the coefficient estimates for the return to education, we find smaller effects than the full-sample estimate (8.1%), ranging from 4.5% to 7.1%. The robust estimates are still statistically significant despite using the larger, bias-corrected variance estimate derived in Jiao (2022). These bias-corrected standard errors account for the false classification and removal of observations as outliers if the sample were not contaminated. We also test whether the robust and the full-sample coefficients on education are different from each other using the Hausman-type test suggested in Jiao (2022). For all three tuning parameters, we obtain p-values below 0.002 and therefore reject the null hypothesis of no difference between the parameters.

Figure 2.8.1 displays the estimated density of the non-outlying, standardised

Figure 2.8.1: Density Comparison of Standardised Residuals



Estimated density of the standardised residuals (blue) of the fixed point model using Algorithm 2.R with $c = 2.58$. The normal density estimate (red) is obtained by drawing $n = 1{,}000{,}000$ normally distributed random errors and truncating them using the cut-off value $c = 2.58$.

residuals for the fixed point of Algorithm 2.R with $c = 2.58$ and compares it to the truncated standard normal density. We see that the fit is reasonable but that the residuals have slightly fatter tails, indicating that there could be outliers left in the sample or that the true distribution is more heavy-tailed than the normal distribution. Sølvsten (2020) also re-examines the results of Angrist and Krueger (1991) but their regression specification of Table VII with many (180) instruments. He also finds that "the structural errors are distributed roughly like a normal distribution contaminated with gross outliers" (Sølvsten 2020, p.495), so our assumption of normality might be justified for the centre of the residuals. Future research could devise a formal statistical test for the normality of 2SLS regression residuals after truncation, similar to Berenguer-Rico and Nielsen (2023b) in the OLS case.

2.9 Conclusion

This chapter proposes statistical tests for the presence of outliers in 2SLS regression models. The test statistics formalise and improve upon the heuristic approach that applied researchers currently use for checking the presence of outliers. The tests evaluate the null hypothesis that there are no outliers by comparing the actual share of detected outliers to the expected share if there is no contamination.

We propose two types of such tests. The simple tests fix a single cut-off value and test whether there are more detected outliers than expected, i.e. it is based on outlier frequency. The scaling tests run the algorithms with a range of cut-off values and therefore also take outlier magnitude into account.

Extensive Monte Carlo simulations show that iterating the algorithms improves the size performance of all tests except the count test. The power of the tests increases with the sample size, outlier magnitude, and outlier frequency. If the outlier frequency becomes too high, the simple tests start to lose power again while the scaling tests tend to keep their high power.

An application to the seminal work of Angrist and Krueger (1991) who estimate the return to education using quarter of birth as an instrument shows evidence for

the presence of outliers. The robust parameter estimates of the return to education tend to be smaller than the original estimates but are still statistically significant.

For future research, it seems especially important and worthwhile to extend the usability of the test statistics to other reference distributions. Several potential avenues exist for doing so.

Instead of deriving and using the asymptotic distribution of the test statistics, one could bootstrap their distribution, which might additionally improve the finite sample performance. However, the bootstrap procedure would have to ensure that we resample from the non-contaminated data because we need the distribution of the test statistic under the null hypothesis of no outliers. Davison and Hinkley (1997) already emphasize this issue: "What if simulated resampling is used when there are outliers in the data? There is no substitute for careful data scrutiny in this or any other statistical context, and if obvious outliers are found, they should be removed or corrected (p.44)." Jiao and Pretis (2022) suggest re-sampling from the subsample of non-outlying observation, i.e. after detecting outliers using one of the algorithms described in Section 2.4. This means the user still needs to assume a certain reference distribution to classify outliers in a first step. Jiao and Pretis (2022) then either record the outlier share or the proportion test statistic across resamples and test for statistical significance against their respective bootstrap distributions. The validity of this resampling procedure relies on having identified the true outliers. The simulations in Jiao and Pretis (2022) show that when the reference distribution is correctly assumed, resampling the outlier share yields oversized tests while resampling the test statistic yields undersized tests. Under contamination, the power of both their bootstrap versions is lower compared to using the asymptotic distribution of the proportion test. When the assumed reference distribution (normal) deviates from the true error distribution (t_3 in their case), bootstrapping helps control size but still yields lower power under contamination.

Alternatively, we might be able to extend the usable class of reference distributions when they have similar simplifications of the asymptotic variance as under normality. For example, it could be that no more or only a few additional moments

need to be estimated compared to the normal reference distribution.

Finally, the theory derived in Chapter 1 is general enough to relax the joint normality assumption of the errors to any reference distribution that satisfies the required assumptions for the asymptotic analysis of the FODR. However, the formulas for the asymptotic variances of the test statistics then contain moments for which estimators need to be devised. These estimators should ideally be robust but still consistent when the sample is not contaminated. Such estimators and potential bias correction factors could be devised using new empirical processes.

Appendices of Chapter 2

2.A Simplifications Under Normality

Under joint normality of the structural and first stage errors, the formula for the asymptotic variance of the FODR $\widehat{\gamma}_c^{(m)}$ simplifies. This section shows how some of the moments in the general asymptotic variance derived in Chapter 1 simplify under normality (Lemma 2.A.1) and then shows the simplified formulas (Lemma 2.A.2).

Lemma 2.A.1 (**Moments under Normality**). *Suppose that the structural and first stage errors are jointly normally distributed as*

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

Then, the following relationships between moments hold

$$\zeta_c^- = 2c\Omega, \quad \zeta_c^+ = 0_{d_x}, \tag{2.A.1}$$

$$\omega_c = \tau_2^c \Omega, \tag{2.A.2}$$

$$\tau_2^c = \psi_c - 2c\mathsf{f}_u(c), \tag{2.A.3}$$

$$\tau_4^c = 3\psi_c - 2c(c^2+3)\mathsf{f}_u(c), \tag{2.A.4}$$

$$\tau_4 = 3, \tag{2.A.5}$$

where the moments are defined in Section 2.2.

Proof of Lemma 2.A.1. We use the general result that the conditional and marginal distribution of a multivariate normal are still normally distributed. Let x and y be two random vectors that are jointly normally distributed

$$\begin{pmatrix} x \\ y \end{pmatrix} \overset{\mathsf{D}}{=} \mathsf{N}\left\{ \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \right\},$$

where $\Sigma_{yx} = \Sigma_{xy}'$ due to the symmetry of the covariance. Then, the following results

about the marginal and conditional distribution hold:

$$y \sim \mathsf{N}(\mu_y, \Sigma_{yy}) \tag{2.A.6}$$

$$y|x \sim \mathsf{N}(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}). \tag{2.A.7}$$

Using property (2.A.6), we obtain the marginal distribution $u_i/\sigma \sim \mathsf{N}(0, 1)$. We obtain the conditional distribution $\Sigma^{-1/2}r_i|(u_i/\sigma) \sim \mathsf{N}(u_i\Omega/\sigma, I_{d_x}^* - \Omega\Omega')$ using property (2.A.7).

ξ_c is defined as $\mathsf{E}(\Sigma^{-1/2}r_i|u_i/\sigma = c)$, hence $\xi_c = c\Omega$ and $\xi_{-c} = -c\Omega$. ζ_c^+ is defined as $\xi_c + \xi_{-c}$, hence $\zeta_c^+ = 0_{d_x}$. ζ_c^- is defined as $\xi_c - \xi_{-c}$, hence $\zeta_c^- = 2c\Omega$. The truncated covariance ω_c is defined as $\mathsf{E}(\Sigma^{-1/2}r_i\frac{u_i}{\sigma}1_{(|u_i|\leq\sigma c)})$ and can be shown to equal $\tau_2^c\Omega$ under normality using the following arguments:

$$\begin{aligned}
\omega_c &= \int\int_{y\in\mathbb{R}, x\in\mathbb{R}^{d_x}} xy1_{(|y|\leq c)}\phi_{u,r}(y,x)dy(dx) \\
&=_{(1)} \int_{-c}^{c}\int_{x\in\mathbb{R}^{d_x}} xy\phi_{r|u}(x|y)\phi_u(y)dy(dx) \\
&=_{(2)} \int_{-c}^{c}\{\int_{x\in\mathbb{R}^{d_x}} x\phi_{r|u}(x|y)(dx)\}y\phi_u(y)dy \\
&=_{(3)} \int_{-c}^{c}\xi_y y\phi_u(y)dy =_{(4)} \Omega\int_{-c}^{c} y^2\phi_u(y)dy =_{(5)} \tau_2^c\Omega,
\end{aligned}$$

where step (1) replaces the indicator function $1_{(|y|\leq c)}$ with the integration limits $[-c, c]$ and decomposes the joint normal density into the product of conditional and marginal density, $\phi_{u,r}(y,x) = \phi_{r|u}(x|y)\phi_u(y)$. Step (2) collects terms that depend on x in order to integrate with respect to x first. Step (3) recognises that the inner integral is the conditional expected value of $\Sigma^{-1/2}r_i|u_i/\sigma$, which we denote by ξ_y. Step (4) uses normality to replace ξ_y with $y\Omega$ and takes the constant Ω outside of the integral. Finally, step (5) recognises the remaining integral as the second truncated moment of a standard normal distribution, which we denote by τ_2^c.

The truncated moments of u_i/σ, which follows a standard normal distribution, are defined as $\tau_k^c = \int_{-c}^{c} y^k\phi_u(y)\,dy$. Under normality, we can make the following

calculations:

$$\begin{aligned}\tau_2^c &= \int_{-c}^{c} y^2 \phi_u(y)\, dy = \int_{-c}^{c} y \times y(2\pi)^{-1/2} \exp(-y^2/2)\, dy \\ &=_{(6)} [-y(2\pi)^{-1/2} \exp(-y^2/2)]_{-c}^{c} + \int_{-c}^{c} (2\pi)^{-1/2} \exp(-y^2/2)\, dy \\ &=_{(7)} [-c\phi_u(c) - c\phi_u(c)] + \psi_c = \psi_c - 2c\phi_u(c),\end{aligned}$$

where step (6) follows from integration by parts and step (7) from the definition of ψ_c as the probability of the scaled structural error lying inside the interval $[-c, c]$. Similarly, we can show

$$\begin{aligned}\tau_4^c &= \int_{-c}^{c} y^4 \phi_u(y)\, dy = \int_{-c}^{c} y^3 \times y(2\pi)^{-1/2} \exp(-y^2/2)\, dy \\ &=_{(8)} [-y^3(2\pi)^{-1/2} \exp(-y^2/2)]_{-c}^{c} + 3\int_{-c}^{c} y^2 (2\pi)^{-1/2} \exp(-y^2/2)\, dy \\ &=_{(9)} [-c^3\phi_u(c) - c^3\phi_u(c)] + 3\tau_2^c = 3\psi_c - 2c\phi_u(c)(c^2+3),\end{aligned}$$

where step (8) again follows from integration by parts and step (9) from recognising that the second term is $3\tau_2^c$. Finally, the kurtosis of a standard normal distribution is 3, such that $\tau_4 = \int_{-\infty}^{\infty} y^4 \phi_u(y)\, dy = 3$. ■

In the next Lemma, we use the results from Lemma 2.A.1 to show that various terms in the asymptotic expansions of the FODR derived in Chapter 1 are zero under normality. Consequently, these terms do not contribute to the asymptotic variance and covariance of the FODR.

Lemma 2.A.2 (**Asymptotic (Co-)Variances of the FODR under Normality**). *Suppose that the structural and first stage errors are jointly normally distributed as in Lemma 2.A.1. Then, the following terms in the asymptotic expansions of the FODR that were derived in Chapter 1 are zero:*

for Theorems 1.3.2, 1.3.3 in Chapter 1:

$$\zeta_c^- - 2c\Omega = 0_{d_x}, \qquad (2.A.8)$$

for Theorem 1.C.1 in Chapter 1:

$$\frac{c^2+\varsigma_c^2}{2}\zeta_c^- - \frac{2c}{\psi_c}\omega_c - c(c^2-\varsigma_c^2)\Omega = 0_{d_x}, \tag{2.A.9}$$

$$\zeta_c^- - \frac{2c}{\tau_2^c}\omega_c = 0_{d_x}, \tag{2.A.10}$$

for Theorem 1.C.2 in Chapter 1:

$$\{\varrho^{(m)}_{\beta\beta,c} + \frac{c\phi_u(c)\varrho^{(m)}_{\sigma\beta,c}(c^2-\varsigma_c^2)}{\tau_2^c}\}\zeta_c^- - \frac{4c^2\phi_u(c)\varrho^{(m)}_{\sigma\beta,c}}{\psi_c\tau_2^c}\omega_c - 2c\varrho^{(m)}_{\sigma\sigma,c}\Omega = 0_{d_x}, \tag{2.A.11}$$

$$\{\varrho^{(m)}_{\beta\tilde{x}u,c} + c\varrho^{(m)}_{\sigma\tilde{x}u,c}(c^2-\varsigma_c^2)\}\zeta_c^- - \frac{2c}{\psi_c}(2c\varrho^{(m)}_{\sigma\tilde{x}u,c} + \varrho^{(m)}_{\sigma uu,c})\omega_c = 0_{d_x}, \tag{2.A.12}$$

for Theorem 1.3.4 in Chapter 1:

$$\{\varrho^{(*)}_{\beta\tilde{x}u,c} + c\varrho^{(*)}_{\sigma\tilde{x}u,c}(c^2-\varsigma_c^2)\}\zeta_c^- - \frac{2c}{\psi_c}(2c\varrho^{(*)}_{\sigma\tilde{x}u,c} + \varrho^{(*)}_{\sigma uu,c})\omega_c = 0_{d_x}. \tag{2.A.13}$$

Proof of Lemma 2.A.2. The proof repeatedly uses the relationships between various moments under normality that are presented in Lemma 2.A.1, as well as the notation $\varsigma_c^2 = \tau_2^c/\psi_c$ that was introduced in Section 2.2.

Equation (2.A.8):

$$\zeta_c^- - 2c\Omega =_{(1)} 2c\Omega - 2c\Omega = 0_{d_x},$$

where step (1) uses equation (2.A.1).

Equation (2.A.9):

$$\begin{aligned}\frac{c^2+\varsigma_c^2}{2}\zeta_c^- - \frac{2c}{\psi_c}\omega_c - c(c^2-\varsigma_c^2)\Omega &=_{(2)} \frac{c^2+\varsigma_c^2}{2}2c\Omega - \frac{2c}{\psi_c}\tau_2^c\Omega - c(c^2-\varsigma_c^2)\Omega \\ &= \{c^2+\varsigma_c^2-2\varsigma_c^2-(c^2-\varsigma_c^2)\}c\Omega \\ &= 0_{d_x},\end{aligned}$$

where step (2) uses equations (2.A.1) and (2.A.2).

Equation (2.A.10):

$$\begin{aligned}\zeta_c^- - \frac{2c}{\tau_2^c}\omega_c &=_{(3)} 2c\Omega - \frac{2c}{\tau_2^c}\tau_2^c\Omega \\ &= 0_{d_x},\end{aligned}$$

where step (3) uses equations (2.A.1) and (2.A.2).

To prove that equations (2.A.11) and (2.A.12) are zero, we need two further algebraic equalities. First, under normality we have

$$2\tau_2^c - \psi_c(c^2 - \varsigma_c^2) = 2\psi_c\varsigma_c^2 + \psi_c\varsigma_c^2 - \psi_c c^2 = \psi_c(3\varsigma_c^2 - c^2). \tag{2.A.14}$$

Second, we can rewrite the coefficient $\varrho^{(m)}_{\sigma\beta,c}$ as follows

$$\begin{aligned}
\varrho^{(m)}_{\sigma\beta,c} &= \sum_{l=0}^{m-1} \{\frac{2c\phi_u(c)}{\psi_c}\}^{m-l-1} \{\frac{c\phi_u(c)(c^2-\varsigma_c^2)}{\tau_2^c}\}^l \\
&= \{\frac{2c\phi_u(c)}{\psi_c}\}^{m-1} \sum_{l=0}^{m-1} \{\frac{\psi_c(c^2-\varsigma_c^2)}{2\tau_2^c}\}^l \\
&=_{(4)} \{\frac{2c\phi_u(c)}{\psi_c}\}^{m-1} [\frac{(2\tau_2^c)^m - \{\psi_c(c^2-\varsigma_c^2)\}^m}{(2\tau_2^c)^{m-1}\{2\tau_2^c - \psi_c(c^2-\varsigma_c^2)\}}] \\
&=_{(5)} \{\frac{2c\phi_u(c)}{\psi_c}\}^{m-1} [\frac{(2\tau_2^c)^m - \{\psi_c(c^2-\varsigma_c^2)\}^m}{(2\tau_2^c)^{m-1}\psi_c(3\varsigma_c^2 - c^2)}],
\end{aligned} \tag{2.A.15}$$

where step (4) applies the formula for a geometric series and step (5) uses equation (2.A.14). We can now show that equation (2.A.11) holds.

Equation (2.A.11):

$$\begin{aligned}
&\{\varrho^{(m)}_{\beta\beta,c} + \frac{c\phi_u(c)\varrho^{(m)}_{\sigma\beta,c}(c^2-\varsigma_c^2)}{\tau_2^c}\}\zeta_c^- - \frac{4c^2\phi_u(c)\varrho^{(m)}_{\sigma\beta,c}}{\psi_c\tau_2^c}\omega_c - 2c\varrho^{(m)}_{\sigma\sigma,c}\Omega \\
&=_{(6)} \{\varrho^{(m)}_{\beta\beta,c} + \frac{\varrho^{(m)}_{\sigma\beta,c}c\phi_u(c)(c^2-3\varsigma_c^2)}{\tau_2^c} - \varrho^{(m)}_{\sigma\sigma,c}\}2c\Omega \\
&=_{(7)} \Big(\frac{c(c^2-3\varsigma_c^2)\phi_u(c)}{\tau_2^c}\{\frac{2c\phi_u(c)}{\psi_c}\}^{m-1}[\frac{(2\tau_2^c)^m - \{\psi_c(c^2-\varsigma_c^2)\}^m}{(2\tau_2^c)^{m-1}\psi_c(3\varsigma_c^2-c^2)}] \\
&+ \{\frac{2c\phi_u(c)}{\psi_c}\}^m - \{\frac{c(c^2-\varsigma_c^2)\phi_u(c)}{\tau_2^c}\}^m\Big)2c\Omega \\
&= \Big(\{\frac{2c\phi_u(c)}{\psi_c}\}^m - \{\frac{c(c^2-\varsigma_c^2\phi_u(c))}{\tau_2^c}\}^m - \{\frac{2c\phi_u(c)}{\psi_c}\}^m + \{\frac{c(c^2-\varsigma_c^2)\phi_u(c)}{\tau_2^c}\}^m\Big)2c\Omega \\
&= 0_{d_x},
\end{aligned}$$

where step (6) uses equations (2.A.1) and (2.A.2), and step (7) replaces the $\varrho^{(m)}_{\cdot,c}$ coefficients with their detailed expressions, including (2.A.15).

Equation (2.A.12):

$$
\begin{aligned}
&\{\varrho^{(m)}_{\beta\tilde{x}u,c} + c\varrho^{(m)}_{\sigma\tilde{x}u,c}(c^2-\varsigma_c^2)\}\zeta_c^- - \frac{2c}{\psi_c}(2c\varrho^{(m)}_{\sigma\tilde{x}u,c} + \varrho^{(m)}_{\sigma uu,c})\omega_c \\
&\quad =_{(8)} \{\varrho^{(m)}_{\beta\tilde{x}u,c} + c\varrho^{(m)}_{\sigma\tilde{x}u,c}(c^2-3\varsigma_c^2) - \varsigma_c^2\varrho^{(m)}_{\sigma uu,c}\}2c\Omega \\
&\quad =_{(9)} \Big(\frac{\psi_c^m - \{2c\phi_u(c)\}^m}{\psi_c^m\tau_2^c} - \varsigma_c^2\frac{(\tau_2^c)^m - \{c(c^2-\varsigma_c^2)\phi_u(c)\}^m}{(\tau_2^c)^m\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}_u(c)\}} + \frac{c(c^2-3\varsigma_c^2)\phi_u(c)}{(\tau_2^c)^2} \\
&\quad \times [\frac{(\tau_2^c)^m - \{c(c^2-\varsigma_c^2)\phi_u(c)\}^m}{(\tau_2^c)^{m-1}\{\tau_2^c - c(c^2-\varsigma_c^2)\phi_u(c)\}} - \{\frac{2c\phi_u(c)}{\psi_c}\}^{m-1}\sum_{l=0}^{m-1}\{\frac{\psi_c(c^2-\varsigma_c^2)}{2\tau_2^c}\}^l]\Big)2c\Omega \\
&\quad =_{(10)} 0_{d_x}
\end{aligned}
$$

where step (8) uses equations (2.A.1) and (2.A.2), and step (9) replaces the $\varrho^{(m)}_{\cdot,c}$ coefficients with their expressions. Step (10) uses the formula for the geometric series, expands the different fractions to have the same denominator, and uses several algebraic manipulations to show that the numerator is zero.

Equation (2.A.13):

$$
\begin{aligned}
&\{\varrho^{(*)}_{\beta\tilde{x}u,c} + c\varrho^{(*)}_{\sigma\tilde{x}u,c}(c^2-\varsigma_c^2)\}\zeta_c^- - \frac{2c}{\psi_c}(2c\varrho^{(*)}_{\sigma\tilde{x}u,c} + \varrho^{(*)}_{\sigma uu,c})\omega_c \\
&\quad =_{(11)} \{\varrho^{(*)}_{\beta\tilde{x}u,c} + c\varrho^{(*)}_{\sigma\tilde{x}u,c}(c^2-3\varsigma_c^2) - \varsigma_c^2\varrho^{(*)}_{\sigma uu,c}\}2c\Omega \\
&\quad =_{(12)} \{\frac{1}{\tau_2^c} + \frac{c(c^2-3\varsigma_c^2)\phi_u(c)}{\tau_2^c\{\tau_2^c - c\phi_u(c)(c^2-\varsigma_c^2)\}} - \frac{\varsigma_c^2}{\tau_2^c - c\phi_u(c)(c^2-\varsigma_c^2)}\}2c\Omega \\
&\quad = \frac{\tau_2^c\varsigma_c^2 - \tau_2^c\varsigma_c^2}{\tau_2^c\{\tau_2^c - c\phi_u(c)(c^2-\varsigma_c^2)\}}2c\Omega \\
&\quad = 0_{d_x},
\end{aligned}
$$

where step (11) uses equations (2.A.1) and (2.A.2), and step (12) replaces the $\varrho^{(*)}_{\cdot,c}$ coefficients with their expressions. ■

2.B Proofs of Asymptotics of Test Statistics

The distributions of the tests statistics presented in Section 2.5.2 are derived from a combination of theorems in Chapter 1.

The proportion test, sum test, and supremum test are all based on weak convergence results of the FODR $\widehat{\gamma}_c^{(m)}$ as a process in the cut-off value c for different iteration steps m and for Algorithms 2.R and 2.S. The proportion test fixes a single

cut-off value c and therefore only requires the univariate finite dimensional convergence result of the FODR for that particular choice of c. The sum and supremum test are based on a scale of cut-off values c and are therefore based on the weak convergence results. In practice, a finite number of K different cut-off values is chosen, such that the distribution of the test statistics follows from the multivariate finite dimensional convergence results of the FODR across different cut-off values $c_1, \ldots, c_K$.

The count test is based on the Poisson approximation to the number of detected outliers $n\widehat{\gamma}_{c_n}^{(m)}$ derived in Chapter 1. It holds for any iteration step $m \in [0, \infty)$ and requires the initial estimators $(\widetilde{\beta}, \widetilde{\sigma}^2)$ to be tight. While Algorithms 2.R and 2.S satisfy this condition, it also allows for other outlier detection algorithms.

Proof of Theorem 2.5.1. In Chapter 1, Theorem 1.3.2 covers $m = 0$, Theorem 1.C.2 covers $m \geq 1$, and Theorem 1.3.4 covers the fixed point $m = *$ when $m \to \infty$ for Algorithm 2.R. For Algorithm 2.S, Theorem 1.C.3 in Chapter 1 covers $m \geq 0$ by proving asymptotic equivalence with Algorithm 2.R. The fixed point $m = *$ when $m \to \infty$ for Algorithm 2.S is also covered in Theorem 1.3.4 in Chapter 1. Overall, the theorems cover both Algorithm 2.R and Algorithm 2.S for general iteration steps $m \in [0, \infty)$, including the fixed point $m = *$.

These theorems require Assumption 1.3.1(ia, iia, iii, iv) in Chapter 1 to hold. The proportion test requires Assumption 2.5.1(i, ii) of the present chapter to hold. The distributional Assumption 2.5.1(i) of joint normality of the errors implies the following, more general assumptions in Chapter 1: Assumption 1.3.1(ia) on the structural error u_i (Johansen and Nielsen 2016a), Assumption 1.3.1(iia) on the first stage errors r_i, and Assumption 1.3.1(iii) on their joint distribution (Johansen and Nielsen 2016b). Assumption 2.5.1(ii) is identical to Assumption 1.3.1(iv) in Chapter 1. Therefore, Assumption 2.5.1(i, ii) implies the Assumption 1.3.1(ia, iia, iii, iv) required to invoke the relevant theorems in Chapter 1.

The proportion test fixes a specific cut-off value c, such that the finite dimensional

convergence result for any $c \in [c_+, \infty)$ suffices. For $m \in [0, \infty)$ and as $n \to \infty$,

$$\mathbb{G}_n^{(m)}(c) = n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c^0) \xrightarrow{\mathsf{D}} \mathsf{N}(0, K_{cc}^{(m)}), \tag{2.B.1}$$

where $K_{cc}^{(m)}$ is the asymptotic variance of the respective theorem in Chapter 1. Dividing by $\sqrt{K_{cc}^{(m)}}$ and using the normality simplifications in Lemma 2.A.2 yields the result. ■

Proof of Theorem 2.5.2. The asymptotic distribution of the count test follows directly from Theorem 1.3.5 in Chapter 1, which derives a Poisson approximation for the number of expected outliers for a certain target expected value $\lambda^0 = n\gamma_{c_n}^0$ and for general iteration steps $m \in [0, \infty)$, i.e. including the fixed point $m = *$.

To apply that theorem, Assumption 1.3.1(i, ii, iii, iv, v) with $\eta = 1/4$ in Chapter 1 must hold. The asymptotic distribution of the count test is derived under Assumption 2.5.1(i, ii, iii) in the present chapter. The distributional Assumption 2.5.1(i) of joint normality of the errors implies the following, more general assumptions in Chapter 1: Assumption 1.3.1(i) on the structural error u_i (Johansen and Nielsen 2016a, 2016b), Assumption 1.3.1(ii) on the first stage errors r_i, and Assumption 1.3.1(iii) on their joint distribution (Johansen and Nielsen 2016b). Assumption 2.5.1(ii) is identical to Assumption 1.3.1(iv) in Chapter 1. Assumption 2.5.1(iii) coincides with Assumption 1.3.1(v) holding with $\eta = 1/4$, which means that the initial estimators for β and σ must be tight. This is fulfilled by but not limited to Algorithms 2.R and 2.S. Together, Assumption 2.5.1(i, ii, iii) implies Assumption 1.3.1(i, ii, iii, iv, v) with $\eta = 1/4$ in Chapter 1 and therefore allows to invoke Theorem 1.3.5 in Chapter 1.

Then, uniformly in $m \in [0, \infty)$ and as $n \to \infty$, the FODR satisfies

$$n\widehat{\gamma}_c^{(m)} \xrightarrow{\mathsf{D}} \text{Poisson}(\lambda^0),$$

where c_n and λ^0 are defined as described in Theorem 2.5.2, that is

$$n\mathsf{P}(|u_i| > \sigma c_n) = n\gamma_{c_n}^0 = \lambda^0.$$

■

Proof of Theorem 2.5.3. The proof relies on the same theorems in Chapter 1 as the proof of Theorem 2.5.1 (Theorems 1.3.2, 1.3.4, and 1.C.2). These theorems provide weak convergence results of the FODR $\widehat{\gamma}_c^{(m)}$ as a process in the cut-off value c for different iteration steps m and for Algorithms 2.R and 2.S. However, since the sum test is based on a scale of K different cut-off values $c_1, \ldots, c_K$ instead of a single cut-off value c, it uses the multivariate finite dimensional convergence results of the FODRs across the cut-off values.

Theorem 2.5.1 and Theorem 2.5.3 both make Assumption 2.5.1(i, ii). As described in the proof of Theorem 2.5.1, this assumption implies Assumption 1.3.1(ia, iia, iii, iv) in Chapter 1, which is required to invoke the relevant theorems.

The sum test chooses an ordered set of K different cut-off values $\mathbb{S}_c = \{c_1, \ldots, c_K\}$ in increasing order $c_1 < \cdots < c_K$ with a corresponding ordered set of hypothesised shares of falsely detected outliers $\mathbb{S}_\gamma = \{\gamma^0_{c_1}, \ldots, \gamma^0_{c_K}\}$ in strictly decreasing order $\gamma^0_{c_1} > \cdots > \gamma^0_{c_K}$. The test also fixes a certain iteration step $m \in [0, \infty)$. From the theorems above, we obtain the multivariate finite dimensional convergence of the centred FODR vector $\widehat{\theta}$, for any $m \in [0, \infty)$, $c \in [c_+, \infty)$, and as $n \to \infty$

$$\widehat{\theta} = n^{1/2} \begin{pmatrix} \widehat{\gamma}^{(m)}_{c_1} - \gamma^0_{c_1} \\ \widehat{\gamma}^{(m)}_{c_2} - \gamma^0_{c_2} \\ \vdots \\ \widehat{\gamma}^{(m)}_{c_K} - \gamma^0_{c_K} \end{pmatrix} \xrightarrow{\mathsf{D}} \mathsf{N}\left\{ \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} K^{(m)}_{c_1c_1} & K^{(m)}_{c_1c_2} & \cdots & K^{(m)}_{c_1c_K} \\ K^{(m)}_{c_1c_2} & K^{(m)}_{c_2c_2} & \cdots & K^{(m)}_{c_2c_K} \\ \vdots & \vdots & \ddots & \vdots \\ K^{(m)}_{c_1c_K} & K^{(m)}_{c_2c_K} & \cdots & K^{(m)}_{c_Kc_K} \end{pmatrix} \right\} = \mathsf{N}\left\{0_{K\times 1}, V\right\},$$

where $K^{(m)}_{st}$ for $c_+ \leq s \leq t < \infty$ is specified in the theorems listed above.

Now, take the sum of the FODR elements, which is equivalent to taking a linear combination of the FODR vector $\iota'\widehat{\theta}$, where $\iota = (1 \cdots 1)'$ is a non-stochastic $K \times 1$ vector of ones. The resulting asymptotic distribution is

$$\iota'\widehat{\theta} = \sum_{k=1}^{K} n^{1/2}(\widehat{\gamma}^{(m)}_{c_k} - \gamma^0_{c_k}) \xrightarrow{\mathsf{D}} \mathsf{N}\left\{\iota' 0_{K\times 1}, \iota' V \iota\right\} = \mathsf{N}\left\{0, \textstyle\sum_{k=1}^{K} K^{(m)}_{c_kc_k} + 2\sum_{1\leq k<l\leq K} K^{(m)}_{c_kc_l}\right\}$$
$$= \mathsf{N}\left\{0, K^{(m)}_{\mathbb{S}_c}\right\}.$$

Dividing by $\sqrt{K^{(m)}_{\mathbb{S}_c}}$ yields the sum test statistic and the formula for $K^{(m)}_{\mathbb{S}_c}$ can be simplified under normality using Lemma 2.A.2. ■

Proof of Theorem 2.5.4. The proof relies on the same theorems in Chapter 1 as the proof of Theorem 2.5.1 (Theorems 1.3.2, 1.3.4, and 1.C.2). These theorems demonstrate weak convergence of the FODR $\widehat{\gamma}_c^{(m)}$ as a process in the cut-off value c for different iteration steps m and for Algorithms 2.R and 2.S to a Gaussian process.

Theorem 2.5.1 and 2.5.4 both make Assumption 2.5.1(i, ii). As described in the proof of Theorem 2.5.1, this assumption implies Assumption 1.3.1(ia, iia, iii, iv) in Chapter 1, which is required to invoke the relevant theorems.

Applying the continuous mapping theorem to the weak convergence results (see Billingsley 1968; van der Vaart and Wellner 1996), we obtain the asymptotic distribution of the supremum test as $n \to \infty$

$$\sup_{1\leq k\leq K} |n^{1/2}(\widehat{\gamma}_{c_k}^{(m)} - \gamma_{c_k}^{0})| \xrightarrow{\mathsf{D}} \sup_{1\leq k\leq K} |\mathsf{GP}(c_k; 0, K^{(m)})|,$$

where $\mathsf{GP}(c; 0, K^{(m)})$ for $c \in [c_+, \infty)$ refers to a Gaussian process with expected value $\mathsf{E}\{\mathsf{GP}(c; 0, K^{(m)})\} = 0$ and covariance $K_{st}^{(m)} = \mathsf{Cov}\{\mathsf{GP}(s; 0, K^{(m)}), \mathsf{GP}(t; 0, K^{(m)})\}$ for any $c_+ \leq s \leq t < \infty$. Using Lemma 2.A.2, we obtain the simplified formula for the variance-covariance structure $K_{st}^{(m)}$ under normality. ■

2.C Comparison to Normality Tests

The theory developed in Chapter 1 that underlies our tests allows for other reference distributions than normality. This chapter restricts the analysis to the normal reference distribution because normality is a common choice in the outlier detection algorithms described in Section 2.4 and in other outlier detection procedures more generally (Barbato et al. 2011). The question then arises whether — in our normality setting — a goodness-of-fit test for normality of the structural errors can also serve as a test for outliers.

Several papers have explored the performance of various normality tests in the presence of outliers. Gel and Gastwirth (2008) argue that the Jarque-Bera (JB) test (Jarque and Bera 1987) can have low power to detect deviations from normality in the presence of outliers. The reason is that the JB test statistic is based on the sample skewness and kurtosis, which feature the sample variance in the denomin-

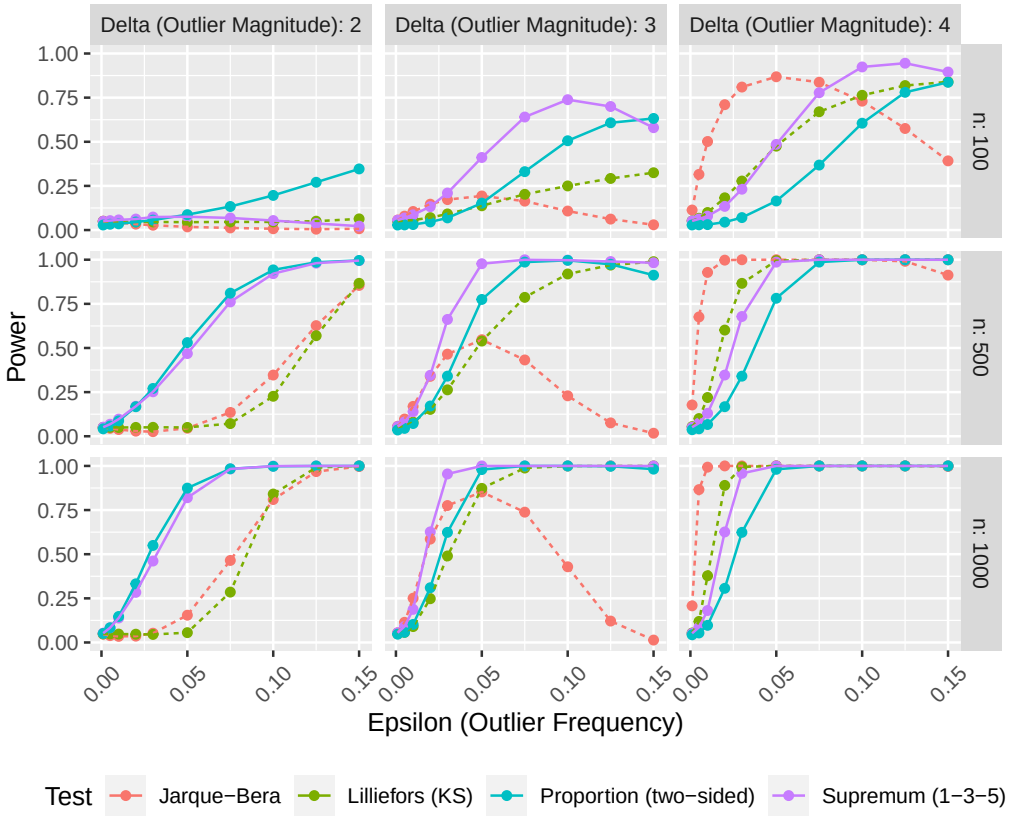
ator and which is easily inflated by outliers. The authors therefore suggest a Robust Jarque-Bera (RJB) test where they replace the variance estimator with the (suitably scaled) average absolute deviation from the median as a more robust measure of the spread of the data. Their simulations confirm higher power of the RJB test compared to non-normal distributions, including a normal distribution contaminated with outliers. Brys, Hubert and Struyf (2004) explain that the JB test, being based on central moments of the data, has a breakdown value of zero, meaning that a single outlier can induce an arbitrary bias in the test statistic. This bias can go in both directions, either making the test statistic very large or very small. Their main concern is that the JB test incorrectly rejects normality in a sample that mostly comes from a normal distribution except for a small share of outliers. Their Monte Carlo simulations implement Huber's (1964) ϵ-contamination where the reference distribution is standard normal. They find that the JB test almost always rejects normality for very large outlier magnitudes (7 or if drawn from a $\mathsf{N}(0, 5)$). For contaminations where less extreme values occur more often than expected (drawn from $\mathsf{N}(0, 0.05)$), the JB test does not reject normality often (0.06 for $\epsilon = 0.01$, 0.162 for $\epsilon = 0.05$). Also using Huber's (1964) ϵ-contamination framework, Saculinggan and Balase (2013) compare the rejection frequencies of six different normality tests with the standard normal reference distribution and $\mathsf{N}(5, 100)$ being the contaminating distribution. They find that the power to reject normality is generally low for small sample sizes ($n \leq 20$) but high for large sample sizes ($n = 100, 1000$), tends to increase in outlier frequency ϵ, and varies with the type of normality test. Cavus, Yazici and Sezer (2021) compare ten different normality tests in terms of size and power in a contamination setting where the contaminating distribution follows a uniform distribution with support in and beyond the realised tails of the "good" data in order to generate heavy tails. The Monte Carlo study finds that the power performance of the different tests varies across the sample size, outlier frequency ϵ, and outlier magnitude (measured by shifting the right bound of the support of the uniform distribution outwards).

Overall, it seems that the normality test may or may not have power to reject

normality depending on the type of contamination and that the type of contamination, outlier frequency, outlier magnitude, and sample size affect which of the normality tests has high power.

To be able to directly compare the performance in the contamination setting presented in Section 2.6.2, Figure 2.C.1 compares the proportion and supremum test to the Jarque-Bera (Jarque and Bera 1987) and the Lilliefors (Lilliefors 1967) test, which is based on the one-sample Kolmogorov-Smirnov (Kolmogorov 1933) test but specifically tests for normality with estimated mean and variance. The Kolmogorov-Smirnov test could also be used as an alternative test when we extend our proposed tests beyond normality and is therefore of special interest.

Figure 2.C.1: Power Comparison, $\alpha = 0.05$



Simulated power (size-adjusted) for different outlier frequencies ϵ, outlier magnitudes δ, and sample sizes n. Significance level $\alpha = 0.05$. Normality tests shown are the Jarque-Bera test, Lilliefors test; tests presented in this chapter are the two-sided proportion test with $c = 1.96$ and the supremum test with $\mathbb{S}_c = \{2.58, 2.17, 1.96\}$ based on Algorithm 2.R iterated to the fixed point. Each setting with $M = 10{,}000$ replications.

We find that neither test dominates the other. For small and medium-sized outliers ($\delta = 2, 3$), our proposed tests have higher power. For larger outliers ($\delta = 4$),

the JB test reaches very high power even for small shares of contamination but loses power if the outliers become too frequent, especially for small sample sizes. The Lilliefors test performs similarly to the supremum test, with power increasing more slowly than that of the JB test but staying high even for large shares of contamination.

Normality tests provide an alternative test for outliers when the reference distribution is the normal distribution but which type of normality test is preferred depends on the outlier magnitude, frequency, and general contamination scheme, which is not known to the researcher a priori. Moreover, a rejection of the normality test does not provide information on whether a specific subset of observations is not compatible with normality and therefore causing the rejection. In contrast, our tests are based on outlier classifications such that the observations that are flagged as outliers can be further investigated.

CHAPTER 3

A Monte Carlo Study of Super Saturation for Detecting Outliers and Location Shifts*

Abstract Unmodelled outliers and location shifts can cause serious problems for empirical modelling. Super saturation, a combination of impulse and step indicator saturation, has been used in the literature to detect outliers and location shifts of unknown magnitude and timing. However, its properties have not yet been systematically assessed. Therefore, this chapter conducts a Monte Carlo study of the performance of super saturation. To account for the fact that many different combinations of indicators can model an outlier or location shift, we suggest new evaluation measures based on the classical measures of gauge and potency. We find that super saturation reaches high detection rates of outliers and location shifts, with benefits to subsequent inference and model selection. However, to control the risk of retaining too many irrelevant indicators and thereby overfitting the model, we recommend using tight significance levels for selection and turning parsimonious encompassing tests off during the indicator search.

∗. We are grateful to participants at CFE-CMStatistics 2019, and the 3rd Oxford Workshop on Global Priorities Research 2019 for helpful comments and discussions. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (Richards 2015).

3.1 Introduction

Outliers and location shifts pose serious problems for empirical modelling. Failure to take them into account can lead to biased coefficient estimates, incorrect inference, distorted model selection, and systematic forecast failure (Hendry and Neal 1991; Castle and Hendry 2014; Hendry and Mizon 2014; Hendry 2018; Hendry and Muellbauer 2018). One approach to handle outliers, location shifts, and other types of structural breaks are indicator saturation methods. The idea is to translate the problem of structural break detection into one of model selection by searching over sets of different indicator types, each specifically designed to capture a certain type of structural break. To jointly detect outliers and location shifts, one can combine impulse indicator saturation (IIS) and step indicator saturation (SIS), which has been termed "super saturation" in Ericsson (2012).

While super saturation has already been applied in the literature many times, its performance has not yet been analysed systematically. This chapter aims to fill this gap and makes the following three contributions to the indicator saturation literature. First, we suggest a new measure for evaluating the performance of super saturation, which is adapted from the classical gauge and potency measures (Castle, Doornik and Hendry 2011). Instead of considering whether a retained indicator is relevant or irrelevant, we focus on whether a retained indicator contributes to capturing an outlier or location shift or not. Second, we use extensive Monte Carlo simulations to assess how well super saturation detects outliers and location shifts, and how many indicators are falsely retained. We provide a response surface of the gauge in terms of different tuning parameters of the selection algorithm, which can be used to calibrate the algorithm under the null. Third, we assess the costs and benefits of using super saturation prior to covariate selection, both in the absence and presence of outliers and location shifts.

Indicator saturation methods are usually implemented in a general-to-specific (GETS) model selection framework (see Hendry 2024, for a history of GETS). We illustrate its implementation using a stylised algorithm, before explaining some im-

portant characteristics of the more sophisticated algorithms which are used in practice, like Autometrics (Doornik 2009) or the R package *gets* (Pretis, Reade and Sucarrat 2018).

This chapter uses Autometrics for all its analyses. We highlight three tuning parameters of Autometrics and assess their impact on the performance of super saturation: the significance level for retaining indicators; backtesting a reduced model against the original model to test whether the former encompasses the latter; and diagnostic testing, such as whether the errors are serially uncorrelated and homoskedastic.

To evaluate the performance of super saturation, we suggest new performance metrics based on the existing metrics of gauge and potency (Castle, Doornik and Hendry 2011). Typically, the gauge represents the share of falsely retained irrelevant indicators and potency the share of correctly retained relevant indicators. However, with super saturation, outliers and location shifts can be modelled using a combination of impulse and step indicators, making the classification into relevant and irrelevant indicators ambiguous. To overcome this ambiguity, our new variants of gauge and potency instead focus on whether a retained indicator contributes to modelling an outlier or location shift — individually or in conjunction with others.

To assess the cost of super saturation, we use extensive Monte Carlo simulations to study the gauge under the null of no outliers or location shifts. We show that the gauge is well-controlled for tight significance levels, diagnostic testing has little effect on the gauge, and backtesting increases the gauge in most cases.

In order to generalise the simulation results beyond the specific settings that we analyse, we estimate response surfaces for the mean and variance of the gauge. In the response surfaces, we include asymptotic theory, which has been derived separately for stylised versions of IIS (Johansen and Nielsen 2016b) and SIS (Nielsen and Qian 2023) — the constituent parts of super saturation. Despite the fact that the asymptotic theory has been developed for IIS and SIS, we find that it approximates the gauge of super saturation well for tight significance levels and when backtesting is turned off.

Based on our simulations under the null, we recommend adopting a stringent significance level to control the gauge. Backtesting should be turned off, while diagnostic testing can be turned on if desired because it has a negligible impact on the gauge in a well-specified model but otherwise has the potential to improve the resulting final model.

Applying super saturation with our recommended tuning parameters in location-scale regression models with outliers, location shifts, or both, our Monte Carlo simulations show that Autometrics reaches a high detection rate of these phenomena while the gauge is still well-controlled. Super saturation is therefore well-suited for its intended purpose.

In practice, super saturation is often used either in combination with an econometric model that represents economic theory or prior to model selection, as discussed in Hendry and Johansen (2015). In a first step, super saturation selects relevant indicators while retaining all the covariates. In a second step, the covariates are selected while retaining the previously selected indicators. To assess the costs and benefits of super saturation in such a setting, we extend the location-scale regression model by adding five relevant and five irrelevant, strongly exogenous covariates.

When the data generating process (DGP) contains no outliers or location shifts, we find that the additional search cost of super saturation is mainly characterised by an underestimate of the error variance, which leads to higher retention rates of both relevant and irrelevant covariates in subsequent model selection. If instead the DGP contains outliers and direct or induced location shifts, it successfully captures these breaks and improves subsequent model selection. A location shift in an irrelevant regressor does not lead to spurious retention of a matching step indicator because indicator saturation is performed prior to covariate selection.

This research is most closely linked to the indicator saturation literature, which has developed various modelling techniques to address different forms of structural breaks and non-stationarity: Impulse indicator saturation (IIS, Hendry, Johansen and Santos 2008) for detecting outliers, step indicator saturation (SIS, Castle et

al. 2015) for location shifts, trend indicator saturation (TIS, Castle et al. 2022) for breaks in trends, and multiplicative indicator saturation (MIS, Castle and Hendry 2019, chapter 5.3.2) for parameter non-constancy. In addition, indicators can be designed to detect specific phenomena using subject-matter knowledge (Pretis et al. 2016; Schneider et al. 2017). Super saturation is the term for using IIS and SIS simultaneously (Ericsson 2012). The idea for IIS originated in Hendry (1999), who used impulse indicators to "dummy out" times of food rationing and other outliers when modelling US food expenditure between 1931 and 1989. Hendry, Johansen and Santos (2008) later formalise IIS and analyse it in a location-scale regression model. The authors show that under the null hypothesis of no outliers, the false retention rate of impulse indicators equals $\gamma \times T$, where γ is the selection significance level and T is the sample size. Furthermore, they derive the probability limit of the variance estimator and the asymptotic distribution of the mean estimator. Johansen and Nielsen (2009) extend the analysis to autoregressions with stationary, trend stationary, and unit root regressors. They also establish a link between IIS and the robust statistics literature. Split-half IIS belongs to the class of one-step M-estimators with the Huber-skip objective function. Johansen and Nielsen (2013) consider the properties of iterated Huber-skip M-estimators, which is closer to the more sophisticated version of IIS implemented in tree-search algorithms like Autometrics (Doornik 2009) or the R package *gets* (Pretis, Reade and Sucarrat 2018). Johansen and Nielsen (2016b) provide an asymptotic normal and Poisson approximation for the gauge using empirical processes derived in Johansen and Nielsen (2016a). Hendry and Doornik (2014) provide extensive Monte Carlo simulations analysing the performance of IIS in terms of gauge and potency, for various different DGPs both under the null and in the presence of outliers or location shifts.

SIS has been introduced in Castle et al. (2015) as a method to detect location shifts at unknown times and of unknown magnitude and duration. Their Monte Carlo simulations show that, as for IIS, the gauge under the null of no shifts is close to $\gamma \times T$ for a selection significance level γ. Recently, Nielsen and Qian (2023) have derived the asymptotic distribution of the gauge for the split-half version of SIS in

the location-scale model as well as regressions with strictly exogenous, stationary, or non-stationary regressors. In contrast to IIS, the asymptotic variance of the gauge does not only depend on the significance level γ but also on the correlation between the forward differences of the regressors and the error term.

Even though the performance of super saturation has not yet been systematically analysed, it has been applied in the literature for different purposes, which can be classified into three different categories. First, super saturation has been used directly in the modelling stage to account for outliers and location shifts, usually while retaining or prior to selecting an econometric model. Second, super saturation has been used to test quantitative forecasts for time-varying, systematic biases. Third, it has been applied as a test for super exogeneity, i.e. whether the regressor in a conditional model is weakly exogenous and whether the parameter in the conditional model is invariant to certain changes in the regressor's marginal process.

Next, we illustrate the different purposes using examples from the literature. These examples are not exhaustive but rather serve to illustrate the method's versatility in different context.

In terms of modelling outliers and location shifts, Castle and Hendry (2020, Section 7) use super saturation to detect outliers and structural breaks in a model of UK annual CO_2 per capita emissions from 1860 to 2017. The retained impulse and step indicators are crucial for obtaining a stable model across such a long time period and the outliers and shifts can be traced back to important events, e.g., a national strike by coal miners in 1912 and the 2008 UK Climate Change Act. Ericsson, Dore and Butt (2022) apply super saturation to detect breaks in precipitation patterns that are driven by climate change. Castle, Hendry and Martinez (2023) use super saturation and other types of indicator saturation to handle structural breaks in models of UK real wage growth, unemployment, prices, and production over a time span of 160 years. Apergis et al. (2023) model Australian electricity prices using indicator saturation methods, including super saturation, to deal with the many outliers and location shifts present in electricity price data due to intermittency of renewable energy generation and non-storability of electricity.

Ericsson (2017) suggests using super saturation to test for the presence of time-varying, systematic biases in forecasts. Unlike the test suggested by Mincer and Zarnowitz (1969), which averages forecast errors across the whole forecast horizon, super saturation allows for the detection of time-varying biases that might otherwise cancel each other. Ericsson (2017) analyses forecasts of one-year-ahead US gross federal debt produced by the US government and Ericsson, Fiallos and Seymour (2015) analyse forecasts of GDP growth produced by the FED (Greenbook forecasts). While the Mincer-Zarnowitz test rarely detects systematic forecast biases, super saturation does detect biases for certain time periods, which seem to coincide with changes in the business cycle.

Finally, Hendry and Santos (2010) and Castle, Hendry and Martinez (2017) propose the use of IIS and super saturation as automatic tests for super exogeneity (Engle, Hendry and Richard 1983). The idea is to test whether outliers and location shifts detected in the marginal equation of a conditioning variable are also significant in the conditional model for the outcome of interest, which would violate super exogeneity. Castle, Hendry and Martinez (2023) use super saturation to test for super exogeneity of the variables in their wage and production equations.

This chapter is structured as follows. Section 3.2 explains how super saturation can be implemented in a general-to-specific model selection framework and introduces some notation. Section 3.3 reviews the existing metrics of gauge and potency used to evaluate automatic model selection, explains why they cannot be directly applied to super saturation, and suggests new, adapted measures of gauge and potency. Section 3.4 uses Monte Carlo simulations to evaluate the performance of super saturation in a location-scale regression model. Section 3.4.1 analyses the gauge under the null of no outliers or location shifts and estimates response surfaces to generalise the Monte Carlo experiments beyond the parameter space that we explored. Sections 3.4.2, 3.4.3, and 3.4.4 assess the performance when the DGP contains outliers, location shifts, or both. Section 3.5 studies the costs and benefits of using super saturation prior to inference on, or model selection of, other covariates included in the regression model. Section 3.6 concludes.

3.2 Super Saturation: Overview and Estimation

Super saturation, the combined application of IIS and SIS, is designed to detect both outliers and location shifts of unknown magnitude and timing. It selects over saturating sets of impulse and step indicators and is often implemented in a general-to-specific modelling framework. The implementation can be illustrated using a stylised "split-half" algorithm but in practice, more sophisticated algorithms like Autometrics (Doornik 2009) in OxMetrics (Doornik 2018a) or the *gets* package (Pretis, Reade and Sucarrat 2018) in R (R Core Team 2022) are used to implement the selection of indicators.

IIS is suitable to agnostically detect outliers of unknown magnitudes at unknown time periods. IIS saturates the regression model with a full set of impulse indicator variables. Impulse indicators each take on the value 1 in a specific time period and are 0 otherwise. Consecutive indicators can also model longer lasting location shifts but SIS is specifically designed to do so in a more parsimonious way with higher detection probabilities.

SIS can detect location shifts of unknown magnitude and duration at unknown time periods. SIS fully saturates the regression model with step indicator variables, which take on the value 1 from the start of the sample until a certain time period, and are 0 thereafter. An equivalent formulation could take on the value 0 until a certain time period and 1 afterwards. However, the first formulation is preferred for interpretability when forecasting because the step indicators then do not extend into the forecast horizon.

We use $1_{(\cdot)}$ to denote the indicator function that takes on the value 1 when the condition in parentheses is fulfilled and 0 otherwise. We can then write an impulse indicator for time period j as $1_{(t=j)}$ and we sometimes refer to it using the variable name "Ij". Similarly, a step indicator for a shift that lasts until time period j can be denoted by $1_{(t\leq j)}$ and we occasionally refer to it as "Sj". The general unrestricted model (GUM) with super saturation can be written as

$$y_t = \beta_0 + \sum_{c=1}^{C} \phi_c x_{c,t} + \sum_{j=2}^{T} \beta_j^I 1_{(t=j)} + \sum_{j=1}^{T-1} \beta_j^S 1_{(t\leq j)} + \epsilon_t, \quad t = 1, \ldots, T, \qquad (3.2.1)$$

where β_0 is the coefficient on the intercept, $\{x_{c,t}\}_{c=1}^{C}$ are explanatory covariates representing economic theory with associated coefficients $\{\phi_c\}_{c=1}^{C}$, and $\{\beta_j^I\}_{j=2}^{T}$ and $\{\beta_j^S\}_{j=1}^{T-1}$ are the coefficients on the impulse and step indicators. The first impulse indicator $1_{(t=1)}$ is omitted because it is identical to the first step indicator $1_{(t\leq 1)}$, and the final step indicator $1_{(t\leq T)}$ is omitted because it coincides with the intercept. For the rest of the chapter, the error will be identically and independently normally distributed, $\epsilon_t \overset{iid}{\sim} \mathsf{N}(0, \sigma_\epsilon^2)$, and the equations hold for $t = 1, \ldots, T$.

Due to the double saturation with indicators, the model includes many more variables than observations by construction. In addition to the intercept, there are $T-1$ impulse and step indicators each, as well as potentially $C \geq 0$ covariates. In total, this yields $1 + 2(T-1) + C$ variables compared to only T time periods.

Since most indicators are likely to be irrelevant, we can assume sparsity of their coefficients and use model selection to find the relevant indicators. In the indicator saturation literature, this is usually done in a general-to-specific modelling framework, as implemented by the algorithm Autometrics in OxMetrics or the *gets* package in R.

To gain an intuition for how these model selection algorithms work, let us consider a simplified implementation of IIS and SIS, which is called the split-half algorithm. It is used as an illustration of the feasibility of the indicator saturation methods and serves as an analytically tractable approximation for the more complicated algorithms when deriving asymptotic theory (Hendry, Johansen and Santos 2008; Johansen and Nielsen 2009, 2016b; Castle et al. 2015; Nielsen and Qian 2023).

For simplicity, we assume that the sample contains an even number of observations T and that $T/2 + C < T$. Choose a significance level γ for testing the significance of an indicator. We start the algorithm with IIS. Split the set of impulse indicators into two halves such that each subset covers one half of the sample. First, include the indicators for the first half $\{1_{(t=s)}\}_{s=2}^{T/2}$ alongside the C covariates, estimate the model, and record which impulse indicators are statistically significant. Second, include the second half of the indicators $\{1_{(t=s)}\}_{s=T/2+1}^{T}$ alongside the C theory variables and again record the significant indicators. Under the null of

no outliers and location shifts, approximately $\gamma \times \frac{T}{2}$ indicators are expected to be significant by chance in each half. Next, add the union of the previously significant indicators to the theory variables and conduct another test of significance. This procedure can be repeated for SIS, yielding a selection of step indicators. Finally, do a last round of selection in a model containing both the previously retained impulse and step indicators. The resulting model is the selection outcome of super saturation.

This chapter uses Autometrics to implement all the simulations, which is a much more sophisticated algorithm than the stylised split-half algorithm described above.[1] Doornik (2009) describes the details of Autometrics. Autometrics uses a multi-path tree-search where insignificant variables are removed in turns, each resulting model in principle representing the starting point for a new removal. In order to keep the computational cost low, Autometrics uses pruning (removing all sub-branches if a model is invalid), bunching (deleting groups of variables if they are highly insignificant instead of considering each separately), and chopping (permanently removing a highly insignificant regressor from all potential models). The algorithm may arrive at several different terminal models and information criteria can be used to choose a final "best" model. For indicator saturation methods in particular, Autometrics uses many more than two splits in the block search and also combines impulse and step indicators within the same block, allowing for both outliers and location shifts to be captured at the same time.

Moreover, Autometrics implements diagnostic testing and backtesting in the reduction process. These tests can be turned on or off by the user. Diagnostic testing includes, but is not limited to, whether the errors are serially uncorrelated, homoskedastic, and normally distributed. A variable may be retained even if it is insignificant if it ensures the passing of the diagnostic tests.

Backtesting is a step that Autometrics uses to test whether a reduction (removing a regressor) is permissible or not. It uses a parsimonious encompassing test,

1. The evaluations and graphs were created using the R packages (R Statistical Software v4.1.2; R Core Team 2022) of the tidyverse (Wickham et al. 2019) and rgl (Murdoch and Adler 2024) in RStudio (Posit Team 2023).

implemented as an F-test, to compare the reduced model against the initial GUM.[2] With super saturation and indicator saturation methods more generally, it is unclear what the appropriate GUM is. Since there are many more regressors than observations, the GUM cannot be directly estimated. In practice, Autometrics runs an initial search for relevant indicators and then compares the reductions to this initially fitted model.

The selection of indicators can be done while retaining (not selecting over) the other regressors $\{x_{c,i}\}_{c=1}^{C}$. As explained in Hendry and Johansen (2015), one can orthogonalise the indicator variables with respect to the theory variables before selection. If all indicators are irrelevant, the asymptotic distributions of the estimators of $\{\phi_c\}_{c=1}^{C}$ are unaffected by the selection process. Under the alternative, however, selection can yield a better model.

We will refer to the explanatory variables $\{x_{c,t}\}_{c=1}^{C}$ as "covariates" to distinguish them from the indicator variables. Similarly, we will use the term "covariate selection" to refer to model selection of these variables in order to distinguish it from super saturation, which uses model selection to select among the indicators.

3.3 New Performance Metrics for Super Saturation

The performance of indicator saturation methods is usually evaluated based on how many relevant and irrelevant indicators have been retained. However, with super saturation, outliers and location shifts can be modelled using a combination of impulse and step indicators, making the classification into relevant or irrelevant indicators ambiguous. To overcome this ambiguity, we adapt the existing metrics to make them phenomenon-based, i.e. a retained indicator counts as "relevant" if it contributes to modelling an outlier or shift in the DGP, and is otherwise "irrelevant".

2. If lagged regressors are included in the GUM and the pre-search option is turned on, it compares the reduced model against initial GUM after insignificant lagged regressors have been removed during the pre-search.

3.3.1 Existing Performance Metrics for Automatic Model Selection

Castle, Doornik and Hendry (2011) propose the metrics gauge and potency to evaluate automatic model selection.[3] These metrics have been widely adopted in the general-to-specific (GETS) literature, not just for evaluating indicator saturation methods but also covariate selection. Before suggesting an extension of the gauge and potency in Section 3.3.3 to evaluate the special case of super saturation, we review the original concepts, which we still use for evaluating covariate selection in Section 3.5.

Definition 1 (Classical Performance Metrics). *Suppose the DGP contains* $N > 0$ *covariates and let* $\mathcal{N} = \{1, \dots, N\}$ *denote their index set. In the model selection, we consider the* N *relevant variables of the DGP and* $C - N > 0$ *additional, irrelevant regressors. Let* $\widehat{\phi}_{c,i}$ *denote the estimated coefficient of regressor* $c \in \{1, \dots, N, N+1, \dots, C\}$ *after model selection in Monte Carlo replication* $i \in \{1, \dots, M\}$.

Definition 1.1 (Classical Potency). *Castle, Doornik and Hendry (2011) define potency as the retention rate of relevant regressors. It is the ratio of the number of correctly retained regressors to the number of relevant regressors in the DGP, averaged across Monte Carlo replications:*

$$\widehat{\rho} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{c \in \mathcal{N}} 1_{(\widehat{\phi}_{c,i} \neq 0)}. \tag{3.3.1}$$

Definition 1.2 (Classical Gauge). *Castle, Doornik and Hendry (2011) define gauge as the retention rate of irrelevant regressors, calculated as the ratio of the number of retained irrelevant regressors to the number of irrelevant regressors considered in the selection set, averaged across Monte Carlo replications:*

$$\widehat{\gamma} = \frac{1}{M(C-N)} \sum_{i=1}^{M} \sum_{c \notin \mathcal{N}} 1_{(\widehat{\phi}_{c,i} \neq 0)}. \tag{3.3.2}$$

These classical measures of gauge and potency are applicable not only for general model selection of covariates but also in the setting of indicator selection in indicator

3. The term gauge has been first used in Hendry and Santos (2010) to distinguish it from size because a variable might be retained due to diagnostic testing or backtesting even if it is insignificant.

saturation methods — as long as it can be uniquely determined whether an indicator is relevant or irrelevant. For example, if step indicator saturation is applied to a DGP with a location shift at a certain point in time, the matching step indicator can be seen as relevant and all the other step indicators as irrelevant. The next section, however, explains why the concept of relevant and irrelevant indicators becomes ambiguous in the case of super saturation.

3.3.2 Shortcomings of Existing Metrics for Super Saturation

The existing metrics of gauge and potency require the set of relevant and irrelevant variables to be clearly defined. With super saturation, however, both location shifts and outliers can be modelled using either impulse indicators, step indicators, or a combination of both. Hence, whether an indicator is relevant or irrelevant becomes ambiguous.

To illustrate this ambiguity, consider a DGP with a location shift of magnitude λ in time period t_s and an outlier of magnitude δ in time period t_o

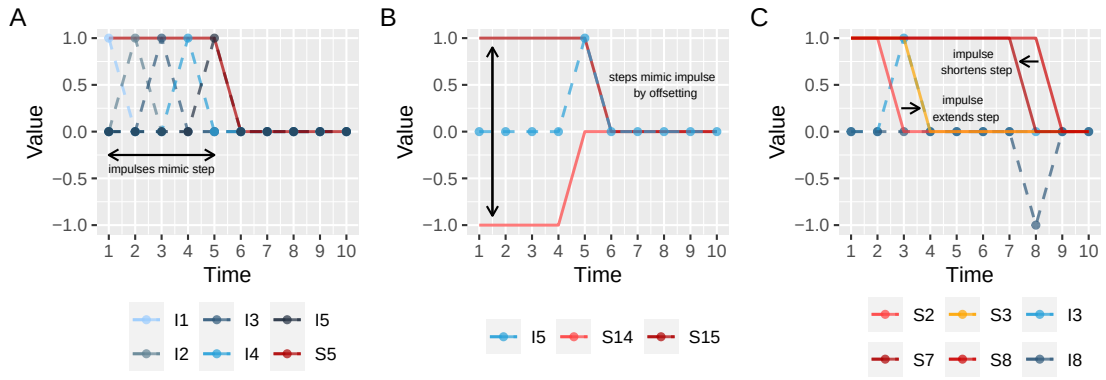
$$y_t = \beta_0 + \lambda 1_{(t \leq t_s)} + \delta 1_{(t=t_o)} + \epsilon_t. \tag{3.3.3}$$

Instead of the step indicator $1_{(t \leq t_s)}$, the location shift can also be expressed and hence modelled by a set of consecutive impulse indicators of equal sign and magnitude. Similarly, the impulse indicator $1_{(t=t_o)}$ can be modelled by two consecutive step indicators of equal magnitude but opposite sign. This yields the following equivalent re-parameterisation of equation (3.3.3)

$$y_t = \beta_0 + \lambda\{1_{(t=1)} + \ldots + 1_{(t=s)}\} + \delta\{1_{(t \leq t_o)} - 1_{(t \leq (t_0-1))}\} + \epsilon_t. \tag{3.3.4}$$

Of course, combinations of step and impulse indicators are equally able to capture the same location shift. Figure 3.3.1 illustrates the three cases of (A) a location shift modelled by a single step indicator or several consecutive impulse indicators; (B) an outlier modelled by a single impulse indicator or two consecutive, opposite-signed step indicators; and (C) a location shift modelled by a step indicator that is extended by a subsequent impulse indicator or shortened by an opposite-signed impulse indicator offsetting its effect in the last time period.

Figure 3.3.1: Illustration of Ambiguity of Indicator Relevance



This ambiguity about an indicator's relevance renders the classical gauge and potency metrics equally ambiguous. The next section therefore suggests adapted versions of the gauge and potency to evaluate the performance of super saturation.

3.3.3 Adapted, Phenomenon-Based Performance Metrics

We propose a phenomenon-based variant of the gauge and potency, which considers whether an indicator contributes to modelling an outlier or location shift. Potency is then defined as the share of detected outliers and location shifts that exist in the DGP, irrespective of whether they are captured by impulse indicators, step indicators, or a combination thereof. Any retained indicator that does not count towards potency instead counts towards the gauge.

In this chapter, we place further, conservative restrictions on the definition of potency. First, we do not simply allow for *any* pair of consecutive step indicators to capture an outlier. Only if their coefficients sum to approximately zero within a tolerance level δ^I do they count towards the potency of the outlier. Second, we do not allow a step shift to be captured purely by a sequence of impulse indicators because a step indicator could do so more parsimoniously, which is preferred. Only a consecutive pair of one impulse and one step indicator are allowed to capture a step shift. Similar to the outlier case, we require their magnitude to be approximately equal within tolerance level δ^S such that the impulse indicator is extending the step indicator by one period or shortening it by one period. This is formalised in the

following definitions.

Definition 2 (Phenomenon-Based Performance Metrics). *Let* $\mathcal{T} = \{1, 2, \ldots, T\}$ *be the set of time periods. Let* $\mathcal{T}^S$ *(*$\mathcal{T}^I$*) be the set of time periods in which a step shift (outlier) occurs. Let* N^S *(*N^I*) be the number of time periods in which a step shift (outlier) occurs, i.e. the cardinality of the set* $\mathcal{T}^S$ *(*$\mathcal{T}^I$*). Let* $\widehat{\beta}^S_{t,i}$ *(*$\widehat{\beta}^I_{t,i}$*) denote the estimated coefficient on the step indicator (impulse indicator) for time period* $t \in \mathcal{T}$ *after model selection in Monte Carlo replication* $i \in \{1, 2, \ldots, M\}$*. The indicator counts as retained if* $\widehat{\beta}^K_{t,i} \neq 0$*, for* $K \in \{S, I\}$*. Let* δ^I *denote the tolerance for two consecutive step indicators to count towards impulse potency and let* δ^S *denote the tolerance for a consecutive pair of impulse and step indicators to count towards the step potency. As in equation* (3.2.1)*, the infeasible GUM includes an intercept that is not selected over, and a set of* $T - 1$ *impulse and step indicators each.*

Definition 2.1 (Outlier Potency). *An outlier counts as detected either when it is matched by a single impulse indicator or two consecutive step indicators with coefficients summing up to approximately zero within a tolerance level* δ^I*:*

$$\widehat{\rho}^I = \frac{1}{MN^I} \sum_{i=1}^{M} \sum_{t \in \mathcal{T}^I} \max(1_{(\widehat{\beta}^I_{t,i} \neq 0)}, 1_{(\widehat{\beta}^S_{t-1,i} \neq 0 \wedge \widehat{\beta}^S_{t,i} \neq 0 \wedge |\widehat{\beta}^S_{t-1,i} + \widehat{\beta}^S_{t,i}| < \delta^I)}).$$

Definition 2.2 (Shift Potency). *A location shift counts as detected either when it is matched by a single step indicator or a step indicator that ends one time period early (late) compensated by an impulse indicator effectively extending (shortening) its duration. The magnitudes of the coefficients must be within a tolerance level* δ^S*:*

$$\widehat{\rho}^S = \frac{1}{MN^S} \sum_{i=1}^{M} \sum_{t \in \mathcal{T}^S} \max(1_{(\widehat{\beta}^S_{t,i} \neq 0)}, 1_{(\widehat{\beta}^S_{t-1,i} \neq 0 \wedge \widehat{\beta}^I_{t,i} \neq 0 \wedge |\widehat{\beta}^I_{t,i} - \widehat{\beta}^S_{t-1,i}| < \delta^S)},$$
$$1_{(\widehat{\beta}^S_{t+1,i} \neq 0 \wedge \widehat{\beta}^I_{t+1,i} \neq 0 \wedge |\widehat{\beta}^S_{t+1,i} + \widehat{\beta}^I_{t+1,i}| < \delta^S)}).$$

Definition 2.3 (Impulse Gauge). *A retained impulse indicator counts towards the impulse gauge if it does not contribute to potency, i.e. if it does not match an*

outlier and if it does not account for a step shift jointly with a step indicator:

$$\widehat{\gamma}^I = \frac{1}{M(T-N^I-1)} \sum_{i=1}^{M} \{ \sum_{t \notin \mathcal{T}^I} 1_{(\widehat{\beta}^I_{t,i} \neq 0)} - \sum_{t \in \mathcal{T}^S} \max(1_{(\widehat{\beta}^S_{t-1,i} \neq 0 \wedge \widehat{\beta}^I_{t,i} \neq 0 \wedge |\widehat{\beta}^I_{t,i} - \widehat{\beta}^S_{t-1,i}| < \delta^S)}, 1_{(\widehat{\beta}^S_{t+1,i} \neq 0 \wedge \widehat{\beta}^I_{t+1,i} \neq 0 \wedge |\widehat{\beta}^S_{t+1,i} + \widehat{\beta}^I_{t+1,i}| < \delta^S)}) \}$$

Definition 2.4 (Step Gauge). *A retained step indicator counts towards the step gauge if it does not contribute to potency, i.e. if it does not match a location shift exactly or jointly with an impulse indicator and if it does not account for an outlier:*

$$\widehat{\gamma}^S = \frac{1}{M(T-N^S-1)} \sum_{i=1}^{M} \{ \sum_{t \notin \mathcal{T}^S} 1_{(\widehat{\beta}^S_{t,i} \neq 0)} - \sum_{t \in \mathcal{T}^S} \max(1_{(\widehat{\beta}^S_{t-1,i} \neq 0 \wedge \widehat{\beta}^I_{t,i} \neq 0 \wedge |\widehat{\beta}^I_{t,i} - \widehat{\beta}^S_{t-1,i}| < \delta^S)}, 1_{(\widehat{\beta}^S_{t+1,i} \neq 0 \wedge \widehat{\beta}^I_{t+1,i} \neq 0 \wedge |\widehat{\beta}^S_{t+1,i} + \widehat{\beta}^I_{t+1,i}| < \delta^S)}) - 2 \sum_{t \in \mathcal{T}^I} 1_{(\widehat{\beta}^S_{t-1,i} \neq 0 \wedge \widehat{\beta}^S_{t,i} \neq 0 \wedge |\widehat{\beta}^S_{t-1,i} + \widehat{\beta}^S_{t,i}| < \delta^I)} \}$$

3.4 Performance in a Location-Scale Regression Model

The simulations under the null without outliers or location shifts show that the gauge can be controlled by choosing a tight significance level for retaining indicators. Backtesting increases the gauge considerably for loose significance levels and might better be turned off, while diagnostic testing has little effect on the gauge. Despite the tight significance level, super saturation successfully detects outliers, location shifts, and combinations of both and is therefore well-suited for its intended purpose.

3.4.1 DGP without Outliers or Location Shifts

To assess the cost of super saturation, we analyse the gauge under the null of no outliers or location shifts. We especially focus on the effect of three tuning parameters of Autometrics: the selection significance level, diagnostic testing, and backtesting.

The DGP for y_t is simply an intercept with an independently and identically distributed error $\epsilon_t \overset{iid}{\sim} \mathsf{N}(0, \sigma^2_\epsilon)$,

$$y_t = \beta_0 + \epsilon_t. \tag{3.4.1}$$

Without loss of generality, we set $\beta_0 = 0$ and $\sigma^2_\epsilon = 1$. The GUM is given by equation (3.2.1), with a fixed intercept and the step and impulse indicators for super saturation but excludes any covariates ($C = 0$).

In order to assess the impact of various tuning parameters on the gauge, we cover a wide variety of settings in the Monte Carlo simulations. We call the parameter space that we explore Θ , which contains variations of the selection significance level $\gamma \in \{0.001, 0.005, 0.01, 0.025, 0.05\}$; the sample size $T \in \{25, 50, 75, 100, 200, 300\}$; backtesting turned off or on, $b \in \{0, 1\}$[4]; and which (if any) diagnostic tests are turned on, $d \in \{None, AR, ARCH, Chow, Heterosk., Normality, Default\}$.

AR tests for autocorrelation of the errors (Godfrey 1978), *ARCH* for autoregressive conditional heteroskedasticity of the errors (Engle 1982), *Chow* for parameter constancy (Chow 1960), *Heterosk.* for heteroskedasticity of the errors (White 1980), and *Normality* for normality of the errors (Doornik and Hansen 2008). *None* does not use any of the tests whereas *Default* uses all five.

In total, this yields 420 different Monte Carlo experiments. We use index j to refer to one of these 420 Monte Carlo experiments and $\theta_j \in \Theta$ is its specific parameter configuration. We use $M = 10{,}000$ replications for each experiment and index $i = 1, \ldots, M$ refers to a specific replication within an experiment.

We analyse the impulse and the step gauge separately because the two indicator types are fundamentally different. Impulse indicators are mutually orthogonal while step indicators are highly correlated, making it more difficult to analyse step indicator saturation theoretically and to find the step indicator that exactly matches the location shift (if present) during the indicator selection.

3.4.1.1 Asymptotic Theory for the Gauge of IIS and SIS

We will compare the simulated mean and variance of the gauge of super saturation to the theoretical asymptotic mean and variance of the gauge of IIS and SIS, which were derived in Johansen and Nielsen (2016b) and Nielsen and Qian (2023), respectively.

Note that the asymptotic results are not directly applicable to our setting. First, the asymptotic theory has been derived for the cases when either IIS or SIS is used but not both. Second, the theoretical derivations are based on the stylised split-half algorithm because of analytical tractability, whereas we present the results from

4. The Ox (Doornik 2018b) command `model.Autometrics("block_method", "reduced");` was used to turn backtesting off in the simulations.

Autometrics. Autometrics not only uses a sophisticated multi-path block search, it also allows the use of diagnostic testing and backtesting during the selection process.

Hence, even for large sample sizes, the simulation results need not be close to the asymptotic theory. Nevertheless, the next sections show that the theory does provide reasonably close approximations for the mean and variance of the gauge in our setting, allowing us to estimate response surfaces of the gauge.

Johansen and Nielsen (2016b) show that the empirical gauge of impulse indicators, $\widehat{\gamma}^I$ is consistent for γ^I, the target significance level for selecting impulse indicators. The asymptotic distribution is shown to be

$$T^{1/2}(\widehat{\gamma}^I - \gamma^I) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \mathsf{avar}(\widehat{\gamma}^I)), \tag{3.4.2}$$

where the formula for the asymptotic variance is provided in Johansen and Nielsen (2016b, Corollary 5, p. 338).

Similarly, Nielsen and Qian (2023) show that

$$T^{1/2}(\widehat{\gamma}^S - \gamma^S) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \mathsf{avar}(\widehat{\gamma}^S)), \tag{3.4.3}$$

where γ^S corresponds to the target significance level for selecting step indicators. The formula for the asymptotic variance can be found in Nielsen and Qian (2023, Theorem 4.5, p. 15).

Even though the true asymptotic distribution of the gauge is unknown, we can use Monte Carlo simulations to learn about its mean and variance as functions of the tuning parameters of Autometrics. See Hendry (1984) and Doornik and Hendry (2018) for a detailed treatment of the use of Monte Carlo experiments in econometrics.

We use $\widehat{\gamma}^K_{j,i}$ to denote the gauge, $\gamma^K_j = \mathsf{E}(\widehat{\gamma}^K_{j,i})$ its mean, and $\omega^K_{j,k} = \mathsf{E}(\widehat{\gamma}^K_{j,i} - \gamma^K_j)^k$ its k-th central moment for indicator type $K \in \{I, S\}$ in replication i of Monte Carlo experiment j. We know that

$$\widehat{\gamma}^K_{j,i} \overset{iid}{\sim} (\gamma^K_j, \omega^K_{j,2}), \tag{3.4.4}$$

where γ^K_j and $\omega^K_{j,2}$ denote the true mean and variance.

We estimate the mean using our simulation results by averaging the simulated gauges across all replications, i.e.

$$\widehat{\gamma}_j^K = \frac{1}{M} \sum_{i=1}^{M} \widehat{\gamma}_{j,i}^K. \tag{3.4.5}$$

Since $\mathsf{E}(\widehat{\gamma}_j^K) = \gamma_j^K$ and $\mathsf{V}(\widehat{\gamma}_j^K) = \omega_{j,2}^K/M$, our estimate of the mean becomes more precise as the number of replications M increases.

The variance is estimated by the variance across replications,

$$\widehat{\omega}_{j,2}^K = \frac{1}{M-1} \sum_{i=1}^{M} (\widehat{\gamma}_{j,i}^K - \widehat{\gamma}_j^K)^2, \tag{3.4.6}$$

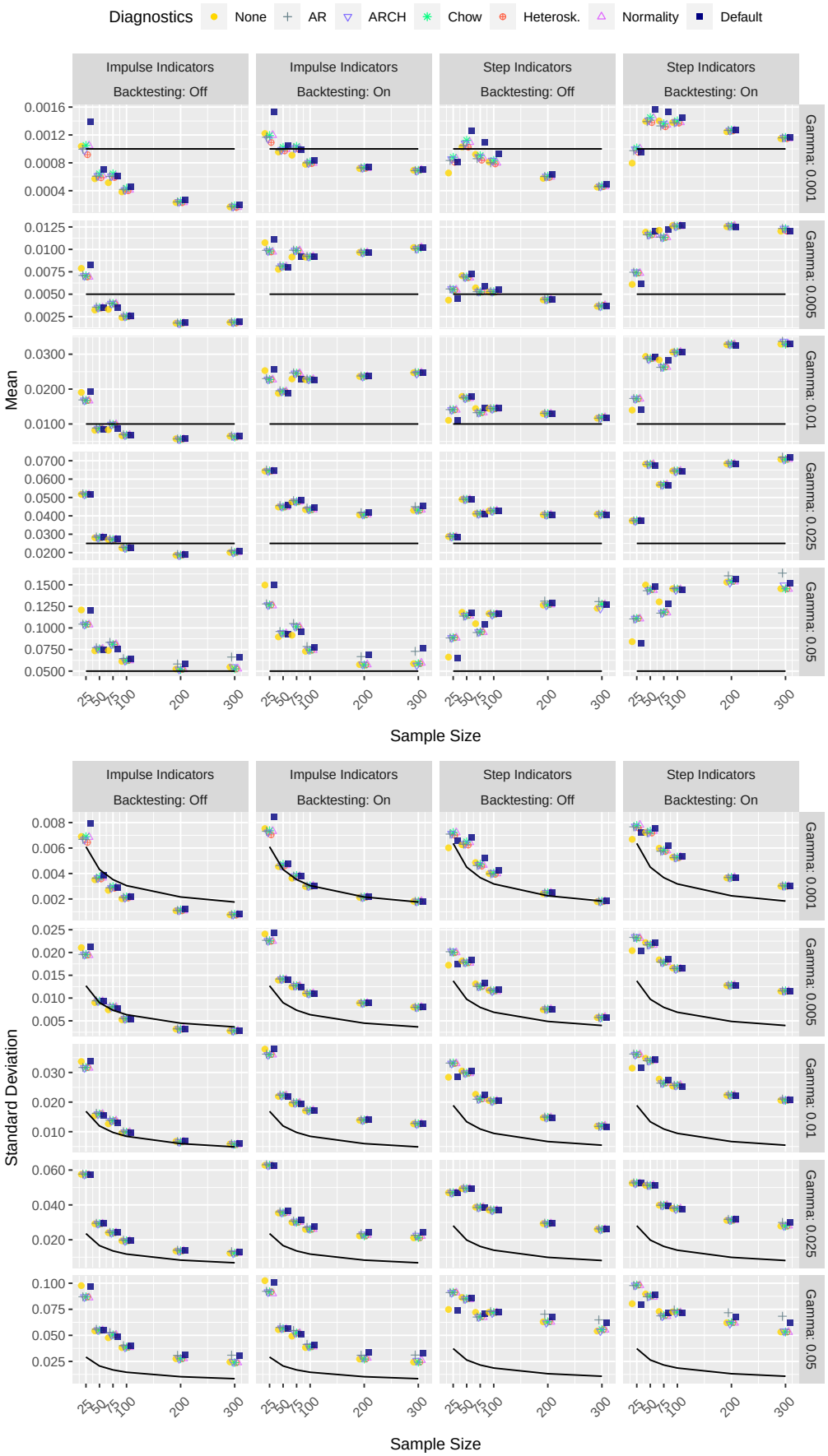
with $\mathsf{E}(\widehat{\omega}_{j,2}^K) = \omega_{j,2}^K$ and $\mathsf{V}(\widehat{\omega}_{j,2}^K) \approx \frac{1}{M}(\omega_{j,4}^K - (\omega_{j,2}^K)^2)$, where the approximation holds for large numbers of replications (Hendry 1984).

In the next sections, we compare the estimated moments from the simulations to their asymptotic counterparts. That is, we assess whether $\widehat{\gamma}_j^K$ is close to γ_j^K and whether $\widehat{\omega}_{j,2}^K$ is close to $\omega_{j,2}^K = \mathsf{avar}(\widehat{\gamma}_j^K)/T_j$. Since the asymptotic theory has been derived for the split-half algorithms without backtesting or diagnostic testing, the values γ_j^K and $\omega_{j,2}^K$ only differ across settings j due to differences in γ and T.

3.4.1.2 Descriptive Statistics of the Gauge

Our simulations show that the gauge is well-controlled for tight significance levels. Backtesting increases the gauge in most cases, especially for relatively loose significance levels. In contrast, the different diagnostic tests as well as their default combination have a negligible effect on the gauge. Figure 3.4.1 plots the mean and the standard deviation of the gauge on the y-axis against the sample size T_j on the x-axis across the 420 different Monte Carlo experiments described in Section 3.4.1. Each row of subplots represents a different significance level γ_j. The first two columns show the results for the impulse indicators with and without backtesting, and the other two columns repeat the results for the step indicators. The different diagnostic settings are represented by different shapes and colours of the markers. The data points are slightly shifted in the x-direction to avoid overlapping and enhance visibility. The black lines indicate the theoretical means γ_j and standard deviations $\sqrt{\omega_{j,2}^K}$.

Figure 3.4.1: Effect of Diagnostic Tests and Backtesting on the Gauge



Diagnostic Testing

We find that the diagnostic tests generally have very little impact on the mean or standard deviation unless the sample size is very small ($T = 25$), and/or when all the default tests are used. Sometimes, using the default diagnostics produces similar results as turning diagnostics completely off. For impulse indicators, this leads to cases where using a single diagnostic test leads to a lower gauge than without any diagnostics at all, whereas for step indicators a single diagnostic test may lead to a higher gauge than the default of using all five tests. Both of these patterns disappear quickly as the sample size increases. The impact of diagnostic testing also does not seem to differ much by whether backtesting is turned on or off.

Backtesting

By comparing adjacent columns, we observe that backtesting increases both the mean and the standard deviation of the gauge for both impulse and step indicators. Except for the tightest significance level $\gamma = 0.001$, the average gauge is closer to the theoretical mean when backtesting is turned off.

For impulse indicators, the average gauge falls as the sample size increases irrespective of backtesting. It approaches the theoretical mean except when γ is very tight, in which case fewer indicators are falsely retained than the theory suggests. For the step indicators, the average gauge only falls with the sample size when backtesting is turned off and the significance level is small enough ($\gamma \leq 0.01$). When γ is too large, the gauge even increases with the sample size before it levels off, suggesting that there is a bias relative to the theory that does not vanish asymptotically. This bias is exacerbated when backtesting is turned on and it is already present for tight significance levels, such as $\gamma = 0.005$.

The standard deviation of the gauge is generally closer to the theoretical value without backtesting than with backtesting, except for impulse indicators with $\gamma = 0.001$. That backtesting increases the standard deviation of the gauge might be driven by the additional variability that is introduced by the initial fitting of a GUM. As explained in Section 3.2, the GUM with super saturation cannot be estimated dir-

ectly because there are many more variables than observations. Therefore, an initial search across indicators is undertaken, which is then used as the GUM to compare subsequent reductions to. If this initial fit happens to retain many indicators, the final selected model will also tend to contain more indicators because backtesting ensures the reduction is valid. If on the other hand the initial fit contains few or no indicators, then the final selected model is unlikely to contain many indicators.

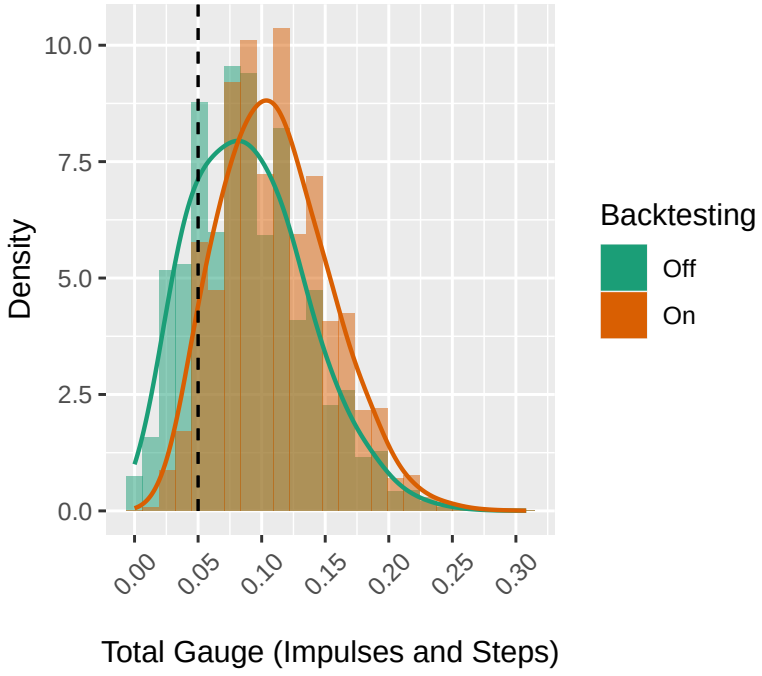
For impulses, the simulated standard deviation is close to the theoretical values when backtesting is off and γ is small. This holds even for small sample sizes of $T = 50$ or $T = 75$. For step indicators, we require slightly tighter significance levels and larger sample sizes for the theory to approximate the simulated values well. For both impulse and step indicators, the difference between the theoretical and simulated standard deviation is quite large for loose significance levels. This entails that individual realisations of the gauge may actually be quite far apart from their mean, making the gauge harder to control.

The fact that the variability of the gauge can be substantial for the larger significance levels is also illustrated in Figure 3.4.2. It shows the histogram and density estimate of the total gauge averaged across impulse and step indicators for $\gamma = 0.05$. Since diagnostic testing makes little difference to the gauge, we only show the results for the default diagnostics. The sample size is $T = 100$.

Even without backtesting, the total gauge is often much larger than the theoretical value indicated by the black, dashed line. It is not uncommon to observe realisations of the gauge that are two or three times as large as its theoretical mean. Backtesting exacerbates this issue, shifting the histogram to the right towards larger realisations. For tight significance levels and without backtesting, this is less of an issue.

In general, indicator saturation methods run the risk of overfitting when applied with loose significance levels. An incorrectly retained indicator in one of the block searches improves the fit of the model, reducing the estimated error variance and thereby making it more likely to falsely retain even more indicators in subsequent blocks. Using both IIS and SIS at the same time increases this risk further because

Figure 3.4.2: Histogram of the Gauge, $\gamma = 0.05$



there are more indicators to be searched over and more block search iterations.

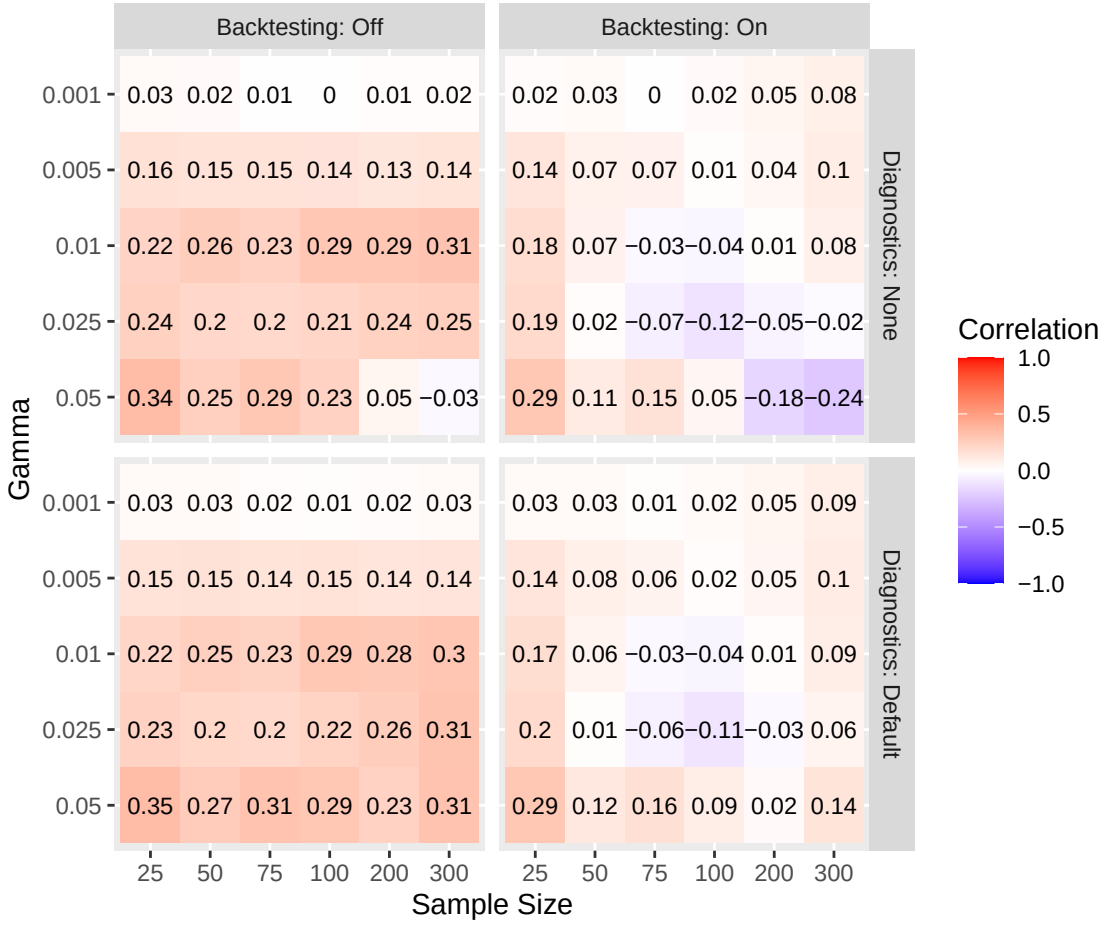
Correlation Between Impulse and Step Gauge

To better understand the joint behaviour of the impulse and step gauge, we assess their correlation. Since the asymptotic theories have been derived for IIS and SIS separately, it is unclear what the correlation structure between the two is when used jointly. On the one hand, we might expect a positive correlation because of the risk of overfitting that was just explained. On the other hand, we might expect a negative correlation because of the substitutability between different types of indicators to model the same phenomenon.

Figure 3.4.3 shows a heat map where each cell represents the correlation between the gauge of impulse and step indicators for a different sample size and significance level. The first row displays the results without diagnostics and the second row with default diagnostics. The first column is without and the second with backtesting.

We find that the majority of correlations is either close to zero or slightly positive. In general, the correlation tends to be weaker for smaller values of γ and close to zero for $\gamma = 0.001$. Comparing the first row to the second row, we find that diagnostic testing again makes little difference to the correlation structure. Dia-

Figure 3.4.3: Correlation Between the Impulse and Step Gauge



gnostic testing only seems to make the correlation more positive for larger values of $\gamma \geq 0.025$. Backtesting tends to decrease the correlation, occasionally making it slightly negative for large significance values. Without backtesting, the correlation is relatively stable across sample sizes with the exception of $\gamma = 0.05$.

3.4.1.3 Response Surface of the Gauge

We estimate response surfaces for the mean and the variance of the gauge for two purposes. First, we want to test how well the asymptotic theory for stylised IIS and SIS approximates the gauge of super saturation. Second, we try to reduce the specificity of the simulation results by finding a model for the simulated values with which we can predict the gauge for combinations of tuning parameters that we have not covered in our experiments. Applied researchers can then use the response surface to calibrate the tuning parameters of Autometrics depending on

their tolerance for the level and variability of the gauge.

We try to find a good approximation for the mean and variance of the gauge across the parameter space Θ that was considered in the Monte Carlo experiments. There exists a general relationship between the specific parameter settings $\theta_j \in \Theta$ and our outcomes of interest, the mean γ_j^K and the variance $\omega_{j,2}^K$ of the gauge for indicator type K. We denote such a relationship by $\mathsf{G}_m^K(\theta_j)$ for indicator type $K \in \{I, S\}$ and moment $m \in \{\gamma, \omega_2\}$. That is, $\gamma_j^K = \mathsf{G}_\gamma^K(\theta_j)$ and $\omega_{j,2}^K = \mathsf{G}_{\omega_2}^K(\theta_j)$.

As argued before, the simulated outcomes from the Monte Carlo simulations are unbiased estimates of the true moments $\mathsf{G}_m^K(\theta_j)$. For instance, the simulated and the true mean can be described by

$$\widehat{\gamma}_j^K = \mathsf{G}_\gamma^K(\theta_j) + u_{\gamma,j}^K, \tag{3.4.7}$$

where $u_{j,\gamma}^K$ represents the Monte Carlo sampling error, which should be mean zero and asymptotically normally distributed with shrinking variance as the number of replications $M \to \infty$. We now approximate the true $\mathsf{G}_\gamma^K(\cdot)$ using functions of the underlying parameters, including the asymptotic theory. This is called the response function $\mathsf{R}_\gamma^K(\cdot)$. We can re-write equation (3.4.7) to obtain

$$\begin{aligned}
\widehat{\gamma}_j^K &= \mathsf{G}_\gamma^K(\theta_j) + u_{\gamma,j}^K && (3.4.8)\\
&= \mathsf{R}_\gamma^K(\theta_j) + [\mathsf{G}_\gamma^K(\theta_j) - \mathsf{R}_\gamma^K(\theta_j) + u_{\gamma,j}^K] && (3.4.9)\\
&= \mathsf{R}_\gamma^K(\theta_j) + v_{\gamma,j}^K, && (3.4.10)
\end{aligned}$$

where $v_{\gamma,j}^K$ is a new error that contains both the sampling error $u_{\gamma,j}^K$ and an approximation error when $\mathsf{R}_\gamma^K(\theta_j) \neq \mathsf{G}_\gamma^K(\theta_j)$. Similarly, we obtain a response surface for the variance as

$$\widehat{\omega}_{j,2}^K = \mathsf{R}_{\omega_2}^K(\theta_j) + v_{\omega_2,j}^K. \tag{3.4.11}$$

Our response surface includes the asymptotic theory values derived for IIS (Johansen and Nielsen 2016b) and SIS (Nielsen and Qian 2023) as regressors. Due to small sample effects for finite T, we also include terms that vanish asymptotically as $T \to \infty$. If the response surface is a good approximation to the true relationship,

$\mathsf{R}(\cdot) \approx \mathsf{G}(\cdot)$, and if the asymptotic theories for stylised IIS and SIS are applicable to super saturation with Autometrics, we obtain three testable implications. First, the coefficient estimates on the asymptotic theory should be close to unity. Second, the intercept should be insignificant because it otherwise represents a bias term that does not vanish asymptotically. Third, the error should be approximately normally distributed provided our choice of $M = 10{,}000$ replications is large enough. Failures of these tests either mean that the response surface is misspecified, or that the asymptotic theory does not extend perfectly to the case of super saturation with a more sophisticated selection algorithm like Autometrics, or both.

We provide response surface estimates for two different samples: (1) the full sample of all 420 Monte Carlo experiments, and (2) a baseline sample that is restricted to the 30 experiments without backtesting or diagnostic testing. The baseline experiments are closest to the asymptotic theory, which has been derived for the stylised split-half algorithm. The regression equations are

$$\widehat{\gamma}_j^K = \beta_0 + \beta_1 \gamma_j^K + \beta_2 \frac{1}{T_j} + \beta_3 \frac{\gamma_j^K}{T_j} + v_{\gamma,j}^K \tag{3.4.12}$$

$$\ln(\widehat{\omega}_{j,2}^K) = \beta_0 + \beta_1 \ln(\omega_{j,2}^K) + \beta_2 \frac{1}{T_j} + \beta_3 \frac{\ln(\omega_{j,2}^K)}{T_j} + v_{\omega_2,j}^K. \tag{3.4.13}$$

In a second exercise, we create many polynomials and interactions between the different parameters and use Autometrics to select a specification of our response surface. The aim is to provide a response surface that practitioners can use to control the mean and variance of the gauge according to their preferences.

For the selection exercise, we add further polynomials of the underlying parameters γ and T, as well as a full set of dummy variables to allow for biases for each of the diagnostic tests and backtesting, interactions of those dummies with the theory variable to allow for slope heterogeneity, and interactions with $1/T$ to allow each of these effects to differ across sample sizes. The full list of regressors can be found in Appendix 3.A. We do not select over the intercept or the theory regressor.

Response Surfaces for the Mean of the Gauge

Table 3.4.1 shows the response surfaces for the mean of the impulse gauge, $\widehat{\gamma}_j^I$.

Table 3.4.1: Response Surface for the Mean of the Impulse Gauge

	Baseline		All	
	Theory	Selection	Theory	Selection
Cons (F)	-0.0041 **	-0.0006	0.0005	-0.0047 ***
	0.0017	0.0004	0.0006	0.0004
γ (F)	0.926 ***	0.264 *	1.076 ***	1.037 ***
	0.065	0.136	0.038	0.027
$1/T$	-0.020		-0.112 ***	
	0.085		0.029	
γ/T	36.908 ***	36.345 ***	35.427 ***	29.379 ***
	3.329	4.315	2.492	1.366
γ^2		12.960 ***		
		2.600		
$1_{(\text{back}=1)}$				0.011 ***
				0.001
$1_{(\text{back}=1)}/T$				-0.283 ***
				0.028
$1_{(\text{back}=1)} \times \gamma/T$				13.729 ***
				2.043
$1_{(\text{AR}=1)} \times \gamma$				0.093 ***
				0.032
Num. Obs.	30	30	420	420
Num. Params.	4	4	4	7
R2 Adj.	0.982	0.993	0.939	0.976
Std.Errors	Standard	Robust	Robust	Robust
Num. Neg. Fitted	6	2	14	49
Heterosk. (p-value)	0.115	0.014 **	0.000 ***	0.000 ***
Normality (p-value)	0.801	0.003 ***	0.000 ***	0.000 ***

Note: Response surface regression estimates. "Baseline" restricts the Monte Carlo experiments to those without backtesting or diagnostic testing, whereas "All" includes all experiments. "Theory" includes the displayed regressors without selection. "Selection" uses Autometrics with normality and heteroskedasticity diagnostics to select among a wide set of regressors with a significance level of 0.05 for the "Baseline" experiments and 0.01 for "All" experiments. The regressors labelled with (F) are fixed and not selected over. Standard errors are reported below the parameter estimates. Significance levels abbreviated as * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

For the baseline case without backtesting or diagnostic testing, the theoretical specification fits well. The coefficient on the theory variable is 0.926, with unity being included in the 95%-confidence interval. The intercept is statistically significant but relatively small compared to the range of the dependent variable considered in the experiments, $[0.001, 0.05]$. The finite sample terms show that the mean of the gauge is larger than the theory predicts for small sample sizes. The theoretical specification can explain 98.2% of the variation in sample but predicts a negative mean gauge in

6 cases, which are all the cases for $\gamma = 0.001$. In practice, such predictions could be replaced with zero or with the theoretical mean γ.[5] We do not reject normality of the error term. Overall, the response surface seems well-specified and could be used to make an informed choice about the significance level γ. The specification chosen by the selection procedure yields an insignificant intercept but now includes the term γ^2 to account for the higher-than-expected gauge for loose significance levels γ that we had observed in Figure 3.4.1.

For the full sample of simulation settings, we find that the theory specification still performs well. The intercept is very small and statistically insignificant, the coefficient on the theory variable is close to unity (borderline rejection at 5% significance level), and the specification still explains more than 90% of the variation in sample. However, the error is not normally distributed. Using model selection across a set of regressors allowing for heterogeneity across the backtesting and diagnostic testing settings, we mostly retain terms allowing for heterogeneity in the backtesting setting: a positive bias term, an additional finite sample correction, and slope heterogeneity that vanishes asymptotically. Note that the finite sample correction term is negative, implying that backtesting increases the mean of the gauge more as the sample size T increases. We also retain a variable modifying the slope of the theory variable when the AR test for autoregressive errors is switched on, yielding a slope of the theory variable of approximately 1.13 compared to 1.04 otherwise.

The response surfaces for the mean of the step gauge are presented in Table 3.4.2. They deviate more from the asymptotic theory than for the impulse indicators. The theory specification in the baseline case estimates the coefficient on the asymptotic theory to be much larger than unity. In small samples, the coefficient is effectively smaller because of the negative coefficient on the interaction term γ/T but is still larger than unity for, e.g., a sample size of $T = 25$ with a coefficient of 1.45 on γ. The model predicts the six simulations with $\gamma = 0.001$ to have a negative mean of the gauge, which could be set to γ in practice when choosing a significance level for selection. The normality test rejects at the 5% but not the 1% significance level.

5. Alternatively, a censored regression model (for example, Tobin 1958) could be used to avoid the problem of negative predicted values.

Table 3.4.2: Response Surface for the Mean of the Step Gauge

	Baseline		All	
	Theory	Selection	Theory	Selection
Cons (F)	-0.0100 ***	-0.0004	-0.0052 ***	-0.0074 ***
	0.0031	0.0008	0.0008	0.0010
γ (F)	2.714 ***	1.112 ***	2.875 ***	1.837 ***
	0.120	0.221	0.056	0.075
$1/T$	0.212		0.048	0.537 ***
	0.155		0.034	0.116
γ/T	-31.685 ***		-24.857 ***	-24.857 ***
	6.093		2.909	1.950
γ^2		30.746 ***		15.521 ***
		4.417		1.318
γ^2/T		-600.549 **		
		291.347		
$(1/T)^2$				-8.127 ***
				2.389
$1_{(\text{back}=1)}$				0.009 ***
				0.001
$1_{(\text{back}=1)}/T$				-0.264 ***
				0.045
$1_{(\text{back}=1)} \times \gamma$				0.468 ***
				0.049
Num. Obs.	30	30	420	420
Num. Params.	4	4	4	9
R2 Adj.	0.969	0.988	0.948	0.984
Std.Errors	Standard	Robust	Robust	Robust
Num. Neg. Fitted	6	0	84	28
Heterosk. (p-value)	0.389	0.017 **	0.000 ***	0.000 ***
Normality (p-value)	0.034 **	0.001 ***	0.000 ***	0.000 ***

Note: Response surface regression estimates. "Baseline" restricts the Monte Carlo experiments to those without backtesting or diagnostic testing, whereas "All" includes all experiments. "Theory" includes the displayed regressors without selection. "Selection" uses Autometrics with normality and heteroskedasticity diagnostics to select among a wide set of regressors with a significance level of 0.05 for the "Baseline" experiments and 0.01 for "All" experiments. The regressors labelled with (F) are fixed and not selected over. Standard errors are reported below the parameter estimates. Significance levels abbreviated as * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

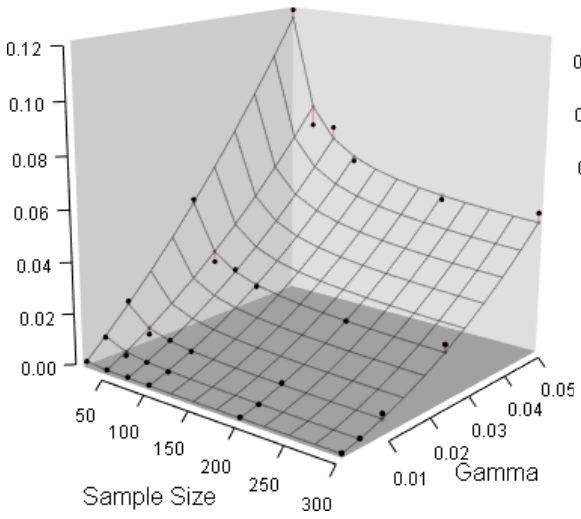
Selecting over higher polynomials of γ yields a specification that includes the square of the theory variable, as was the case for the impulse indicators. This allows for values of the mean gauge that are higher than what the theory predicts. The negative coefficient on the additional term γ^2/T allows the gauge to increase asymptotically, which was observed for larger values of γ. Due to the square of γ, these terms only have a small influence on the mean gauge for small values of γ. Hence, for small values of γ, the asymptotic theory seems to be a good approximation with a coefficient close to unity, which is also contained in the 95%-confidence interval. The constant is small in magnitude and statistically insignificant. The selected model explains above 95% of the variation across our experiments.

Extending the response surface to the whole set of Monte Carlo simulations, we obtain similar estimates in the theory specification as in the baseline case. The selection also retains the term γ^2 but not γ^2/T. None of the terms with dummy variables for any of the diagnostic tests are retained. There is evidence of bias from backtesting, which is slightly negative for sample sizes below 30 and then increases to 0.009 asymptotically. In addition, the slope coefficient on the theory variable is larger by almost 25% when backtesting is turned on.

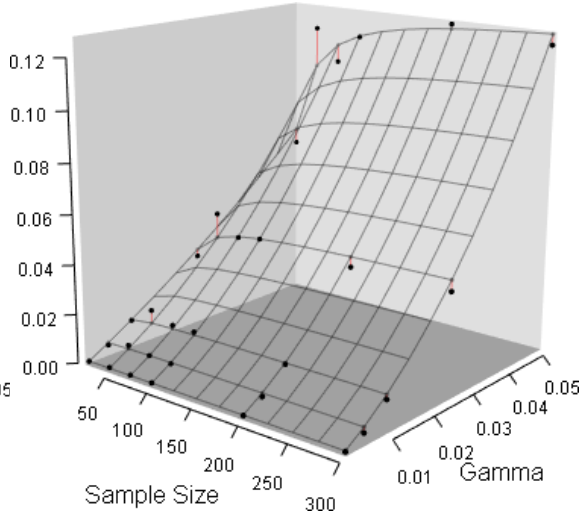
Figure 3.4.4 shows the selection baseline response surfaces for the mean of the impulse and step gauge, where negative predictions have been replaced with zero.

Figure 3.4.4: Response Surfaces for the Mean of the Gauge

(a) Impulse Gauge



(b) Step Gauge



The grid represents the response surface while the black points are the actual simulated values. We see that the predictions work very well for tight significance levels of $\gamma \leq 0.01$, even for relatively small sample sizes of $T \leq 75$. For larger significance levels, the approximation seems still reasonable for impulse indicators but more erratic for step indicators. Figure 3.B.1 in Appendix 3.B shows the response surfaces that include all simulation settings.

Response Surfaces for the Variance of the Gauge

To model the variances of the impulse and step gauge, we transform them using the natural logarithm to ensure the predicted variances are always positive. The results for the impulse gauge are presented in Table 3.4.3. In the baseline theory specification, only the intercept and the theory regressor are highly significant. The finite sample terms are both insignificant. When adding additional polynomial terms in the search, we retain two terms that allow the effect of the theory variable to vary by the significance level and a bias term that vanishes at rate T^2. In both the theoretical and the selection specification, the intercept is highly significant and large in magnitude, suggesting that the true, simulated variance of the gauge is often much higher than what the theory predicts. This phenomenon was also visible in Figure 3.4.1. Nevertheless, both baseline specifications explain a large share of the variation in the logarithm of the variance of the impulse gauge and we cannot reject normality of the errors. Figure 3.4.5a shows the response surface converted back to levels. Visually, the fit is good, especially for relatively tight levels of $\gamma \leq 0.025$ and moderate to large sample sizes of $T \geq 75$.

Extending the analysis to all simulation settings, we find that the theoretical specification fits similarly well as in the baseline case. The selection now yields a specification where the effect of the theory variable not only depends on γ but also γ^2, allowing both of these effects to differ in finite samples. For small values of γ, these terms are again negligible. Ignoring the finite sample corrections, we obtain sensible coefficients on the theory variable for the values of γ in our analysed parameter space. For example, for $\gamma = 0.001$, the exponent of ω_2^I is effectively 1.63 and for $\gamma = 0.05$ it

Table 3.4.3: Response Surface for the Variance of the Impulse Gauge

	Baseline		All	
	Theory	Selection	Theory	Selection
Cons (F)	8.6915 ***	3.6552 ***	7.5869 ***	6.6457 ***
	1.0976	0.5245	0.3796	0.4924
$\ln(\omega_2^I)$ (F)	1.809 ***	1.403 ***	1.637 ***	1.641 ***
	0.108	0.047	0.040	0.038
$1/T$	-22.950		-0.311	-38.745 ***
	50.665		14.826	9.310
$\ln(\omega_2^I)/T$	-0.312		3.283 *	
	5.607		1.733	
$(1/T)^2$		497.707 ***		1124.670 ***
		97.969		146.658
$\gamma \times \ln(\omega_2^I)$		-5.034 ***		-6.134 ***
		0.401		0.738
$\gamma \times \ln(\omega_2^I)/T$		79.173 ***		281.061 ***
		20.964		26.431
$\gamma^2 \times \ln(\omega_2^I)$				61.972 ***
				11.522
$\gamma^2 \times \ln(\omega_2^I)/T$				-3897.397 ***
				479.555
$1_{(\text{back}=1)}$				-2.659 ***
				0.325
$1_{(\text{back}=1)}/T$				80.505 ***
				11.985
$1_{(\text{back}=1)} \times \ln(\omega_2^I)$				-0.377 ***
				0.033
$1_{(\text{back}=1)} \times \ln(\omega_2^I)/T$				10.565 ***
				1.309
Num. Obs.	30	30	420	420
Num. Params.	4	5	4	12
R2 Adj.	0.960	0.994	0.925	0.987
Std.Errors	Standard	Standard	Robust	Robust
Heterosk. (p-value)	0.252	0.337	0.000 ***	0.000 ***
Normality (p-value)	0.329	0.220	0.135	0.237

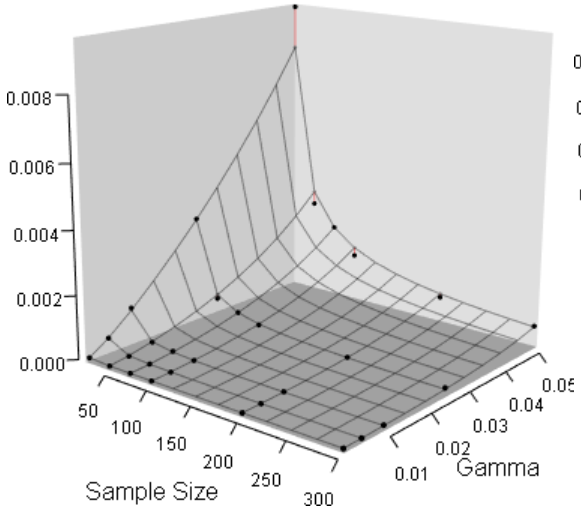
Note: Response surface regression estimates. "Baseline" restricts the Monte Carlo experiments to those without backtesting or diagnostic testing, whereas "All" includes all experiments. "Theory" includes the displayed regressors without selection. "Selection" uses Autometrics with normality and heteroskedasticity diagnostics to select among a wide set of regressors with a significance level of 0.05 for the "Baseline" experiments and 0.01 for "All" experiments. The regressors labelled with (F) are fixed and not selected over. Standard errors are reported below the parameter estimates. Significance levels abbreviated as * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

is 1.49. While none of the diagnostic testing terms are retained, we retained several terms related to backtesting. Backtesting asymptotically increases the variance of the impulse gauge. Both the theoretical and the selection specification fail to reject normality of the errors.

The results are similar for the response surfaces of the variance of the step gauge in Table 3.4.4. In all specifications, we have highly statistically significant intercepts of large magnitudes and the coefficients on the theory regressor are significantly larger than unity. The selected models for the impulse and step indicators are very similar. The same regressors are selected in the baseline sample and the only difference in the full sample is that $(1/T)^2$ is not retained for the variance of the step gauge. The effect of the theory regressor again varies by the significance level γ and its square. As was the case for the impulses, backtesting again asymptotically increases the variance of the gauge. The normality test is only clearly insignificant for the baseline sample with selection. The corresponding response surface in levels is shown in Figure 3.4.5b. Compared to the other response surfaces, we can clearly see that it is much less smooth and the residuals are considerably larger for values of $\gamma \geq 0.025$, even for moderate sample sizes of $T = 100$. For tight significance levels and medium sample sizes, however, the approximation seems to perform well.

Figure 3.4.5: Response Surfaces for the Variance of the Gauge

(a) Impulse Gauge



(b) Step Gauge

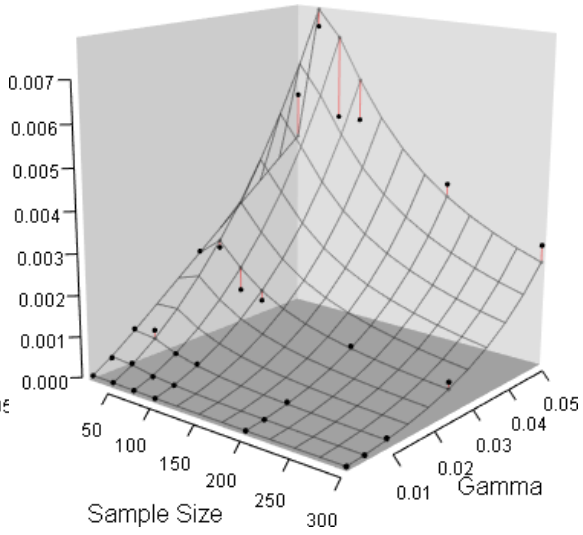


Table 3.4.4: Response Surface for the Variance of the Step Gauge

	Baseline		All	
	Theory	Selection	Theory	Selection
Cons (F)	8.8794 ***	4.3903 ***	7.4565 ***	6.0401 ***
	0.6352	0.4932	0.2393	0.2274
$\ln(\omega_2^S)$ (F)	1.701 ***	1.340 ***	1.533 ***	1.463 ***
	0.064	0.045	0.025	0.019
$1/T$	-122.334 ***		-77.080 ***	60.998 ***
	29.236		9.081	14.486
$\ln(\omega_2^S)/T$	-7.129 **		-2.527 **	8.563 ***
	3.316		1.021	1.483
$(1/T)^2$		-559.520 ***		
		94.134		
$\gamma \times \ln(\omega_2^S)$		-4.302 ***		-4.798 ***
		0.429		0.496
$\gamma \times \ln(\omega_2^S)/T$		113.787 ***		497.677 ***
		21.825		40.327
$\gamma^2 \times \ln(\omega_2^S)$				30.179 ***
				8.577
$\gamma^2 \times \ln(\omega_2^S)/T$				-5796.684 ***
				569.854
$1_{(\text{back}=1)}$				-2.284 ***
				0.247
$1_{(\text{back}=1)}$ / T				69.556 ***
				10.425
$1_{(\text{back}=1)} \times \ln(\omega_2^S)$				-0.289 ***
				0.025
$1_{(\text{back}=1)} \times \ln(\omega_2^S)/T$				8.129 ***
				1.180
Num. Obs.	30	30	420	420
Num. Params.	4	5	4	12
R2 Adj.	0.979	0.993	0.958	0.989
Std.Errors	Standard	Standard	Robust	Robust
Heterosk. (p-value)	0.074 *	0.668	0.000 ***	0.000 ***
Normality (p-value)	0.057 *	0.895	0.000 ***	0.055 *

Note: Response surface regression estimates. "Baseline" restricts the Monte Carlo experiments to those without backtesting or diagnostic testing, whereas "All" includes all experiments. "Theory" includes the displayed regressors without selection. "Selection" uses Autometrics with normality and heteroskedasticity diagnostics to select among a wide set of regressors with a significance level of 0.05 for the "Baseline" experiments and 0.01 for "All" experiments. The regressors labelled with (F) are fixed and not selected over. Standard errors are reported below the parameter estimates. Significance levels abbreviated as * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

3.4.1.4 Guidance for Controlling the Gauge

Our recommendations are based on the following insights. First, the mean of both the impulse and the step gauge can be controlled by choosing tight significance levels of at most $\gamma = 0.01$, yielding results close to or even below what the theory predicts. Small sample effects vanish quickly and are negligible for sample sizes of $T \geq 100$. For loose significance levels, the gauge may be well above its theoretical mean, especially for step indicators. Backtesting increases the mean of the gauge, even for relatively tight significance levels. For sample sizes of $T \geq 50$, the effect of diagnostic testing is negligible under the null and therefore relatively costless in well-specified models, while yielding potential improvements under the alternative. Only the AR test was found to significantly increase the mean of the impulse gauge.

Second, the variance of the impulse and step gauge is close to their theoretical values when the chosen significance level is tight and when backtesting is turned off. The theoretical approximation is closer for impulse than for step indicators. Backtesting increases the variability in the gauge considerably. None of the variables allowing for biases or slope heterogeneity for diagnostic testing were statistically significant, again confirming the limited role of diagnostic testing for the gauge when the model is well-specified.

As explained in Section 3.2, the use of backtesting is controversial with super saturation. Since the GUM cannot be estimated, it becomes unclear against which initial model the reductions should be tested. Its theoretical justification is therefore unclear and the simulations have shown that backtesting might lead to overfitting when used with super saturation.

In terms of theoretical considerations, the advantage of choosing a tight significance level γ is that the false retention of indicators is inefficient. Dummying out an observations using impulse indicators effectively means losing this observation for the estimation of the other parameters. For step indicators, the problem can be even worse. Nielsen and Qian (2023) provide simulation evidence that the false retention of step indicators leads to biased estimates of covariates included in the model if they

are not strictly exogenous. This affects, for example, lagged dependent variables. The bias increases in the selection significance level γ. They conjecture that this bias might not vanish asymptotically for a fixed significance level γ. Increasing the sample size means that we also expect to falsely retain more step indicators, leaving the effective sample size for estimating each "block" with a different intercept approximately constant ($\gamma \times T$ falsely retained step indicators, yielding "blocks" of lengths of approximately $1/\gamma$ even as $T \to \infty$). Their explicit recommendation is to use $\gamma = 0.01$ for a sample size of $T = 100$ and choosing tighter levels as $T \to \infty$, targeting a number instead of a share of falsely retained step indicators.

Because of these theoretical considerations and the observed patterns in our simulations, we recommend choosing a tight significance level of at most 1% and turning backtesting off during indicator selection. Under the null of no outliers or location shifts, this yields an expected value of the impulse and step gauge close to what the theory predicts. Moreover, this limits the variability of the gauge such that it is unlikely to obtain values of the gauge that are far from its mean. For efficiency reasons and to limit the bias induced by falsely retained step indicators (Nielsen and Qian 2023), one could choose even tighter significance levels, for example using a rule of $\min(0.01, 1/T)$. The response surfaces fit the simulated moments well for such tight significance levels and when backtesting is turned off, so they can be used to guide the choice of significance level in finite samples.

The recommendation to turn backtesting off deviates from the recommendation in Doornik (2008), which — inter alia — assessed the impact of backtesting on the gauge and potency. The simulations showed that turning backtesting off reduces the gauge below its expected level and also decreases the potency. Instead, using backtesting resulted in a gauge closer to the target while also improving potency. However, those simulation settings did not cover indicator saturation methods, only covariate selection. Our simulations have shown that for super saturation, the gauge behaves better when backtesting is turned off. If the user prefers to keep backtesting on, we recommend choosing an even tighter significance level of at most 0.5%. Also, note that the selection of indicators and of the theory model is recommended to

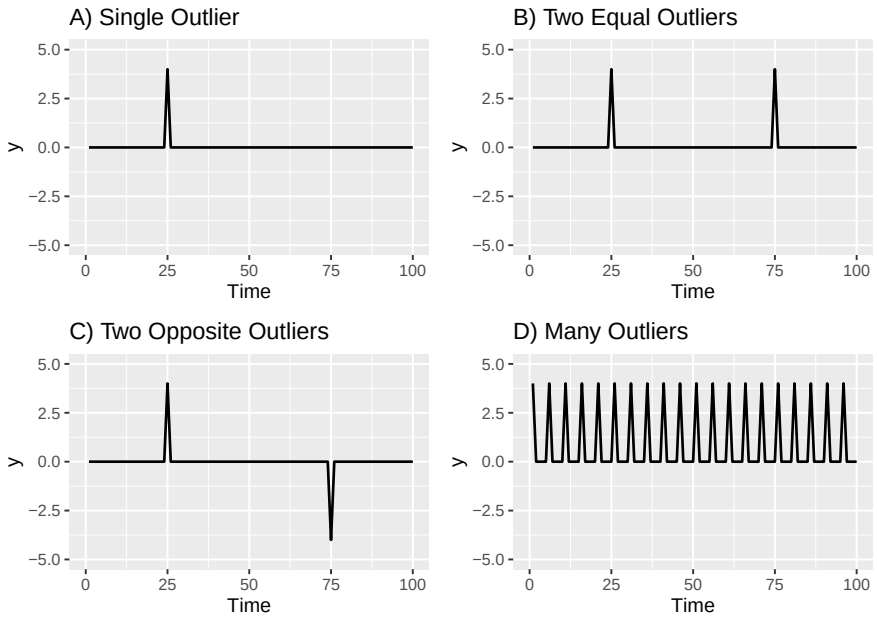
be done in two consecutive steps (Hendry and Johansen 2015; Castle and Hendry 2019). This allows for different settings of backtesting for the two steps. One could turn backtesting off when selecting indicators while retaining the theory covariates, and then turn backtesting on when selecting the theory covariates while retaining the previously selected indicators.

3.4.2 DGP with Outliers

With our recommended tuning parameters of Autometrics, super saturation reaches a high detection rate of outliers, as long as outliers do not occur too often. The step gauge tends to be higher than the impulse gauge but they are on average close to the chosen significance level.

We first assess the performance of super saturation for DGPs with only outliers, which are illustrated in Figure 3.4.6. The outliers are implemented as deterministic one-off impulse dummies of large magnitude ($3\sigma_\epsilon$ and $4\sigma_\epsilon$) relative to the normal distribution of the errors ϵ_t. Following our recommendation from the previous section, we use Autometrics with a significance level $\gamma = 0.01$, no backtesting, and default diagnostic tests. We will use the new gauge and potency measures that we suggested in Section 3.3.3 to evaluate the performance of super saturation.

Figure 3.4.6: Stylised Illustrations of the DGPs with Outliers



We first analyse performance in a DGP with a single outlier

$$y_t = \beta_0 + \lambda 1_{(t=s)} + \epsilon_t. \tag{3.4.14}$$

As a reference that represents doing inference, we can compute the power of a t-test on an impulse indicator that exactly matches the outlier. A t-test for $H_0 : \lambda = 0$ in DGP (3.4.14) has an approximate non-centrality of λ/σ_ϵ, resulting in a power of 0.66 for $\lambda = 3\sigma_\epsilon$ and 0.92 for $\lambda = 4\sigma_\epsilon$. We can compare the power to the achieved potency, which includes the additional cost of searching for outliers across the whole sample.

Table 3.4.5 shows the potency of super saturation and compares it to using IIS only. We vary when the outlier occurs to assess whether the potency is sensitive to the timing.

We find that IIS reaches a potency that is close to the theoretical power for a

Table 3.4.5: Single Outlier

		$\lambda = 3$				$\lambda = 4$			
		Potency		Gauge		Potency		Gauge	
s	Type	Super	IIS	Super	IIS	Super	IIS	Super	IIS
5	I	0.439	0.635	0.007	0.011	0.701	0.895	0.007	0.011
5	S	0.093	–	0.014	–	0.106	–	0.014	–
5	total	0.532	0.635	0.011	0.011	0.807	0.895	0.010	0.011
25	I	0.198	0.643	0.006	0.011	0.182	0.898	0.006	0.011
25	S	0.331	–	0.015	–	0.607	–	0.014	–
25	total	0.529	0.643	0.011	0.011	0.789	0.898	0.010	0.011
50	I	0.101	0.636	0.006	0.011	0.079	0.904	0.006	0.011
50	S	0.413	–	0.015	–	0.708	–	0.015	–
50	total	0.514	0.636	0.011	0.011	0.788	0.904	0.011	0.011
75	I	0.177	0.629	0.007	0.011	0.157	0.896	0.007	0.011
75	S	0.350	–	0.015	–	0.630	–	0.015	–
75	total	0.527	0.629	0.011	0.011	0.787	0.896	0.011	0.011
95	I	0.071	0.648	0.007	0.011	0.058	0.902	0.007	0.011
95	S	0.405	–	0.015	–	0.688	–	0.014	–
95	total	0.475	0.648	0.011	0.011	0.746	0.902	0.011	0.011

Note: Monte Carlo experiment with a single outlier of magnitude λ in period s. Columns "Super" refer to super saturation and columns "IIS" refer to impulse indicator saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance level $\gamma = 0.01$, backtesting off, default diagnostic testing, and tolerance for detecting outliers with step shifts $\delta^I = 1$.

known outlier, hence the search cost is small. The potency of super saturation is lower by approximately 10 percentage points for both outlier magnitudes λ. This is in part due to the strict criteria for counting an outlier as detected by two consecutive step indicators only if the coefficients sum to approximately 0 with a tolerance of only $\delta^I = 1$. Empirically, the detection of outliers using step indicators is very common and underlines why a phenomenon-based adaption of the potency measure is required. Detection via step indicators is highest towards the middle of the sample. This is presumably the case because two consecutive step indicators whose coefficients do not perfectly sum to zero also improve the fit more generally because it allows for slightly different means before and after. This improved fit has the strongest effect in the middle of the sample. For IIS, the potency is independent of when the outlier occurs while it tends to be slightly less stable across time when using super saturation. For both IIS and super saturation, the gauge is well-controlled, yielding an average gauge across both impulse and step indicators close to the theoretical value of $\gamma = 0.01$.

The next DGP shown in equation (3.4.15) contains one outlier in each half of the sample, which we choose to be at time periods $T_1 = 25$ and $T_2 = 75$. We consider two experiments where the outliers are of equal magnitude but either have the same ($\lambda_1 = \lambda_2$) or opposite sign ($\lambda_1 = -\lambda_2$).

$$y_t = \beta_0 + \lambda_1 1_{(t=T_1)} + \lambda_2 1_{(t=T_2)} + \epsilon_t. \tag{3.4.15}$$

The results in Table 3.4.6 show slightly lower potencies than before, which is expected because the second outlier causes the estimated error variance to increase, in turn making it harder to detect them. As before, the detection of outliers via step indicators is practically very relevant and the average gauge is still well-controlled.

Finally, we consider a DGP with 20 evenly spread outliers of magnitude $\lambda = 4\sigma_\epsilon$, which can be alternatively interpreted as unmodelled seasonality,

$$y_t = \beta_0 + \lambda \sum_{j=0}^{19} 1_{(t=5j+1)} + \epsilon_t. \tag{3.4.16}$$

Their abundance makes it hard for the algorithm to detect them, which is confirmed by the potency results in Table 3.4.7.

Table 3.4.6: Two Outliers, Equal and Opposite

			Potency $T_1 = 25$		Potency $T_2 = 75$		Gauge	
Experiment	$\|\lambda_1\|$	Type	Super	IIS	Super	IIS	Super	IIS
$\lambda_1 = \lambda_2$	3	I	0.187	0.599	0.208	0.592	0.006	0.010
	3	S	0.324	–	0.314	–	0.015	–
	3	total	0.510	0.599	0.522	0.592	0.011	0.010
$\lambda_1 = \lambda_2$	4	I	0.184	0.861	0.270	0.857	0.006	0.010
	4	S	0.604	–	0.530	–	0.014	–
	4	total	0.789	0.861	0.800	0.857	0.010	0.010
$\lambda_1 = -\lambda_2$	3	I	0.184	0.606	0.210	0.620	0.006	0.010
	3	S	0.331	–	0.326	–	0.015	–
	3	total	0.515	0.606	0.536	0.620	0.011	0.010
$\lambda_1 = -\lambda_2$	4	I	0.181	0.862	0.265	0.868	0.006	0.010
	4	S	0.608	–	0.548	–	0.014	–
	4	total	0.788	0.862	0.813	0.868	0.010	0.010

Note: Monte Carlo experiment with two equal or opposite outliers of magnitude $|\lambda_1|$ in periods $T_1 = 25$ and $T_2 = 75$. Columns "Super" refer to super saturation and columns "IIS" refer to impulse indicator saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance level $\gamma = 0.01$, backtesting off, default diagnostic testing, and tolerance for detecting outliers with step shifts $\delta^I = 1$.

In our usual setting without backtesting, only about 25% of the outliers are detected using super saturation or IIS. The average gauge is slightly lower than the expected $\gamma = 0.01$, because the outliers mask as a higher error variance, making it not only harder to detect the true outliers but also less likely to falsely retain indicators. Table 3.4.7 also shows the results with backtesting, which only yields

Table 3.4.7: Many Outliers

		Potency		Gauge	
Backtesting	Type	Super	IIS	Super	IIS
Off	I	0.166	0.259	0.004	0.009
Off	S	0.073	–	0.010	–
Off	total	0.239	0.259	0.007	0.009
On	I	0.193	0.319	0.004	0.009
On	S	0.082	–	0.011	–
On	total	0.276	0.319	0.008	0.009

Note: Monte Carlo experiment with 19 outliers of magnitude $\lambda = 4\sigma_\epsilon^2$. Columns "Super" refer to super saturation and columns "IIS" refer to impulse indicator saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance level $\gamma = 0.01$, backtesting on and off, default diagnostic testing, and tolerance for detecting outliers with step shifts $\delta^I = 1$.

slightly higher potency.

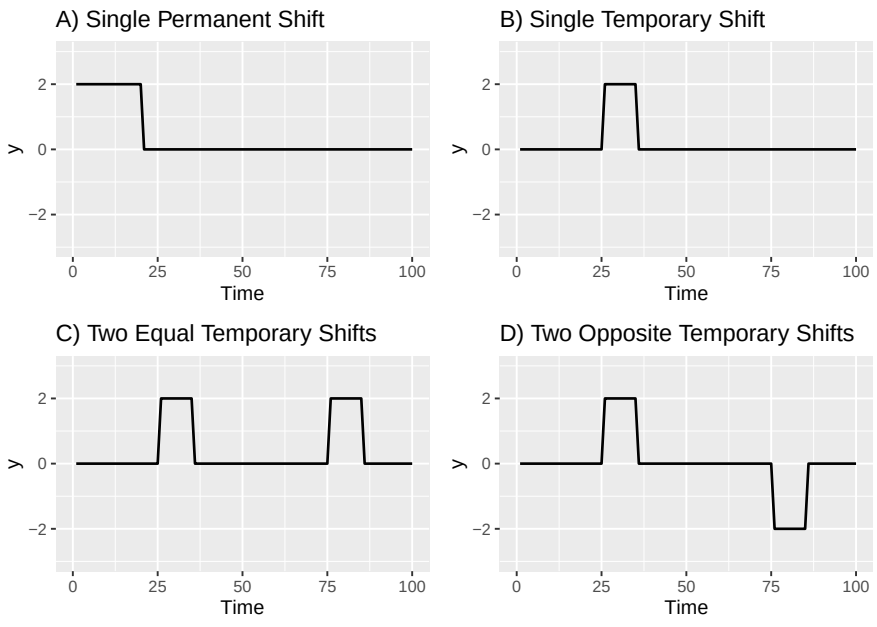
3.4.3 DGP with Location Shifts

Using our recommendations for the tuning parameters of Autometrics, super saturation generally reaches a high detection rate of location shifts. However, we find evidence of drops in the detection rate for specific time periods, presumably an artefact of the algorithmic nature of Autometrics. For small step sizes, the step gauge exceeds the chosen significance level, partly due to mistimed indicators capturing the location shift in its vicinity rather than precisely at its occurrence.

The different settings with location shifts are illustrated in Figure 3.4.7.

As for the outliers, we can compare the potency when searching for a location shift to the power of a t-test when doing inference on a single matching step indicator. The formula for the non-centrality is derived in Castle et al. (2015), who show that it depends on where in the sample the shift occurs. The non-centrality is smaller for shifts occurring towards the beginning or the end of the sample. For a shift in time period 5 of 100, a shift magnitude of $\lambda = 2\sigma_\epsilon$, and a significance level of $\gamma = 0.01$, the power of the t-test is already 0.96. In the simulations, we choose shift magnitudes of $\lambda = 2\sigma_\epsilon$ and $\lambda = 4\sigma_\epsilon$, which yield theoretical powers of virtually unity even for

Figure 3.4.7: Stylised Illustrations of the DGPs with Location Shifts



short shift lengths of $s = 5$.

First, we consider a DGP with a single location shift that lasts until time s, as shown in equation (3.4.17)

$$y_t = \beta_0 + \lambda 1_{(t \leq s)} + \epsilon_t. \tag{3.4.17}$$

We vary the timing s of the location shift to examine how it affects potency. Table 3.4.8 shows that for the smaller shift magnitude, potency is generally high (> 0.5) but substantially lower than the theoretical power. However, for the larger shift magnitude, potency is consistently above 0.9 and therefore close to the corresponding power of unity.

There is almost no difference in potency between super saturation and SIS —

Table 3.4.8: Single Permanent Shift

		$\lambda = 2$				$\lambda = 4$			
		Potency		Gauge		Potency		Gauge	
s	Type	Super	SIS	Super	SIS	Super	SIS	Super	SIS
5	S	0.461	0.544	0.018	0.022	0.915	0.934	0.013	0.015
5	I	0.002	–	0.008	–	0.002	–	0.007	–
5	total	0.463	0.544	0.013	0.022	0.918	0.934	0.010	0.015
10	S	0.525	0.559	0.018	0.022	0.921	0.935	0.012	0.014
10	I	0.002	–	0.007	–	0.003	–	0.006	–
10	total	0.526	0.559	0.013	0.022	0.923	0.935	0.009	0.014
15	S	0.594	0.568	0.019	0.022	0.936	0.932	0.013	0.015
15	I	0.002	–	0.007	–	0.002	–	0.007	–
15	total	0.596	0.568	0.013	0.022	0.938	0.932	0.010	0.015
20	S	0.536	0.608	0.019	0.023	0.925	0.951	0.010	0.017
20	I	0.001	–	0.007	–	0.002	–	0.005	–
20	total	0.538	0.608	0.013	0.023	0.926	0.951	0.008	0.017
25	S	0.557	0.569	0.018	0.022	0.924	0.932	0.013	0.017
25	I	0.002	–	0.007	–	0.003	–	0.006	–
25	total	0.559	0.569	0.012	0.022	0.926	0.932	0.010	0.017
35	S	0.605	0.567	0.019	0.022	0.948	0.934	0.014	0.017
35	I	0.002	–	0.007	–	0.001	–	0.006	–
35	total	0.607	0.567	0.013	0.022	0.949	0.934	0.010	0.017

Note: Monte Carlo experiment with a single step shift of magnitude λ ending in period s. Columns "Super" refer to super saturation and columns "SIS" refer to step indicator saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance level $\gamma = 0.01$, backtesting off, default diagnostic testing, and tolerance for detecting step shifts with two consecutive step and impulse indicators $\delta^S = 2$.

sometimes one is better than the other and vice versa. Only for the small shift magnitude at time $s = 5$ does the potency of super saturation seem considerably lower than that of SIS. Even though the power increases towards the middle of the sample, we do not find such a pattern for the potency, which does not systematically differ across shift timings. For super saturation, most location shifts are indeed modelled by a single matching step indicator instead of a pair of one step and one impulse indicator. The latter combination plays a negligible role in capturing the location shifts.

Next, we consider potential mistiming of step indicators, i.e. that the indicators do not exactly match the actual shift time but are retained in its vicinity. In practice, error draws that are approximately the size of the shift in absolute value may make the shift seem longer or shorter in the data than it actually is in the DGP. Moreover, there might be algorithmic reasons for the mistiming, such as a coarse block search leading to the retention of a nearby step indicator. Table 3.4.9 shows the potency allowing for a mistiming of up to 3 time periods before or after the actual shift. Since potency was already almost unity for $\lambda = 4\sigma_\epsilon$, the results are only shown for $\lambda = 2\sigma_\epsilon$.

We find that a tolerance of a single time period already increases potency by approximately 0.15 to 0.2 across all different lengths. The absolute increase is similar for super saturation and SIS. With a tolerance of three time periods, we reach very

Table 3.4.9: Single Permanent Shift Tolerating Misspecified Timing

	Super: Retained Step Indicator				SIS: Retained Step Indicator			
s	s	$s \pm 1$	$s \pm 2$	$s \pm 3$	s	$s \pm 1$	$s \pm 2$	$s \pm 3$
5	0.461	0.627	0.709	0.753	0.544	0.735	0.824	0.869
10	0.525	0.714	0.815	0.872	0.559	0.767	0.868	0.921
15	0.594	0.829	0.904	0.942	0.568	0.778	0.873	0.926
20	0.536	0.729	0.824	0.880	0.608	0.837	0.918	0.953
25	0.557	0.759	0.856	0.911	0.569	0.770	0.867	0.920
35	0.605	0.840	0.913	0.952	0.567	0.766	0.863	0.919

Note: Monte Carlo experiment with a single step shift of magnitude $\lambda = 2$ ending in period s. Columns "Super" refer to super saturation and columns "SIS" refer to step indicator saturation. $s \pm x$ reports whether a step indicator has been retained within x time periods of the true shift time s . Settings: M = 10,000 replications, significance level $\gamma = 0.01$, backtesting off, default diagnostic testing, and tolerance for detecting step shifts with two consecutive step and impulse indicators $\delta^S = 2$.

high potencies, usually well above 80% except for the shortest shift length $s = 5$. This is reassuring because even a mistimed step indicator can mitigate much of the negative impacts that unmodelled location shifts may cause.

The next DGP shown in equation (3.4.18) represents a temporary shift, which might be harder to detect. The shift begins at time period T_1 and lasts until T_2. We will consider two versions of this DGP

$$y_t = \beta_0 + \lambda(1_{(t \leq T_2)} - 1_{(t \leq T_1)}) + \epsilon_t. \tag{3.4.18}$$

First, we consider the case where the temporary shift is fully contained in one half of the sample, i.e. $T_1 < T_2 < T/2$. Following Castle et al. (2015), we choose $T_1 = 25$ and $T_2 = 35$. The results are presented in Table 3.4.10.

For SIS, we find a potency of the two step indicators modelling the temporary shift that is similar to the potency of a single, permanent shift. Moreover, the potency is similar for both location shifts. In contrast, super saturation yields much lower potency for the second location shift for $\lambda = 2\sigma_\epsilon$ — both compared to SIS and to its previous performance of capturing a single location shift at time $s = 35$. For the second shift, we also find that the location shift is often not modelled using a single matching step indicator but a combination of one step and one impulse indicator, which we allowed for in our new definition of potency in Section 3.3.3. For SIS, the gauge is slightly higher than the targeted 1%, while the average gauge

Table 3.4.10: Single Temporary Shift

		Potency T_1		Potency T_2		Gauge	
λ	Type	Super	SIS	Super	SIS	Super	SIS
2	I	0.002	–	0.108	–	0.008	–
2	S	0.520	0.571	0.193	0.564	0.023	0.024
2	total	0.522	0.571	0.301	0.564	0.015	0.024
4	I	0.004	–	0.658	–	0.007	–
4	S	0.911	0.934	0.191	0.932	0.012	0.014
4	total	0.915	0.934	0.849	0.932	0.010	0.014

Note: Monte Carlo experiment with a single temporary shift of magnitude λ starting in period $T_1 = 25$ and ending in $T_2 = 35$. Columns "Super" refer to super saturation and columns "SIS" refer to step indicator saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance level $\gamma = 0.01$, backtesting off, default diagnostic testing, and tolerance for detecting step shifts with two consecutive step and impulse indicators $\delta^S = 2$.

of super saturation is closer to the target. This, however, could be due to the fact that super saturation misses the location shifts more often, thereby overestimating the error variance and hence making it less likely to retain "irrelevant" indicators.

In a second version of the DGP in equation (3.4.18), we consider the case when the shift extends across both halves of the sample, which might be harder to detect algorithmically since one location shift occurs in each half of the sample. We choose $T_1 = 35$ and $T_2 = 65$.

However, Table 3.4.11 shows that super saturation performs very well in this setting. Its potency is higher than that of SIS, for $T_1 = 35$ and $\lambda = 2$ even considerably so. The previous phenomenon of low potency for a shift at $T = 35$ has completely vanished. Also, the higher potency of super saturation does not come at the cost of a higher gauge. The gauge of super saturation is close to the target of 1% and slightly higher for SIS.

Table 3.4.11: Single Temporary Shift Across Both Halves

		Potency T_1		Potency T_2		Gauge	
λ	Type	Super	SIS	Super	SIS	Super	SIS
2	I	0.000	–	0.000	–	0.006	–
2	S	0.687	0.510	0.554	0.537	0.019	0.024
2	total	0.687	0.510	0.555	0.537	0.013	0.024
4	I	0.000	–	0.002	–	0.006	–
4	S	0.966	0.933	0.927	0.928	0.012	0.013
4	total	0.967	0.933	0.929	0.928	0.009	0.013

Note: Monte Carlo experiment with a single temporary shift of magnitude λ spanning across both halves of the sample, starting in period $T_1 = 35$ and ending in $T_2 = 65$. Columns "Super" refer to super saturation and columns "SIS" refer to step indicator saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance level $\gamma = 0.01$, backtesting off, default diagnostic testing, and tolerance for detecting step shifts with two consecutive step and impulse indicators $\delta^S = 2$.

The next DGP (3.4.19) contains two temporary shifts, one in each half. Specifically, we choose $T_1 = 25$, $T_2 = 35$, $T_3 = 75$, and $T_4 = 85$ in

$$y_t = \beta_0 + \lambda\{(1_{(t \leq T_2)} - 1_{(t \leq T_1)}) + (1_{(t \leq T_4)} - 1_{(t \leq T_3)})\} + \epsilon_t. \tag{3.4.19}$$

Table 3.4.12 shows the results. For SIS, we find mostly similar potencies as before, across both shift magnitudes and timings. For super saturation, we observe the

Table 3.4.12: Two Equal Temporary Shifts

		Potency T_1		Potency T_2		Potency T_3		Potency T_4		Gauge	
λ	Type	Super	SIS	Super	SIS	Super	SIS	Super	SIS	Super	SIS
2	I	0.002	–	0.065	–	0.002	–	0.001	–	0.007	–
2	S	0.504	0.542	0.236	0.546	0.529	0.438	0.508	0.500	0.029	0.031
2	total	0.506	0.542	0.300	0.546	0.532	0.438	0.509	0.500	0.018	0.031
4	I	0.008	–	0.158	–	0.005	–	0.004	–	0.007	–
4	S	0.897	0.923	0.670	0.922	0.911	0.931	0.911	0.926	0.013	0.014
4	total	0.906	0.923	0.829	0.922	0.916	0.931	0.914	0.926	0.010	0.014

Note: Monte Carlo experiment with two equal temporary shifts of magnitude λ from period $T_1 = 25$ to $T_2 = 35$ and from $T_3 = 75$ to $T_4 = 85$. Columns "Super" refer to super saturation and columns "SIS" refer to step indicator saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance level $\gamma = 0.01$, backtesting off, default diagnostic testing, and tolerance for detecting step shifts with two consecutive step and impulse indicators $\delta^S = 2$.

same drop in potency for time period $T_2 = 35$ as before in Table 3.4.10, which is accompanied by a higher likelihood of modelling the location shift with a pair of one step and impulse indicator. Apart from this time period, however, the potency of super saturation is similar to that of SIS. The average gauge is better controlled for super saturation than for SIS but they are both higher than the target for $\lambda = 2$, again probably in part due to some mistiming of the retained indicators.

Finally, we modify the previous setup by having two opposite, temporary location shifts instead of two equal location shifts in the DGP

$$y_t = \beta_0 + \lambda(1_{(t \leq T_2)} - 1_{(t \leq T_1)}) - \lambda(1_{(t \leq T_4)} - 1_{(t \leq T_3)}) + \epsilon_t. \tag{3.4.20}$$

This implies that the average mean is constant across the sample, making it potentially harder for the algorithm to detect the intermediate location shifts.

However, our simulation results in Table 3.4.13 show the same high potencies and patterns as for DGP equation (3.4.19) in Table 3.4.12.

3.4.4 DGP with Both Location Shift and Outliers

Super saturation successfully captures both outliers and location shifts when they are both present in the DGP. Their detection rates are similar to the previous cases with only outliers or only location shifts and the gauge is well-controlled for tight significance levels without backtesting.

Table 3.4.13: Two Opposite Temporary Shifts

		Potency T_1		Potency T_2		Potency T_3		Potency T_4		Gauge	
λ	Type	Super	SIS	Super	SIS	Super	SIS	Super	SIS	Super	SIS
2	I	0.002	–	0.070	–	0.002	–	0.002	–	0.008	–
2	S	0.514	0.549	0.277	0.548	0.514	0.468	0.516	0.520	0.030	0.032
2	total	0.517	0.549	0.346	0.548	0.516	0.468	0.518	0.520	0.019	0.032
4	I	0.005	–	0.316	–	0.004	–	0.005	–	0.007	–
4	S	0.908	0.927	0.540	0.925	0.907	0.926	0.912	0.929	0.014	0.015
4	total	0.913	0.927	0.857	0.925	0.912	0.926	0.916	0.929	0.010	0.015

Note: Monte Carlo experiment with two opposite-signed temporary shifts of magnitude λ from period $T_1 = 25$ to $T_2 = 35$ and from $T_3 = 75$ to $T_4 = 85$. Columns "Super" refer to super saturation and columns "SIS" refer to step indicator saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance level $\gamma = 0.01$, backtesting off, default diagnostic testing, and tolerance for detecting step shifts with two consecutive step and impulse indicators $\delta^S = 2$.

The DGP we analyse is

$$y_t = \beta_0 + \lambda(1_{(t \leq T_1)} + 1_{(t=T_2)} + 1_{(t=T_3)} + 1_{(t=T_4)} + 1_{(t=T_5)}) + \epsilon_t, \tag{3.4.21}$$

where the location shift ends in time period $T_1 = 25$ and the four outliers occur in the second half of the sample in time periods $T_2 = 60$, $T_3 = 70$, $T_4 = 80$, and $T_5 = 90$. Both the location shift and the outliers are of magnitude $\lambda = 4\sigma_\epsilon$.

Figure 3.4.8: Stylised Illustration of a DGP with Location Shift and Outliers

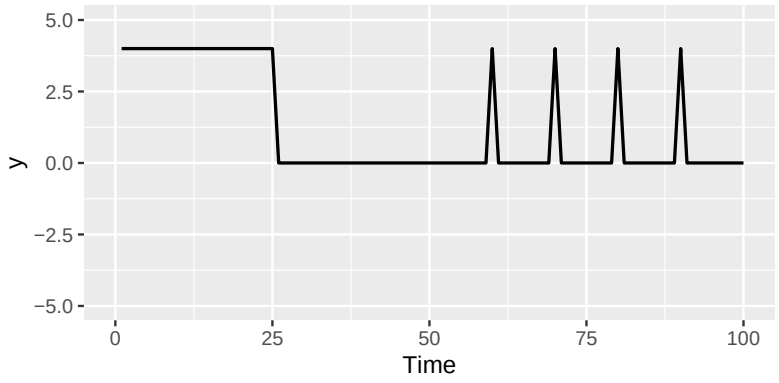


Table 3.4.14 shows the potency and gauge for two different significance levels $\gamma \in \{0.005, 0.01\}$. For both, the potency of the location shift is above 90%. The outliers are detected most of the time, with detection rates averaging approximately 70% and 77% for the two significance levels, respectively. The detection rate is stable across the different time periods. For both significance levels, we find that the average gauge is close to the target γ.

Table 3.4.14: Single Shift and Multiple Outliers

	$\gamma = 0.005$						$\gamma = 0.01$					
	Potency						Potency					
Type	T_1	T_2	T_3	T_4	T_5	Gauge	T_1	T_2	T_3	T_4	T_5	Gauge
I	0.001	0.169	0.207	0.176	0.338	0.002	0.003	0.168	0.211	0.178	0.331	0.004
S	0.904	0.538	0.494	0.527	0.382	0.008	0.914	0.599	0.549	0.590	0.446	0.013
total	0.905	0.708	0.700	0.703	0.720	0.005	0.916	0.768	0.760	0.769	0.776	0.009

Note: Monte Carlo experiment with one step shift in time period $T_1 = 25$ and four outliers in time periods $T_2 = 60$, $T_3 = 70$, $T_4 = 80$, $T_5 = 90$, each of magnitude $\lambda = 4\sigma_\epsilon^2$. Results shown are for super saturation. The column "Type" indicates the potency and gauge for each type of indicator separately and jointly. Settings: $M = 10{,}000$ replications, significance levels $\gamma \in \{0.005, 0.01\}$, backtesting off, default diagnostic testing, tolerance for detecting step shifts with two consecutive step and impulse indicators $\delta^S = 2$, and tolerance for detecting outliers with two consecutive step indicators $\delta^I = 1$.

3.5 Performance Prior to Model Selection in a Regression Model with Covariates

When there are no outliers or location shifts, we find that the additional search cost of super saturation is mainly characterised by an underestimate of the error variance, which leads to higher retention rates of both relevant and irrelevant covariates in subsequent model selection. If instead the dependent variable contains outliers and a location shift, super saturation successfully captures these breaks and improves subsequent model selection. This location shift may occur either directly or as a result of an omitted relevant regressor that itself undergoes a shift. A shift in an irrelevant regressor does not lead to spurious retention of a matching step indicator because indicator saturation is performed prior to covariate selection.

Hendry and Johansen (2015) discuss how extending a theoretical model specification with additional, potentially relevant variables can both improve the assessment of the theoretical model, which may otherwise be incorrect or incomplete, and allows for the discovery of new empirical relationships. Super saturation is a special case where the added indicators can model outliers and location shifts of unknown magnitude and timings.

The following simulations assess the costs of using super saturation when the DGP contains no outliers or location shifts, as well as its benefits when the data is affected by such phenomena. The simulation setup we use is similar to Krolzig and

Hendry (2001) but we do not include any lags of the regressors or of the dependent variable. Our general DGP setup is

$$\begin{aligned} y_t &= \beta_0 + \boldsymbol{\phi}_r' \mathbf{x}_t^r + \boldsymbol{\phi}_i' \mathbf{x}_t^i + \lambda_{y,S} 1_{(t \leq 25)} + \lambda_{y,O}(1_{(t=60)} + 1_{(t=70)} + 1_{(t=80)} + 1_{(t=90)}) + \epsilon_t, \\ x_{c,t} &= \lambda_c 1_{(t \leq 25)} + r_{c,t}, \quad \forall c = 1, \ldots, 10, \\ (r_{1,t}, \ldots, r_{10,t}, \epsilon_t)' &\overset{iid}{\sim} \mathsf{N}(\mathbf{0}_{11}, \boldsymbol{I}_{11}), \end{aligned} \tag{3.5.1}$$

where β_0 is the intercept, $\mathbf{x}_t^r = (x_{1,t}, \ldots, x_{5,t})'$ is the vector of relevant covariates with non-zero coefficient vector $\boldsymbol{\phi}_r$, and $\mathbf{x}_t^i = (x_{6,t}, \ldots, x_{10,t})'$ are irrelevant covariates with coefficient vector $\boldsymbol{\phi}_i = \mathbf{0}_5$. The coefficients $\{\lambda_{y,O}, \lambda_{y,S}, \lambda_1, \ldots, \lambda_{10}\}$ control the presence and size of outliers or location shifts in the dependent variable and the covariates. In our simulations, we set $\beta_0 = 0$, $T = 100$, and $\boldsymbol{\phi}_r = (.2, .3, .4, .6, .8)'$. Without breaks, our chosen settings yield non-centralities of the relevant regressors of $\boldsymbol{\xi} = (2, 3, 4, 6, 8)'$, which causes variation in the retention probabilities.

Table 3.5.1: Overview Over DGPs

Description	Non-Zero λ-Parameters	Section
No Outliers or Location Shifts	– (all zero)	3.5.1
Outliers and Location Shifts in y	$\lambda_{y,O} = \lambda_{y,S} = 4\sigma_\epsilon^2$	3.5.2
Outliers in y, Location Shift in x_1	$\lambda_{y,O} = 4\sigma_\epsilon^2$, $\lambda_1 = 4\sigma_{r_1}^2$	3.5.3
Outliers in y, Location Shift in x_3 (but x_3 omitted from GUMs)	$\lambda_{y,O} = 4\sigma_\epsilon^2$, $\lambda_3 = 4\sigma_{r_3}^2$	3.5.4
Outliers in y, Location Shift in x_6	$\lambda_{y,O} = 4\sigma_\epsilon^2$, $\lambda_6 = 4\sigma_{r_6}^2$	3.5.5

For each DGP, we consider four different GUMs, all of which include an intercept term that is always retained. GUM0 includes the relevant regressors $\boldsymbol{x}_t^r$. GUM1 commences from a wider set of regressors, including both the relevant regressors $\boldsymbol{x}_t^r$ and the irrelevant regressors $\boldsymbol{x}_t^i$. In variants of these GUMs, we also apply super saturation, adding a full set of impulse and step indicators to the GUM as shown in equation (3.2.1). We refer to them as GUM0-S and GUM1-S.

For GUM0-S and GUM1-S, we implement super saturation using Autometrics with a tight significance level of $\gamma^{\text{indicators}} = 0.01$, no backtesting, and default diagnostic testing. This step is not applicable to GUM0 and GUM1. Next, the resulting

Table 3.5.2: Overview of Terms Included in the GUMs

Abbreviation	Intercept	$\boldsymbol{x}_t^r$	$\boldsymbol{x}_t^i$	Super Saturation
GUM0	Yes	Yes	No	No
GUM1	Yes	Yes	Yes	No
GUM0-S	Yes	Yes	No	Yes
GUM1-S	Yes	Yes	Yes	Yes

model is either estimated or followed by covariate selection, where any previously retained indicators from super saturation are not selected over again. For covariate selection, we use a looser significance level of $\gamma^{\text{covariates}} = 0.05$. We either use Autometrics for the selection or the 1-cut approach. In the 1-cut approach, the model is estimated and any regressors that are deemed insignificant based on their individual t-test are removed in a single step. The 1-cut approach resembles doing inference in a theoretical model, where the researcher removes insignificant variables. Since the covariates in the DGP setup (3.5.1) are uncorrelated, the 1-cut and multi-path procedures will only differ slightly because of finite sample correlations.

The analysis we present is similar to Hendry and Krolzig (2005), who study the properties of automatic general-to-specific modelling but without indicator saturation methods. We present the average bias, average standard error (SE), and the standard deviation (SD) and root mean squared error (RMSE) across the $M =$ 10,000 replications for the different GUMs and selection strategies. Each metric is calculated in two different versions: the unconditional version of the statistics includes the values of zero when a variable has not been retained, whereas the conditional version is computed only for the subset of replications in which the variable has been retained. The test statistics are scaled by the true asymptotic standard error of the coefficient estimators $\widehat{\boldsymbol{\phi}}$. In addition, we show the average estimated standard deviation of the error term.

Where applicable, we also report the gauge of irrelevant covariates ($\widehat{\gamma}^{\text{covariates}}$), the impulse gauge ($\widehat{\gamma}^{I}$), and the step gauge ($\widehat{\gamma}^{S}$). Similarly, we report the impulse potency ($\widehat{\rho}^{I}$), step potency ($\widehat{\rho}^{S}$), as well as the potency of relevant covariates, both averaged across all covariates ($\widehat{\rho}^{\text{covariates}}$) and separately for each individual covariate ($\widehat{\rho}^{x_c}$).

3.5.1 DGP without Outliers or Location Shifts

Under the null of no outliers or location shifts, the use of super saturation runs the risk of overfitting the model, which can be controlled by choosing a tight significance level. Overfitting affects subsequent model selection through higher retention rates of irrelevant and relevant covariates alike.

Table 3.5.3 shows the results for the DGP without outliers or location shifts. GUM0 coincides with the DGP, so can serve as a benchmark. As expected, the average bias is small, standard errors are on average close to the theoretical asymptotic standard errors and close to the Monte Carlo standard deviations, which represent an estimate of the true variability of the estimators. The average estimated standard deviation of the error is close to the true standard error of 1.

When selection is undertaken in any of the models, unconditional coefficient estimates are downward biased because they are zero when not retained, and $\widehat{\phi}_c$ when retained. Conditional coefficient estimates are upward biased because larger absolute values of $\widehat{\phi}_c$ are needed for retention. Both the conditional and the unconditional bias decrease in the non-centrality because it increases their retention probabilities. In fact, the covariates with the largest non-centralities of $\xi_4 = 6$ and $\xi_5 = 8$ are almost always retained.

Selection affects the RMSEs of the estimated coefficients. We find increased unconditional RMSEs for the relevant regressors because the selection induces bias, which is not compensated by a reduction in variance. The changes in RMSEs are similar for 1-cut selection and Autometrics.

3.5.1.1 Cost of Search versus Cost of Inference

The cost of search when starting from a more general model than the GUM is similar to the cost of doing inference when starting from the DGP. The average potency of relevant covariates is high in both cases.

We can interpret reductions from the DGP as doing inference. The researcher might estimate a correct model and subsequently remove insignificant variables. 1-cut selection is close to the practice of considering variables' statistical significance

Table 3.5.3: Comparison of Reduction, No Outliers and No Mean Shifts

Model:	GUM0					GUM0-S					GUM1					GUM1-S				
Selection:	None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto	
Statistic:		U	C	U	C		U	C	U	C		U	C	U	C		U	C	U	C
Bias/SE_{Th}																				
$x_1(\xi = 2)$	0.001	-0.624	0.844	-0.600	0.812	0.006	-0.509	0.842	-0.492	0.815	-0.001	-0.679	0.870	-0.609	0.831	0.006	-0.534	0.863	-0.496	0.842
$x_2(\xi = 3)$	-0.030	-0.314	0.301	-0.301	0.291	-0.030	-0.276	0.316	-0.265	0.306	-0.028	-0.364	0.327	-0.318	0.307	-0.021	-0.304	0.347	-0.260	0.323
$x_3(\xi = 4)$	0.008	-0.053	0.089	-0.049	0.086	0.011	-0.046	0.102	-0.042	0.100	0.010	-0.072	0.107	-0.056	0.098	0.016	-0.062	0.125	-0.048	0.115
$x_4(\xi = 6)$	0.029	0.030	0.031	0.030	0.031	0.029	0.028	0.031	0.028	0.031	0.030	0.032	0.033	0.032	0.033	0.036	0.038	0.041	0.037	0.039
$x_5(\xi = 8)$	0.003	0.002	0.002	0.001	0.001	-0.001	-0.003	-0.003	-0.003	-0.003	0.004	0.001	0.001	0.001	0.001	0.003	0.000	0.000	0.000	0.000
SE/SE_{Th}																				
$x_1(\xi = 2)$	1.033	0.492	1.017	0.507	1.018	0.975	0.503	0.959	0.514	0.960	1.061	0.466	1.012	0.497	1.012	0.998	0.487	0.952	0.503	0.952
$x_2(\xi = 3)$	1.031	0.831	1.022	0.838	1.022	0.974	0.793	0.965	0.799	0.966	1.060	0.805	1.016	0.824	1.016	0.996	0.772	0.959	0.790	0.959
$x_3(\xi = 4)$	1.031	0.992	1.028	0.994	1.028	0.974	0.936	0.971	0.937	0.971	1.060	0.979	1.024	0.985	1.023	0.996	0.921	0.965	0.926	0.964
$x_4(\xi = 6)$	1.031	1.030	1.030	1.030	1.030	0.974	0.973	0.973	0.973	0.973	1.060	1.027	1.027	1.025	1.025	0.996	0.967	0.967	0.966	0.966
$x_5(\xi = 8)$	1.031	1.030	1.030	1.030	1.030	0.973	0.973	0.973	0.973	0.973	1.059	1.027	1.027	1.025	1.025	0.995	0.967	0.967	0.966	0.966
SD/SE_{Th}																				
$x_1(\xi = 2)$	1.042	1.492	0.650	1.485	0.675	1.124	1.515	0.733	1.509	0.755	1.074	1.501	0.670	1.492	0.672	1.173	1.534	0.771	1.525	0.770
$x_2(\xi = 3)$	1.029	1.484	0.823	1.470	0.828	1.115	1.504	0.889	1.491	0.894	1.065	1.538	0.830	1.494	0.828	1.161	1.555	0.906	1.509	0.908
$x_3(\xi = 4)$	1.028	1.206	0.963	1.196	0.964	1.121	1.278	1.042	1.270	1.043	1.056	1.260	0.962	1.226	0.965	1.155	1.329	1.040	1.302	1.045
$x_4(\xi = 6)$	1.038	1.055	1.052	1.054	1.051	1.130	1.148	1.141	1.146	1.140	1.067	1.063	1.058	1.063	1.060	1.161	1.155	1.148	1.154	1.148
$x_5(\xi = 8)$	1.035	1.052	1.052	1.052	1.052	1.126	1.140	1.140	1.140	1.140	1.064	1.060	1.060	1.060	1.060	1.167	1.157	1.157	1.158	1.158
RMSE/SE_{Th}																				
$x_1(\xi = 2)$	1.042	1.617	1.066	1.601	1.056	1.124	1.598	1.116	1.587	1.111	1.074	1.647	1.098	1.611	1.069	1.173	1.624	1.157	1.604	1.141
$x_2(\xi = 3)$	1.030	1.517	0.876	1.500	0.877	1.116	1.529	0.944	1.514	0.945	1.066	1.581	0.892	1.528	0.883	1.161	1.584	0.970	1.531	0.964
$x_3(\xi = 4)$	1.028	1.207	0.967	1.197	0.968	1.121	1.279	1.047	1.271	1.048	1.056	1.262	0.968	1.227	0.970	1.155	1.331	1.048	1.302	1.051
$x_4(\xi = 6)$	1.039	1.056	1.052	1.055	1.051	1.130	1.148	1.142	1.146	1.140	1.068	1.064	1.059	1.064	1.061	1.162	1.156	1.148	1.155	1.149
$x_5(\xi = 8)$	1.035	1.052	1.052	1.051	1.051	1.126	1.140	1.140	1.140	1.140	1.064	1.060	1.060	1.060	1.060	1.167	1.157	1.157	1.158	1.158
$\widehat{\sigma}_u$	0.998	1.001		1.001		0.931	0.934		0.934		0.998	0.997		0.995		0.926	0.927		0.925	
Retention																				
$\widehat{\gamma}^{\text{covariates}}$	–	–		–		–	–		–		–	0.049		0.061		–	0.093		0.103	
$\widehat{\rho}^{\text{covariates}}$	–	0.853		0.857		–	0.862		0.866		–	0.842		0.853		–	0.854		0.863	
$\widehat{\rho}^{x1}$	–	0.484		0.498		–	0.525		0.536		–	0.460		0.491		–	0.512		0.529	
$\widehat{\rho}^{x2}$	–	0.814		0.820		–	0.822		0.827		–	0.792		0.811		–	0.805		0.825	
$\widehat{\rho}^{x3}$	–	0.965		0.967		–	0.964		0.965		–	0.957		0.963		–	0.955		0.960	
$\widehat{\rho}^{x4}$	–	1.000		1.000		–	1.000		1.000		–	1.000		1.000		–	1.000		1.000	
$\widehat{\rho}^{x5}$	–	1.000		1.000		–	1.000		1.000		–	1.000		1.000		–	1.000		1.000	
$\widehat{\gamma}^{I}$	–	–		–		0.008	0.008		0.008		–	–		–		0.009	0.009		0.009	
$\widehat{\gamma}^{S}$	–	–		–		0.015	0.015		0.015		–	–		–		0.015	0.015		0.015	

Note: Monte Carlo experiment for a DGP with covariates but without outliers or mean shifts. GUM0 includes the covariates $x_1 - x_5$, GUM1 in addition contains the covariates $x_6 - x_{10}$. GUM0-S and GUM1-S apply super saturation before selecting over covariates. Super saturation has been implemented using Autometrics and there are three types of covariate selection: "None" without selection, "1-Cut" drops insignificant covariates in a single step, and "Auto" uses general-to-specific model selection as implemented in Autometrics. The columns denoted by "U" show the unconditional version of the statistics, including replications where the covariate has been removed, and "C" shows the conditional version of the statistics excluding replications where the variable has been removed. "Bias" refers to the average bias of the estimated coefficient relative to its true coefficient, "SE" to the average standard error of the coefficient estimate, "SD" to the Monte Carlo standard deviation of the estimated coefficient across replications, and "RMSE" to the Monte Carlo root mean square error across replications. All these statistics are scaled by the theoretical standard error, "SE_{th}". The average estimated standard deviation of the error term is denoted by $\widehat{\sigma}_u$. The second part of the table reports the retention rates of relevant and irrelevant covariates and indicators. $\widehat{\gamma}$ refers to the average gauge and $\widehat{\rho}$ to the average potency across replications. The superscript denotes the variable or category of variables for which the gauge and potency have been calculated. Settings: $M = 10{,}000$ replications, selection significance level for indicators $\gamma^{\text{indicators}} = 0.01$, selection significance level for covariates $\gamma^{\text{covariates}} = 0.05$, backtesting off, and default diagnostic testing.

and removing them when insignificant. Autometrics could also be used, taking diagnostics into account and using a multi-path search for the reductions. In both cases, even when the estimated model coincides with the DGP (as is the case with GUM0), relevant regressors may be incorrectly removed when they are insignificant.

The cost of inference can be compared to the cost of selection when starting from a more general GUM (for example GUM1). Model selection is used to arrive at a terminal model. Again, both 1-cut and Autometrics could be used for that purpose.

By comparing the results for GUM1 with selection to those with GUM0 with selection, we can compare the cost of search plus inference to the cost of inference alone.

First, we note that the 1-cut algorithm and Autometrics yield very similar results. 1-cut selection is appropriate for orthogonal regressors as in our case, but will perform much worse than Autometrics when the regressors are mutually correlated. Even in our case, however, Autometrics performs slightly better on all metrics except the gauge, where it produces a gauge of irrelevant covariates of 6% instead of 5% of the 1-cut approach. The results are very similar despite the fact that Autometrics uses multi-path search with diagnostic testing.

Second, the cost of search is similar to the cost of inference when the GUM nests the DGP. The average potency of relevant covariates is high in both cases, irrespective of whether one starts from the DGP (GUM0) or a larger GUM (GUM1). The average potency reaches approximately 85% in both cases. The potencies of the individual covariates are close to the theoretical power on their t-tests when doing inference on them. For example, for non-centralities $\xi = (2, 3, 4)$, the probability of a significant t-test is $(0.51, 0.84, 0.98)$, which is close to the observed corresponding potencies $(0.49, 0.81, 0.96)$ of Autometrics when starting from GUM1. The conditional and unconditional bias, standard errors, and the RMSEs are also similar for GUM0 and GUM1. The standard deviation of the error is only slightly underestimated when starting from GUM1 (0.995 compared to the true value of 1). This is because the gauge of irrelevant covariates is well-controlled, being close to the nominal significance level of $\gamma^{\text{covariates}} = 0.05$.

3.5.1.2 Additional Cost of Search of Super Saturation

We now consider the additional cost that is introduced by using super saturation when there are no outliers or location shifts. Super saturation has the risk of incorrectly retaining impulse and step indicators and thereby underestimating the error variance. This yields higher retention rates of both relevant and irrelevant covariates.

First, consider the cases without subsequent model selection. Comparing the results of GUM0-S and GUM1-S to their counterparts without super saturation, we find that the cost of additionally searching over $2T - 2$ indicators is low. The impulse gauge is slightly below our target of $\gamma^{\text{indicators}} = 0.01$, while the step gauge is slightly above. There is also no difference between GUM0-S and GUM1-S, the additional irrelevant covariates do not noticeably increase the retention probability of indicators. No systematic biases of the coefficient estimates are introduced, and their Monte Carlo standard deviation and RMSE are only slightly increased. Because of the false retention of indicators, we underestimate the standard deviation of the error and hence also the average standard errors of the coefficients.

Second, super saturation followed by covariate selection slightly improves the outcomes for GUM0-S compared to GUM0. While in practice it is extremely unlikely to start from the DGP, the overfitting due to the adventitiously retained indicators helps with retaining more of the relevant covariates. This tends to improve the average bias induced by selection but the RMSE increases slightly. The results are similar for comparing GUM1-S to GUM1. The most noticeable difference is that super saturation increases the gauge of irrelevant covariates because the underestimated error variance makes it more likely to retain irrelevant regressors. The gauge increases from 0.06 (0.05) to 0.1 (0.09) for Autometrics (1-cut). Again, there is little difference between Autometrics and 1-cut, highlighting again that a more sophisticated search need not cause significant repeated testing problems.

If the false retention of covariates is a major concern, one can either tighten the significance level for selecting indicators in the first step, or tighten the significance

level for selecting covariates in the second step.

For example, using super saturation with $\gamma^{\text{indicators}} = 0.005$ yields an average indicator gauge of 0.004, mitigates the underestimation of the standard deviation of the error (0.97), and results in a gauge of irrelevant covariates of only 0.07 with the same target level of $\gamma^{\text{covariates}} = 0.05$ as before.

3.5.2 DGP with Location Shift and Outliers in Regressand

Undetected outliers and location shifts in the dependent variable substantially reduce the retention rate of relevant covariates and increase their RMSEs. Super saturation has a high chance of detecting the outliers and shifts, which increases the potency of relevant covariates and thereby improves their RMSEs and estimated standard errors.

In this section, the DGP contains a location shift and four outliers in the dependent variable y. Note, however, that the GUMs have not changed. In particular, this means that GUM0 and GUM1 only include the covariates and therefore, GUM0 does not coincide with the DGP any more. Only GUM0-S and GUM1-S can capture the outliers and location shifts using super saturation. The results are shown in Table 3.5.4.

As expected, the standard deviation of the error is overestimated for GUM0 and GUM1. The average standard errors, the standard deviation, and the RMSE increase substantially. The biases of the coefficient estimates are similar to before, which is because the covariates are mean zero, have no dynamics, and are uncorrelated to the outliers and the location shift. Otherwise, the breaks in y could induce major biases.

Doing covariate selection in GUM0 or GUM1 yields much lower potency than before, reaching only approximately 0.55 compared to 0.85 under the null. This is especially true for the covariates with low non-centralities. Autometrics finds more relevant covariates than the 1-cut approach but also retains more irrelevant covariates. This is likely due to the diagnostic testing, which Autometrics tries to improve by retaining more covariates. For both GUMs and selection methods,

Table 3.5.4: Comparison of Reduction, Outliers and Mean Shift in y

Model:	GUM0					GUM0-S					GUM1					GUM1-S				
Selection:	None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto	
Statistic:		U	C	U	C		U	C	U	C		U	C	U	C		U	C	U	C
Bias/SE$_{Th}$																				
$x_1(\xi=2)$	-0.007	-1.210	3.157	-1.058	2.101	-0.002	-0.588	0.985	-0.570	0.957	-0.005	-1.247	3.170	-1.068	2.148	-0.009	-0.639	1.033	-0.578	0.992
$x_2(\xi=3)$	-0.017	-1.462	2.507	-1.324	2.016	-0.026	-0.357	0.443	-0.337	0.428	-0.022	-1.528	2.543	-1.337	2.139	-0.019	-0.396	0.481	-0.350	0.460
$x_3(\xi=4)$	0.014	-1.338	1.841	-1.214	1.641	0.018	-0.078	0.164	-0.073	0.160	0.007	-1.453	1.896	-1.264	1.741	0.010	-0.123	0.195	-0.100	0.183
$x_4(\xi=6)$	0.018	-0.641	0.793	-0.608	0.778	0.026	0.025	0.032	0.025	0.032	0.014	-0.747	0.857	-0.681	0.839	0.026	0.022	0.034	0.023	0.034
$x_5(\xi=8)$	0.033	-0.127	0.240	-0.125	0.239	0.007	0.003	0.003	0.003	0.003	0.034	-0.174	0.277	-0.157	0.265	0.006	0.004	0.004	0.005	0.005
SE/SE$_{Th}$																				
$x_1(\xi=2)$	2.151	0.322	2.103	0.486	2.115	1.050	0.487	1.029	0.499	1.030	2.212	0.305	2.096	0.473	2.107	1.086	0.461	1.028	0.489	1.029
$x_2(\xi=3)$	2.149	0.587	2.102	0.705	2.111	1.049	0.795	1.036	0.805	1.037	2.209	0.556	2.094	0.681	2.104	1.084	0.775	1.036	0.793	1.036
$x_3(\xi=4)$	2.148	0.960	2.106	1.042	2.109	1.049	0.982	1.043	0.985	1.043	2.208	0.907	2.099	1.002	2.102	1.084	0.965	1.044	0.973	1.044
$x_4(\xi=6)$	2.148	1.672	2.120	1.686	2.120	1.049	1.046	1.047	1.046	1.047	2.208	1.617	2.111	1.640	2.109	1.084	1.048	1.051	1.047	1.049
$x_5(\xi=8)$	2.147	2.034	2.129	2.034	2.128	1.049	1.047	1.047	1.047	1.047	2.207	2.006	2.121	2.010	2.118	1.083	1.051	1.051	1.049	1.049
SD/SE$_{Th}$																				
$x_1(\xi=2)$	2.162	1.954	1.551	2.093	2.472	1.201	1.580	0.763	1.575	0.783	2.225	1.924	1.603	2.105	2.528	1.268	1.603	0.810	1.596	0.815
$x_2(\xi=3)$	2.146	2.560	1.268	2.600	1.865	1.204	1.664	0.922	1.645	0.929	2.203	2.543	1.335	2.610	1.784	1.253	1.717	0.942	1.682	0.943
$x_3(\xi=4)$	2.141	3.047	1.342	3.030	1.575	1.196	1.429	1.077	1.419	1.079	2.197	3.058	1.379	3.050	1.503	1.253	1.525	1.087	1.487	1.089
$x_4(\xi=6)$	2.154	3.148	1.679	3.117	1.678	1.210	1.233	1.218	1.233	1.217	2.214	3.258	1.690	3.209	1.685	1.276	1.286	1.258	1.283	1.257
$x_5(\xi=8)$	2.167	2.599	2.012	2.591	2.007	1.207	1.223	1.223	1.223	1.223	2.218	2.708	2.005	2.671	2.008	1.267	1.258	1.258	1.257	1.257
RMSE/SE$_{Th}$																				
$x_1(\xi=2)$	2.162	2.298	3.517	2.345	3.244	1.201	1.686	1.246	1.675	1.236	2.225	2.292	3.551	2.361	3.317	1.268	1.726	1.312	1.698	1.284
$x_2(\xi=3)$	2.146	2.948	2.810	2.918	2.746	1.204	1.701	1.023	1.680	1.022	2.203	2.967	2.872	2.932	2.785	1.253	1.762	1.058	1.718	1.049
$x_3(\xi=4)$	2.141	3.328	2.278	3.264	2.274	1.196	1.431	1.089	1.421	1.090	2.197	3.385	2.344	3.301	2.300	1.253	1.530	1.104	1.490	1.105
$x_4(\xi=6)$	2.154	3.212	1.857	3.175	1.849	1.210	1.234	1.218	1.233	1.217	2.213	3.342	1.894	3.280	1.883	1.276	1.287	1.258	1.283	1.258
$x_5(\xi=8)$	2.167	2.602	2.026	2.594	2.021	1.207	1.223	1.223	1.223	1.223	2.218	2.713	2.024	2.675	2.025	1.267	1.258	1.258	1.257	1.257
$\widehat{\sigma}_u$	2.079	2.091		2.088		0.978	0.982		0.981		2.079	2.083		2.072		0.983	0.984		0.982	
Retention																				
$\widehat{\gamma}^{\text{covariates}}$	–	–		–		–	–		–		–	0.049		0.129		–	0.088		0.101	
$\widehat{\rho}^{\text{covariates}}$	–	0.527		0.562		–	0.836		0.841		–	0.511		0.550		–	0.824		0.834	
$\widehat{\rho}^{x_1}$	–	0.153		0.230		–	0.473		0.484		–	0.146		0.225		–	0.449		0.475	
$\widehat{\rho}^{x_2}$	–	0.279		0.334		–	0.768		0.777		–	0.266		0.324		–	0.748		0.766	
$\widehat{\rho}^{x_3}$	–	0.456		0.494		–	0.942		0.944		–	0.432		0.477		–	0.924		0.932	
$\widehat{\rho}^{x_4}$	–	0.789		0.795		–	0.999		0.999		–	0.766		0.778		–	0.998		0.998	
$\widehat{\rho}^{x_5}$	–	0.956		0.956		–	1.000		1.000		–	0.946		0.949		–	1.000		1.000	
$\widehat{\gamma}^{I}$	–	–		–		0.005	0.005		0.005		–	–		–		0.005	0.005		0.005	
$\widehat{\gamma}^{S}$	–	–		–		0.012	0.012		0.012		–	–		–		0.012	0.012		0.012	
$\widehat{\rho}^{I}$	–	–		–		0.760	0.760		0.760		–	–		–		0.721	0.722		0.721	
$\widehat{\rho}^{S}$	–	–	–	–	–	0.901	0.901	0.901	0.901	0.901	–	–	–	–	–	0.878	0.878	0.878	0.878	0.878
$\widehat{\rho}^{S25}+\widehat{\gamma}^{S24,S26}$	–	–	–	–	–	0.973	0.973	0.973	0.973	0.973	–	–	–	–	–	0.963	0.963	0.963	0.963	0.963

Note: Monte Carlo experiment for a DGP with covariates, a mean shift in time period $T_1 = 25$, and four outliers in time periods $T_2 = 60, T_3 = 70, T_4 = 80, T_5 = 90$. GUM0 includes the covariates $x_1 - x_5$, GUM1 in addition contains the covariates $x_6 - x_{10}$. GUM0-S and GUM1-S apply super saturation before selecting over covariates. Super saturation has been implemented using Autometrics and there are three types of covariate selection: "None" without selection, "1-Cut" drops insignificant covariates in a single step, and "Auto" uses general-to-specific model selection as implemented in Autometrics. The columns denoted by "U" show the unconditional version of the statistics, including replications where the covariate has been removed, and "C" shows the conditional version of the statistics excluding replications where the variable has been removed. "Bias" refers to the average bias of the estimated coefficient relative to its true coefficient, "SE" to the average standard error of the coefficient estimate, "SD" to the Monte Carlo standard deviation of the estimated coefficient across replications, and "RMSE" to the Monte Carlo root mean square error across replications. All these statistics are scaled by the theoretical standard error, "SE$_{th}$". The average estimated standard deviation of the error term is denoted by $\widehat{\sigma}_u$. The second part of the table reports the retention rates of relevant and irrelevant covariates and indicators. $\widehat{\gamma}$ refers to the average gauge and $\widehat{\rho}$ to the average potency across replications. The superscript denotes the variable or category of variables for which the gauge and potency have been calculated. Settings: $M = 10{,}000$ replications, selection significance level for indicators $\gamma^{\text{indicators}} = 0.01$, selection significance level for covariates $\gamma^{\text{covariates}} = 0.05$, backtesting off, and default diagnostic testing.

selection induces a stronger bias and a higher RMSE than under the null.

Super saturation successfully detects the outliers and the location shift in both GUM0-S and GUM1-S. The average potency for outliers is higher than 0.72 and the average potency of the location shift is higher than 0.96 in both cases when allowing for mistiming of plus or minus one time period. The average gauge of the indicators is well-controlled.

The application of super saturation substantially improves subsequent model selection. The successful detection of the outliers and the shift results in a much better estimate of the error variance, higher potencies of the relevant covariates that are close to the potencies under the null, improved biases, and lower RMSEs across all coefficient estimates. In GUM1-S, the gauge of irrelevant covariates is still relatively high (0.1 for Autometrics) but lower than in GUM1 (0.13). The failure to correct for the structural breaks leads to a higher false retention of covariates in GUM1, presumably in an attempt to improve the diagnostic testing.

3.5.3 DGP with Location Shift in Relevant Regressor and Outliers in Regressand

A location shift in a relevant regressor does not cause super saturation to spuriously retain more indicators because super saturation is done prior to covariate selection. Super saturation reaches a high outlier potency in the dependent variable, with improvements to subsequent covariate selection.

Now, the relevant regressor x_1 undergoes a shift of magnitude $\lambda_1 = 4\sigma_{r_1} = 4$, inducing a mean shift in y of magnitude $\phi_1 \times \lambda_1 = 0.2 \times 4 = 0.8$. The mean shift in x_1 changes its non-centrality from previously 2 to ≈ 4. The DGP still contains outliers in the dependent variable.

Table 3.5.5 shows the results. The results are qualitatively similar to Table 3.5.4. However, the biases, overestimates of the error variance, and RMSE without super saturation are not as pronounced because the location shift is accounted for when x_1 is retained in the final specification. Even if it is missed, its induced shift in y is much smaller than in the previous experiment (0.8 instead of 4).

Table 3.5.5: Comparison of Reduction, Outliers in y and Mean Shift in x_1

Model:	GUM0					GUM0-S					GUM1					GUM1-S				
Selection:	None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto	
Statistic:		U	C	U	C		U	C	U	C		U	C	U	C		U	C	U	C
Bias/SE$_{Th}$																				
$x_1(\xi \approx 4)$	-0.793	-1.345	-0.252	-1.315	-0.274	-0.031	-0.183	0.268	-0.178	0.262	-0.793	-1.427	-0.221	-1.366	-0.237	-0.035	-0.222	0.290	-0.196	0.273
$x_2(\xi = 3)$	-0.025	-0.690	0.801	-0.630	0.730	-0.019	-0.330	0.419	-0.321	0.408	-0.022	-0.754	0.837	-0.654	0.772	-0.020	-0.374	0.455	-0.331	0.435
$x_3(\xi = 4)$	0.007	-0.279	0.359	-0.252	0.337	0.010	-0.074	0.142	-0.070	0.138	0.006	-0.333	0.389	-0.282	0.362	0.007	-0.112	0.177	-0.088	0.160
$x_4(\xi = 6)$	0.021	0.004	0.044	0.003	0.044	0.026	0.024	0.029	0.025	0.028	0.022	-0.003	0.049	-0.002	0.052	0.024	0.021	0.027	0.020	0.027
$x_5(\xi = 8)$	0.003	0.002	0.002	0.001	0.001	0.004	0.003	0.003	0.003	0.003	0.005	0.005	0.005	0.005	0.005	0.011	0.007	0.007	0.007	0.007
SE/SE$_{Th}$																				
$x_1(\xi \approx 4)$	1.296	0.908	1.285	0.925	1.286	1.132	0.992	1.109	0.995	1.110	1.332	0.869	1.279	0.894	1.280	1.162	0.975	1.108	0.985	1.107
$x_2(\xi = 3)$	1.312	0.788	1.297	0.825	1.299	1.026	0.792	1.014	0.798	1.015	1.349	0.756	1.291	0.804	1.292	1.056	0.770	1.013	0.787	1.013
$x_3(\xi = 4)$	1.312	1.113	1.303	1.127	1.304	1.027	0.969	1.023	0.971	1.023	1.349	1.084	1.297	1.106	1.297	1.056	0.950	1.020	0.959	1.019
$x_4(\xi = 6)$	1.312	1.303	1.312	1.302	1.311	1.026	1.025	1.026	1.025	1.026	1.348	1.296	1.307	1.294	1.305	1.056	1.024	1.025	1.023	1.024
$x_5(\xi = 8)$	1.311	1.312	1.312	1.312	1.312	1.026	1.026	1.026	1.026	1.026	1.348	1.308	1.308	1.306	1.306	1.056	1.025	1.025	1.024	1.024
SD/SE$_{Th}$																				
$x_1(\xi \approx 4)$	1.108	1.830	0.814	1.806	0.823	1.496	1.765	1.254	1.759	1.259	1.149	1.882	0.827	1.851	0.826	1.560	1.830	1.273	1.795	1.277
$x_2(\xi = 3)$	1.315	1.984	0.899	1.951	0.956	1.189	1.632	0.921	1.622	0.929	1.354	2.017	0.918	1.976	0.949	1.246	1.690	0.945	1.654	0.944
$x_3(\xi = 4)$	1.313	1.836	1.083	1.800	1.092	1.185	1.390	1.071	1.384	1.074	1.341	1.905	1.083	1.847	1.090	1.237	1.483	1.076	1.441	1.086
$x_4(\xi = 6)$	1.316	1.396	1.310	1.394	1.308	1.196	1.214	1.204	1.213	1.204	1.352	1.426	1.318	1.427	1.315	1.248	1.247	1.233	1.249	1.232
$x_5(\xi = 8)$	1.319	1.348	1.348	1.345	1.345	1.198	1.208	1.208	1.208	1.208	1.355	1.361	1.361	1.362	1.362	1.246	1.234	1.234	1.232	1.232
RMSE/SE$_{Th}$																				
$x_1(\xi \approx 4)$	1.362	2.271	0.852	2.234	0.867	1.496	1.775	1.282	1.768	1.286	1.396	2.362	0.856	2.301	0.859	1.561	1.843	1.305	1.805	1.305
$x_2(\xi = 3)$	1.315	2.101	1.204	2.050	1.203	1.189	1.665	1.012	1.653	1.015	1.354	2.153	1.243	2.082	1.223	1.246	1.731	1.049	1.687	1.039
$x_3(\xi = 4)$	1.312	1.857	1.141	1.817	1.143	1.185	1.392	1.080	1.385	1.082	1.341	1.934	1.150	1.868	1.149	1.236	1.487	1.090	1.444	1.098
$x_4(\xi = 6)$	1.316	1.396	1.311	1.394	1.309	1.197	1.214	1.204	1.213	1.204	1.352	1.426	1.318	1.427	1.315	1.248	1.248	1.234	1.249	1.232
$x_5(\xi = 8)$	1.319	1.348	1.348	1.345	1.345	1.198	1.208	1.208	1.208	1.208	1.355	1.361	1.361	1.362	1.362	1.246	1.234	1.234	1.232	1.232
$\widehat{\sigma}_u$	1.269	1.276		1.275		0.959	0.961		0.961		1.269	1.270		1.266		0.959	0.959		0.958	
Retention																				
$\widehat{\gamma}^{\text{covariates}}$	–	–		–		–	–		–		–	0.050		0.105		–	0.092		0.104	
$\widehat{\rho}^{\text{covariates}}$	–	0.832		0.842		–	0.924		0.926		–	0.818		0.833		–	0.914		0.921	
$\widehat{\rho}^{x1}$	–	0.707		0.719		–	0.894		0.896		–	0.679		0.698		–	0.880		0.890	
$\widehat{\rho}^{x2}$	–	0.608		0.635		–	0.781		0.786		–	0.585		0.622		–	0.760		0.777	
$\widehat{\rho}^{x3}$	–	0.854		0.864		–	0.948		0.950		–	0.835		0.852		–	0.931		0.941	
$\widehat{\rho}^{x4}$	–	0.993		0.993		–	0.999		0.999		–	0.991		0.991		–	0.999		0.999	
$\widehat{\rho}^{x5}$	–	1.000		1.000		–	1.000		1.000		–	1.000		1.000		–	1.000		1.000	
$\widehat{\gamma}^{I}$	–	–		–		0.006	0.006		0.006		–	–		–		0.006	0.006		0.006	
$\widehat{\gamma}^{S}$	–	–		–		0.014	0.014		0.014		–	–		–		0.015	0.015		0.015	
$\widehat{\gamma}^{S25}$	–	–		–		0.009	0.009		0.009		–	–		–		0.009	0.009		0.009	
$\widehat{\rho}^{I}$	–	–		–		0.785	0.784		0.784		–	–		–		0.750	0.751		0.751	

Note: Monte Carlo experiment for a DGP with covariates, a mean shift in the relevant covariate x_1 in time period $T_1 = 25$, and four outliers in y in time periods $T_2 = 60, T_3 = 70, T_4 = 80, T_5 = 90$. GUM0 includes the covariates $x_1 - x_5$, GUM1 in addition contains the covariates $x_6 - x_{10}$. GUM0-S and GUM1-S apply super saturation before selecting over covariates. Super saturation has been implemented using Autometrics and there are three types of covariate selection: "None" without selection, "1-Cut" drops insignificant covariates in a single step, and "Auto" uses general-to-specific model selection as implemented in Autometrics. The columns denoted by "U" show the unconditional version of the statistics, including replications where the covariate has been removed, and "C" shows the conditional version of the statistics excluding replications where the variable has been removed. "Bias" refers to the average bias of the estimated coefficient relative to its true coefficient, "SE" to the average standard error of the coefficient estimate, "SD" to the Monte Carlo standard deviation of the estimated coefficient across replications, and "RMSE" to the Monte Carlo root mean square error across replications. All these statistics are scaled by the theoretical standard error, "SE$_{th}$". The average estimated standard deviation of the error term is denoted by $\widehat{\sigma}_u$. The second part of the table reports the retention rates of relevant and irrelevant covariates and indicators. $\widehat{\gamma}$ refers to the average gauge and $\widehat{\rho}$ to the average potency across replications. The superscript denotes the variable or category of variables for which the gauge and potency have been calculated. Settings: $M = 10{,}000$ replications, selection significance level for indicators $\gamma^{\text{indicators}} = 0.01$, selection significance level for covariates $\gamma^{\text{covariates}} = 0.05$, backtesting off, and default diagnostic testing.

Super saturation again reaches high potency of more than 0.75 for the outliers, which yields better estimates of the standard deviation of the error, with the usual benefit of higher potencies of the relevant covariates. The average indicator gauge is close to our target of 1%. Theoretically, the induced location shift at time period 25 could lead to the false retention of a corresponding step indicator. However, this is not a concern in practice. We find that the gauge for step indicator S25 closely matches the expected gauge and is even below the average step gauge. The reason why a step indicator does not incorrectly proxy for the regressor x_1 lies in the step-wise procedure recommended in Hendry and Johansen (2015) and Castle and Hendry (2019). The covariates are included and not selected over when applying super saturation in a first step. Since x_1 already explains the induced location shift in y, there is no need to retain the step indicator S25.

3.5.4 DGP with Location Shift in Omitted Relevant Regressor and Outliers in Regressand

Super saturation successfully corrects for shifts in omitted but relevant covariates, which induce a shift in the dependent variable. This improves model selection and estimation of the coefficients in the model.

This experiment continues with outliers in y and a location shift in a relevant regressor, but this time, the relevant regressor with the location shift is excluded from the estimation. To investigate how this affects covariates with small and large non-centralities alike, we impose the location shift in regressor x_3. The GUMs all omit x_3 in their estimation.

The results are presented in Table 3.5.6. Estimating GUM0 and GUM1 without selection produces qualitatively similar results as when the location shift occurred in y. This is because the shift in x_3 induces a shift in y, albeit smaller than in Section 3.5.2. The error variance is overestimated, which then leads to a loss in potency and increased biases and RMSEs for the relevant regressors. For GUM1, Autometrics again yields higher potencies of relevant covariates but also a higher gauge of irrelevant covariates compared to the 1-cut procedure.

Table 3.5.6: Comparison of Reduction, Outliers in y and Mean Shift in x_3 but Omitted in Estimation

Model:	GUM0					GUM0-S					GUM1					GUM1-S				
Selection:	None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto	
Statistic:		U	C	U	C		U	C	U	C		U	C	U	C		U	C	U	C
Bias/SE_{Th}																				
$x_1(\xi=2)$	0.000	-1.031	1.997	-0.926	1.642	0.010	-0.615	1.180	-0.593	1.140	0.000	-1.081	2.020	-0.955	1.763	0.001	-0.656	1.224	-0.597	1.192
$x_2(\xi=3)$	-0.019	-1.006	1.307	-0.909	1.147	-0.025	-0.445	0.617	-0.428	0.599	-0.021	-1.076	1.339	-0.938	1.205	-0.016	-0.478	0.667	-0.425	0.645
x_3(omitted)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
$x_4(\xi=6)$	0.018	-0.091	0.165	-0.081	0.158	0.022	0.011	0.039	0.011	0.039	0.019	-0.127	0.189	-0.097	0.172	0.026	0.008	0.051	0.012	0.047
$x_5(\xi=8)$	0.002	-0.008	0.007	-0.007	0.007	-0.015	-0.018	-0.017	-0.018	-0.017	0.004	-0.013	0.012	-0.008	0.011	0.003	-0.001	0.000	-0.001	-0.001
SE/SE_{Th}																				
$x_1(\xi=2)$	1.585	0.376	1.553	0.460	1.560	1.128	0.480	1.102	0.495	1.104	1.629	0.353	1.545	0.431	1.552	1.162	0.459	1.101	0.485	1.103
$x_2(\xi=3)$	1.583	0.720	1.556	0.787	1.561	1.126	0.783	1.108	0.793	1.109	1.626	0.686	1.548	0.761	1.552	1.160	0.762	1.108	0.783	1.109
x_3(omitted)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
$x_4(\xi=6)$	1.582	1.508	1.573	1.512	1.573	1.126	1.118	1.123	1.118	1.123	1.625	1.486	1.566	1.497	1.565	1.160	1.117	1.124	1.116	1.123
$x_5(\xi=8)$	1.582	1.573	1.576	1.573	1.575	1.126	1.123	1.123	1.123	1.123	1.625	1.565	1.570	1.564	1.568	1.160	1.125	1.125	1.124	1.124
SD/SE_{Th}																				
$x_1(\xi=2)$	1.587	1.771	0.913	1.796	1.259	1.336	1.675	0.857	1.672	0.893	1.633	1.746	0.935	1.793	1.163	1.406	1.699	0.930	1.696	0.914
$x_2(\xi=3)$	1.583	2.247	0.973	2.223	1.130	1.347	1.844	0.987	1.831	0.997	1.625	2.257	1.004	2.242	1.112	1.414	1.901	1.026	1.869	1.022
x_3(omitted)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
$x_4(\xi=6)$	1.594	1.892	1.468	1.870	1.472	1.357	1.406	1.347	1.405	1.347	1.641	1.976	1.468	1.917	1.477	1.424	1.464	1.379	1.453	1.381
$x_5(\xi=8)$	1.602	1.646	1.610	1.643	1.609	1.341	1.355	1.353	1.354	1.352	1.640	1.673	1.615	1.661	1.618	1.401	1.389	1.387	1.387	1.387
$\text{RMSE}/\text{SE}_{Th}$																				
$x_1(\xi=2)$	1.587	2.049	2.196	2.021	2.069	1.336	1.784	1.458	1.774	1.448	1.633	2.054	2.226	2.032	2.112	1.406	1.821	1.537	1.798	1.502
$x_2(\xi=3)$	1.583	2.462	1.630	2.402	1.610	1.347	1.897	1.164	1.880	1.163	1.625	2.500	1.674	2.430	1.639	1.414	1.960	1.224	1.917	1.208
x_3(omitted)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
$x_4(\xi=6)$	1.594	1.894	1.477	1.871	1.480	1.357	1.406	1.348	1.405	1.347	1.641	1.980	1.480	1.920	1.487	1.424	1.463	1.380	1.453	1.382
$x_5(\xi=8)$	1.602	1.646	1.610	1.643	1.609	1.341	1.355	1.353	1.354	1.352	1.639	1.673	1.615	1.661	1.618	1.401	1.389	1.387	1.387	1.387
$\widehat{\sigma}_u$	1.539	1.545		1.544		1.056	1.060		1.059		1.539	1.537		1.533		1.059	1.060		1.058	
Retention																				
$\widehat{\gamma}^{\text{covariates}}$	–	–		–		–	–		–		–	0.050		0.091		–	0.102		0.113	
$\widehat{\rho}^{\text{covariates}}$	–	0.665		0.690		–	0.784		0.789		–	0.654		0.681		–	0.774		0.785	
$\widehat{\rho}^{x1}$	–	0.242		0.295		–	0.436		0.448		–	0.228		0.278		–	0.417		0.440	
$\widehat{\rho}^{x2}$	–	0.463		0.504		–	0.707		0.715		–	0.443		0.490		–	0.688		0.706	
$\widehat{\rho}^{x3}$	–	–		–		–	–		–		–	–		–		–	–		–	
$\widehat{\rho}^{x4}$	–	0.959		0.961		–	0.995		0.995		–	0.949		0.956		–	0.993		0.994	
$\widehat{\rho}^{x5}$	–	0.998		0.998		–	1.000		1.000		–	0.997		0.998		–	1.000		1.000	
$\widehat{\gamma}^{I}$	–	–		–		0.005	0.005		0.005		–	–		–		0.005	0.005		0.005	
$\widehat{\gamma}^{S}$	–	–		–		0.019	0.019		0.019		–	–		–		0.020	0.020		0.020	
$\widehat{\rho}^{I}$	–	–		–		0.677	0.676		0.676		–	–		–		0.633	0.633		0.634	
$\widehat{\rho}^{S}$	–	–	–	–	–	0.452	0.452	0.452	0.452	0.452	–	–	–	–	–	0.424	0.425	0.425	0.425	0.425
$\widehat{\rho}^{S25}+\widehat{\gamma}^{S24,S26}$	–	–	–	–	–	0.641	0.641	0.641	0.641	0.641	–	–	–	–	–	0.618	0.618	0.618	0.618	0.618
$\widehat{\rho}^{S25}+\widehat{\gamma}^{S23,S24,S26,S27}$	–	–	–	–	–	0.751	0.751	0.751	0.751	0.751	–	–	–	–	–	0.729	0.729	0.729	0.729	0.729

Note: Monte Carlo experiment for a DGP with covariates, a mean shift in the relevant but omitted covariate x_3 in time period $T_1 = 25$, and four outliers in y in time periods $T_2 = 60, T_3 = 70, T_4 = 80, T_5 = 90$. GUM0 includes the covariates x_1, x_2, x_4, x_5, GUM1 in addition contains the covariates $x_6 - x_{10}$. GUM0-S and GUM1-S apply super saturation before selecting over covariates. Super saturation has been implemented using Autometrics and there are three types of covariate selection: "None" without selection, "1-Cut" drops insignificant covariates in a single step, and "Auto" uses general-to-specific model selection as implemented in Autometrics. The columns denoted by "U" show the unconditional version of the statistics, including replications where the covariate has been removed, and "C" shows the conditional version of the statistics excluding replications where the variable has been removed. "Bias" refers to the average bias of the estimated coefficient relative to its true coefficient, "SE" to the average standard error of the coefficient estimate, "SD" to the Monte Carlo standard deviation of the estimated coefficient across replications, and "RMSE" to the Monte Carlo root mean square error across replications. All these statistics are scaled by the theoretical standard error, "SE_{th}". The average estimated standard deviation of the error term is denoted by $\widehat{\sigma}_u$. The second part of the table reports the retention rates of relevant and irrelevant covariates and indicators. $\widehat{\gamma}$ refers to the average gauge and $\widehat{\rho}$ to the average potency across replications. The superscript denotes the variable or category of variables for which the gauge and potency have been calculated. Settings: $M = 10{,}000$ replications, selection significance level for indicators $\gamma^{\text{indicators}} = 0.01$, selection significance level for covariates $\gamma^{\text{covariates}} = 0.05$, backtesting off, and default diagnostic testing.

Applying super saturation in GUM0-S and GUM1-S allows us to correct for the omission of x_3. In GUM0-S, a step indicator that matches time period $t = 25$ exactly is retained in 45% of the replications, and within a neighbourhood of ± 2 time periods in 75% of the replications. These potencies are only slightly lower when starting from GUM1-S. The outliers in y are also detected in more than 60% of the cases, resulting in estimates of the error variance much closer to its true value of unity. The indicator gauge is well-controlled. Our evaluation of the gauge is very conservative in that if the step indicator (or a pair of step indicator and impulse indicator) does not exactly end in time period 25, it counts towards the gauge. So the step gauge of 0.019 in GUM0-S and of 0.02 in GUM1-S already includes the slightly mistimed step indicators in the neighbourhood of time period $t = 25$, even though their retention is still desirable compared to leaving the induced location shift unmodelled.

Taking the induced location shift into account improves all metrics across all coefficient estimates. The improved estimate of the error variance also increases the potency of relevant covariates, while only slightly increasing the gauge of irrelevant covariates.

3.5.5 DGP with Location Shift in Irrelevant Regressor and Outliers in Regressand

A location shift in an irrelevant covariate that is considered in the model selection does not cause super saturation to adventitiously retain more irrelevant indicators. Super saturation can also help take outliers and location shifts into account, which may otherwise lead to the incorrect retention of nonlinear covariates that otherwise proxy for undetected outliers and location shifts.

We modify the DGP to contain outliers in y and a location shift in x_6, one of the irrelevant covariates. Since x_6 is irrelevant, it does not shift y but it does shift the model if it is falsely retained.

We now also report the average bias, standard error, standard deviation, and RMSE for the coefficient estimates of x_6 in Table 3.5.7. Without super saturation

Table 3.5.7: Comparison of Reduction, Outliers in y and Mean Shift in x_6

Model:	GUM0					GUM0-S					GUM1					GUM1-S				
Selection:	None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto		None	1-Cut		Auto	
Statistic:		U	C	U	C		U	C	U	C		U	C	U	C		U	C	U	C
Bias/SE$_{Th}$																				
$x_1(\xi=2)$	-0.001	-0.887	1.442	-0.809	1.211	0.000	-0.563	0.961	-0.543	0.932	-0.005	-0.939	1.467	-0.813	1.260	0.002	-0.598	0.995	-0.547	0.969
$x_2(\xi=3)$	-0.028	-0.705	0.802	-0.647	0.731	-0.022	-0.323	0.412	-0.308	0.399	-0.023	-0.758	0.832	-0.667	0.769	-0.021	-0.384	0.458	-0.337	0.433
$x_3(\xi=4)$	0.007	-0.277	0.354	-0.250	0.332	0.009	-0.078	0.147	-0.075	0.144	0.004	-0.333	0.384	-0.285	0.360	0.011	-0.103	0.178	-0.085	0.166
$x_4(\xi=6)$	0.021	0.004	0.046	0.003	0.048	0.027	0.027	0.032	0.027	0.032	0.023	-0.001	0.052	0.001	0.051	0.024	0.022	0.030	0.024	0.031
$x_5(\xi=8)$	0.003	-0.001	-0.001	-0.001	-0.001	0.007	0.007	0.007	0.006	0.007	0.005	0.003	0.003	0.004	0.004	0.007	0.003	0.004	0.003	0.003
$x_6(\xi=0)$	–	–	–	–	–	–	–	–	–	–	-0.782	-0.148	-2.873	-0.255	-2.165	-0.030	-0.013	-0.124	-0.013	-0.110
SE/SE$_{Th}$																				
$x_1(\xi=2)$	1.314	0.418	1.291	0.480	1.294	1.028	0.488	1.006	0.500	1.007	1.351	0.393	1.285	0.469	1.289	1.058	0.470	1.005	0.492	1.006
$x_2(\xi=3)$	1.313	0.781	1.294	0.817	1.295	1.027	0.796	1.014	0.804	1.015	1.349	0.754	1.288	0.798	1.290	1.057	0.766	1.012	0.785	1.012
$x_3(\xi=4)$	1.313	1.112	1.301	1.127	1.302	1.027	0.966	1.021	0.967	1.021	1.349	1.083	1.295	1.104	1.295	1.057	0.952	1.020	0.958	1.019
$x_4(\xi=6)$	1.312	1.300	1.309	1.299	1.309	1.026	1.024	1.025	1.024	1.025	1.348	1.293	1.304	1.292	1.303	1.056	1.024	1.025	1.022	1.023
$x_5(\xi=8)$	1.312	1.310	1.310	1.309	1.309	1.026	1.025	1.025	1.025	1.025	1.348	1.305	1.305	1.304	1.304	1.056	1.025	1.025	1.023	1.023
$x_6(\xi=0)$	–	–	–	–	–	–	–	–	–	–	1.331	0.065	1.261	0.150	1.276	1.162	0.132	1.216	0.142	1.204
SD/SE$_{Th}$																				
$x_1(\xi=2)$	1.320	1.669	0.769	1.673	1.030	1.194	1.572	0.761	1.567	0.784	1.353	1.659	0.811	1.677	0.984	1.245	1.594	0.810	1.586	0.801
$x_2(\xi=3)$	1.318	1.987	0.901	1.954	0.955	1.192	1.626	0.929	1.611	0.936	1.356	2.015	0.919	1.978	0.952	1.249	1.697	0.946	1.659	0.950
$x_3(\xi=4)$	1.313	1.829	1.080	1.792	1.091	1.187	1.404	1.073	1.397	1.075	1.342	1.899	1.081	1.846	1.090	1.242	1.480	1.083	1.447	1.087
$x_4(\xi=6)$	1.319	1.396	1.308	1.397	1.302	1.199	1.221	1.209	1.221	1.208	1.351	1.424	1.314	1.416	1.311	1.249	1.254	1.236	1.252	1.237
$x_5(\xi=8)$	1.320	1.344	1.344	1.342	1.342	1.198	1.213	1.213	1.214	1.211	1.358	1.358	1.358	1.358	1.358	1.254	1.246	1.243	1.243	1.243
$x_6(\xi=0)$	–	–	–	–	–	–	–	–	–	–	1.138	0.680	1.065	0.810	1.201	1.555	1.131	3.426	1.152	3.357
RMSE/SE$_{Th}$																				
$x_1(\xi=2)$	1.320	1.890	1.635	1.858	1.590	1.194	1.670	1.226	1.658	1.218	1.352	1.907	1.676	1.864	1.598	1.245	1.703	1.283	1.678	1.257
$x_2(\xi=3)$	1.318	2.108	1.206	2.058	1.203	1.192	1.658	1.016	1.640	1.017	1.357	2.153	1.239	2.087	1.223	1.250	1.739	1.051	1.693	1.044
$x_3(\xi=4)$	1.313	1.850	1.136	1.809	1.140	1.187	1.406	1.083	1.399	1.084	1.342	1.928	1.147	1.867	1.147	1.242	1.483	1.097	1.449	1.100
$x_4(\xi=6)$	1.319	1.396	1.308	1.397	1.303	1.200	1.221	1.210	1.221	1.209	1.351	1.424	1.315	1.416	1.312	1.250	1.255	1.237	1.253	1.237
$x_5(\xi=8)$	1.320	1.344	1.344	1.342	1.342	1.198	1.213	1.213	1.214	1.211	1.358	1.358	1.358	1.358	1.358	1.254	1.246	1.243	1.243	1.243
$x_6(\xi=0)$	–	–	–	–	–	–	–	–	–	–	1.380	0.696	3.063	0.849	2.475	1.555	1.131	3.426	1.152	3.358
$\widehat{\sigma}_u$	1.270	1.276		1.275		0.959	0.962		0.962		1.270	1.270		1.266		0.960	0.961		0.959	
Retention																				
$\widehat{\gamma}^{\text{covariates}}$	–	–		–		–	–		–		–	0.050		0.110		–	0.096		0.107	
$\widehat{\gamma}^{x6}$	–	–		–		–	–		–		–	0.052		0.118		–	0.109		0.118	
$\widehat{\rho}^{\text{covariates}}$	–	0.755		0.772		–	0.843		0.847		–	0.744		0.765		–	0.831		0.841	
$\widehat{\rho}^{x1}$	–	0.323		0.371		–	0.485		0.497		–	0.306		0.364		–	0.468		0.490	
$\widehat{\rho}^{x2}$	–	0.604		0.631		–	0.785		0.792		–	0.585		0.619		–	0.757		0.776	
$\widehat{\rho}^{x3}$	–	0.855		0.866		–	0.946		0.947		–	0.837		0.852		–	0.933		0.940	
$\widehat{\rho}^{x4}$	–	0.993		0.993		–	0.999		0.999		–	0.991		0.992		–	0.999		0.999	
$\widehat{\rho}^{x5}$	–	1.000		1.000		–	1.000		1.000		–	1.000		1.000		–	1.000		1.000	
$\widehat{\gamma}^{I}$	–	–		–		0.006	0.006		0.006		–	–		–		0.006	0.006		0.006	
$\widehat{\gamma}^{S}$	–	–		–		0.015	0.015		0.015		–	–		–		0.015	0.015		0.015	
$\widehat{\gamma}^{S25}$	–	–		–		0.011	0.011		0.011		–	–		–		0.009	0.009		0.009	
$\widehat{\rho}^{I}$	–	–		–		0.781	0.781		0.781		–	–		–		0.752	0.753		0.753	

Note: Monte Carlo experiment for a DGP with covariates, a mean shift in the irrelevant covariate x_6 in time period $T_1 = 25$, and four outliers in y in time periods $T_2 = 60, T_3 = 70, T_4 = 80, T_5 = 90$. GUM0 includes the covariates $x_1 - x_5$, GUM1 in addition contains the covariates $x_6 - x_{10}$. GUM0-S and GUM1-S apply super saturation before selecting over covariates. Super saturation has been implemented using Autometrics and there are three types of covariate selection: "None" without selection, "1-Cut" drops insignificant covariates in a single step, and "Auto" uses general-to-specific model selection as implemented in Autometrics. The columns denoted by "U" show the unconditional version of the statistics, including replications where the covariate has been removed, and "C" shows the conditional version of the statistics excluding replications where the variable has been removed. "Bias" refers to the average bias of the estimated coefficient relative to its true coefficient, "SE" to the average standard error of the coefficient estimate, "SD" to the Monte Carlo standard deviation of the estimated coefficient across replications, and "RMSE" to the Monte Carlo root mean square error across replications. All these statistics are scaled by the theoretical standard error, "SE$_{th}$". The average estimated standard deviation of the error term is denoted by $\widehat{\sigma}_u$. The second part of the table reports the retention rates of relevant and irrelevant covariates and indicators. $\widehat{\gamma}$ refers to the average gauge and $\widehat{\rho}$ to the average potency across replications. The superscript denotes the variable or category of variables for which the gauge and potency have been calculated. Settings: $M = 10{,}000$ replications, selection significance level for indicators $\gamma^{\text{indicators}} = 0.01$, selection significance level for covariates $\gamma^{\text{covariates}} = 0.05$, backtesting off, and default diagnostic testing.

(GUM1), we see that the coefficient of x_6 is negatively biased. This is true even without selection, since x_6 can partially proxy for the outliers in y because they are negatively correlated. Super saturation corrects for this bias to a large degree by capturing the outliers in y. With model selection, x_6 is now retained much more randomly, that is, equally for large positive and large negative values of its t-statistic, as we would expect for an irrelevant regressor.

In general, x_6 is not retained much more often than the other irrelevant regressors. In GUM1 with 1-cut selection, the gauge of irrelevant covariates is 0.05 overall and 0.052 for x_6. With Autometrics, the corresponding gauges are higher, 0.11 and 0.118. As before, Autometrics probably retains more irrelevant regressors to improve the diagnostic tests. This is supported by the fact that in GUM1-S, super saturation does not increase the gauge of covariate x_6 despite reducing $\widehat{\sigma}_\epsilon$ by almost 25%, bringing it closer to its true value of 1. In contrast, the better estimate of the error variance increases the gauge of irrelevant covariates for 1-cut selection — but again not substantially more for x_6 than the other insignificant covariates.

3.6 Conclusion

In this chapter, we used an extensive set of Monte Carlo simulations to evaluate the performance of super saturation, a combination of impulse and step indicator saturation intended to detect outliers and location shifts of unknown magnitude and timing, respectively. Both phenomena can be present in time series data, potentially distorting inference, model selection, and forecasting and hence making it important to be able to detect them simultaneously. While IIS and SIS are each flexible enough to detect either of the two phenomena, they have higher power to detect the phenomenon that they are explicitly designed for and do not model the other one parsimoniously. It is therefore preferable to use IIS and SIS jointly via super saturation.

To evaluate the performance of super saturation, we generalise the existing metrics of gauge and potency to allow for combinations of impulse and step indicators

to model outliers and location shifts. We find that under the null of no outliers or location shifts, the gauge is well-controlled when using a tight significance level. While diagnostic testing has a negligible impact on the gauge when the model is well-specified, backtesting tends to increase the gauge substantially. We therefore recommend choosing a stringent significance level for selecting indicators of at most 1%, turning backtesting off, and turning any desired diagnostic tests on since they are costless in a well-specified model but might otherwise yield a better terminal model. If super saturation is followed by model selection or inference, one might want to choose an even tighter significance level to avoid overfitting and hence underestimating the error variance.

Overall, we conclude that super saturation is well-suited for its intended purpose. In DGPs with outliers, location shifts, or both, we find that super saturation reaches high detection rates of such breaks. It successfully detects outliers and location shifts, including when they are indirectly induced by omitted relevant covariates, and improves subsequent model selection outcomes and parameter estimates. A location shift in an irrelevant covariate does not lead to spurious retention of a matching step indicator.

As with any Monte Carlo study, a major limitation of our analysis is that the results are applicable to the specific parameter settings that we included and might not generalise. Our estimated response surfaces for the gauge attempt to reduce this specificity. They are well-behaved for our recommendation of using tight significance levels and no backtesting, so they can be used to set the gauge according to the researcher's tolerance for the mean and variability of the gauge under the null. But ideally we would like to have an asymptotic theory for the gauge when using super saturation, analogous to the separate analyses for IIS in Johansen and Nielsen (2016b) and SIS in Nielsen and Qian (2023). However, analytical treatment has proven to be difficult already in the isolated cases of IIS and SIS. Therefore, Monte Carlo simulations will likely remain an integral part of analysing the performance of indicator saturation methods in the future.

As an extension, it would be interesting to compare the performance of different

ways of implementing super saturation. We use Autometrics (Doornik 2009) for all our analyses, but super saturation can also be implemented in other ways, for example in the R package *gets* (Pretis, Reade and Sucarrat 2018) or using the Lasso estimator (Tibshirani 1996).

Appendices of Chapter 3

3.A Regressors in Response Surface Modelling

We use general-to-specific modelling to estimate a response surface for the mean and variance of both impulse and step indicators. The sets of variables included in the search are as follows.

Mean of Impulse and Step Gauge

For the baseline response surface without backtesting or diagnostic testing:

- terms to test theory (fixed): Intercept, γ
- polynomial: γ^2
- finite sample corrections: $1/T$, $(1/T)^2$, γ/T, γ^2/T

Additional terms for the full set of simulations:

- intercept heterogeneity: $1_{(\text{back=1})}$, $1_{(\text{AR=1})}$, $1_{(\text{ARCH=1})}$, $1_{(\text{Chow=1})}$, $1_{(\text{Heterosk.=1})}$, $1_{(\text{Normality=1})}$
- slope heterogeneity: $1_{(\text{back=1})} \times \gamma$, $1_{(\text{AR=1})} \times \gamma$, $1_{(\text{ARCH=1})} \times \gamma$, $1_{(\text{Chow=1})} \times \gamma$, $1_{(\text{Heterosk.=1})} \times \gamma$, $1_{(\text{Normality=1})} \times \gamma$
- finite sample corrections: $1_{(\text{back=1})}/T$, $1_{(\text{AR=1})}/T$, $1_{(\text{ARCH=1})}/T$, $1_{(\text{Chow=1})}/T$, $1_{(\text{Heterosk.=1})}/T$, $1_{(\text{Normality=1})}/T$, $1_{(\text{back=1})} \times \gamma/T$, $1_{(\text{AR=1})} \times \gamma/T$, $1_{(\text{ARCH=1})} \times \gamma/T$, $1_{(\text{Chow=1})} \times \gamma/T$, $1_{(\text{Heterosk.=1})} \times \gamma/T$, $1_{(\text{Normality=1})} \times \gamma/T$

Variance of Impulse and Step Gauge

For the baseline response surface without backtesting or diagnostic testing:

- terms to test theory (fixed): Intercept, $\ln(\omega_{j,2}^K)$
- polynomial: $\gamma \times \ln(\omega_{j,2}^K)$, $\gamma^2 \times \ln(\omega_{j,2}^K)$
- finite sample corrections: $1/T$, $(1/T)^2$, $\ln(\omega_{j,2}^K)/T$, $\gamma \times \ln(\omega_{j,2}^K)/T$, $\gamma^2 \times \ln(\omega_{j,2}^K)/T$

Additional terms for the full set of simulations:

- intercept heterogeneity: $1_{(\text{back=1})}$, $1_{(\text{AR=1})}$, $1_{(\text{ARCH=1})}$, $1_{(\text{Chow=1})}$, $1_{(\text{Heterosk.=1})}$, $1_{(\text{Normality=1})}$
- slope heterogeneity: $1_{(\text{back=1})} \times \ln(\omega_{j,2}^K)$, $1_{(\text{AR=1})} \times \ln(\omega_{j,2}^K)$, $1_{(\text{ARCH=1})} \times \ln(\omega_{j,2}^K)$, $1_{(\text{Chow=1})} \times \ln(\omega_{j,2}^K)$, $1_{(\text{Heterosk.=1})} \times \ln(\omega_{j,2}^K)$, $1_{(\text{Normality=1})} \times \ln(\omega_{j,2}^K)$
- finite sample corrections: $1_{(\text{back=1})}/T$, $1_{(\text{AR=1})}/T$, $1_{(\text{ARCH=1})}/T$, $1_{(\text{Chow=1})}/T$, $1_{(\text{Heterosk.=1})}/T$, $1_{(\text{Normality=1})}/T$, $1_{(\text{back=1})} \times \ln(\omega_{j,2}^K)/T$, $1_{(\text{AR=1})} \times \ln(\omega_{j,2}^K)/T$, $1_{(\text{ARCH=1})} \times \ln(\omega_{j,2}^K)/T$, $1_{(\text{Chow=1})} \times \ln(\omega_{j,2}^K)/T$, $1_{(\text{Heterosk.=1})} \times \ln(\omega_{j,2}^K)/T$, $1_{(\text{Normality=1})} \times \ln(\omega_{j,2}^K)/T$

3.B Response Surfaces for All Simulations

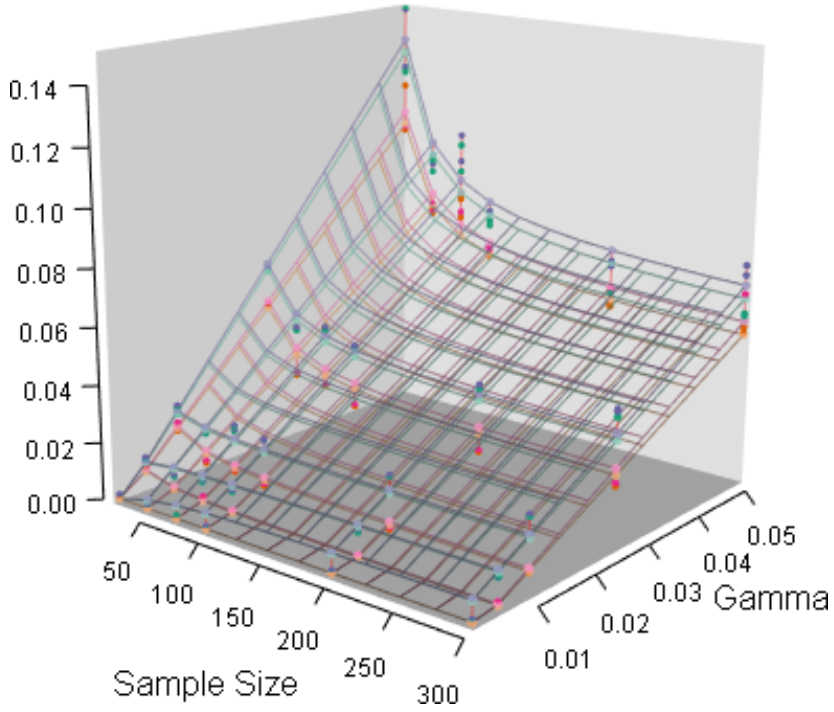
Figure 3.B.1 (3.B.2) displays the response surface with selection for the mean (variance) of the gauge using all 420 Monte Carlo simulations, where negative predictions of the mean were replaced with zero. Since the response surfaces depend on more parameters than just the significance level γ and the sample size T, we need to plot them separately for the other parameter values.

The response surfaces for the mean of the step gauge and for the variance of both the impulse and the step gauge contain the indicator variable $1_{(\text{back=1})}$. Therefore, two separate grids need to be plotted. The grid without backtesting is shown in orange, whereas the grid with backtesting is shown in green.

The response surface for the mean of the impulse gauge additionally contains the indicator variable $1_{(\text{AR=1})}$ for the AR test (Godfrey 1978). Hence, there are four different variations of the grid, which are shown in separate colours. The orange grid is without backtesting and without the AR test. The green grid is with backtesting but without the AR test. The pink grid is without backtesting but with the AR test. And finally, the purple grid is with both backtesting and the AR test.

Figure 3.B.1: Response Surfaces for the Mean of the Gauge

(a) Impulse Gauge



(b) Step Gauge

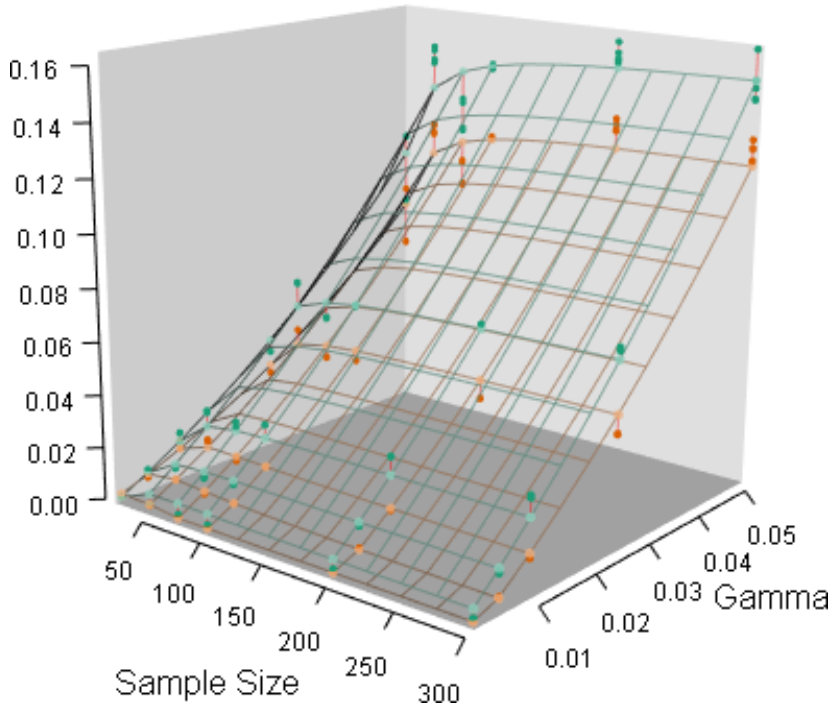
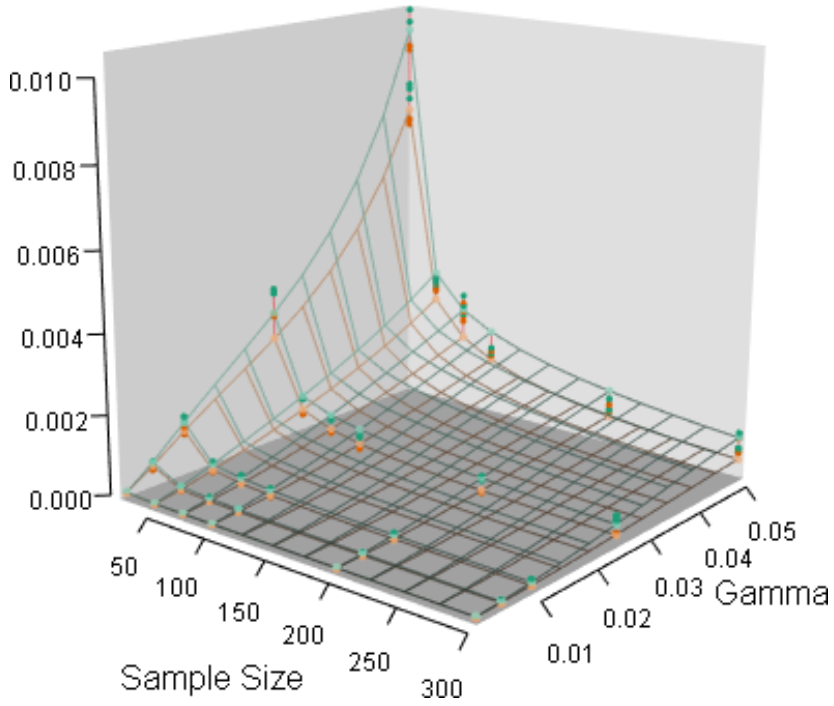
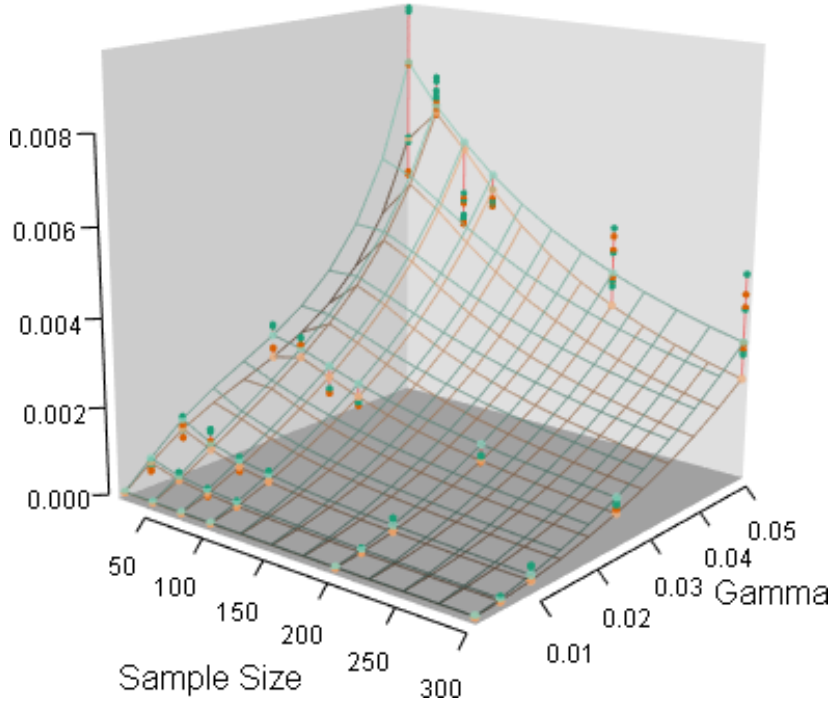


Figure 3.B.2: Response Surfaces for the Variance of the Gauge

(a) Impulse Gauge



(b) Step Gauge



Conclusion

This thesis analysed methods for detecting outliers and location shifts in econometric models, which can otherwise distort coefficient estimates and statistical inference as well as cause systematic forecast failure and non-constancy of the model parameters.

In Chapter 1, we introduced outlier detection algorithms in the cross-sectional, instrumental variables regression setting along with the false outlier detection rate (FODR), which we then studied asymptotically. We showed the tightness of the FODR, uniformly both in the cut-off value and the iteration step, and discussed the importance of bias correcting the variance estimator when the algorithms are iterated because the probability limit of the FODR will otherwise be higher than the researcher's target. Considering the FODR as a stochastic process with varying cut-off value and iteration step, we derived a Gaussian approximation for the share and a Poisson approximation for the number of falsely detected outliers in the sample.

Chapter 2 built on the asymptotic theory derived in the previous chapter and proposed statistical tests for testing the null hypothesis that there are no outliers in the sample. Monte Carlo simulations showed that in general the size can be well-controlled and that the power tends to increase in the outlier frequency, outlier magnitude, and sample size. We recommended using scaling tests across a range of cut-off values that are based on iterated versions of the outlier detection algorithms in order to control size while achieving high power. To illustrate the usage of the tests, we re-analysed the seminal paper by Angrist and Krueger (1991) on the impact of schooling on earnings and found that the returns to education might be slightly lower than suggested in the original analysis when excluding outliers. The algorithms and statistical tests from Chapters 1 and 2 are readily available for applied researchers from the R package *robust2sls* (Kurle 2022b).

In Chapter 3, we conducted a Monte Carlo study of super saturation as a method for jointly detecting outliers and location shifts in time series regression models. We

estimated response surfaces for the false detection rate (gauge) as a function of several tuning parameters of the selection algorithm Autometrics (Doornik 2009). Our results suggest that in order to control the gauge, researchers should choose tight significance levels of less than 1%, turn parsimonious encompassing testing (backtesting) off, and diagnostic testing on. Super saturation achieves high detection rates (potency) of outliers and location shifts. Furthermore, it can improve subsequent covariate selection in the presence of such phenomena by improving estimates of the error variance and by preventing the false retention of irrelevant covariates that might otherwise proxy for unmodelled outliers and location shifts.

The limitations of this thesis provide avenues for future research. The results in Chapter 1 rely on specific assumptions about the distribution of the errors, which is further restricted to normality in Chapter 2. The tests can be generalised to more distributions by finding (robust) estimators for the parameters which appear in the formulas of Chapter 1 but cancel under normality. This might require the derivation of new empirical processes. Moreover, the algorithms analysed in the first two chapters only use the magnitude of the structural error for their outlier classification. It would be important to suggest and analyse other outlier detection algorithms that also consider the magnitude of the first stage errors. In relation to this, it would be interesting to analyse the algorithms in terms of their influence function or breakdown point in order to formalise their robustness properties.

Finally, future work could compare the performance of super saturation implemented in Autometrics to other model selection methods, such as the R package *gets* (Pretis, Reade and Sucarrat 2018) or the LASSO estimator (Tibshirani 1996). A common set of simulations could not only serve as a benchmark for performance comparisons across methods but also to monitor any changes in performance in case the selection methods change, such as alterations in the search algorithm.

Overall, this thesis advanced our understanding of different methods that are commonly used to detect and address outliers and location shifts through both theoretical, asymptotic analysis and Monte Carlo studies. It is hoped that these insights improve the robustness and accuracy of empirical research.

References

Acemoglu, Daron, Simon Johnson and James A Robinson. 2001. 'The Colonial Origins of Comparative Development: An Empirical Investigation'. *American Economic Review* 91, no. 5 (December): 1369–1401. https://doi.org/10.1257/aer.91.5.1369.

———. 2012. 'The Colonial Origins of Comparative Development: An Empirical Investigation: Reply'. *American Economic Review* 102, no. 6 (October): 3077–3110. https://doi.org/10.1257/aer.102.6.3077.

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo and James A Robinson. 2019. 'Democracy Does Cause Growth'. *Journal of Political Economy* 127, no. 1 (February): 47–100. https://doi.org/10.1086/700936.

Albouy, David Y. 2012. 'The Colonial Origins of Comparative Development: An Empirical Investigation: Comment'. *American Economic Review* 102, no. 6 (October): 3059–3076. https://doi.org/10.1257/aer.102.6.3059.

Anderson, T. W. and Herman Rubin. 1949. 'Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations'. *The Annals of Mathematical Statistics* 20, no. 1 (March): 46–63. https://doi.org/10.1214/aoms/1177730090.

Angrist, Joshua D. and Alan B. Krueger. 1991. 'Does Compulsory School Attendance Affect Schooling and Earnings?' *The Quarterly Journal of Economics* 106, no. 4 (November): 979–1014. https://doi.org/10.2307/2937954.

Apergis, Nicholas, Wei-Fong Pan, James Reade and Shixuan Wang. 2023. 'Modelling Australian Electricity Prices Using Indicator Saturation'. *Energy Economics* 120 (April): 106616. https://doi.org/10.1016/j.eneco.2023.106616.

Bancroft, T. A. 1944. 'On Biases in Estimation Due to the Use of Preliminary Tests of Significance'. *The Annals of Mathematical Statistics* 15, no. 2 (June): 190–204. https://doi.org/10.1214/aoms/1177731284.

Barbato, G., E. M. Barini, G. Genta and R. Levi. 2011. 'Features and Performance of Some Outlier Detection Methods'. *Journal of Applied Statistics* 38, no. 10 (October): 2133–2149. https://doi.org/10.1080/02664763.2010.545119.

Barnett, Vic. 1978. 'The Study of Outliers: Purpose and Model'. *Journal of the Royal Statistical Society Series C: Applied Statistics* 27, no. 3 (November): 242–250. https://doi.org/10.2307/2347159.

Benjamini, Yoav and Yosef Hochberg. 1995. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, no. 1 (January): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Berenguer-Rico, Vanessa, Søren Johansen and Bent Nielsen. 2023. 'A Model Where the Least Trimmed Squares Estimator Is Maximum Likelihood'. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85, no. 3 (July): 886–912. https://doi.org/10.1093/jrsssb/qkad028.

Berenguer-Rico, Vanessa and Bent Nielsen. 2023a. 'Least Trimmed Squares Asymptotics: Regression with Leverage', Nuffield College Economics Discussion Papers, no. 2023-W01 (May).

———. 2023b. 'Normality Testing after Outlier Removal'. *Econometrics and Statistics* (June). https://doi.org/10.1016/j.ecosta.2023.06.001.

Berenguer-Rico, Vanessa and Ines Wilms. 2021. 'Heteroscedasticity Testing after Outlier Removal'. *Econometric Reviews* 40, no. 1 (January): 51–85. https://doi.org/10.1080/07474938.2020.1735749.

Billingsley, Patrick. 1968. *Convergence of Probability Measures.* New York: John Wiley & Sons.

Broderick, Tamara, Ryan Giordano and Rachael Meager. 2023. 'An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?', Working Paper (July). https://doi.org/10.48550/arXiv.2011.14999.

Brys, Guy, Mia Hubert and Anja Struyf. 2004. 'A Robustification of the Jarque-Bera Test of Normality'. In *COMPSTAT 2004 - Proceedings in Computational Statistics,* edited by Jaromir Antoch, 753–760. Heidelberg: Physica-Verlag. ISBN: 978-3-7908-1554-2.

Castle, Jennifer L., Jurgen A. Doornik and David F. Hendry. 2011. 'Evaluating Automatic Model Selection'. *Journal of Time Series Econometrics* 3, no. 1 (January): 8. https://doi.org/10.2202/1941-1928.1097.

Castle, Jennifer L., Jurgen A. Doornik, David F. Hendry and Felix Pretis. 2015. 'Detecting Location Shifts during Model Selection by Step-Indicator Saturation'. *Econometrics* 3, no. 2 (April): 240–264. https://doi.org/10.3390/econometrics3020240.

Castle, Jennifer L., Jurgen A. Doornik, David F. Hendry and Felix Pretis. 2022. 'Detecting Breaks in Trends by Trend-indicator Saturation', Mimeo (September).

Castle, Jennifer L. and David F. Hendry. 2009. 'The Long-Run Determinants of UK Wages, 1860–2004'. *Journal of Macroeconomics* 31, no. 1 (March): 5–28. https://doi.org/10.1016/j.jmacro.2007.08.018.

———. 2014. 'Model Selection in Under-Specified Equations Facing Breaks'. *Journal of Econometrics* 178 (January): 286–293. https://doi.org/10.1016/j.jeconom.2013.08.028.

———. 2019. *Modelling Our Changing World.* 1st ed. Edited by Michael P. Clements. Palgrave Texts in Econometrics. Cham: Palgrave Pivot. ISBN: 978-3-030-21431-9. https://doi.org/10.1007/978-3-030-21432-6.

———. 2020. 'Climate Econometrics: An Overview'. *Foundations and Trends® in Econometrics* 10, nos. 3-4 (August): 145–322. https://doi.org/10.1561/0800000037.

Castle, Jennifer L., David F. Hendry and Andrew B. Martinez. 2017. 'Evaluating Forecasts, Narratives and Policy Using a Test of Invariance'. *Econometrics* 5, no. 3 (September): 39. https://doi.org/10.3390/econometrics5030039.

———. 2023. 'The Historical Role of Energy in UK Inflation and Productivity with Implications for Price Inflation'. *Energy Economics* 126 (October): 106947. https://doi.org/10.1016/j.eneco.2023.106947.

Cavus, Mustafa, Berna Yazici and Ahmet Sezer. 2021. 'Penalized Power Properties of the Normality Tests in the Presence of Outliers'. *Communications in Statistics - Simulation and Computation* 52, no. 8 (July): 3568–3580. https://doi.org/10.1080/03610918.2021.1938124.

Chmelarova, Viera and R. Carter Hill. 2010. 'The Hausman Pretest Estimator'. *Economics Letters* 108, no. 1 (July): 96–99. https://doi.org/10.1016/j.econlet.2010.04.027.

Chow, Gregory C. 1960. 'Tests of Equality Between Sets of Coefficients in Two Linear Regressions'. *Econometrica* 28, no. 3 (July): 591–605. https://doi.org/10.2307/1910133.

Clements, Michael and David Hendry. 1998. *Forecasting Economic Time Series.* 1st ed. Cambridge University Press, October. ISBN: 978-0-511-59928-6. https://doi.org/10.1017/CBO9780511599286.

Danilov, Dmitry and Jan R Magnus. 2004. 'On the Harm That Ignoring Pretesting Can Cause'. *Journal of Econometrics* 122, no. 1 (September): 27–46. https://doi.org/10.1016/j.jeconom.2003.10.018.

Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application.* 1st ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. ISBN: 978-0-511-80284-3. https://doi.org/10.1017/CBO9780511802843.

Donoho, David L. and Peter J. Huber. 1983. 'The Notion of Breakdown Point'. In *A Festschrift for Eric Lehmann: In Honour of His Sixty-Fifth Birthday,* edited by P. Bickel, K. Doksum and J. Hodges, 157–184. Belmont, California: Wadsworth.

Doornik, Jurgen A. 2008. 'Encompassing and Automatic Model Selection'. *Oxford Bulletin of Economics and Statistics* 70, no. s1 (December): 915–925. https://doi.org/10.1111/j.1468-0084.2008.00536.x.

———. 2009. 'Autometrics'. In *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry,* edited by Jennifer L. Castle and Neil Shephard, 88–121. Oxford: Oxford University Press. ISBN: 978-0-19-923719-7. https://doi.org/10.1093/acprof:oso/9780199237197.001.0001.

———. 2018a. *An Introduction to OxMetrics™ 8. A Software System for Data Analysis and Forecasting.* Richmond, UK: Timberlake Consultants Ltd. ISBN: 978-0-9932738-0-3.

———. 2018b. *An Object-oriented Matrix Programming Language - Ox™ 8.* Richmond, UK: Timberlake Consultants Ltd. ISBN: 978-0-9932738-3-4.

Doornik, Jurgen A. and Henrik Hansen. 2008. 'An Omnibus Test for Univariate and Multivariate Normality'. *Oxford Bulletin of Economics and Statistics* 70, no. s1 (December): 927–939. https://doi.org/10.1111/j.1468-0084.2008.00537.x.

Doornik, Jurgen A. and David F. Hendry. 2018. *Interactive Monte Carlo Experimentation in Econometrics - PcNaive™ 6.* Richmond, UK: Timberlake Consultants Ltd. ISBN: 978-0-9932738-7-2.

Engle, Robert F. 1982. 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation'. *Econometrica* 50, no. 4 (July): 987–1007. https://doi.org/10.2307/1912773.

Engle, Robert F., David F. Hendry and Jean-Francois Richard. 1983. 'Exogeneity'. *Econometrica* 51, no. 2 (March): 277–304. https://doi.org/10.2307/1911990.

Ericsson, Neil R. 2012. 'Detecting Crises, Jumps, and Changes in Regime'. (Washington, DC), Working Paper, Board of Governors of the Federal Reserve System (November).

———. 2017. 'How Biased Are U.S. Government Forecasts of the Federal Debt?' *International Journal of Forecasting* 33, no. 2 (April): 543–559. https://doi.org/10.1016/j.ijforecast.2016.09.001.

Ericsson, Neil R., Mohammed H. I. Dore and Hassan Butt. 2022. 'Detecting and Quantifying Structural Breaks in Climate'. *Econometrics* 10, no. 4 (November): 33. https://doi.org/10.3390/econometrics10040033.

Ericsson, Neil R., Emilio J Fiallos and J E Seymour. 2015. 'Detecting Time-Dependent Bias in the Fed's Greenbook Forecasts of Foreign GDP Growth'. In *JSM Proceedings, Business and Economic Statistics Section,* 858–872. Alexandria, VA: American Statistical Association. Revised version from February 2017: https://unassumingeconomist.com/wp-content/uploads/2017/04/Ericsson.pdf. Accessed on 2023-12-19.

Fay, Michael, P. 2010. 'Two-Sided Exact Tests and Matching Confidence Intervals for Discrete Data'. *The R Journal* 2, no. 1 (June): 53–58. https://doi.org/10.32614/RJ-2010-008.

Gel, Yulia R. and Joseph L. Gastwirth. 2008. 'A Robust Modification of the Jarque–Bera Test of Normality'. *Economics Letters* 99, no. 1 (April): 30–32. https://doi.org/10.1016/j.econlet.2007.05.022.

Godfrey, L. G. 1978. 'Testing for Higher Order Serial Correlation in Regression Equations When the Regressors Include Lagged Dependent Variables'. *Econometrica* 46, no. 6 (November): 1303–1310. https://doi.org/10.2307/1913830.

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw and Werner A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions.* 1st ed. Wiley Series in Probability and Statistics. New York: Wiley. ISBN: 0-471-82921-8. https://doi.org/10.1002/9781118186435.

Hampel, Frank Rudolf. 1968. 'Contributions to the Theory of Robust Estimation'. PhD diss., University of California, Berkeley.

———. 1974. 'The Influence Curve and Its Role in Robust Estimation'. *Journal of the American Statistical Association* 69, no. 346 (June): 383–393. https://doi.org/10.1080/01621459.1974.10482962.

Hendry, David F. 1984. 'Monte Carlo Experimentation in Econometrics'. In *Handbook of Econometrics,* edited by Z. Griliches and M.D. Intriligator, vol. II. Elsevier. https://doi.org/10.1016/S1573-4412(84)02008-0.

———. 1999. 'An Econometric Analysis of US Food Expenditure, 1931-1989'. In *Methodology and Tacit Knowledge: Two Experiments in Econometrics,* edited by J. R. Magnus and M. S. Morgan, 341–361. Wiley Series in Applied Econometrics. Chichester: John Wiley & Sons.

Hendry, David F. 2018. 'Deciding between Alternative Approaches in Macroeconomics'. *International Journal of Forecasting* 34, no. 1 (January): 119–135. https://doi.org/10.1016/j.ijforecast.2017.09.003.

———. 2024. 'A Brief History of General-to-specific Modelling'. *Oxford Bulletin of Economics and Statistics* 86, no. 1 (February): 1–20. https://doi.org/10.1111/obes.12578.

Hendry, David F. and Jurgen A. Doornik. 2014. *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics.* Cambridge, MA: The MIT Press. ISBN: 978-0-262-32441-0.

Hendry, David F. and Søren Johansen. 2015. 'Model Discovery And Trygve Haavelmo's Legacy'. *Econometric Theory* 31, no. 1 (February): 93–114. https://doi.org/10.1017/S0266466614000218.

Hendry, David F., Søren Johansen and Carlos Santos. 2008. 'Automatic Selection of Indicators in a Fully Saturated Regression'. *Computational Statistics* 23 (April): 317–335. https://doi.org/10.1007/s00180-007-0054-z. Erratum in *Computational Statistics* 23 (April): 337-339. https://doi.org/10.1007/s00180-008-0112-1.

Hendry, David F. and Hans-Martin Krolzig. 2005. 'The Properties of Automatic *Gets* Modelling'. *The Economic Journal* 115, no. 502 (March): C32–C61. https://doi.org/10.1111/j.0013-0133.2005.00979.x.

Hendry, David F. and Grayham E. Mizon. 2011. 'Econometric Modelling of Time Series with Outlying Observations'. *Journal of Time Series Econometrics* 3, no. 1 (January): 6. https://doi.org/10.2202/1941-1928.1100.

———. 2014. 'Unpredictability in Economic Analysis, Econometric Modeling and Forecasting'. *Journal of Econometrics* 182, no. 1 (September): 186–195. https://doi.org/10.1016/j.jeconom.2014.04.017.

Hendry, David F. and John N J Muellbauer. 2018. 'The Future of Macroeconomics: Macro Theory and Models at the Bank of England'. *Oxford Review of Economic Policy* 34, nos. 1-2 (January): 287–328. https://doi.org/10.1093/oxrep/grx055.

Hendry, David F. and Adrian J. Neal. 1991. 'A Monte Carlo Study of the Effects of Structural Breaks on Tests for Unit Roots'. In *Economic Structural Change: Analysis and Forecasting,* edited by Peter Hackl and Anders H. Westlund. Berlin: Springer. ISBN: 978-3-662-06824-3. https://doi.org/10.1007/978-3-662-06824-3_8.

Hendry, David F. and Carlos Santos. 2010. 'An Automatic Test of Super Exogeneity'. In *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle,*

edited by Tim Bollerslev, Jeffrey Russell and Mark Watson, 164–193. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199549498.003.0009.

Hirji, Karim F. 2006. *Exact Analysis of Discrete Data.* Boca Raton, Fl.: Chapman & Hall/CRC. ISBN: 978-1-58488-070-7.

Huber, Peter J. 1964. 'Robust Estimation of a Location Parameter'. *The Annals of Mathematical Statistics* 35, no. 1 (March): 73–101. https://doi.org/10.1214/aoms/1177703732.

Jarque, Carlos M. and Anil K. Bera. 1987. 'A Test for Normality of Observations and Regression Residuals'. *International Statistical Review / Revue Internationale de Statistique* 55, no. 2 (August): 163–172. https://doi.org/10.2307/1403192.

Jiao, Xiyu. 2022. 'A Simple Robust Procedure in Instrumental Variables Regression', Working Paper.

Jiao, Xiyu and Bent Nielsen. 2017. 'Asymptotic Analysis of Iterated 1-Step Huber-Skip M-Estimators with Varying Cut-Offs'. In *Analytical Methods in Statistics,* edited by Jaromír Antoch, Jana Jurečková, Matúš Maciak and Michal Pešta, 193:23–52. Cham: Springer International Publishing. ISBN: 978-3-319-51313-3. https://doi.org/10.1007/978-3-319-51313-3_2.

Jiao, Xiyu and Felix Pretis. 2022. 'Testing the Presence of Outliers in Regression Models'. *Oxford Bulletin of Economics and Statistics* 84, no. 6 (December): 1452–1484. https://doi.org/10.1111/obes.12511.

Jiao, Xiyu, Felix Pretis and Moritz Schwarz. 2024. 'Testing for Coefficient Distortion Due to Outliers with an Application to the Economic Impacts of Climate Change'. *Journal of Econometrics* 239, no. 1 (February): 105547. https://doi.org/10.1016/j.jeconom.2023.105547.

Johansen, Søren and Bent Nielsen. 2009. 'An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator'. In *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry,* edited by Jennifer L. Castle and Neil Shephard, 1–36. https://doi.org/10.1093/acprof:oso/9780199237197.003.0001.

———. 2013. 'Outlier Detection in Regression Using an Iterated One-Step Approximation to the Huber-Skip Estimator'. *Econometrics* 1, no. 1 (May): 53–70. https://doi.org/10.3390/econometrics1010053.

———. 2016a. 'Analysis of the Forward Search Using Some New Results for Martingales and Empirical Processes'. *Bernoulli* 22, no. 2 (May): 1131–1183. https://doi.org/10.3150/14-BEJ689.

Johansen, Søren and Bent Nielsen. 2016b. 'Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models'. *Scandinavian Journal of Statistics* 43, no. 2 (June): 321–348. https://doi.org/10.1111/sjos.12174.

———. 2019. 'Boundedness of M-Estimators for Linear Regression in Time Series'. *Econometric Theory* 35, no. 3 (June): 653–683. https://doi.org/10.1017/S0266466618000257.

Kaji, Tetsuya. 2018. 'Switching to the New Norm: From Heuristics to Formal Tests Using Integrable Empirical Processes', Working Paper (November).

Klooster, Jens and Mikhail Zhelonkin. 2024. 'Outlier Robust Inference in the Instrumental Variable Model with Applications to Causal Effects'. *Journal of Applied Econometrics* 39, no. 1 (January): 86–106. https://doi.org/10.1002/jae.3012.

Kolmogorov, A. N. 1933. 'Sulla Determinazione Empirica Di Una Legge Di Distribuzione'. *Giornale dell'instituto Italiano Degli Attuari* 4:83–91.

Krolzig, Hans-Martin and David F. Hendry. 2001. 'Computer Automation of General-to-Specific Model Selection Procedures'. *Journal of Economic Dynamics and Control* 25, no. 6 (June): 831–866. https://doi.org/10.1016/S0165-1889(00)00058-0.

Kurle, Jonas Kai. 2022a. *Ivgets: General to Specific Modeling and Indicator Saturation in 2SLS Models.* https://cran.r-project.org/package=ivgets.

———. 2022b. *Robust2sls: Outlier Robust Two-Stage Least Squares Inference and Testing.* https://cran.r-project.org/package=robust2sls.

Leeb, Hannes and Benedikt M. Pötscher. 2005. 'Model Selection and Inference: Facts and Fiction'. *Econometric Theory* 21, no. 1 (February): 21–59. https://doi.org/10.1017/S0266466605050036.

———. 2008. 'Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator'. *Journal of Econometrics* 142, no. 1 (January): 201–211. https://doi.org/10.1016/j.jeconom.2007.05.017.

Lilliefors, Hubert W. 1967. 'On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown'. *Journal of the American Statistical Association* 62, no. 318 (June): 399–402. https://doi.org/10.2307/2283970.

Mincer, Jacob A. and Zarnowitz. 1969. 'The Evaluation of Economic Forecasts'. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance,* edited by Jacob A. Mincer, 3–46. New York: National Bureau of Economic Research (NBER). ISBN: 0-87014-202-X.

Muirhead, Robb J. 1982. *Aspects of Multivariate Statistical Theory.* 1st ed. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons. ISBN: 978-0-470-31655-9. https://doi.org/10.1002/9780470316559.

Murdoch, Duncan and Daniel Adler. 2024. *Rgl: 3D Visualization Using OpenGL.* https://cran.r-project.org/package=rgl.

Nielsen, Bent and Matthias Qian. 2023. 'Asymptotic Properties of the Gauge and Power of Step-Indicator Saturation', Nuffield College Economics Discussion Papers, no. 2023-W03 (November).

Posit Team. 2023. *RStudio: Integrated Development Environment for R.* Posit Software, PBC. Boston, MA. https://www.posit.co/.

Pretis, Felix, J. James Reade and Genaro Sucarrat. 2018. 'Automated General-to-Specific (GETS) Regression Modeling and Indicator Saturation for Outliers and Structural Breaks'. *Journal of Statistical Software* 86, no. 3 (September): 1–44. https://doi.org/10.18637/jss.v086.i03.

Pretis, Felix, Lea Schneider, Jason E. Smerdon and David F. Hendry. 2016. 'Detecting Volcanic Eruptions in Temperature Reconstructions by Designed Break-Indicator Saturation'. *Journal of Economic Surveys* 30, no. 3 (July): 403–429. https://doi.org/10.1111/joes.12148.

Pretis, Felix, Moritz Schwarz, Kevin Tang, Karsten Haustein and Myles R. Allen. 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: Mathematical, Physical and Engineering Sciences* 376, no. 2119 (May): 20160460. https://doi.org/10.1098/rsta.2016.0460.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Richards, Andrew. 2015. 'University of Oxford Advanced Research Computing' (August). https://doi.org/10.5281/zenodo.22558.

Rousseeuw, Peter J. 1984. 'Least Median of Squares Regression'. *Journal of the American Statistical Association* 79, no. 388 (December): 871–880. https://doi.org/10.2307/2288718.

Ruppert, David and Raymond J Carroll. 1980. 'Trimmed Least Squares Estimation in the Linear Model'. *Journal of the American Statistical Association* 75, no. 372 (December): 828–838. https://doi.org/10.2307/2287169.

Saculinggan, Mayette and Emily Amor Balase. 2013. 'Empirical Power Comparison Of Goodness of Fit Tests for Normality In The Presence of Outliers'. *Journal of*

Physics: Conference Series 435 (April): 012041. https://doi.org/10.1088/1742-6596/435/1/012041.

Schneider, L, J E Smerdon, F Pretis, C Hartl-Meier and J Esper. 2017. 'A New Archive of Large Volcanic Events over the Past Millennium Derived from Reconstructed Summer Temperatures'. *Environmental Research Letters* 12, no. 9 (September): 094005. https://doi.org/10.1088/1748-9326/aa7a1b.

Simes, R J. 1986. 'An Improved Bonferroni Procedure for Multiple Tests of Significance'. *Biometrika* 73, no. 3 (December): 751–754. https://doi.org/10.2307/2336545.

Sølvsten, Mikkel. 2020. 'Robust Estimation with Many Instruments'. *Journal of Econometrics* 214, no. 2 (February): 495–512. https://doi.org/10.1016/j.jeconom.2019.04.040.

Tibshirani, Robert. 1996. 'Regression Shrinkage and Selection via the Lasso'. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, no. 1 (January): 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tobin, James. 1958. 'Estimation of Relationships for Limited Dependent Variables'. *Econometrica* 26, no. 1 (January): 24–36. https://doi.org/10.2307/1907382.

van der Vaart, Aad W. and Jon A. Wellner. 1996. 'Weak Convergence'. In *Weak Convergence and Empirical Processes: With Applications to Statistics,* 16–28. New York: Springer. ISBN: 978-1-4757-2545-2. https://doi.org/10.1007/978-1-4757-2545-2_3.

Wallace, T. Dudley. 1977. 'Pretest Estimation in Regression: A Survey'. *American Journal of Agricultural Economics* 59, no. 3 (August): 431–443. https://doi.org/10.2307/1239645.

White, Halbert. 1980. 'A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity'. *Econometrica* 48, no. 4 (May): 817–838. https://doi.org/10.2307/1912934.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund et al. 2019. 'Welcome to the Tidyverse'. *Journal of Open Source Software* 4, no. 43 (November): 1686. https://doi.org/10.21105/joss.01686.

Young, Alwyn. 2022. 'Consistency without Inference: Instrumental Variables in Practical Application'. *European Economic Review* 147 (August): 104112. https://doi.org/10.1016/j.euroecorev.2022.104112.

Zhelonkin, Mikhail, Marc G. Genton and Elvezio Ronchetti. 2012. 'On the Robustness of Two-Stage Estimators'. *Statistics & Probability Letters* 82, no. 4 (April): 726–732. https://doi.org/10.1016/j.spl.2011.12.014.