Essays on Empirical Econometrics



Olga Ewa Gdula
Nuffield College

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2021

Essays in Empirical Econometrics

Olga E. Gdula

Nuffield College, University of Oxford

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

Michaelmas Term 2021

Abstract

The thesis consists of three main chapters.

Chapter 1 studies methods of identification, model selection, and heterogeneity analysis in discrete choice settings. The chapter considers a general class of revealed preference models and shows that, without further structure imposed on the model, inequality restrictions from revealed prefence allow only for the identification of a compact set of parameters. I show how the principle of maximum entropy can be applied in this setting to identify a unique vector of parameters for each agent, by choosing the solution which is maximally agnostic about missing information. For the purposes of model selection, I apply Selten's measure of predictive success to this setting and I suggest an alternative model selection method based on the Kullback-Leibler divergence. I propose multiple methods for analysing parameter heterogeneity.

The section originally presented here cannot currrently be made freely available via ORA.

Chapter 2 focuses on the practical and computational aspects of implementing partial identification methods in practice. The aim of the chapter is to bridge the gap between econometric theory and its applications. The chapter focuses on binary-outcome models with random coefficients and endogeneity, and a set-identifying approach to this model developed by Chesher

and Rosen (2014). To facilitate practical implementations of the model, I illustrate the mathematical results of Chesher and Rosen (2014) with simple graphical examples and I contrast their model with point-identifying alternatives. To show how the method can be implemented, I provide step-by step pseudo-code and Matlab code. I analyse the increases in the computational complexity of the model arising from increasing the cardinality of the instruments, the endogenous variables, and exogenous explanatory variables; and I discuss computational and practical issues arising from the inclusion of continuous variables. I analyse the empirical practicability of the method by applying it to data from the British Household Panel Survey in the context of analysing female labour supply decisions. The empirical application shows that insufficient structure imposed on the model may lead to unbounded sets, while small cell sizes may result in empty sets. Empirical results derived for a fixed value of the variance of the coefficient on the endogenous variable show that, under this restriction, the set of identified parameters appears bounded and that it has a non-trivial shape. The results also show that the volume of the set is reduced significantly when more instruments are used.

In Chapter 3, the focus shifts from partial identification, analysed in the previous two chapters, to the econometric theory of non-linear regressions. I derive asymptotic results for marked and weighted empirical processes of residuals from non-linear regressions with outlier detection. I show that, under a number of conditions imposed on the regressors, marks, weights, the non-linear function, and the distribution of the errors, the existing asymptotic theory developed for marked and weighted functions of residuals from linear models with outlier detection can be extended to non-linear models with outlier detection. This allows me to apply the theory developed for empirical processes of residuals from linear models, to non-linear models; and to show the asymptotic equivalence of the processes analysed in the chapter and marked and weighted empirical processes based on the true innovations, without estimation errors. My results do not depend on unverifiable assumptions. To illustrate this, I show through an example how each of the assumptions can be checked. I suggest several applications of the theory developed in the chapter.

In loving memory of Adam Pulchny [1937 - 2011],

whose life of compassion and perseverance taught me more than any degree.

Thank you for giving me hope for a better tomorrow.

Acknowledgements

I would like to wholeheartedly thank my supervisors: Vanessa Berenguer-Rico, Bent Nielsen and Ian Crawford, for their invaluable guidance, patience, and support throughout the process of writing this thesis, with its many highs and lows.

I would like to thank Sophocles Mavroeidis for his guidance in the early stages of preparing Chapter 2 of this thesis, and Steve Bond for his valuable comments on my confirmation of status presentation and submission.

Research undertaken in this thesis has been partially supported by the Economic and Social Research Council, who funded a scholarship which covered my MPhil studies and the first two years of my DPhil degree. I am very grateful for their support, without which I would not have been able to study at Oxford.

Teaching at Oxford enabled me to continue this degree after my funding period came to an end. I am very grateful to the Department of Economics and the colleges I have taught for. I would particularly like to thank the Economics tutors at Hertford, LMH and St Catz, who acted as my line managers at various points during this degree and became my teaching mentors: Elizabeth Baldwin, Natalie Quinn, Frank Di Traglia and Alex Teytelboym.

I am eternally grateful to Tom for, well, everything.

Contents

Thesis Introduction **1**

1 Maximum-Entropy Based Identification in Revealed-Preference Models with Heterogeneity **14**

1.1 Introduction and Literature Review 15

1.1.1 Outline of Existing Approaches 15

1.1.2 Outline of the Chapter and Contributions 18

1.1.3 Summary of Empirical Findings 21

1.2 Part A: Methods 22

1.2.1 The Model 22

1.2.2 Set Identification 23

1.2.3 Point Identification 25

1.2.4 Model Selection 30

1.2.5 Analysis of Parameter Heterogeneity 33

1.3 Part B: Empirical Application to Models of Other-Regarding Preferences 40

1.3.1 Literature Review 40

1.3.2 Data and Experimental Design 45

1.3.3 Structural Model 48

1.3.4 Empirical Results 53

1.4 Conclusion 71

1.A Appendix 82

2 Computational Aspects of Partial Identification in Practice **83**

2.1 Introduction 83

2.1.1 Related Literature and Motivation . . . 84
2.2 The Model . . . 86
2.2.1 Formal Model . . . 86
2.2.2 Empirical Motivation . . . 88
2.3 Methods . . . 89
2.3.1 Point-Identifying Methods . . . 89
2.3.2 Model of Chesher and Rosen (2014) . . . 93
2.4 Empirical Application . . . 99
2.4.1 Data . . . 99
2.4.2 Estimated Models . . . 106
2.5 Implementation Guide . . . 110
2.5.1 Full Estimation and the Projection Method . . . 110
2.5.2 Example Implementation Code . . . 112
2.6 Computational Considerations . . . 118
2.6.1 Computational Time . . . 118
2.6.2 Effects of Adding Further Endogenous Variables . . . 122
2.6.3 Continuous Variables . . . 123
2.7 Illustration of Results . . . 126
2.7.1 Point-identifying Approach to Models A and B . . . 126
2.7.2 Set-identifying Method: Binary Instrument (Model A) . . . 129
2.7.3 Set-identifying Method: Adding More Instruments (Model B) . . . 135
2.7.4 Models A and B: The Issue of an Increasing Variance . . . 138
2.7.5 Model C: Adding Exogenous Variables . . . 140
2.8 Conclusion and Discussion . . . 143
2.A Appendix to Computational Aspects of Partial Identification in Practice . . . 145

3 Marked and Weighted Empirical Processes of Residuals from Non-Linear Regressions 151

3.1 Introduction . . . 151
3.2 Related Literature and Contributions . . . 152
3.3 Marked and Weighted Empirical Processes of Residuals . . . 156

3.3.1 Linearisation of the Residuals . . . 157
3.3.2 Replacing Stochastic Errors with Non-Stochastic Terms . . . 158
3.3.3 Other Notation . . . 159
3.4 Key Results . . . 159
3.5 Verification of Assumptions: Example . . . 168
3.5.1 Example: Wu (1981) model . . . 168
3.6 Applications . . . 179
3.7 Conclusion . . . 180
3.A Appendix 3.A: Details and Proofs of Lemmas . . . 181
3.A.1 Structure of the Appendix . . . 181
3.A.2 Notation and Definitions . . . 181
3.A.3 Preliminary Lemmas . . . 181
3.A.4 Summary of Key Lemmas . . . 183
3.A.5 Main Lemmas . . . 185
3.B Appendix 3.B: Proofs of Key Theorems . . . 218
3.B.1 Structure of the Appendix . . . 219
3.B.2 Outline of the Approach . . . 219
3.B.3 Outline of the Proof of Result (3.19) . . . 219
3.B.4 Outline of the Proof of Result (3.21) . . . 221
3.B.5 Theorems Required for Result (3.19) . . . 222
3.B.6 Theorems Required for Result (3.21) . . . 234

Thesis Conclusion **241**

Bibliography **242**

List of Figures

1.1 Indifference curves in the case of Fehr-Schmidt preferences, for different sets of parameters . . . 52
1.2 Distribution of estimates of $\hat{\theta}_1$ . . . 57
1.3 Distribution of estimates of $\hat{\theta}_2$ . . . 57
1.4 Distribution of estimates of $\hat{\theta}_3$ . . . 58
1.5 Distribution of estimates of $\hat{\theta}_4$ . . . 58
1.6 Bootstrap distributions of the convex hull measure h . . . 64
1.7 Bootstrap distributions of the ellipsoid measure v . . . 65
1.8 Distributions of parameter estimates for the Isanga . . . 67
1.9 Distributions of offers and acceptance decisions for selected societies . . . 82

2.1 Partitions of the parameters space corresponding to different combinations of observables . . . 97
2.2 Estimated Set; Var(β_i) set to 10 . . . 130
2.3 Projection of the Estimated Set, Showing the Mean of α_i; Var(β_i) set to 10 . . . 130
2.4 Estimated Set - Var(β_i) set to 100 . . . 131
2.5 Estimated Set - Var(β_i) set to 10 . . . 135
2.6 Estimated Set - Var(β_i) set to 10 . . . 137

List of Tables

1.1 Information on societies included in the Human Sociality Project 45
1.2 Society averages of key demographic variables 46
1.3 Model selection measures 54
1.4 Summary statistics of parameter estimates 59
1.5 Principal component analysis of $\hat{\Theta}$ 60
1.6 Correlation structure of parameter estimates 61
1.7 Measures of heterogeneity and clustering 62
1.8 Summary statistics of bootstrap distributions of h and v 63
1.9 Measures of clustering - percentiles 66
1.10 OLS analysis of determinants of $\hat{\theta}_1$ 74
1.11 OLS analysis of determinants of $\hat{\theta}_2$ 76
1.12 OLS analysis of determinants of $\hat{\theta}_3$ 78
1.13 OLS analysis of determinants of $\hat{\theta}_4$ 80

2.1 Distribution of Instruments: Smoking and 2-way Cross-Products 102
2.2 Distribution of Instruments: 3-way Cross-Products 102
2.3 Strength of Instruments 103
2.4 Strength of Instruments 104
2.5 Number of Calculations in the Case of One Endogenous Variable 122
2.6 Point-identifying Model - Results 127
2.7 Proportions of the Sample who are Working and in Education 129
2.8 Extreme Points of the Estimated Sets - Binary Instrument 132
2.9 Extreme Points of the (Enlargement of the) Estimated Intervals — Triangular Distribution 134

2.10 Extreme Points of the Estimated Sets - Smoking + Father's FE 136
2.11 Extreme Points of the Estimated Sets - Smoking + Mother's FE 136
2.12 Strengthening the Instruments and the Volume of the Identified Set 137
2.13 Point-identifying Model - Results with Exogenous Variables 141
2.14 Cell Sizes — Combinations of Data 142

Thesis Introduction

We live in an age of data: as reported by Vopson (2021), in 2018 the world produced 33 zetabytes (or 33 trillion gigabytes) of data. The Forbes magazine predicts that by 2025 over 150 zetabytes of data will need analysis (Kulkarni, 2019). The growing availability and the sheer diversity of data being produced make this an exciting time to be an econometrician due to the constant challenge to develop new data analysis methods, and to apply existing methods to new scenarios. This thesis presents my humble contribution to the field of analysing data in Economics: the theory and practice of Econometrics.

The thesis is comprised of three chapters which focus on different aspects of econometric theory and its applications. The chapters cover a wide range of topics and, consequently, each chapter is presented as a self-contained unit with its own introduction and motivation, a literature review relevant to the specific chapter, separate analysis, conclusions and chapter abstract. The purpose of this additional thesis introduction is to provide a brief overview of the work presented in each chapter, while the details of the motivation, the relevant literature and results, are deferred to individual chapters.

The first two chapters of the thesis focus on two approaches to partial identification. Partial identification arises when the modelling assumptions and the data are not sufficient to uniquely identify the parameter vector of interest, for example, due to inequality, as opposed to equality, constraints. Despite the shared theme, the focus of the two chapters is different. In Chapter 1 I analyse partial identification in the context of revealed preference, where I aim to recover a unique parameter vector in the identified set on the basis of the maximum entropy principle. In Chapter 2, I analyse the computational aspects of implementing in practice set-identifying methods based on random set theory, in the context of binary-outcome random-coefficient models with endogeneity.

The final chapter has a different focus to the first two. While the first two chapters discuss both theory and applications, the final chapter focuses entirely on econometric theory. I also move away from the theme of partial identification, and focus on non-linear regressions. In Chapter 3, I formally derive asymptotic results which capture the large-sample properties of marked and weighted empirical processes of residuals from non-linear regressions with outlier detection.

The remaining part of this introduction contains three sections, which provide a detailed overview of each chapter.

Chapter 1

In Chapter 1, I consider scenarios in which an agent is observed making one or more choices which, by revealed preference, impose inequality constraints on the admissible values of the parameters of the agent's payoff function. Examples of applicable settings include game theory, consumer choice and profit maximisation. Formally, an agent i is assumed to be acting so as to maximise a payoff function $u_i(\cdot)$, which is a function of her own choice, and possibly the choices made by other agents in the same choice setting. As detailed in Chapter 1, the modelling framework allows for both strategies chosen by the agent, and for choices made by other agents involved in the same setting, as well as for possible uncertainty. Unlike the assumptions of random utility theory, my model does not assume any errors in the agent's utility function.

Importantly, the payoff function is assumed to be linear in parameters, i.e. $u_i(x) = \theta_i' f_i(x)$, where x is a vector of outcomes, $f_i(x)$ is a function known to the researcher and θ_i is a vector of the preference parameters of interest.

In this setting, the principles of revealed preference impose only inequality restrictions on the parameter vector of interest. Chapter 1 shows that under these circumstances the researcher can only recover a compact subset of the parameter space, but not necessarily a unique point. To overcome this limitation, I propose that the principle of maximum entropy be applied to select that point within the identified set which is is maximally agnostic about missing information. This approach allows me to estimate a specific parameter vector for each individual.

A researcher may be uncertain as to the functional form of the payoff function which best fits the data. Since the choice may be made among non-nested models, and given the lack of an error term about which functional form assumptions can be made, usual approaches such as

Wald or likelihood ratio tests are not applicable. Instead, I propose that either Selten's measure of predictive success, or the Kullback-Leibler divergence, be used for model selection. Both measures are based on comparing the proportion of the sample whose behaviour is rationalisable by a given model, with the proportion of all behaviours which are rationalisable. I discuss the relative merits of the two measures.

Since the point-identification method proposed in Chapter 1 allows a researcher to recover a separate parameter vector for each agent, I discuss methods which can be used to analyse parameter heterogeneity. I show how principal component analysis of the matrix of parameters for a group of observations can be applied to analyse the main directions of variation in parameters. I propose two methods for quantifying the amount of heterogeneity in a group compared to the sample as a whole, and for testing whether the amount of heterogeneity in a group is the same as that in the entire sample. I also show how Seemingly Unrelated Regression can be used to analyse the determinants of preferences by using the individual-specific parameters as regressors.

The section originally presented here cannot currrently be made freely available via ORA.

The section originally presented here cannot currrently be made freely available via ORA.

Chapter 2

Chapter 2 aims to bridge the gap between the rapid advances in econometric theory in the area of partial identification, and the relative sparsity of empirical applications of the theory. In the spirit of Nevo (2000), the aim of the chapter is to make partial identification methods more accessible to applied researchers. To this aim, in Chapter 2 I analyse the computational aspects of implementing partial identification in practice. My focus is on binary-outcome random-coefficients models with endogeneity.

Formally, I assume that the researcher is interested in a binary outcome variable, Y_i, which takes the value of 1 if and only if an underlying latent variable, Y_i^*, is non-negative:

$$Y_i = 1.\left[Y_i^* \geq 0\right]. \tag{1}$$

The latent variable Y_i^* is a function of observable variables, which can be endogenous (E_i) or exogenous (W_i). Importantly, the intercept and the slope coefficients are allowed to be fully heterogeneous across observations. The formal equation determining Y_i can be expressed as

$$Y_i = 1.\left[Y_i^* \geq 0\right] = 1.\left[\alpha_i + \beta_i E_i + \gamma_i W_i \geq 0\right]. \tag{2}$$

Furthermore, I assume that the researcher has at their disposal a vector of instruments, Z_i, which are not included in the outcome equation (2) and which are independent of the heterogeneous coefficients (α_i, β_i).

To motivate my interest in this model, in Chapter 2 I review empirical literature on the returns to education. On the basis of this literature, I argue that the random-coefficients binary-outcome model is particularly well-suited to modelling female labour force participation decisions, due to the heterogeneity in the returns to education and the possibility of the returns to education being correlated with education itself.

To aid researchers with selecting the most appropriate method for estimating the parameters of this model, in Section 2.3 I discuss restrictions under which point-identification of some versions of equation (2) can be achieved. These are applicable in cases such as: the absence of heterogeneity; the absence of endogenous regressors; or when both heterogeneity and endogeneity are present but the interest lies only in estimating the mean of the distribution of the random coefficients, rather than its entire shape. I contrast the point-identifying approaches with the method of Chesher and Rosen (2014) which allows for all aspects of the model but at the cost of achieving only set identification.

Since the method of Chesher and Rosen (2014) is based on random set theory, which is not widely taught to economists, I devote a portion of Section 2.3 to interpreting the key equations of Chesher and Rosen (2014) which are needed in order to implement the method in pracice; and I illustrate the key mathematical results using simple graphical examples.

Having presented the details of the estimation procedure, in Section 2.4 I discuss the details of the empirical application to analysing female labour force participation decisions based on data from the British Household Panel Survey. I discuss the choice of variables and, in particular, the instrumental variables in view of the existing literature and the availability of data. I also analyse the strength and variability of the instruments.

In Section 2.4 I also present three versions of the model described by equation (2), which differ in the instruments and explanatory variables used. As detailed in Section 2.4, all three versions are estimated under the assumption that the random coefficients are jointly normally distributed. Additionally, one of the models is also estimated under the assumption that the coefficients are independent and follow a triangular distribution. In the final part of the section, I discuss the impact of the limitations of the sample size and the computation complexity required on my variable choices.

Having established the model which is to be estimated, in Section 2.5 I present an implementation guide for the method of Chesher and Rosen (2014) under the assumption that the

family of distributions of random coefficients that the researcher is interested in can be fully specified by the vector of means and the variance-covariance matrix. A special case of this family of distributions is the multivariate normal distribution.

Section 2.5 begins by presenting two approaches to estimation: the 'full estimation' method, which recovers the full set of parameters of the distribution of the random coefficients; and the 'projection method', which recovers a subset of that set, by focusing only on selected parameters of interest. The two methods are contrasted in view of their applications and computational time.

To aid implementation of the method of Chesher and Rosen (2014) by those who are new to it, Section 2.5 presents a step-by-step guide for implementing the projection method using informal pseudo-code. Additionally, the actual Matlab code used for empirical applications in this chapter is provided in the Appendix.

Section 2.6 focuses on computational considerations when applying the method of Chesher and Rosen (2014) in practice. The computational issues arise from the specifics of the identification method. Identification in this model relies on 'core-determining sets', which are subsets of the parameter space such that if the identifying inequality holds for all those sets for a given conjectured distribution $\tilde{F}$, then the conjectured distribution $\tilde{F}$ belongs to the identified set.

In the case of the random coefficients being normally distributed, for a given conjectured distribution an integral of the multivariate normal distribution must be calculated for each core-determining set. Such integrals are time consuming and if coupled with testing a large number of conjectured distributions – e.g. due to testing a large number of parameters defining those distributions – this may lead to prohibitively high computational time costs.

Consequently, in Section 2.6 I analyse mathematically how the computational complexity of the model varies with the support of the instrumental variable, the endogenous variable, as well as the support and number of exogenous variables. I show that even for a coarse grid of values of tested parameters, the computational complexity of the model soon becomes prohibitive, particularly when exogenous variables are added. In Section 2.6 I also consider the implications of adding more endogenous variables to the model, as well as difficulties arising from the inclusion of continuous variables.

In Section 2.7 the method is applied to a model of female labour force participation based on Wave 8 of BHPS data. The model is chosen to be as simple as possible while maintaining all key features of the random-coefficient binary-outcome model with endogeneity. The reason for this is

that the focus of my empirical application is on evaluating the practicability of the method and illustrating the type of results that it can produce, rather than focusing on the interpretation and policy implications of the estimates, while computational complexity restricts the richness of the feasible specification.

The outcome variable of interest is a binary indicator of whether the woman in question works: it takes a value of 1 if the woman is employed or self-employed, and it takes a value of 0 otherwise. The simplest version of the estimated model has one explanatory variable which is an endogenous indicator for university education. Following Dickson (2013), I use smoking at the age of 16 as a binary instrument for education. In further specifications, I add more instruments in the form of indicators for whether each of one's parents' obtained further education. In the final specification, I also include age categories as exogenous explanatory variables.

To obtain a set of baseline results for comparisons with the set-identifying approach, I begin by estimating all specifications through the point-identifying bivariate probit, which imposes that all coefficients are constant. I find that all instruments are strong in all specifications. However, the effect of education on the probability of working is very imprecisely estimated and is not statistically significant in any of the specifications, which likely stems from the relatively small size of the dataset (672 observations) and low variability of the education variable in the sample.

In the next step of the empirical application, the random-coefficients binary-outcome model with endogeneity is estimated using the method of Chesher and Rosen (2014). The results give rise to several interesting findings. Firstly, in the estimated set, the variance of the parameter β_i appears to be unbounded, and the mean of β_i increases with the variance, leading to the set of estimates of the mean of β_i to also be unbounded.

However, if the variance of β_i is fixed at a particular value, the method produces bounded, and interesting, sets. In particular, the mean of the random intercept α_i appears to be relatively precisely estimated. Moreover, the shape of the estimated set, which is presented graphically in Section 2.7, is non-trivial and shows that certain combinations of the means of α_i and β_i and certain values of the correlations between those random coefficient are rejected by the data.

Moreover, results obtained when the variance of β_i is fixed show that the volume of the identified set reduces dramatically when further instruments are added: it falls by a half when father's further education indicator is used as an instrument in addition to smoking when young;

and it falls by a factor of 4 when mother's higher education indicator is used in addition to the smoking indicator. Moreover, the more acute angles in the identified set show that more combinations of the parameters are rejected when further instruments are added. Thus, the results illustrate the identifying strength of the instruments.

Due to the issue of an unbounded variance of β_i, I re-estimate the baseline specification under the assumption that the random coefficients α_i and β_i are independent and that their follow the triangular distribution since, unlike the normal distribution, it has bounded support. However, the problem of the infinite variance persists, suggesting that the outcome equation with only a binary outcome variable and a binary explanatory variable does not partition the parameter space into sufficiently varied core-determining sets to obtain a bounded set of estimates. This issue can be addressed by adding exogenous explanatory variables to the equation.

In the final empirical specification, I use an exogenous age variable in the outcome equation in addition to the endogenous education indicator. However, the estimated set appears to be empty. This is most likely caused by the limitations of the data due to the small cell sizes. The set-identifying method of Chesher and Rosen (2014) relies on comparing the probability of core-determining sets arising under the hypothesised distribution, with probabilities implied by the empirical distribution of observable variables. If the latter is very unreliably estimated due to small cell sizes, the entire estimation method is affected.

Since the focus of the chapter is on analysing computational issues of implementing partial identification in practice, the finding of a small, and possibly empty, set is in itself valuable as it highlights the importance of not only sufficiently large sample sizes, but also sample sizes which achieve a sufficient coverage of all cells to make the estimation of the distribution of observables reliable, even when variables with non-trivial cardinality are included.

Chapter 3

The final chapter of the thesis is the most analytical chapter and its relative brevity reflects its technical contribution, supported by the chapter's appendices with mathematical proofs, which span some 60 pages.

Unlike the first two chapters, the final chapter focuses entirely on econometric theory. I no longer assume that parameters can only be set-identified. Instead, I focus on marked and weighted empirical processes of residuals from non-linear regressions with outlier detection and

I derive their asymptotic properties.

Formally, I assume that the researcher wishes to estimate the model

$$y_i = g(x_i, \beta_0) + \varepsilon_i, \tag{3}$$

where y_i is a scalar variable of interest, x_i is a vector of explanatory variables, $g(x_i, \beta)$ is a possibly non-linear function whose form is known except for the parameter vector β, and β_0 is the true underlying value of the vector of unobserved parameters. While the regressors need not be stationary, I assume that the innovations ε_i/σ are independently and identically distributed with some unknown distribution function $F(c) = P(\varepsilon_i \leq \sigma c)$, where the scale σ is unknown.

My interest lies in non-linear regressions in which outlier detection and deselection is applied, such as using the Huber-skip or forward search algorithm. Formally, I assume that the researcher obtains some initial estimates $(\tilde{\beta}, \tilde{\sigma})$, which are used to calculate residuals $\tilde{\varepsilon}_i = y_i - g(x_i, \tilde{\beta})$. Those observations whose residuals are greater than a specified cut-off c, i.e. those for which $\tilde{\varepsilon}_i/\tilde{\sigma} \geq c$, are deselected. The final set of estimates $(\hat{\beta}, \hat{\sigma})$ is obtained based only on those observations which are not classified as outliers. The final set of residuals $\hat{\varepsilon}_i$ is obtained based on the final set of parameter estimates: $\hat{\varepsilon}_i = y_i - g(x_i, \hat{\beta})$.

In this context, my interest lies in marked and weighted functions of residuals. I begin my analysis by considering the marked and weighted empirical distribution function

$$\hat{\mathsf{F}}_n^{w,m,g}(c) = n^{-1} \sum_{i=1}^{n} w_{in} m\left(\frac{\hat{\varepsilon}_i}{\hat{\sigma}}\right) 1_{(\tilde{\varepsilon}_i \leq \tilde{\sigma} c)}, \tag{4}$$

where w_{in} are the weights, which will normally be functions of the regressors, and $m(\cdot)$ are the marks, which are functions of the residuals. This framework can be applied, for example, to tests of heteroskedasticity or tests of normality.

To develop asymptotic theory of processes of marked and weighted residuals, I linearise the residuals and replace the stochastic estimation errors $\tilde{b}_1 = N^{-1}(\tilde{\beta} - \beta)$, $\tilde{a}_1 = n^{1/2}(\tilde{\sigma} - \sigma)$, $\tilde{b}_m = N^{-1}(\hat{\beta} - \beta)$, $\tilde{a}_m = n^{1/2}(\hat{\sigma} - \sigma)$, where N is a normalisation matrix, with their non-stochastic components $\theta := (b_1, b_m, a_1, a_m)$, the mathematical details of which are deferred until Chapter 3. This allows me to obtain a form of the empirical distribution function which depends on

non-stochastic estimation errors θ and linearised residuals:

$$F_n^{w,m,g}(\theta,c) = n^{-1}\sum_{i=1}^{n} w_{in}\left\{ m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right) 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2}a_1 c + b_1'\dot{g}_{in}^0 + \frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1\right)}\right\}, \tag{5}$$

where $\dot{g}_{in}^0$ is the normalised first derivative of the non-linear function $g(\cdot)$ evaluated at the true value of β, while $\ddot{g}_{in}^{1*}$ and $\ddot{g}_{in}^{m*}$ are normalised second derivatives of the non-linear function evaluated at certain intermediate values of the parameter vector β.

In order to be able to apply martingale theory, in Chapter 3 I further define the compensator, which is the conditional expectation of the linearised empirical distribution function:

$$\bar{F}_n^{w,m,g}(\theta,c) = n^{-1}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1}\left\{ m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right) 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2}a_1 c + b_1'\dot{g}_{in}^0 + \frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1\right)}\right\}. \tag{6}$$

Finally, I define the empirical process of interest as:

$$\mathbb{F}_n^{w,m,g}(\theta,c) = n^{1/2}\left\{F_n^{w,m,g}(\theta,c) - \bar{F}_n^{w,m,g}(\theta,c)\right\}. \tag{7}$$

The key results in Chapter 3 prove that, as the sample size n tends to infinity, the importance of the second derivative terms $\ddot{g}_{in}^{1*}$ and $\ddot{g}_{in}^{m*}$ in $F_n^{w,m,g}(\theta,c)$, $\bar{F}_n^{w,m,g}(\theta,c)$ and $\mathbb{F}_n^{w,m,g}(\theta,c)$ vanishes asymptotically. To state this more formally, I prove that, as $n \to \infty$:

$$\sup_{\theta}\sup_{c}\left| n^{1/2}\{F_n^{w,m,g}(\theta,c) - F_n^{w,m}(\theta,c)\}\right| = o_p(1), \tag{8}$$

where the term $F_n^{w,m}(\theta,c)$ is defined analogously to $F_n^{w,m,g}(\theta,c)$ but without the second derivative terms $\ddot{g}_{in}^{1*}$ and $\ddot{g}_{in}^{m*}$:

$$F_n^{w,m}(\theta,c) = n^{-1}\sum_{i=1}^{n} w_{in} m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma + n^{-1/2}a_m}\right) 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2}a_1 c + b_1'\dot{g}_{in}^0\right)}. \tag{9}$$

Analogously, for the compensator term I show that as $n \to \infty$:

$$\sup_{\theta}\sup_{c}\left| n^{1/2}\{\bar{F}_n^{w,m,g}(\theta,c) - \bar{F}_n^{w,m}(\theta,c)\}\right| = o_p(1), \tag{10}$$

where the compensator term $\bar{F}_n^{w,m}(\theta,c)$ is defined analogously to $\bar{F}_n^{w,m,g}(\theta,c)$ but without the

second derivatives $\ddot{g}_{in}^{1*}$ and $\ddot{g}_{in}^{m*}$:

$$\bar{F}_n^{w,m}(\theta, c) = n^{-1} \sum_{i=1}^{n} w_{in} \mathsf{E}_{i-1} \left\{ m \left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0\right)} \right\}. \tag{11}$$

Together, the results imply that

$$\sup_{\theta} \sup_{c} |\mathbb{F}_n^{w,m,g}(\theta, c) - \mathbb{F}_n^{w,m}(\theta, c)| = o_p(1), \tag{12}$$

where $\mathbb{F}_n^{w,m}(\theta, c)$ is defined analogously to $\mathbb{F}_n^{w,m,g}(\theta, c)$ as:

$$\mathbb{F}_n^{w,m,g}(\theta, c) = n^{1/2} \left\{ F_n^{w,m}(\theta, c) - \bar{F}_n^{w,m}(\theta, c) \right\}. \tag{13}$$

The implications of these results are profound. The reason is that asymptotic properties of marked and weighted empirical processes of residuals from linear, as opposed to non-linear, regressions with outlier detection are well-established. By showing that the empirical process in the non-linear case is asymptotically equivalent its counterpart in linear regressions, I show that existing theory for linear regressions in this class can also be applied to non-linear regressions. In particular, Berenguer-Rico and Nielsen (2017) derive asymptotic theory for an empirical distribution, a compensator and an empirical process, which are analogous to $F_n^{w,m,g}(\theta, c)$, $\bar{F}_n^{w,m,g}(\theta, c)$ and $\mathbb{F}_n^{w,m,g}(\theta, c)$ respectively, but are based on linear, rather than non-linear regressions. By using their results with the term $\dot{g}_{in}^0$ used as the vector of regressors, one can obtain asymptotic distribution results for $F_n^{w,m}(\theta, c)$, $\bar{F}_n^{w,m}(\theta, c)$ and $\mathbb{F}_n^{w,m}(\theta, c)$. This application of existing theory to terms $F_n^{w,m}(\theta, c)$, $\bar{F}_n^{w,m}(\theta, c)$ and $\mathbb{F}_n^{w,m}(\theta, c)$ combined with the results detailed above for $F_n^{w,m,g}(\theta, c)$, $\bar{F}_n^{w,m,g}(\theta, c)$ and $\mathbb{F}_n^{w,m,g}(\theta, c)$, implies that

$$\sup_{\theta} \sup_{c} |\mathbb{F}_n^{w,m,g}(\theta, c) - \mathbb{F}_n^{w,m}(0, c)| = o_p(1). \tag{14}$$

This last result shows that, as $n \to \infty$, the empirical process $\mathbb{F}_n^{w,m,g}(\theta, c)$, which is based on residuals from non-linear regressions, exhibits the same asymptotic behaviour as the empirical process $\mathbb{F}_n^{w,m}(0, c)$, in which the estimation errors are zero and which, therefore, is based on the true innovations ε_i:

$$\mathbb{F}_n^{w,m}(0, c) = n^{1/2} \left\{ F_n^{w,m}(0, c) - \bar{F}_n^{w,m}(0, c) \right\}, \tag{15}$$

where the terms $F_n^{w,m}(0,c)$ and $\bar{F}_n^{w,m}(0,c)$ are based on $F_n^{w,m}(\theta,c)$ and $\bar{F}_n^{w,m}(\theta,c)$ respectively, but with all estimation errors set to zero:

$$F_n^{w,m}(0,c) = n^{-1}\sum_{i=1}^{n} w_{in} m\left(\frac{\varepsilon_i}{\sigma}\right) 1_{(\varepsilon_i \leq \sigma c)} \tag{16}$$

and

$$\bar{F}_n^{w,m}(0,c) = n^{-1}\sum_{i=1}^{n} w_{in} \mathsf{E}_{i-1}\left\{m\left(\frac{\varepsilon_i}{\sigma}\right) 1_{(\varepsilon_i \leq \sigma c)}\right\}. \tag{17}$$

Due to the lack of estimation errors, the behaviour of $\mathbb{F}_n^{w,m}(0,c)$ can be obtained using standard theory.

The results presented in Chapter 3 rely on a set of assumptions on the non-linear function $g(\cdot)$, marks $m(\cdot)$, weights w_{in}, and the properties of the underlying distribution of the innovations ε_i. However, theorems derived in the chapter do not require any assumptions which cannot be verified. To illustrate this, I show through an example how each of the assumptions made can be verified. For my example, I use the Wu (1981) model, which I choose as the simplest example with a non-trivial second derivative. For the purposes of the example, I assume that the regressors are stationary with 12 finite moments, the marks reflect third and fourth moments, and the distribution of the innovations is standard normal. I formally show how each of the assumptions made can be verified.

Since my derivations focus on a general class of marked and weighted empirical processes of residuals, they have a wide range of applications, which I discuss in the chapter, such as the Kolmogorov-Smirnov test, moments-based tests of normality, heteroskedasticity tests and analysing the properties of the Huber-skip estimator when applied to non-linear regressions. However, these applications are left for future research.

The chapter originally presented here cannot currrently be made freely available via ORA.

Chapter 2

Computational Aspects of Partial Identification in Practice

Chapter Abstract

This chapter analyses practical and computational aspects of implementing partial identification in practice, with focus on binary outcome models with random coefficients and endogeneity, following the binary IV approach of Chesher and Rosen (2014). The model is contrasted with point-identifying alternatives and applied to the analysis of female labour supply decisions. I analyse the effect of increasing the cardinality of the instrument vector, and the strength of the instrument, on the size of the identified set. I further analyse the computational time complexity required to implement the model in the baseline case of a single endogenous variable and no exogenous variables, and in the presence of exogenous variables, which increase the time complexity of the problem multiplicatively in the cardinality of the support of each exogenous variable. I discuss the computational difficulties associated with adding more sources of unobserved heterogeneity, and the inclusion of continuous variables. Given the novelty of the technique, pseudocode outlining the implementation, as well as the actual Matlab code, are provided.

2.1 Introduction

Over the last decade partial identification literature has become a rapidly growing branch of theoretical Econometrics. Newly-developed methods allow for set-identification of objects of interest under weaker assumptions, shedding light on parameters in settings in which point identification is not possible, as well as offering methods of inference on the identified sets. Yet, despite the methodological advances, partial identification is yet to gain popularity among applied researchers. This chapter aims to bridge the gap between theoretical literature and empirical applications by exploring the computational aspects of implementing partial identifi-

cation in practice. The chapter focuses on binary-outcome models with random coefficients and the partial-identifying approach to those proposed by Chesher and Rosen (2014). The chapter compares the partial-identifying method with point-identifying alternatives applicable under more restrictive assumptions. Since the key identification results of Chesher and Rosen (2014) are based on random set theory — a branch of mathematics which is not commonly known among economists — the chapter discusses them in detail, and illustrates the intuition behind them with a simple graphical example. The chapter offers an empirical application of the method to modelling female labour force participation decisions and it presents example pseudo-code, as well as Matlab code (in the Appendix), for implementing the method. It also discusses computational time issues which affect the method's applicability. Finally, based on the empirical application, the chapter shows an illustration of the type of results that a user of the method can expect to obtain as well as discussing issues that can arise during estimation.

2.1.1 Related Literature and Motivation

In the study of identification, the focus is shifted from sampling error to what can be inferred from the model and the data under ideal conditions, which is achieved by treating the sample as if one observed the population values of observables, e.g. their actual distribution. As noted by Manski (2008), identification has traditionally been treated as binary: parameters could either be identified — if the model and the data allow the researcher to uniquely pinpoint the value of the object of interest; otherwise they are not identified. Partial identification offers a middle ground: it answers the question of what, if anything, one can learn about the parameters of interest by focusing on the set of parameter values which are equally compatible with the model and data ('observationally equivalent').

Partial identification is not a new concept in Econometrics - in fact, its roots date as far back as 1930s and 40s, to the early work of Frisch (1934) and Reiersol (1941) who developed sharp bounds on the slope parameter in a linear regression with errors in variables. However, it was only in the last 15 years that the literature gained its current momentum. Recently-developed estimation methods draw both on established econometric techniques, such as the extension of GMM to partial identification proposed by Chernozhukov, Hong and Tamer (2007); and on less known areas of mathematics, such as methods based on random set theory proposed by Beresteanu and Molinari (2008) and Chesher and Rosen (2017). State-of-the-art methods

allow for conducting inference on the estimated sets, including Andrews and Guggenberger (2009) and the extension proposed by Andrews and Soares (2010), who develop a generalized moment selection procedure; non-parametric inference proposed by Andrews and Shi (2014); and inference in the case of many moment inequalities, as proposed by Chernozhukov, Chetverikov and Kato (2015).

Yet, compared to the vast body of theoretical literature, partial identification methods are yet to gain popularity among applied researchers. There are a number of possible reasons for this scarcity of applications. Firstly, as noted by Ho and Rosen (2017), many applied researchers are simply unfamiliar with these methods. Their article aims to popularise the use of partial identification methods by showing their existing applications in a variety of settings. However, even after familiarising themselves with the methods, applied researchers face further obstacles. In the absence of user-friendly programs[1] researchers must first understand the mathematics of the model — a task particularly challenging in the case of methods based on random set theory which rely on concepts rarely encountered elsewhere in Econometrics. The next step is to translate the mathematics into usable code. The resulting time cost may prove prohibitive, especially if there is no certainty that the calculated bounds will be narrow enough to provide useful information. Finally, for the results to be informative, the data used must satisfy particular requirements, such as sufficient variation in instruments, a large enough dataset or complying with exogeneity conditions. With no prior experience with these methods, researchers may be unable to gauge *a priori* whether their dataset is 'good enough' and may not find out whether this is the case until only after the costly implementation procedure has been performed.

In the spirit of Nevo (2000), the aim of this chapter is to lower the marginal cost of using partial idenfitication methods with a particular focus on the model of Chesher and Rosen (2014) which is applicable to settings with binary outcome variables, random coefficients and endogeneity. In this chapter, I present the method and compare it with commonly used alternatives which allow for point identification. I discuss key identification results and illustrate them with a simple example, to aid understanding. I provide a detailed description of how the method is implemented in practice, including pseudo-code and example Matlab code (in the Appendix). I discuss problems relating to computational time as well as different types of variables (e.g. con-

[1] With the notable exception of a Stata program (Chernozhukov, Lee and Rosen; 2015) implementing the model of Chernozhukov, Lee and Rosen; 2013)

tinuous variables). I apply the method to the analysis of female labour participation decisions, using data from the British Household Panel Survey, presenting the type of results that a user can expect, showing the impact of adding more instruments on the estimated set, and discussing limitations of the method.

A recent contribution by Chesher and Rosen (2020), published after this chapter was written, applies the Generalized Instrumental Variables model of Chesher and Rosen (2017) to a similar setting of female labour supply decisions. However, their research differs from the approach taken in this chapter, as they only allow for a single source of unobserved heterogeneity. By contrast, this chapter implements the binary IV model of Chesher and Rosen (2014), which allows for multiple such sources.

The chapter is organised as follows. Section 2.2 presents the binary-outcome, random- coefficients model with endogeneity and argues that it is particularly suited to the analysis of female labour force participation decisions. Section 2.3 presents the set-identifying model of Chesher and Rosen (2014) and shows its relation to alternative, point-identifying, methods. Section 2.4 discusses the data used in the empirical illustration as well as motivating the choice of instruments. Section 2.5 shows how the model can be implemented in practice. This section also examines computational time issues, as well as discussing estimation in the presence of continuous variables and multiple endogenous variables. Section 2.7 illustrates points from the previous section with results from the empirical application. Section 2.8 concludes.

2.2 The Model

2.2.1 Formal Model

The focus is on binary-outcome models with random coefficients and endogeneity. The binary outcome variable, $Y_i \in \{0, 1\}$, where i denotes a cross-sectional observation, is modelled using a threshold-crossing approach: it is assumed that $Y_i = 1$ if and only if an underlying latent variable, Y_i^*, is non-negative:

$$Y_i = 1.\left[Y_i^* \geq 0\right]. \tag{2.1}$$

The latent variable itself is a linear-in-parameters function of two types of vectors of characteristics observable to the researcher: endogenous variables, E_i, and exogenous variables, W_i.

As shown in equation (2.2) the intercept and the slope coefficients on both types of variables are allowed to be fully heterogeneous[2]

$$Y_i = 1.\left[Y_i^* \geq 0\right] = 1.\left[\alpha_i + \beta_i E_i + \gamma_i W_i \geq 0\right]. \tag{2.2}$$

Crucially, in a further complication compared to standard random-coefficient models, all random coefficients are allowed to be arbitrarily correlated with the vector of endogenous variables, E_i. By contrast, the joint distribution of the random coefficients $(\alpha_i, \beta_i, \gamma_i)$ is stochastically independent of the vector of exogenous variables, W_i, as well as being stochastically independent from a vector of observed instruments, Z_i, which are excluded from the outcome equation.

While accounting for correlation between endogenous variables and an additive error term (here captured by α_i) is common in Econometrics, models which allow for correlation between endogenous variables and random slope coefficients (β_i and γ_i) are rare, despite a strong rationale for why such a relationship may arise. Following Wooldridge (2010), such models include the case in which there is an omitted variable in the model which enters interactively with an included variable. Using his example, in a wage offer equation, unobserved ability may enter interactively with education due to wages increasing disproportionately for those whose ability is higher.

Importantly, the model of Chesher and Rosen (2014) does not require formally modelling the relationship between the endogenous variables, E_i, and the instruments, Z_i. This is contrast to point-identifying control function approaches of Hoderlein (2009) and Hoderlein and Sherman (2015) discussed in Section 2.3. As emphasised in Chesher and Rosen (2014, p. S3), the absence of an equation which would determine the value of E_i conditional on (Z_i, W_i) and unobserved heterogeneity $(\alpha_i, \beta_i, \gamma_i)$, results in a model which is incomplete.

Although equation (2.2) (or its simplified variants) is common to all models discussed in this chapter, a key feature of the approach of Chesher and Rosen (2014) and one which distinguishes it from others is its focus: the approach aims to set-identify the entire marginal distribution of the vector of random coefficients, $(\alpha_i, \beta_i, \gamma_i)$, as opposed to, for example, identifying only its mean.

Whether using point- or set- identifying approaches, the aim of identifying the parameters

[2]Since in this chapter I will mostly be working with a single endogenous variable and a single included exogenous variable, I do not include transposes in the model. However, the notation is consistent with including multiple variables under the convention that all vectors of parameters are row-vectors, while the vectors of variables are column vectors.

can only be achieved up to scale. This is due to the threshold-crossing nature of the model. Formally, for any real positive constant $c > 0$, multiplying all coefficients in (2.2) by c does not affect the decision rule. Usual identification approaches rely on normalising the variance of one of the random coefficients to 1, or normalising the length of the coefficient vector to 1: $\|(\alpha_i, \beta_i, \gamma_i)\| = 1$. Following Chesher and Rosen (2014), I adopt the former approach.

2.2.2 Empirical Motivation

The model presented in equation (2.2) is particularly well suited for studying female labour force participation decisions. Firstly, the outcome variable is binary, in line with a woman being either in ($Y_i = 1$) or out of ($Y_i = 0$) the labour force. The threshold-crossing model is widely used in studying binary labour supply decisions as it corresponds to the assumption that an individual is working if and only if her wage offer exceeds her reservation wage. Here, Y_i^* can be interpreted as the difference between person i's wage offer and her unobserved reservation wage.

Secondly, a key variable determining one's wage offer and, therefore, labour force participation is education, which is likely to be correlated with individual characteristics captured in the additive error term, α_i. Informally, individuals with a higher 'taste' for hard work or higher levels of unobservable ability, are likely to choose more education as well as being able to command higher wages given the same level of education and other observable characteristics.

Furthermore, there is wide evidence of heterogeneity in the returns to education which, given the threshold-crossing approach, are captured by the vector of random coefficients, β_i. Evidence of heterogeneity in returns to education includes, but is not limited to, Blundell, Dearden and Sianesi (2001) who use a random-coefficients model on UK NCDS data; Harmon, Oosterbeek and Walker (2003) using quantile regression on UK FES data; O'Leary and Sloane (2005) as well as Walker and Zhu (2011), both based on OLS using UK LFS data; Chevalier (2011) using the Longitudinal Destination of Leavers of Higher Education dataset linked to administrative data; and for the US, for example, Henderson, Polachek and Wang (2011) who find significant heterogeneity in returns to education within and across groups, based on a kernel regression analysis using US Census data. Additionally, heterogeneity in the returns to education is understood to be the reason why estimates of the Local Average Treatment Effect (LATE) vary substantially across different instruments and datasets (e.g. Blundell, Dearden and Sianesi, 2001; Harmon, Oosterbeek and Walker, 2003).

Crucially, as pointed out by Hoderlein and Sherman (2015), returns to education and the chosen level of education are likely to be correlated. Using their example, we can consider an individual with a high taste for learning. The person would be more likely to obtain more education even if the returns are low, than a person who is otherwise identical but enjoys learning less. This endogeneity leads to inconsistency of the standard OLS estimator of the mean of the distribution of random coefficients (Heckman and Vytlacil, 1998).

Therefore, since the model analysed in this chapter allows for the endogenous variables to be arbitrarily correlated with the vector of unobservable heterogeneity, $(\alpha_i, \beta_i, \gamma_i)$, it is particularly well-suited for use in analysing female labour force participation decisions.

While the focus of the empirical application in the current chapter is on female labour supply decisions, the binary-outcome random-coefficients model can also be applied to other settings in which the endogenous regressors may be correlated with the heterogenous slope coefficient in a binary-outcome model. For example, Hoderlein and Sherman (2015) study the occurrence of abnormal birth conditions in newborns as a function of smoking during pregnancy. The authors argue that the degree to which smoking increases the risk of abnormal birth conditions (i.e. the slope coefficient) is correlated with smoking itself, due to unobservable risk factors correlated with smoking, such as poor diet or drug use.

2.3 Methods

The first part of this section presents methods alternative to Chesher and Rosen (2014), which achieve point identification but at the cost of imposing stronger assumptions. The second part discusses and illustrates key identification results of Chesher and Rosen (2014).

2.3.1 Point-Identifying Methods

2.3.1.1 Binary Choice Models with Endogeneity and Constant Coefficients

The class of models discussed in this section treats all slope coefficients as constant. Formally:

$$\forall i : \beta_i = \beta, \gamma_i = \gamma. \tag{2.3}$$

This model is applicable when the researcher has good grounds to believe that there is no

heterogeneity in the slope coefficients among observations. Clearly, the only type of endogeneity that can arise in this model is the correlation between the endogenous variables E_i and the random intercept, α_i.

Let us consider the case of a single endogenous regressor. Following Wooldridge (2010), among parametric methods the appropriate approach to this model depends on whether the endogenous variable is continuous or discrete. In both cases the relationship between the endogenous variable and the instruments is modelled explicitly.

In the case of a continuous variable E_i, the following first-stage equation is estimated:

$$E_i = \pi Z_i + \delta W_i + \nu_i. \tag{2.4}$$

If the assumption of normality of ν_i can be imposed, then the model can be consistently estimated using the control function approach: equation (2.4) is estimated by OLS and the residuals from the equation, $\hat{\nu}_i$ are added to equation (2.2) which is then estimated by probit, with an appropriate correction made to the standard errors. An alternative is to use an 'instrumental variable' version of the probit model, in which equation (2.4) is estimated by OLS and used to obtain fitted values of the endogenous variable, $\hat{E}_i$. Contrary to the control function approach, this method does not provide a simple test of whether E_i is indeed endogenous. Yet another option is to estimate the model through conditional maximum likelihood, which is more efficient than the two-step procedures.

When the endogenous variable is discrete, its dependence on the instrumental variables must be modelled using a limited dependent variable approach. Some two-step procedures, such as the analogue to two-stage least squares, no longer produce consistent results due to the 'forbidden regression' problem: the endogenous variable depends nonlinearly on the parameters and the expectation of a nonlinear function of an estimated vector of parameters is not equal to the function of the expectation. Instead, full-information methods, such as maximum likelihood, should be used. In the case of a binary endogenous regressor, this can be achieved using the bivariate probit model, in which both the outcome variable and the endogenous variable are modelled jointly.

Both of the above cases are easy to estimate in practice — in fact, Stata commands (`ivprobit` and `biprobit` respectively) allow the researcher to obtain results within seconds. As such, if the

researcher has grounds to believe that the slope coefficients are constant across the population, the above models may be preferable to the remaining approaches given the time investment involved in estimation.

In addition to the above parametric approaches, the binary-outcome model with endogeneity but with constant coefficients has also been analysed by Blundell and Powell (2004), using a semiparametric approach, and by Vytlacil and Yildiz (2007) in a nonparametric approach.

2.3.1.2 Binary-Outcome Random-Coefficient Models with Exogenous Regressors

This class of models corresponds to imposing the following restriction in equation (2.2):

$$\forall i : \beta_i = 0. \tag{2.5}$$

In words: all explanatory variables are exogenous and thus, included in the vector W_i.

In the more general case of multinomial choice with exogenous regressors, which encompasses binary choice, both probit and logit models have been extended to allow for random coefficients. In both models the distribution of random parameters is specified *a priori*. In the probit case all random parameters are assumed to be normally distributed, which has limited applicability due to the support of the normal distribution: in some cases, such as consumer choice, economic theory may dictate that the coefficients lie on one side of zero. The alternative model, mixed logit, imposes that the random intercept is independently and identically extreme-value distributed. However, the model allows more flexibility in the choice of distribution of the random slope coefficients, including the longnormal and Rayleigh distributions which force coefficients to be of the same sign, and the triangular distribution which can achieve the same goal by imposing additional restrictions on its parameters. In both the mixed probit and mixed logit models, estimation is performed using simulation algorithms.

Studies which identify and estimate the distribution of random coefficients in binary-outcome models with exclusively exogenous regressors include Bajari et. al. (2012), Gautier and Kitamura (2013) and, in a non-linear setting, Fox and Ghandi (2012).

While the methods are relatively easy to implement, they are only applicable in the case of exogenous explanatory variables. Although in a number of empirical settings this assumption may hold true, it is likely to be violated in many others, such as in analysing female labour force

participation decisions.

2.3.1.3 Binary Choice Models with Random Coefficients and Endogenous Regressors: Estimating the Mean of the Distribution of Random Coefficients

Point-identifying approaches which allow for binary-choice models with random coefficients and arbitrary endogeneity are rare in the literature. This section discusses three such methods. All three are limited in that they allow to identify only the mean of the random coefficients.

Lewbel's (2000) 'special regressor' method can be applied to obtain the mean of the distribution of random coefficients but relies on the existence of a particular type of regressor: one which both has heavy tails and, conditional on the instruments, is independent of the structural error and other regressors in the model.

Hoderlein and Sherman (2015) developed a $\sqrt{n}$ - consistent (under a particular conditional median restriction) and asymptotically normal estimator of the trimmed mean of the distribution of random coefficients. The focus on the trimmed mean rather than the conditional expectation is due to the latter suffering from a curse of dimensionality but the authors argue that the trimmed mean can be made arbitrarily close to the actual mean by selecting a large enough trimming constant. The approach uses a version of the smoothed maximum score estimator of Horowitz (1992).

In a more restricted approach, Hoderlein (2009) identifies the mean of the vector of heterogeneous coefficients in a binary-outcome model with endogeneity. However, his approach explicitly precludes correlation between unobservable preferences, which determine random slope coefficients, and explanatory variables - whether included or not. Thus, the only correlation which is admissible is a correlation between the explanatory variables and an additive error term. The article proposes a method of estimation of the mean of the vector of random slope coefficients, proves its consistency and derives the asymptotic distribution. Moreover, tests for heteroskedasticity, endogeneity and overidentification are proposed. The identification method uses a control-function approach and relies on a median exclusion restriction and the assumption that the random slope coefficients and random intercept are jointly symmetrically distributed about an unknown mean.

Methods presented in this section have the advantage of achieving point identification of the mean of random coefficients and, by deriving the asymptotic distribution of the estimators,

allow for conducting inference. By contrast, the method of Chesher and Rosen (2014), to be discussed shortly, allows the researcher to recover the entire distribution of random parameters, however, at the cost of obtaining only set-estimates.

The choice between methods will depend on the application at hand and should be driven by the researcher's object of interest: whether it is the mean effect of a variable across the population, or the entire marginal distribution of the effect. For example, in the case of returns to education, concerns regarding social mobility or pricing of tuition fees may lead the researcher to want to find out not only the mean of the returns to education but also their variance or, indeed, their entire marginal distribution.

2.3.2 Model of Chesher and Rosen (2014)

2.3.2.1 Outline and Key Results

The approach of Chesher and Rosen (2014) accounts for all aspects of the model of interest: all slope coefficients are fully heterogeneous and can be arbitrarily correlated with the endogenous variables. As emphasised before, the focus of estimation is on the entire marginal distribution of random coefficients. The approach relies on the presence of instruments which are exogenous with respect to random coefficients, but idenfitication does not require specifying a functional form for the relationship between endogenous variables and instruments. This flexibility comes at a cost: the distribution of random parameters can only be set-identified and the size of the set depends on such factors as the distributional assumptions made and the support and variability of the instruments.

Formally, the model posits one equation - the outcome equation (2.2):

$$Y_i = 1.\left[Y_i^* \geq 0\right] = 1.\left[\alpha_i + \beta_i E_i + \gamma_i W_i \geq 0\right].$$

Additionally, it is assumed that there exists a vector of instruments which are exogenous with respect to all random coefficients:

$$Z_i \perp (\alpha_i, \beta_i, \gamma_i). \tag{2.6}$$

An explicit relationship between the endogenous variables, E_i, and the instruments, Z_i, can be added for strength of identification but is not necessary. Moreover, no rank or order conditions

for identification are imposed.

It is assumed that, under simple random sampling, the joint distribution of observables, $F_{obs}(Y, E, W, Z)$, is identified. The method requires that under F_{obs} both outcomes: $Y_i = 1$ and $Y_i = 0$ arise with a strictly positive probability for every possible combination of (E, W, Z).

Let $\theta_i = (\alpha_i, \beta_i, \gamma_i)$ be the vector of random coefficients for observation i and let θ be a matrix of those coefficients for the population. The aim of the analysis is to identify the entire marginal distribution of θ: F_θ. F_θ belongs to a class of distribution functions chosen by the researcher: this can, in principle, be parametrically, semi-parametrically, or non-parametrically specified. The choice of distribution involves a trade-off: stronger *a priori* assumptions (e.g. a parametric distribution) lead to greater identification strength but at the cost of risking mis-specification of the model.

Let Θ be the entire parameter space to which θ belongs. The key identification result of Chesher and Rosen (2014, equation 3.2) states that a given conjectured distribution $\tilde{F}$ of random parameters satisfies the requirements of the model and the data if and only if the following holds for all *closed* subsets $S \subseteq \Theta$:

$$\tilde{F}(S) \geq \mathbb{P}\left[\mathcal{T}(W, E, Y) \subseteq S | (w, z)\right], \tag{2.7}$$

where $\mathcal{T}(W, E, Y)$ is the closure of the set of all values of θ which are compatible with a given combination of (W, E, Y); and $\mathbb{P}(.)$ is the probability of the set $\mathcal{T}(W, E, Y)$ being contained in the given test set S - this probability is based on the identified distribution of observables, F_{obs}.

Intuitively, the inequality requires that, for each test set S, which is a closed subset of the parameter space, the probability that this set of parameter values arises under the hypothesised distribution $\tilde{F}(.)$ is weakly greater than the probability of drawing a combination of observables corresponding to the set S, as captured by the probability $\mathbb{P}\left[\mathcal{T}(W, E, Y) \subseteq S\right]$.

Obtaining probability $\mathbb{P}\left[\mathcal{T}(W, E, Y) \subseteq S\right]$ requires, for every set S, checking which combinations of observables, (W, E, Y) are compatible with the set $\mathcal{T}(W, E, Y)$ being fully contained in S, then adding the probabilities of those combinations arising under F_{obs}. An example illustrating the concepts in practice will be provided in what follows.

The key result (2.7) is derived from applying Artstein's inequality (Artstein, 1983; see also Molchanov, 2005 and 2017; and Beresteanu, Molchanov, Molinari, 2012) which provides nec-

essary and sufficient conditions for the event that a given random vector belongs to a given random set, to occur with probability 1.

The use of Artstein's inequality simplifies the problem of searching for the marginal distribution of random coefficients to the easier task of checking whether a set of conjectured distributions satisfies a set of moment inequalities given by equation (2.7). The task is further simplified by applying methods of Chesher, Rosen and Smolinski (2013) and Chesher and Rosen (2012, 2013) whose aim is to find a set of sufficient test sets S such that, if (2.7) holds for those, then it holds for all closed sets $S \subseteq \Theta$. Those are referred to as 'core-determining sets'. The collection of core-determining sets depends on the support of the random set under consideration. Chesher and Rosen (2014) argue that, following analogous reasoning as in the above articles, in the specific case of a binary-outcome random-coefficients model with endogeneity, the collection of core-determining sets is the collection of unions of sets that belong to the support of $\mathcal{T}(W, E, Y)$ conditional on the realization (w, z) of exogenous variables. The support is defined as:

$$\mathsf{T}(w, z) \equiv \{\mathcal{T}(w, e, y) : y \in \{0, 1\} \wedge e \in Supp(E|w, z)\}. \tag{2.8}$$

Equation (2.8) states that the conditional support, $\mathsf{T}(w, z)$, of $\mathcal{T}(W, E, Y)$ is the collection[3] of all sets of values of the parameters θ which are compatible with some combination of (Y, E) arising with non-zero probability conditional on values (w, z) of the exogenous variables.

2.3.2.2 Illustration of Key Identification Results

This section illustrates equations (2.7) and (2.8) with an example. To make the model as simple as possible while preserving all key features of the model, let us assume that there is only one endogenous variable, E, and no included exogenous variables. Thus, the outcome equation (2.2) simplifies to:

$$Y_i = 1.\left[\alpha_i + \beta_i E_i \geq 0\right]. \tag{2.9}$$

Furthermore, for ease of exposition, let us assume that there is only one instrument, Z, and that both E and Z are binary:

$$\forall i : E_i \in \{0, 1\}, Z_i \in \{0, 1\} \quad \text{and} \quad Z_i \perp (\alpha_i, \beta_i, \gamma_i).$$

[3]To be precise, it is the collections of closures of such sets of θ. The notion of closure is omitted to simplify exposition.

Conditional on a given value of Z, there are only 4 combinations of (Y, E): (1,1), (1,0), (0,1), (0,0). The requirements of identification are such that, conditional on any vector of values of (E, Z), both outcomes: $Y = 1$ and $Y = 0$ arise with a positive probability. In our simple model this requires that conditional on both $Z = 0$ and $Z = 1$, all 4 outcomes arise with positive probabilities under F_{obs}.

The 4 outcomes partition the parameter space Θ in 4 ways, based on areas which satisfy equation (2.9). In particular, if $(Y_i, E_i) = (1, 1)$ then, in line with equation (2.9) and the definition[4] of $\mathcal{T}(W, E, Y)$, we have:

$$\begin{aligned} \mathcal{T}(1,1) &= cl\{(\alpha_i, \beta_i) : \alpha_i + 1 \times \beta_i \geq 0\} \\ &= \{(\alpha_i, \beta_i) : \alpha_i \geq -\beta_i\}. \end{aligned} \tag{2.10}$$

The above set is illustrated in Figure 2.1a. Multiplication by 1 in the first line is due to $X_i = 1$. The reason the left hand side is greater or equal than 0 follows from $Y_i = 1$.

By contrast, the set of parameters compatible with the outcome $(Y_i, E_i) = (0, 0)$, illustrated in Figure 2.1d, is the following closure of a set:

$$\begin{aligned} \mathcal{T}(0,0) &= cl\{(\alpha_i, \beta_i) : \alpha_i + 0 \times \beta_i < 0\} \\ &= cl\{(\alpha_i, \beta_i) : \alpha_i < 0\}. \end{aligned} \tag{2.11}$$

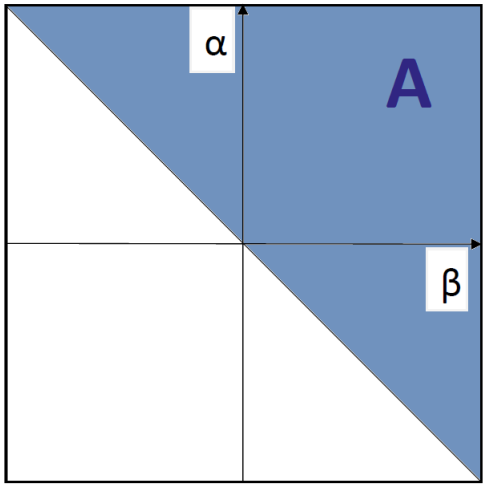
In this case, the parameter β_i is multiplied by 0 since $X_i = 0$. The left hand side is negative in line with equation (2.9) and the value taken by the dependent variable: $Y_i = 0$.

The partition of the parameter space corresponding to the outcome $(Y_i, E_i) = (1, 0)$ is illustrated in Figure 2.1b and, by analogy to the examples above, is given by the following expression:

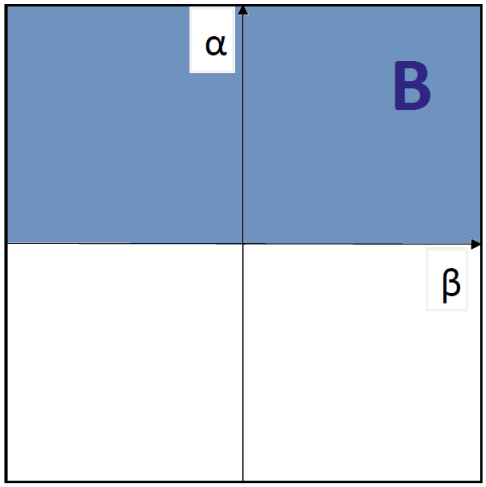
$$\begin{aligned} \mathcal{T}(1,0) &= cl\{(\alpha_i, \beta_i) : \alpha_i + 0 \times \beta_i \geq 0\} \\ &= \{(\alpha_i, \beta_i) : \alpha_i \geq 0\}. \end{aligned} \tag{2.12}$$

Finally, Figure 2.1c illustrates the subspace of the parameter space corresponding to the

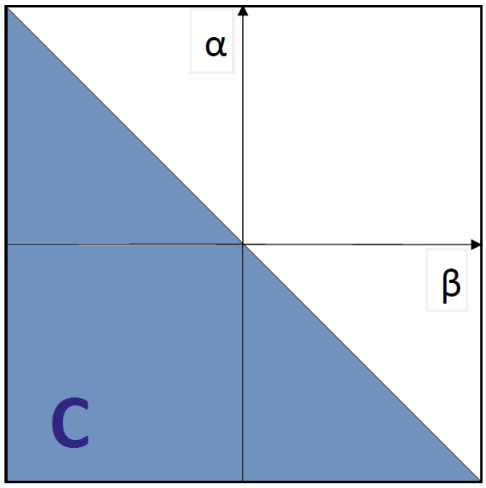
[4]Formally, we should further take the closure of the set. However, in this particular example, the set is closed since the inequality which defines the set is weak.



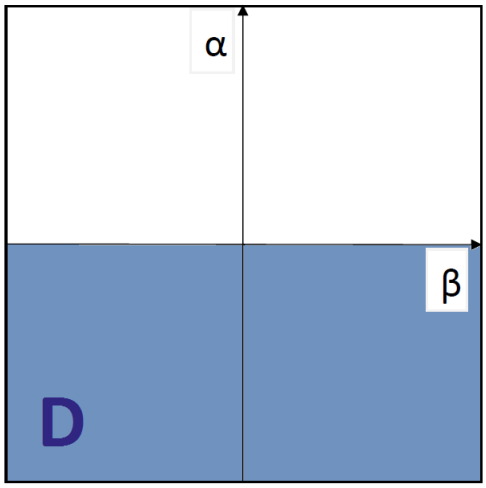
(a) $Y_i = 1, X_i = 1$



(b) $Y_i = 1, X_i = 0$



(c) $Y_i = 0, X_i = 1$



(d) $Y_i = 0, X_i = 0$

Figure 2.1: Partitions of the parameters space corresponding to different combinations of observables

remaining possible outcome, $(Y_i, E_i) = (0, 1)$, in which case the set $\mathcal{T}$ is given by:

$$\begin{aligned} \mathcal{T}(0,1) &= cl\{(\alpha_i, \beta_i) : \alpha_i + 1 \times \beta_i < 0\} \\ &= cl\{(\alpha_i, \beta_i) : \alpha_i < -\beta_i\}. \end{aligned} \tag{2.13}$$

In order to illustrate the key identifying equation, (2.7) using the above sets, let us denote the sets in Figures 2.1a, 2.1b, 2.1c, 2.1d by A, B, C, D respectively. Since all 4 outcomes occur with positive probability for both $Z = 0$ and $Z = 1$, the conditional support $\mathsf{T}(z)$ is the same for both values of Z: $\mathsf{T}(0) = \mathsf{T}(1) = \{A, B, C, D\} \equiv \mathsf{T}$.

Suppose we wish to test whether a conjectured distribution $\tilde{F}$ of parameters (α_i, β_i) satisfies the sufficient and necessary condition stated in inequality (2.7). The first step in applying the inequality consists of determining the collection of tests sets S for which the inequality must hold. As defined in the previous section, the collection of core-determining sets is the collection of all unions of elements of T. Omitting trivial unions (i.e. those which cover the entire space), the collection of those sets is, therefore: $\mathcal{S} = \{A, B, C, D, A \cup B, A \cup D, B \cup C, C \cup D\}$.

The conjectured marginal distribution of random parameters, $\tilde{F}$, belongs to the identified set of distributions if and only if for every set $S \in \mathcal{S}$, inequality (2.7) holds for $\tilde{F}$, i.e. for every set $S \in \mathcal{S}$ and conditional on every value of the instrument $Z = z$ we have:

$$\tilde{F}(S) \geq \mathbb{P}\left[\mathcal{T}(E, Y) \subseteq S | (z)\right]. \tag{2.14}$$

Since S is a subset of the parameter space Θ, and $\tilde{F}$ is the distribution of parameters, the left hand side of the inequality is computed by integrating the probability density function corresponding to $\tilde{F}$ over the region defined by the given S. For example, if S is chosen to be the second element of the set $\mathcal{S}$, i.e. $S = B$, then the left hand side corresponds to the probability placed by distribution $\tilde{F}$ on the parameter α_i being nonnegative, as illustrated in Figure 2.1b. Probabilities for all other elements of set $\mathcal{S}$ are calculated analogously. Note that, since (α_i, β_i) are independent of Z, the left hand side of the inequality is the same for a given set S irrespective of the value of the instrument. In the case of more complex models, in which the cardinality of $\mathcal{S}$ is much higher, this observation can save a substantial amount of computational time.

Probabilities on the right hand side are based on the identified distribution of observables, F_{obs}, which, under simple random sampling, is estimated by the empirical distribution of the data. For example, if the test set is $S = B$, then the corresponding right-hand-side probability is the relative frequency of the pair $(Y_i, E_i) = (1, 0)$ arising in the data, conditional on $Z = 1$ and then on $Z = 0$. In the case of tests sets which are unions of simpler sets, such as $S = A \cup B$, the corresponding right-hand-side probability is the probability of the event that either of the combinations of data corresponding to sets A or B arises, i.e. the event $\{(Y_i, E_i) = (1, 1) \vee (Y_i, E_i) = (1, 0) | Z = z\}$.

Since the right-hand-side probabilities are conditional on a given value of $Z = z$, they will vary across values of the instrument. This also shows why increasing the cardinality of the

support of Z is likely to increase the identifying power of the model: since inequality (2.7) must hold for every value of exogenous variables, the more values the exogenous variables take (this includes both Z and W), the more moment inequalities a given conjectured distribution has to satisfy.

The identified set of distributions is obtained by first selecting a class $\mathcal{F}$ of distributions that the researcher believes the actual marginal distribution to belong to; and then testing, for each member F of the class $\mathcal{F}$ whether inequality (2.7) is satisfied for each core-determining set and conditional on each realisation of exogenous variables arising with non-zero probability. The set of distributions F which satisfy those criteria is the identified set.

For example, in a parametric case, $\mathcal{F}$ will be specified up to a vector of parameters. In such case, the set of identified distributions will be determined by the set of all such vectors of parameters of the distribution F for which inequalities in (2.7) are satisfied. The testing will normally involve initial tests over a coarse grid of parameters, followed by testing over a narrower but finer grid, or using more sophisticated search methods, such as bisection algorithms or testing locally.

2.4 Empirical Application

This section discusses the details of an empirical application of the Chesher and Rosen (2014) approach to estimating a model of female labour force participation. The aim of this application is to serve as an illustration of the type of results that a user of the method can expect and the type of issues they may encounter when implementing the method in practice.

2.4.1 Data

The dataset[5] used in this chapter is Wave 8 of the British Household Panel Survey (BHPS)[6]. BHPS is a data series started in 1991, which consists of annual interviews of the same households. The questionnaire collects detailed information on each household member, including key variables relating to labour force participation decisions, such as the number of children under the person's care, the age of the youngest child, caring responsibilities, education (as a

[5]The author would like to thank Jonathan Shaw at the Institute for Fiscal Studies for allowing me to use the IFS' Stata code for cleaning BHPS data which creates measures of BHPS variables which are consistent across waves of the survey.

[6]A public version of the dataset was obtained from the UK Data Service (University of Essex).

categorical variable), information on the person's parents (e.g. their level of education, parents' employment status when the person was 14 years old), marital status, etc. Wave 9 included additional questions relating to past smoking habits, which allowed for the construction of the particular instrument used. The choice of Wave 8 was made on the basis of it being closest to Wave 9 but without including 'boost samples' for Scotland and Wales which were introduced in Wave 9.

The analysis uses only one wave of the panel, treating it as a cross-section. Only women aged 25-45 are included. Since comparison results will involve estimating an instrumental variable probit model for a specification with a large set of covariates, for consistency across specifications, observations with missing information on any of the explanatory variables or instruments used in the richest specification[7] are included. The resulting sample has 672 observations.

The outcome variable in the model is an indicator of working status and takes value of 1 if the woman is employed or self-employed, and value of 0 otherwise. The BHPS does not contain reliable information on years of education, instead focusing on school leaving age and qualifications. Inferring a close-to-continuous measure of education from those would be prone to measurement error, which could contaminate the results. Additionally, although the approach of Chesher and Rosen (2014) does not require rank or order conditions, those are still necessary for the comparison 'IV probit' model. In the presence of only a limited number of instruments, including several dummies for different levels of education is therefore infeasible. As a result, the measure of education used is an indicator of university education, taking the value of 1 if a person's has a degree (first degree or higher degree) and value of 0 otherwise.

Following Dickson (2013) the instrument used is an indicator of whether the person was smoking when young. Specifically, it takes the value of 1 if the person started smoking cigarettes regularly at or before the age of 16, and takes the value of 0 if the person started smoking after the age of 16, or if she never smoked regularly. The distribution of the instrument is shown in Table 2.1.

The variable was first proposed as an instrument for schooling by Evans and Montgomery

[7]Those include, among explanatory variables: education, partner's education group, partner's employment status, region, age, council tax band, tenure type, household type, marital status, whether ever married, number of rooms in the house, number of children under 12 the respondent is responsible for, age of youngest child in household; and, among instruments: whether lived with parents at the age of 16, smoking instrument, mother's and father's education group, whether an only or eldest child, whether there were a lot of books in childhood home.

(1994). Their initial argumentation was based on the relationship between education and better health habits. Empirical evidence of such relationship was found, for example, by Grossman (2006) and Oreopoulos (2006). An stronger justification, and one specific to smoking, was proposed by Fuchs (1982) who highlighted the importance of rates of time preference in both the decision whether to smoke and whether to go to school - both involve current costs (going to school, abstaining from smoking) traded against future benefits (higher wages, better health). Since individuals are observed at the age of 25 or above, smoking at 16 is a pre-determined variable. 16 is also the age at which students in the UK are allowed to leave school or alternatively start studying for A-levels as a route to university. Thus, information about their smoking habits at the age of 16 sheds light on their rate of time preference at the age when they were making important education decisions. Dickson (2013) estimates a probit model of current smoking and of smoking at the age of 16, including years of completed education among the explanatory variables. He concludes that it is not education that determines whether someone smokes, but that both early smoking and education are driven by the same underlying factor — time preference.

Burgess and Propper (1998) find that economic outcomes 10 years later for those who used alcohol or soft drugs in adolescence were unaffected, supporting the assumption of the exogeneity of the instrument. Moreover, as argued by Dickson (2013), there is a substantial degree of movement across the groups of early smokers and early non-smokers: with 41.0% of men who smoked when young ceasing to do so by the time they were interviewed and 37.5% moving in the opposite direction. Furthermore, contrary to instruments such as raising of the school leaving age which affect only those at the bottom of the ability distribution, smoking when young is present across all quintiles of the distribution.

One of the aims of this chapter is to present the impact of increasing the cardinality of the support of the instrument vector on the size of the identified set. To achieve this, two further instruments are added: indicators for the mother's and the father's further education. The variables take value of 1 if the relevant parent has a degree (first degree or higher degree) or if the parent has other further education qualifications (e.g. teaching, nursing, or other further qualifications). The variables take value of 0 otherwise. The choice of the parent's further education rather than university education is motivated by higher variability of the instrument. Given the timing of interviews conducted in 1998 and the age of individuals in the sample (25-45 years old), only 24 women in the sample (3.6%) were born to a mother with university education

and only 39 women (5.8%) to a father with university education. As a result, the more variable measure of further education was used, as well as interactions of the three instruments. The distribution of values of the instruments in the sample, as well as the probability distribution of their cross-products, is shown in Tables 2.1 and 2.2.

Table 2.1: Distribution of Instruments: Smoking and 2-way Cross-Products

Smoking	Parent's FE	Smoking	Smoking × Dad's FE	Smoking × Mum's FE
1	1	38.24 %	12.50 %	6.55 %
1	0		25.74 %	31.70 %
0	1	61.76 %	25.45 %	11.31 %
0	0		36.31 %	50.45 %

Table 2.2: Distribution of Instruments: 3-way Cross-Products

Smoking	Mum's FE	Dad's FE	Frequency
1	1	1	4.46 %
1	1	0	2.08 %
1	0	1	8.04 %
1	0	0	23.66 %
0	1	1	7.44 %
0	1	0	3.87 %
0	0	1	18.01 %
0	0	0	32.44 %

Increasing the cardinality of the support of the instrument vector — in this case by increasing the number of combinations of the binary instruments which occur with non-zero probability — reduces the percentage of observations falling into each category. As explained in Section 2.3.2.2, adding more values of the instrument weakly increases identifying power of the model by increasing the number of moment inequalities. However, for a fixed sample size, the more values the instrument takes, the smaller the sizes of each of the cells. For example, as shown in Table 2.2, the case of a person having smoked when young, her mother having a further education qualification but her father not having one, occurs only in 2.08% of the sample, i.e. 14 individuals. This number is then further broken down into categories based on values of the outcome variable and education. Problems due to estimation based on such small cell sizes are discussed in Section 2.7.

Tables 2.3 and 2.4 illustrate the strength of the instruments, using probit regressions with the education binary as the dependent variable. Table 2.3 uses the same instruments and exogenous variables which are used in the partially-identified models. Section 2.4.2 provides a justification

Table 2.3: Strength of Instruments

	(1)	(2)	(3)	(4)	(5)	(6)
	University Education (0/1)					
Smoking when young	-0.651***	-0.655***	-0.617***	-0.632***	-0.655***	-0.648***
	(0.151)	(0.153)	(0.153)	(0.155)	(0.154)	(0.159)
Mum's FE		0.594***		0.476**		0.522**
		(0.150)		(0.157)		(0.161)
Dad's FE			0.476***	0.366**		0.360**
			(0.131)	(0.137)		(0.139)
Age 31-35					0.115	0.107
					(0.212)	(0.218)
Age 36-40					0.359	0.345
					(0.207)	(0.212)
Age 41-45					0.592**	0.641**
					(0.217)	(0.223)
Intercept	-0.989***	-1.124***	-1.214***	-1.269***	-1.270***	-1.561***
	(0.0738)	(0.0837)	(0.0997)	(0.102)	(0.177)	(0.196)
N	672	672	672	672	672	672

Standard errors in parentheses. Age group 25-30 dropped due to collinearity.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

for this choice of specifications. In specification (5) the fourth age category (age 25-30) is omitted to avoid collinearity.

Table 2.4 shows probit regressions of the education variable on a full set of variables available in BHPS which can be used in modelling female labour force participation decisions. The aim of these two specifications is to mimic the first stage regression of a two-stage model which could be estimated in the absence of random coefficients and, therefore, in the absence of a correlation between education and the returns to education, yet still correcting for the correlation between education and the additive error term. In addition to the variables shown, specification (1) contains regional dummies. Specification (2) additionally controls for the number of children under 12 the respondent is responsible for, the age of youngest child in household, number of rooms in the house as well as dummy variables for: region, council tax band, tenure type, household type, partner's education group, partner's employment status, whether married at the time of the interview, and whether ever married.

As results in Table 2.3 illustrate, in the specifications used in partially identified models,

Table 2.4: Strength of Instruments

	(1)	(2)
	University Education (0/1)	
Smoking when young	-0.662***	-0.511*
	(-4.04)	(-2.55)
Eldest or only child	0.163	0.186
	(1.16)	(1.10)
Lived with parents at 16	-0.0510	0.0673
	(-0.26)	(0.28)
Mum's FE	0.493**	0.378
	(2.94)	(1.91)
Dad's FE	0.353*	0.242
	(2.35)	(1.33)
Books - childhood home	0.406*	0.363
	(2.06)	(1.54)
Age	0.159	-0.0939
	(0.78)	(-0.38)
Age squared	-0.00143	0.00240
	(-0.51)	(0.70)
Number of roooms in the house		0.103
		(1.35)
Currently married		-0.297
		(-0.83)
Partner's employment status		0.218
		(0.53)
Age of respondent's youngest child		-0.0598*
		(-2.15)
Number of children under 12 respondent is responsible for		-0.0171
		(-0.15)
Ever married		-0.354
		(-0.76)
Partner's education: GCSEs		0.249
		(0.67)
Partner's education: A Levels		0.598
		(1.60)
Partner's education: Vocational Higher		0.308
		(0.89)
Partner's education: University		1.376***
		(4.04)
Intercept	-5.464	-1.118
	(-1.49)	(-0.26)
N	672	654

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Both specification include regional dummy variables, which, for ease of exposition, are not shown. Additionally, specification (2) includes dummies for household type, tenure type and council tax band.

all three instruments are strong. There is a strong and statistically significant negative partial correlation between smoking when young and attending university. The correlation remains significant after accounting for mother's further education indicator, father's further education indicator, and those two jointly. While each parent's further education increases the chances of their daughter having gone to university, mother's further education is a stronger predictor than that of the father. This finding will be corroborated by estimates of the partially identified model, in which using mother's further education status leads to a greater reduction in the size of the identified set. Results in column (5) suggest that, after controlling for one's age group with the same variables as those used in the partially identified model to follow, smoking is still a statistically significant predictor of university education, which provides additional evidence for using the instrument in set-estimation of Model C, which is detailed below.

Results in Table 2.4 show that, even in the richest specification, smoking when young remains a strong predictor of university education. Specification in column (1) controls non-linearly for the respondent's age, includes regional dummies, as well as including other predetermined variables which can plausibly affect the person's education decisions — such as whether there was a large number of books in her childhood home, whether she lived with her parents at the age at which she was legally allowed to leave school, and whether she is the eldest or only child. Yet, even after controlling for all those factors, both smoking when young and parent's further education indicators, are statistically significant predictors of attending higher education. Crucially, in this specification, mother's further education has a larger effect on daughter's education, than father's education does. In the richest specification (2) the coefficients on all three instruments fall in absolute value by 23-31% compared to specification (1) and coefficients on parents' education lose statistical significance, though the coefficient on smoking when young remains large and statistically significant at the 5% level. The reduced significance of these variables can be attributed to the inclusion of partner's education. The effect is consistent with assortative matching in the marriage market, as evidenced by Mare (1991), Pencavel (1998) and Qian (1998). Additionally, the age of a woman's youngest child seems to be statistically significantly correlated with attending education, though the effect is small in absolute value.

2.4.2 Estimated Models

The computational complexity of the model and its data requirements greatly reduce the set of models which can be feasibly estimated. Since the main aim of this chapter is to provide a 'User Guide' to the method, this chapter uses simplified versions of the model yet preserving its key features.

2.4.2.1 Models Estimated under the Assumption of Normality

Three types of models are estimated in this chapter. In all of them the outcome variable is a binary indicator of working at the time of the interview, and the measure of education is a binary indicator of having obtained a university degree. In all three specifications there are only two random parameters: α_i - a random intercept and β_i - a heterogeneous slope coefficient on education. The three models are estimated under the assumption that the distribution of the vector of random parameters, (α_i, β_i), is bivariate normal with an arbitrary covariance. The use of the normal distribution simplifies the infinitely-dimensional estimation problem to a finite problem of finding a set of values of 4 unknown parameters which are compatible with the model and the data: $\theta = (\bar{\alpha}, \bar{\beta}, \sigma_{\alpha\beta}, \sigma^2_{\beta})$, which are the mean of α_i, the mean of β_i, the covariance between α_i and β_i and the variance of β_i. Note that the variance of α_i is normalised to 1, as discussed in Section 2.2.1.

The first estimated specification, Model A, is the simplest version of the model: it does not include any exogenous variables in the model and relies only on the simplest of instruments: a binary indicator. As a result, the outcome equation becomes:

$$Y_i = 1.\left[Y_i^* \geq 0\right] = 1.\left[\alpha_i + \beta_i E_i \geq 0\right]. \tag{2.15}$$

In Model A only one, binary, instrument is used: an indicator for having smoked regularly at or before the age of 16. Hence, $Z_i \in \{0, 1\}$. The instrument was selected among the three proposed above on account of being the strongest.

The second specification, Model B, uses the same outcome equation (2.15) as the first, however, the cardinality of the instrument vector is increased: $Z_i \in \mathcal{Z}$ where $\#\mathcal{Z} > 2$. The vector of instruments is formed by defining Z_i as a categorical variable, based on categories presented in Tables 2.1 and 2.2. Specifically, in two versions of Model B the instrument takes 4

values ($\#\mathcal{Z} = 4$): in the first version the indicator of early smoking is interacted with the father's further education indicator, and in the second with the mother's further education indicator, in line with the columns of Table 2.1. In the third version of Model B, all categories from Table 2.2 are used to form an instrument with 8 different values ($\#\mathcal{Z} = 8$), by taking into account the smoking instrument, father's further education and mother's further education simultaneously. The aim of using these specifications is to show how the volume and shape of the estimated set of distributions of random parameters change as the cardinality of the support of the instrument vector increases and how it depends on the strength of the instruments.

The choice of one categorical variable with multiple values, rather than using the relevant binary variables in their original form, is for convenience of coding only and does not affect the results, as can be seen in equation (2.7). This is in contrast to methods which model the relationship between the endogenous variable and the instrument explicitly, in which case results would differ depending on whether a set of dummy variables or a single categorical variable is used.

The third, specification, Model C, adds a categorical exogenous variable to the outcome equation, with a coefficient on the variable constrained to be constant, leading to the following outcome equation:

$$Y_i = 1.\left[Y_i^* \geq 0\right] = 1.\left[\alpha_i + \beta_i E_i + \gamma W_i \geq 0\right]. \tag{2.16}$$

Adding exogenous variables to the model poses two types of challenges. As shown in equation (2.7), the identifying moment inequalities must hold for every value of the vector of exogenous variable and instruments. The first challenge is in the form of greatly increasing the number of inequalities that need to be checked and, in particular, increasing the number of evaluations of the left hand side. Recall that in the absence of exogenous variables the left hand side of the inequality remains unchanged since it corresponds to the partitioning of the parameter space by the outcome equation. Since instruments do not enter the outcome equation, adding more instruments does not affect the number of evaluations of the left hand side that need to be performed. On the contrary, adding exogenous variables changes the partitioning of the parameter space since the partition lines are given by the following equalities:

$$\alpha_i + \beta_i E_i + \gamma W_i = 0; \qquad E_i \in \{0, 1\}. \tag{2.17}$$

Hence, for every value of the constant parameter γ and for every value w of the exogenous variable W_i, the partitioning line presented in Figure 2.1 above, is shifted down by $-\gamma w$. Thus, for every combination of the parameter γ and every value w of the exogenous variable, there is a different collection of core-determining sets which need to be tested in inequality (2.7).

2.4.2.2 Model Estimated under the Assumption of a Triangular Distribution

In addition to the above, Model A is also estimated under the assumption that the random parameters are independent and that each has a symmetric triangular distribution. Contrary to the normal distribution, whose support is unbounded and on both sides of zero, the support of the triangular distribution is bounded and can be limited to being fully contained in a given subset of the parameter space. For example, if the spread of the distribution is smaller in absolute value than the mean of the distribution, all values arising with non-zero probability are positive.

The additional set of estimates using the triangular distribution with a finite support is obtained in order to shed some light on the sensitivity of the estimates based on the normal distribution to the infinite support of the normal, and to the additional complexity arising from allowing for arbitrary correlation of the heterogenous parameters.

The symmetric triangular distribution has 2 parameters: the mean and the spread. With the additional assumption of independence of α_i and β_i, and with the necessary scale normalisation discussed in Section 2.2.1, there remain only 3 parameters to be estimated: $\theta_{triang} = (\bar{m_\alpha}, \bar{m_\beta}, s_\beta)$: the first two are the means of α_i and β_i respectively, while the last one is the spread of β_i.

2.4.2.3 Model and Variable Choice Considerations

Section 2.6 discusses computational complexity in the general case. To illustrate the complexity considerations arising in our specific models, suppose that we wish to estimate Model C under the assumption of normality with $\theta = (\bar{\alpha}, \bar{\beta}, \sigma_{\alpha\beta}, \sigma^2_\beta)$ being the vector of unknown parameters governing the distribution of random parameters.

For the sake of this example, suppose that if we were to estimate the model with no included exogenous variables, we would do so by testing a coarse grid of parameters with T values overall. For example, with 100 tested values for each mean, and 10 values for the covariance and for the variance each, the grid has 1 million points. If, for each value of the vector instruments

the number of core-determining sets to be tested is 8 (as is the case with a binary outcome variable and a binary instrument) then the researcher must evaluate $1,000,000 \times 8$ values of the left-hand-side.

If, as is the case in Model C, the researcher wishes to add exogenous variables with the cardinality of the support of W_i being K_W; and with g hypothesised values of the parameter γ, since each pair of w and γ leads to a different partitioning of the parameter space, and since the number of test sets increases nonlinearly in the number of partitions due to the need to take unions of the sets, the overall number of evaluations of inequality (2.7) increases by a factor greater than $K_W \times g$. In fact, as shown in Section 2.6, in the presence of exogenous variables the number of evaluations of inequality (2.7) is a fourth-order polynomial in the cardinality of the endogenous variable, rather than being a second-order polynomial, as is the case in the absence of exogenous variables.

For instance, if the exogenous variable is age in years which, with the sample being restricted to women aged 25-45, takes on 21 values; if the parameter γ is to be tested on a grid of 100 variables; and with all other parameters tested on the same grid as above, the number of evaluations of inequality (2.7), is $100 \times 10^6 \times 21 \times \frac{2^4+2^3+2^2}{2} = 294 \times 10^8$, which is a number 3,675 times larger than the original value of 8 million evaluations. Since each of these evaluations is an integral of a multivariate normal distribution over an appropriate subset of the parameter space, the computational time requirements are prohibitive.

The second challenge due to adding exogenous variables arises on the right-hand side of Artstein's inequality (2.7). The inequality requires calculating the empirical distribution of observables. However, due to a finite sample size, increasing the number of possible cells reduces the reliability of the estimates. For example, in the current dataset with a binary outcome variable, a binary education indicator and a binary instrument, there are $2 \times 2 \times 2 = 8$ possible combinations of data. Adding the age variable, with 21 distinct values, increases the number of cells to $8 \times 21 = 168$ and, with additional 2 binary instruments, to 672. With a sample size of 672, the empirical distribution of the data is unlikely to be a good approximation to the actual population distribution and using it would lead to unreliable estimates.

In view of the above and given the limitations of computational power and dataset size, in the empirical illustration a variable including 4 age categories is used. These correspond to the following age brackets: 25-30 years old, 31-35, 36-40 and 41-45. The aim of using these is

to illustrate an implementation of the Chesher and Rosen (2014) approach to a model with exogenous variables, while respecting the shortcomings of the data. In this specification, the binary indicator of smoking regularly at or before the age of 16 is used as an instrument. As Table 2.3 shows, the instrument is strong in this setting.

2.5 Implementation Guide

2.5.1 Full Estimation and the Projection Method

There are two types of approaches to estimating the set of distributions of random parameters based on the method of Chesher and Rosen (2014). The first approach recovers the full set of distributions which are compatible with the model and the data i.e. which satisfy the key identifying inequality (2.7). The second approach focuses only on a subset of this set.

To illustrate the difference between the two approaches, consider the parametric case of the family $\mathcal{F}$ of distributions of random parameters being specified as normal, with the model as discussed in Section 2.4.2. In this case, with only one endogenous variable with a random slope coefficient as well as a random intercept, the aim is to estimate the full set of distributions of (α_i, β_i) compatible with the model and the data. Since the bivariate normal distribution is fully specified by 5 parameters and taking into account the scale normalisation consisting of setting the variance of α_i to 1, estimating the set of distributions is equivalent to estimating the set of parameter vectors $\theta = (\bar{\alpha}, \bar{\beta}, \sigma_{\alpha\beta}, \sigma^2_{\beta})$: the means of α_i and β_i, respectively, the covariance between α_i and β_i and the variance of β_i, which satisfy the key identifying inequality (2.7).

In the parametric case, the first approach, which I will refer to as 'full estimation', can be implemented, for example, by first testing a coarse grid of the full parameter vector θ and collecting the values of the parameter vector which satisfy the identifying inequality (2.7). To obtain a better approximation to the shape of the identified set, one can then estimate through simulation: by drawing a large number of random values of the parameters locally to the points which belong to the identified set based on the coarse grid, and then testing whether those satisfy inequality (2.7).

By contrast, the second approach, which I refer to as the 'projection method', focuses on identification over only a subset of the parameter space. In the parametric case above this corresponds to estimating a set of sub-vectors of θ. For example, a researcher may be interested

only in the set of means $(\bar{\alpha}, \bar{\beta})$ or only in the variance σ^2_{β}, with less concern about the other parameters.

The second approach is implemented as follows. Firstly, a coarse grid of only the parameters of interest is selected - for example, a coarse grid of values of the means $(\bar{\alpha}, \bar{\beta})$. Each value $(\bar{\alpha}_g, \bar{\beta}_g)$ on the grid is then checked for compatibility with the model and the data by checking whether there exists any combination of the other parameters - in this case some combination of a covariance and a variance, $(\tilde{\sigma}_{\alpha\beta}, \tilde{\sigma}^2_{\beta})$ - such that the full vector $(\bar{\alpha}_g, \bar{\beta}_g, \tilde{\sigma}_{\alpha\beta}, \tilde{\sigma}^2_{\beta})$ satisfies inequality (2.7). The search for a compatible combination of the remaining parameters is performed by using an optimisation method, as will be shown in the example code in the following section.

The choice between the two methods involves a trade-off and depends on the researcher's object of interest. The first approach leads to more informative estimates since it recovers the full identified set. For example, calculating the hypervolume of the estimated set and comparing it across specifications provides a measure of the strength of the identifying power of the model and allows for comparing the identifying strength of different specifications. Moreover, full estimates provide additional information on the dimensions over which the set is most spread out and can give an indication on what kind of additional data or modelling restrictions can lead to greatest gains in reducing the size of the identified set. For example, as will be illustrated in Section 2.7, the model may be insufficient to provide an upper bound on the variance of the random slope coefficient, σ^2_{β}, which is closely related to there being no upper bound on the mean of the slope coefficients, $\bar{\beta}$. Having uncovered this relationship empirically, the researcher can then think of adding plausible modelling restrictions on the variance or by changing the specification in a way which bounds the mean of the random slope coefficients - for example, by adding exogenous variables to the outcome equation.

However, estimating the full set has the disadvantage of being more demanding computationally: since the grid search is performed over all dimensions rather than just selected few, the number of values of the parameter vector to be tested can differ by several orders of magnitude and this difference is unlikely to be offset by the fact that the full estimation method is less computationally intensive per vector than the projection method. In addition to the computational time advantages, the projection method also allows for easier interpretation of results, especially if the focus is only on a 2- or 3-dimensional subset of the parameter space, in which case the estimates can be presented graphically. Moreover, the projection method can be used to

quickly gauge where the bounds in each direction are. In the example discussed in the previous paragraph, with no bound on the variance σ^2_β, the researcher can calculate the projection of the identified set onto the variance axis only, to see if the problem of potentially infinite variance arises under different specifications. By looking for only one combination of the remaining parameters which is compatible with the grid of variances, rather than testing a full grid of 4 parameters, the results can be obtained without as large a computational time investment.

2.5.2 Example Implementation Code

This section presents example pseudo-code which can be used in order to implement the method of Chesher and Rosen (2014) through the projection approach. The alternative approach — obtaining the full estimated set — is easier to implement and uses components of the code implementing the projection approach. It is briefly discussed at the end of this chapter. Executable Matlab code implementing the projection approach as well as a step-by-step guide to the code can be found in the Appendix.

The code is applicable to Models A and B specified in Section 2.4.2: the outcome equation contains only one endogenous variable, a binary measure of education, and no exogenous variables. The vector of instruments is a categorical variable of arbitrary cardinality. The distribution of parameters, F, belongs to a family of distributions which are fully specified by the vector of means and variance-covariance matrix, such as the normal distribution. The code is written assuming that the researcher is interested in estimating bounds on the means $(\bar{\alpha}, \bar{\beta})$ only, without focusing on the remaining 2 parameters. However, the code can be easily altered to add exogenous variables.

The general principle behind the code is as follows. An objective function is specified, which treats the means $(\bar{\alpha}, \bar{\beta})$ as parameters and uses $(\sigma_{\alpha\beta}, \sigma^2_\beta)$ as arguments. The function is defined in such a way that it is nonpositive if and only if the full vector of parameters $\theta = (\bar{\alpha}, \bar{\beta}, \sigma_{\alpha\beta}, \sigma^2_\beta)$ satisfies the identifying inequality (2.7). A grid of values of the means $(\bar{\alpha}, \bar{\beta})$ is tested for compatibility with the model by checking whether there is any combination of $(\sigma_{\alpha\beta}, \sigma^2_\beta)$ for which inequality (2.7) is satisfied, i.e. for which the objective function is nonpositive. To achieve this, the objective function is minimised over the joint support of $(\sigma_{\alpha\beta}, \sigma^2_\beta)$. If, for a given value of $(\bar{\alpha}_g, \bar{\beta}_g)$ the minimum of the objective function is nonpositive, the point is deemed as belonging to the estimated set. Otherwise, it is excluded from the estimated set.

Description of Pseudo-code in Box 2.1

In the pseudo-code in Box 2.1 as well as in the Matlab code in the Appendix, the estimation method is implemented by minimising the objective function `S_min`. For ease of exposition, the function is defined later on in the current section.

In the pseudo-code, presented in Box 2.1, Step 1 loads the relevant dataset into memory. In Step 2 the algorithm requires the user to specify a grid of means $(\bar{\alpha}, \bar{\beta})$ which will be tested for whether they satisfy inequality (2.7). A finer grid increases the precision of estimates, but at the cost of increasing computational time. Furthermore, additional information about the parameters can be used in order to exclude impossible values.

In Step 3 conditional probabilities corresponding to the right hand side of inequality (2.7) are calculated conditional on each value of the instrument (z)[8]. For each test set S there is a combination of values of (Y, E) corresponding to that set. For example, in the case of set A in Figure 2.1. the values are $(Y, E) = (1, 1)$. As another example, for the test set $S = A \cup B$ the values are $(Y, E) = (1, 1) \vee (Y, E) = (1, 0)$ and so the associated probability is the conditional probability of $Y = 1$ conditional on each value of Z. For a large enough sample size these probabilities can be calculated based on the empirical probabilities of the occurrence of each combination of observables in the data. Since inequality (2.7) must hold for every value of the instrument vector and since, in the absence of exogenous variables in the model, the sets S do not change across values of Z, it is sufficient to check that for each set S the inequality holds for the largest of the right-hand-side probabilities, hence taking the maximum over the values of Z in Step 3.

Box 2.1 Main Code

Inputs: Y,X,Z — data, `MU` — grid matrix of means of random coefficients $(\bar{\alpha}, \bar{\beta})$;

Outputs: Outcome of testing each mean $(\bar{\alpha}, \bar{\beta})$ against inequality (2.7) with possible refinements;

Uses: S_min;

Algorithm:

[8] In the presence of exogenous variables, the probabilities are conditional on the vector of exogenous variables and instruments (w, z) and the combinations of observables refer to combinations of (Y, E, W) rather than just (Y, E).

1. Load data: Y, E, Z.

2. Define a grid matrix `MU`.

   - Contains a grid of values of the means of random coefficients: $(\bar{\alpha}, \bar{\beta})$ with each pair stored in a separate row of `MU`.

3. Calculate probabilities on the right-hand-side of inequality (2.7).

   - For each core-determining set S calculate the conditional probability of the occurrence of combinations of observables corresponding to the set S.
   - The probabilities are calculated conditional on each value of the instrument vector Z and stored in a matrix `P` with each row corresponding to a particular core-determining set S, and each column corresponding to a value of the vector of instruments Z.
   - Define matrix `P_max` by taking the maximum value of `P` in each row.

4. For each row `p` of `MU` test whether there exists any variance-covariance matrix for which inequality (2.7) holds.

   - Define constraints of the optimisation problem:
     - σ_β^2 bounded away from zero.
     - The determinant of the variance-covariance matrix formed by a combination of $(\sigma_{\alpha\beta}, \sigma_\beta^2)$ and the normalised variance $\sigma_\alpha = 1$ must be positive.
     - $\sigma_{\alpha\beta} \in [-1, 1]$.
   - For each row `p` of `MU` minimise the function `S_min` with respect to ($\sigma_{\alpha\beta}$ and σ_β^2), taking `P_max` as an input into the objective function. See Box 2.2 for details on the function `S_min`.

5. Collect rows of `MU` for which the minimum of `S_min` is nonpositive - these belong to the estimated set.

6. (Optional) Refine the search.

- Based on points which currently belong to the estimated set define a new grid matrix `MU`.
  - E.g. points local to identified points.
  - E.g. drawing random points within the enlarged extremes of the set identified in Step 5.
- Repeat steps 2-6 for the new grid matrix `MU`.
- Stop when the researcher is satisfied with the precision of the resulting estimates of the identified set.

Step 4 tests, for each row `p` of the grid of means `MU` whether there exists a variance-covariance matrix such that the full vector of parameters specifying the hypothesised distribution $\tilde{F}$ of random coeffcients satisfies inequality (2.7). The test is implemented by minimising, for a given row `MU(p,:)` a function `S_min` (whose precise definition will be provided below) which takes nonpositive values if and only if inequality (2.7) is satisfied. For a given mean to belong to the projection of the identified set, it is sufficient that there is only one variance-covariance matrix compatible with (2.7). Equally, for the vector to be excluded from the identified set, it must be the case that there is no such variance-covariance matrix, in which case all values of the function `S_min` are positive. Thus, it is sufficient to check whether the minimum of the function `S_min` is nonpositive.

The optimisation step further involves specifying constraints on parameters of the variance-covariance matrix. Importantly, the determinant of the variance-covariance matrix must be positive: sufficient conditions for a matrix to be positive definite are that both its trace and determinant are positive. In the case of a variance-covariance matrix the trace is necessarily positive, leaving only the determinant condition to be checked. Singularity (or near-singularity) may lead to computational issues when the function is integrated within `S_min`.

In Step 5 points belonging to the estimated set are collected. If the original matrix `MU` contains a fine grid of points, these can give a good illustration of where the identified set lies. Otherwise, the search can be refined by creating a new matrix `MU` and repeating steps 2-5. The matrix can be formed, for example by drawing random points within the enlargement of the set identified

based on the coarse grid[9].

There exists an alternative method for attaining refinement based on the initial results: it is known as the bisection method and is suitable for convex or near-convex sets. It requires drawing rays at different angles, passing through a point M_{in} known to belong to the set, then identifying, for each ray, a point M_{out} along the ray which does not belong to the set and then determining the boundary of the set by finding the mid-point between M_{in} and M_{out} and testing whether it belongs to the identified set. Depending on the outcome of the test, the midpoint becomes the 'new' M_{in} or M_{out} point and the procedure is repeated until a sufficient approximation to the boundary is found. The description of the bisection method is provided here for completeness, but its implementation is beyond the scope of this chapter.

Description of Pseudo-code in Box 2.2

Box 2.2 presents an algorithm for defining the function `S_min`. The function takes as arguments the grid of means `MU` as well as the matrix of maximal right-hand-side probabilities, `P_max`, against which inequality (2.7) is evaluated. The function's variable arguments are: `g` - the covariance, and `s2` - the variance of β. These are the arguments over whose domains a minimum is found.

The function takes the vector of parameters consisting of `MU(p,:)` as well as the arguments `g` and `s2` and uses them to form the distribution of random coefficients, $\tilde{F}$. For each core-determining set S, the probability assigned to the set by $\tilde{F}$ is calculated by integrating $\tilde{F}$ over the area of the parameter space defined by S. For each set S this probability is subtracted from the maximal empirical probability corresponding to the set, as captured by `P_max`. The difference is non-positive if for the given set S inequality (2.7) is satisfied; and it is positive if the inequality is violated. The function `S_min` returns the maximum of those differences across all core-determining sets S since, if the largest difference is nonpositive then other differences are also. If the value returned by the function is nonpositive, the distribution $\tilde{F}$ determined by `MU(p,:)`, `g` and `s2` satisfies inequality (2.7), otherwise the inequality is violated. The optimisation stage defined in Box 2.1 minimises function `S_min` over the joint support of $(\sigma_{\alpha\beta}, \sigma^2_{\beta})$.

[9]The enlargement of the set, by one grid step, is taken in order to account for points which belong to the identified set but which fall between the last point on the grid which belongs to the identified set, and the first point which does not.

Box 2.2 S_min

Input arguments:

- Constant arguments: `MU(:)`, `p`, `P_max`.
- Variable arguments with respect to which the minimisation is performed: `g` - covariance, `s2` - variance of β.

Algorithm:

1. Define the distribution $\tilde{F}$ of random coefficients based on the vector of parameters:
   - `MU(p,:)` as a mean.
   - Variance-covariance matrix formed from `g`, `s2` and $\sigma_\alpha^2 = 1$.
2. Check inequality (2.7) for each core-determining set S.
   - For each test set S calculate the probability that the distribution $\tilde{F}$ places on that set (left hand side of inequality (2.7)).
     - Integrate the distribution $\tilde{F}$ over the area of the parameters space defined by S.
   - Subtract the above probability from the maximal empirical probability associated with the same set S, as summarised in the matrix `P_max`.
   - Take the maximum f of the difference between `P_max` and $\tilde{F}(S)$ for each set S.
3. The function returns f.
   - The function takes a nonpositive value if and only if the distribution $\tilde{F}$ defined by parameters `MU(p,:)`, `P_max`, `g`, `s2` satisfies inequality (2.7) for each core-determining set.

Implementation of the alternative approach — estimating the full identified set of parameters rather than its projection onto only one of the axes — is very similar to the code above. In particular, it requires calculating, for each test vector of parameters on a coarse grid, an

analogous set of integrals of a multivariate normal distribution as those used in the definition of the objective function `S_min`. Those values are then compared with the empirical probabilities, in line with inequality (2.7). The estimated set contains of those vectors of parameters which satisfy the inequality.

2.6 Computational Considerations

2.6.1 Computational Time

Implementing the method of Chesher and Rosen (2014) requires a substantial investment of computational time. For example, in obtaining the results in Section 2.7.2.1, evaluating whether 100,000 points satisfy inequality (2.7) took over 17 hours of computations even when implemented through 24 parallel processes on a dedicated computational server[10]. The 100,000 points whose calculations were timed were a small subset of the points tested. This computational burden is caused by the need to evaluate, for each value of the vector of parameters on a tested grid, a number of integrals of the hypothesised distribution such as multivariate normal, where the number of integrals corresponds to the number of core-determining sets.

The aim of this section is to show how the computational complexity of the model varies with observable variables. Importantly, the key factors in determining computational complexity are not only the number of variables but also the cardinality of their support. For ease of exposition, all variables are assumed to be discrete. Section 2.6.3 discusses estimating the model in the presence of continuous variables. Additionally, in the following calculations it is assumed that there is only one endogenous variable and any exogenous variables enter with constant coefficients, so that the space of random parameters is 2-dimensional.

Instruments, endogenous variables and exogenous variables affect the complexity of the model differently and thus their effect is analysed in turn. The following notational conventions are used: K_Z is the cardinality of the support of the vector of instruments (which, in the case of multiple instruments, is the number of all combinations of instruments arising with non-zero probability); K_E is the cardinality of the support of the endogenous variable; K_W is the cardinality of support of the entire vector of exogenous variables and K_{W1}, ..., K_{WM} are the numbers of values that

[10]The machine had 24 cores (2 processors Intel Xeon(R) — Family 6 Model 44) and was conducted through parallel processing in Matlab with 24 parallel processes, using approximately 90% of the CPU power.

each of the M exogenous variables can take with positive probability. Each of the exogenous variables, W_j, enters with a constant coefficient γ_j. It is assumed that the distribution of random parameters is specified parametrically in which case estimation requires testing each *finite* vector of parameters on a grid for whether it satisfies inequality (2.7). For example, in the case of the normal distribution the parameter vector is 4-dimensional, containing the two means of random parameters, their covariance and one free variance. In the case of other distributions the vector may have an arbitrary number of t dimensions. It is assumed that the vector of parameters specifying the distribution of random coefficients is tested on a grid of T points. Moreover, each of the constant coefficients γ_j on exogenous variables is tested on a grid of g_j points.

Importantly, the highest computational burden in implementing the method of Chesher and Rosen (2014) stems from the need to evaluate the left hand side of inequality (2.7) since this requires, for each vector of tested parameters, integrating a multivariate distribution over each of the core-determining sets S. By contrast, in the case of discrete variables evaluating the right hand side takes a negligible amount of time and is implemented by calculating the frequency of each combination of data occurring in the sample.

2.6.1.1 Computational Complexity due to Instrumental Variables

Firstly, let us consider the effect that the dimensionality of the vector of instruments has on computational complexity. As detailed in Section 2.2, the collection of core-determining sets is defined as the collection of all elements and unions of elements of the collection $\mathsf{T}(w, z)$ which consists of subsets of the parameter space corresponding to each combination of values of variables in the outcome equation. Since instruments do not enter the outcome equation, the definition of core-determining sets does not depend on instruments. Thus, while increasing the cardinality of the support of the vector of instruments increases the number of times that inequality (2.7) has to be evaluated, it does not significantly affect computational time since the left hand side of the inequality remains unchanged for each value of the instrument while the computational time involved in evaluating the right hand side is negligible. In particular, the number of integrations that need to be performed is determined by K_E and K_W, in a way to be shown in the following paragraphs, but it does not depend on K_Z.

2.6.1.2 Computational Complexity due to the Endogenous Variable

To focus on the effects of the endogenous variable, for the purposes of this section suppose that no exogenous variables are present in the outcome equation and only one endogenous variable is present. The contribution of exogenous variables to computational time is presented in the following subsection.

The analysis is more complex in the case of the endogenous variable. As is clear from equation (2.18), increasing the cardinality of the support of the endogenous variable, K_E, increases the number of partitions of the parameter space which, together with unions of such partitions, form the collection of core-determining sets. Following Chesher and Rosen (2014), in the special case of one discrete endogenous variable and no exogenous variables it is possible to derive a simple expression for the upper bound on the number of core-determining sets.

In the more general case of one endogenous variable and an arbitrary vector of exogenous variables, core-determining sets are either halfspaces or unions of halfspaces, all determined by the following line based on the outcome equation (for each $e \in \mathcal{E}$ and $w \in \mathcal{W}$):

$$\alpha_i + \beta_i e + \gamma w = 0. \tag{2.18}$$

However, in the special case of no exogenous variables, all such lines pass through the origin. In $\mathbb{R}^2$ an arbitrary union of such halfspaces can be expressed as a union of at most two half-spaces. Since for each of the K_E points of support of E_i there are two possible values of the outcome variable Y_i, there are $2K_E$ original halfspaces. Thus, excluding unions of a set with itself there are $\frac{2K_E(2K_E-1)}{2}$ unions of any 2 sets, with division by 2 due to each set being counted twice, and the overall number of core-determining sets is $2K_E$ (number of single sets) + $\frac{2K_E(2K_E-1)}{2}$ (number of unions of pairs), for a total of $(K_E + 2K_E^2)$ core-determining sets. Thus, with one endogenous variable and no exogenous variables the number of core-determining sets is $O(K_E^2)$, irrespective of the number of instruments. If the vector of parameters governing the distribution of random coefficients is tested on a grid of T points, overall, there are $T \times (K_E + 2K_E^2)$ integrations of a multivariate distribution involved in checking whether inequality (2.7) holds[11]

As a numerical example, suppose that E_i has $K_E = 5$ values (e.g. education groups) and the distribution of parameters is normal with 4 free parameters, each of which is evaluated on

[11] Formally, $(K_E + 2K_E^2)$ is the upper bound as some unions may be complementary (spanning the entire space).

a grid of 100 values, so that $T = 10^8$. Then the overall number of integrals of the multivariate normal distribution is $10^8 \times (5 + 2 \times 5^2) = 5.5 \times 10^9$.

2.6.1.3 Computational Complexity in the Presence of Exogenous Variables

In the presence of exogenous variables, although the line in (2.18) still divides the parameter space into halfplanes, its intercept is shifted and so halfplanes underlying the collection of core-determining sets need not pass through the origin and their unions take more complex shapes. As shown by Chesher and Rosen (2014) in the discrete case with one endogenous variable and in the presence of exogenous variables one only has to consider core-determining sets of the form, for each (w, z):

$$S = \mathcal{T}(w, e_1, 1) \cup \mathcal{T}(w, e_2, 1) \cup \mathcal{T}(w, e_1', 0) \cup \mathcal{T}(w, e_2', 0), \quad where \quad e_2 \geq e_1 \wedge e_2' \geq e_1'. \quad (2.19)$$

Holding w fixed, with K_E points of support of E_i there are $\frac{K_E(K_E+1)}{2}$ combinations of e_1, e_2 such that $e_2 \geq e_1$, since for every value of e_2 the compatible e_1 is either e_2 itself or one of the $e_2 - 1$ numbers smaller than e_2. Equally, there are $\frac{K_E(K_E+1)}{2}$ combinations of e_1', e_2' such that $e_2' \geq e_1'$. Note that the pairs (e_1, e_2) and (e_1', e_2') do not enter symmetrically — they are associated with different values of the outcome variable. As a result, the total number of combinations of pairs (e_1, e_2) and (e_1', e_2') satisfying the requirements of (2.19) is $\frac{K_E(K_E+1)}{2} \times \frac{K_E(K_E+1)}{2} = \frac{K_E^4+2K_E^3+K_E^2}{4}$. Thus, for a given value of (w, z), the number of core-determining sets is of the order $O(K_E^4)$.

Since inequality (2.7) must hold conditional on each value of the vector of exogenous variables (w, z), the overall number of core-determining sets is $K_W \times \frac{K_E^4+2K_E^3+K_E^2}{4}$. Each of these sets must then be integrated over for each of the T grid points.

Importantly, contrary to the case of instruments, for which it is irrelevant whether one categorical variable is used or a set of binary variables, in the case of exogenous explanatory variables computational complexity depends on the number of variables even if the overall cardinality of the support of W, K_W, is held constant. To illustrate, suppose that there is only one exogenous variable with K_W points of support and that the researcher wishes to test the constant coefficient on the variable, γ, over a grid of g values. With T grid points for the vector of parameters governing the distribution of random coefficients, the number of evaluations of the left hand side of inequality (2.7) is $g \times T \times K_W \times \frac{K_E^4+2K_E^3+K_E^2}{4}$. If, instead, there are J exogenous variables with

the same total cardinality K_W and each coefficient γ_j is tested on a grid of g_j values, the number of evaluations of the left-hand side is $g_1 \times ... \times g_J \times T \times K_W \times \frac{K_E^4+2K_E^3+K_E^2}{4}$. In order to keep the same precision of estimates on the constant parameters, the researcher would have to set $g_1 = ... = g_J = g$ in which case the number of evaluations is $g^J \times T \times K_W \times \frac{K_E^4+2K_E^3+K_E^2}{4}$. Thus, even when the cardinality of the support of the vector of exogenous variables is kept constant, increasing the number of variables increases the number of evaluations by orders of magnitude.

As a numerical illustration, extend the above example with $K_E = 5, T = 10^8$, and assume that there are 2 exogenous variables: the former is binary, the latter has 20 values. These variables could correspond to, for example, an indicator of a child of pre-school age and age in years, respectively. Assume further that the two constant parameters on those are tested on a grid of 100 values each. This gives $K_W = 2 \times 20 = 40$, $g_1 \times g_2 = 10^4$ and the total number of integrals to evaluate is: $10^4 \times 10^8 \times 40 \times \frac{5^4+2\times5^3+5^2}{4} = 10^{12} \times 40 \times 225 = 9 \times 10^{15}$, even in the case of a very coarse grid of 100 values per each parameter.

Table 2.5 summarises information on the number of integrations of the underlying multivariate distribution which need to be performed in order to obtain the estimated set.

Table 2.5: Number of Calculations in the Case of One Endogenous Variable

Case	Number of Calculations
No Exogenous Variables	$T \times (K_E + 2K_E^2)$
1 Exogenous Variable	$g \times T \times K_W \times \frac{K_E^4+2K_E^3+K_E^2}{4}$
J Variables	$g^J \times T \times K_W \times \frac{K_E^4+2K_E^3+K_E^2}{4}$

2.6.2 Effects of Adding Further Endogenous Variables

Set identification in the model does not require that the number of instruments be greater than the number of endogenous variables and does not impose conditions on the correlation between the instruments, contrary to rank and order conditions in point-identifying approaches such as 2SLS. It is, however, worth considering the impact, in theory, of adding an endogenous variable to the model without increasing the number of instruments.

To consider this problem formally, suppose that originally there are m random coefficients in equation (2.2) and that the new endogenous variable enters with a random coefficient. This addition increases the complexity of the estimation problem by extending by another dimension the domain of estimation, i.e. the space of functions of random parameters. Even in the sim-

ple parametric case of a normal distribution, this addition will require set-estimating $(2 + m)$ additional parameters: a new mean, a new variance, and a vector of m covariances. In estimation based on grid search, the addition may pose substantial challenges to the feasibility of the method given limited computational time. For example, in the case of a normal distribution with each of the new parameters tested on a very coarse[12] grid of 100 values, the number T of points on the grid which are to be tested will increase by a factor of 100^{2+m}. The overall number of calculations will further be affected by the increased cardinality of E in a non-linear way since the higher dimensionality of the vector of endogenous variables implies that unions of halfspaces into which the parameter space is divided by the outcome equation will take more complex shapes.

To consider the impact of the addition of an endogeous variable on the strength of identification, one has to refer to the key identifying inequality, (2.7). On the left hand side, increasing the number of endogenous variables increases the dimension of the parameter space by 1 random parameter, as well as increasing the number of partitions of the parameter space, which increases the number of identifying inequalities, since those must hold for every core-determining set S. The overall effect due to left hand side on identification in the model depends on the cardinality of the support of the endogenous variables and the assumed type of the distribution of random parameters (parametric/nonparametric).

On the right hand side, without increasing the cardinality of the support of the instrument vector Z (for example, by adding an instrument), the number of inequalities arising due to Z does not change. Thus, when endogenous variables are added to the model without increasing the cardinality of the support of the instrument vector, complexity of the model increases but it is not matched by increased identifying power due to more instruments. The only offsetting force in identification strength arises from an increase in the number of core-determining sets.

2.6.3 Continuous Variables

The method of Chesher and Rosen (2014) admits continuous variables. Let us consider each type of variables: the instrument, the endogeous variable and the exogenous included variables, in turn and analyse the impact of their inclusion on inequality (2.7). For simplicity, the presence

[12]Note that the support of the normal distribution is infinite so a grid of 100 values is a very conservative example.

of only one continuous variable in each category is assumed. The analysis extends naturally to the case of more continuous variables.

Continuous Instrument

If the only continuous variable in the model is the instrument, the left hand side of inequality (2.7) still contains only a countably finite number of core-determining sets. In the case of a single endogenous variable and an arbitrary vector of exogenous variables, these core-determining sets are defined in equation (2.19) above. On the right hand side, the conditional probabilities are not affected by the continuity of Z since it is the variable which is conditioned on and the remaining variables (W, E, Y) are discrete.

However, since inequality (2.7) must hold for every value of the exogenous variables and instruments arising with positive probability, the continuity of Z gives rise to a continuum of moment inequalities which need to be satisfied for a given conjectured distribution to belong to the identified set.

Continuous Endogenous Variable

If the endogenous variable is continuous then, as pointed out in footnote 8 of Chesher and Rosen (2014), the collection of all test sets includes uncountable unions. Consider the simplest case of one endogenous variable and no exogenous variables, with partitions of the parameter space defined by the following line:

$$\alpha_i + \beta_i E_i = 0 \iff \alpha_i = -\beta_i E_i. \tag{2.20}$$

In this case, suppose that the initial partitioning of the parameter space into 2 halfspaces is done with $E_i = e \in Supp(E)$ where $Supp(E)$ is the support of E_i. If the value of E_i is altered, even slightly, the slope of the line in (2.20) changes and the new line partitions the parameter space into 2 new halfspaces, neither of which is fully contained in the original two. Thus, the continuous support gives rise to a continuity of non-mutually-inclusive partitions of the parameter space. Since the collection of core-determining sets is the collection of sets which are unions of those halfspaces, there are uncountably many core-determining test sets to be checked in inequality (2.7) for each (discrete) value of the exogenous variables (w, z).

The inclusion of a continuous endogenous variable additionally requires a different estimation approach for obtaining the joint distribution of observables, such as kernel-based methods for mixtures of discrete and continuous data proposed by Li and Racine (2003).

Because the joint distribution of observables is continuous in the endogenous variable, the probability that $E_i = e$ for any $e \in Supp(E)$, which forms the probability on the right hand side of inequality (2.7), is necessarily zero. This does not, however, imply that the model lacks identifying power: continuous unions of halfspaces correspond to intervals of values of E_i, which can arise with strictly positive probability.

Continuous Exogenous Variables

The case of continuous included exogenous variables gives rise to both the issues present in the case of continuous instruments and continuous endogenous variables. Since changing the value of W shifts the line partitioning the parameter space, on the left hand side of inequality (2.7) there are infinitely many partitions of the parameter space, leading to infinitely many non-mutually-inclusive core-determining sets. As an illustration, in the case of one discrete endogenous variable and arbitrary exogenous variables the collection of sufficient test sets is defined in equation (2.19) above. The uncountability of the support of W leads to an uncountable collection of core-determining sets $\mathcal{S}$.

Additionally, inequality (2.7) has to hold for each value of the continuous variable W, analogously to the previous case of a continuous instrument, which leads to a continuum of inequalities. Furthermore, as in the case of endogenous variables, alternative methods of estimation must be used in order to recover the joint distribution of observables.

Set Estimation in the Presence of Continuous Variables

Chesher and Rosen (2014) do not offer a clear guide on implementation of their methods in the case of continuous variables. The above analysis suggests that the correct way to proceed will depend on the specifics of the application at hand and, in particular, on which variables are continuous.

In all cases the problem can be simplified by discretising the continuous variable, with the usual trade-off: a finer grid of the continuous variable would provide a more accurate approximation to the true identified set. However, limitations due to a finite sample sizes, especially in

the presence of other variables with a high cardinality of support, as well as excessive demands on computational time, as discussed earlier in this section, impose restrictions on how finely the continuous variable can be discretised.

2.7 Illustration of Results

This section presents an application of the method of Chesher and Rosen (2014) to analyse female labour force participation using the three models detailed in Section 2.4.2. The main aim is to familiarise the prospective user with the method by illustrating the identified sets, showing the impact of strengthening the instruments or introducing exogenous variables, and to present problems that the user can encounter, such as the lack of an upper bound on a subset of the parameters, or small cell sizes.

To compare results obtained using the set-identifying method of Chesher and Rosen (2014) with point-identifying alternatives, this section begins with presenting estimates based on the point-identifying bivariate probit model, before applying set-identifying techniques.

2.7.1 Point-identifying Approach to Models A and B

This sub-section uses a point-identifying bivariate probit model to estimate coefficients in constant-coefficients versions of Models A and B of Section 2.4.2. These are presented for comparison with the partially-identifying estimates in the following section. As detailed in Section 2.3 the bivariate probit model accounts for endogeneity in the form of correlation between an explanatory variable and the error term, but not between the explanatory variable and the slope coefficient since the latter is constant. Additionally, it does so at the cost of specifying explicitly the relationship between the endogenous variable and the instrument(s). Formally, the estimated specification is:

$$\begin{aligned} Y_i &= 1.\left[\bar{\alpha} + \beta E_i + \varepsilon_i \geq 0\right], \\ E_i &= 1.\left[\pi_0 + \pi_1 Z_i + \nu_i \geq 0\right], \end{aligned} \tag{2.21}$$

where (ε_i, ν_i) are mean-zero error terms which are jointly normal, with an arbitrary correlation coefficient. The results of estimation through bivariate probit are presented in Table 2.6.

There are a few features of the estimated results which are worth pointing out. Firstly, in the

Table 2.6: Point-identifying Model - Results

	(1)	(2)	(3)	(4)
	Working (0/1)			
University Education (0/1)	0.359	0.453	0.659	0.570
	(0.783)	(0.546)	(0.494)	(0.460)
Intercept	0.497***	0.484***	0.453***	0.467***
	(0.121)	(0.0936)	(0.0914)	(0.0854)
	University Education (0/1)			
Smoking when young	-0.649***	-0.654***	-0.598***	-0.624***
	(0.151)	(0.152)	(0.153)	(0.153)
Mum's FE		0.598***		0.468**
		(0.149)		(0.156)
Dad's FE			0.490***	0.380**
			(0.128)	(0.135)
Intercept	-0.988***	-1.123***	-1.223***	-1.273***
	(0.0738)	(0.0831)	(0.0988)	(0.101)
Athrho				
Constant	-0.217	-0.284	-0.422	-0.367
	(0.458)	(0.338)	(0.340)	(0.304)
N	672	672	672	672

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

second equation, which relates the endogenous variable to the instrument(s), the estimates in all specifications are numerically almost identical to results in Table 2.3 which shows the strength of the instruments by relating the education variable to instruments through a single-equation probit model. The closeness of the two sets of results shows that the joint estimation of the full, two-equation, model does not significantly affect the results of the second equation.

As before, all three instruments are strong. Smoking when young negatively impacts the probability of attending university and is the strongest of the three instruments. Each parent's further education indicator has a strong and statistically significant partial effect on their daughter attending university. Mother's further education indicator is a stronger instrument — a finding which will be mirrored by partial-identification results in the following sections.

In the outcome equation, the intercept is very precisely estimated and stable across specifications. However, the slope coefficient varies substantially across the specifications and is noisily estimated. The variation in point estimates can be explained if the true model in fact has ran-

dom, rather than constant, coefficients. In the case of heterogeneous treatment effects estimators based on instrumental variables identify only the Local Average Treatment Effect (LATE), i.e. the average of the coefficient on the endogenous variable on the particular sub-population whose choice of the endogenous variable was affected by the instrument. Thus, if the three instruments identify different sub-populations: of those whose education decision is affected at the margin by time preference (young smokers) and those whose mothers and fathers attended further education respectively, the resulting estimates reflect means of the random slope coefficient estimated for those different sub-populations. This explanation is further supported by the size of the estimates themselves: the average coefficient is the highest when smoking and father's further education indicators are used and lowest in the case of using only early smoking as an instrument. The results for using mother's further education indicator and smoking, as well as using both parents' further education and smoking, are between the two extremes. The distribution of values of the instruments sheds some light on the results. Specifically, the majority (66.7%) of the subsample of women whose mothers attended further education also have a father who attended further education. By contrast, out of 417 women whose fathers did *not* attend further education, only 40 (9.6%) have a mother with further education qualifications. Thus, the two instruments generally distinguish between different subgroups: father's further education indicator differentiates, roughly, between those whose one or both parents have further education qualifications, *versus* those whose parents do not have such qualifications. On the contrary, mother's further education indicator differentiates largely between those whose both parents have further education qualifications *versus* those in whose case only the father has such a qualification. Thus, the results suggest that the 'returns' to education[13] for the sub-group of individuals who were induced into going to university through at least one of their parents having further education are higher than for the sub-group of individuals whose decision at the margin was affected by whether their mothers also had further education qualifications, in addition to their fathers already having those.

[13]Naturally, since the model under consideration is a binary-outcome model rather than a wage equation, the 'returns' to education are expressed as relative weight placed on education in the choice whether to work.

Table 2.7: Proportions of the Sample who are Working and in Education

	Working	University Education
Not smoking when young	44.3%	10.0%
Smoking when young	26.3%	1.9%
Mum without further education	57.6%	7.7%
Mum has further education	13.1%	4.2%
Dad without further education	42.9%	4.9%
Dad has further education	27.8%	7.0%

2.7.2 Set-identifying Method: Binary Instrument (Model A)

This section presents empirical results from estimating Model A using the approach of Chesher and Rosen (2014), whose outcome equation (2.15) is re-stated here for convenience:

$$Y_i = 1.\left[Y_i^* \geq 0\right] = 1.\left[\alpha_i + \beta_i E_i \geq 0\right].$$

As detailed in Section 2.4.2, no exogenous variables are included in the outcome equation and the only instrument is a binary indicator of early smoking. The baseline case is estimated with the assumption that the joint distribution of the random coefficients is bivariate normal, with an arbitrary covariance.

2.7.2.1 Normal Distribution

In the case of a normal distribution, results were obtained for three fixed levels of the variance of β_i, which is set to 10, 100 and 1000. This choice stems from the finding, discussed later on in this section, that the model does not seem to impose an upper bound on the variance of β_i and the size of the estimated set of estimates of the mean of β_i increases with the variance. Thus, if the variance was allowed to vary freely the identified set for $\bar{\beta}$ would span the whole support of β_i. This dimension reduction also allows for presentation and analysis of the results in 3 dimensions, which helps illustrate empirical regularities and, in particular, allows the reader to see changes in the estimated set stemming from changes of instruments used and assumptions.

Figures 2.2 - 2.3 present the estimated sets for $\sigma_\beta^2 = 10$ and Figure 2.4 shows how the set increases when $\sigma_\beta^2 = 100$. Table 2.8 shows 'extreme values', measured as the furthest lying points of the identified sets, of the means of the two coefficients, for each of the three levels of the variance. The estimates were obtained as follows. At first, the model was estimated based

Figure 2.2: Estimated Set; Var(β_i) set to 10

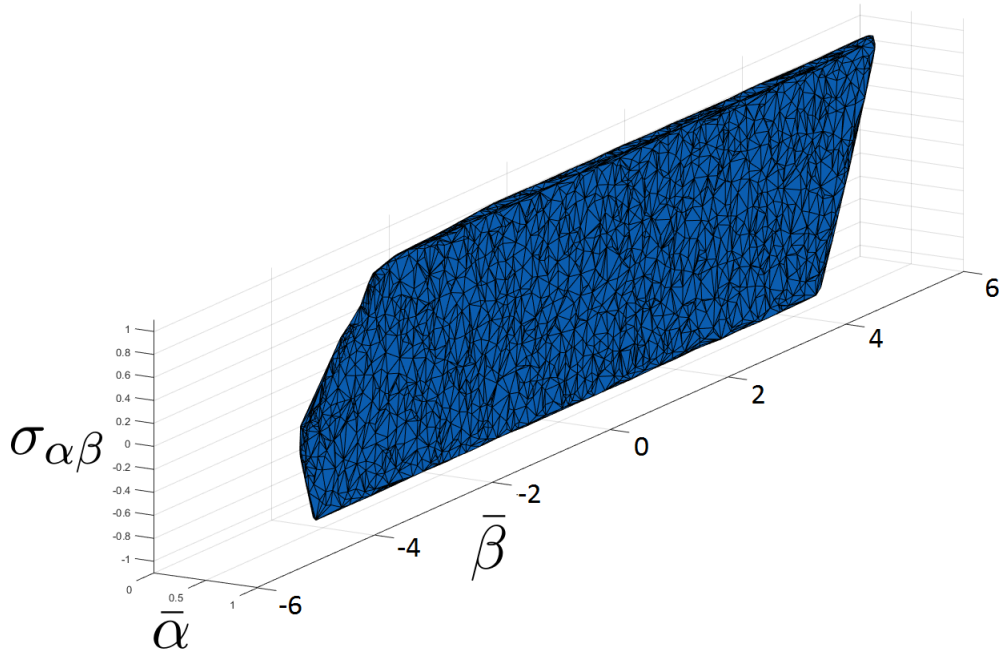
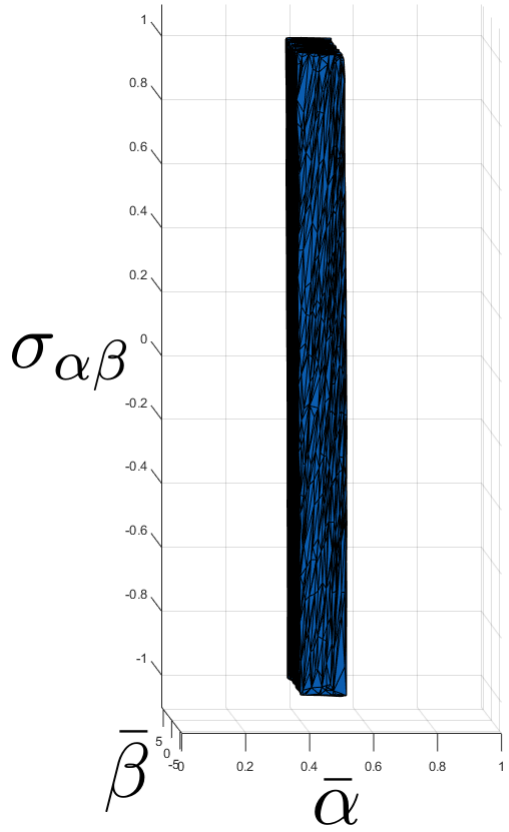


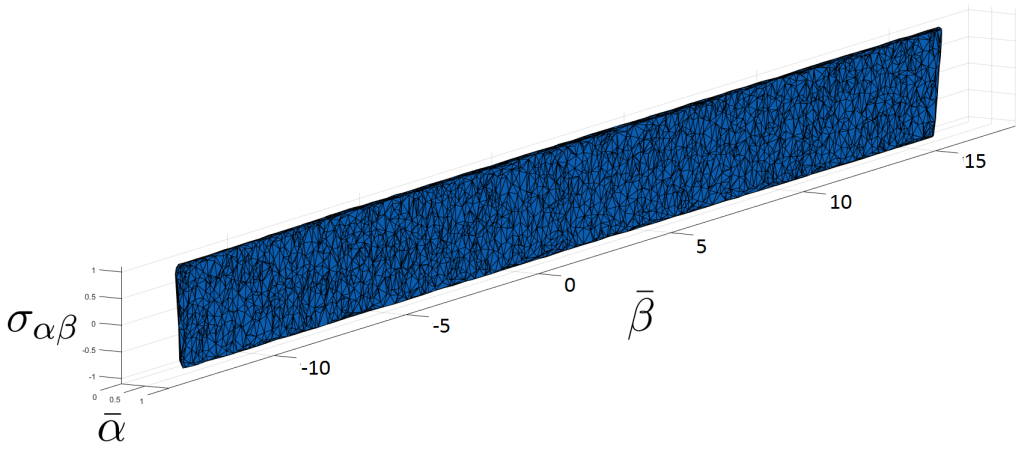
Figure 2.3: Projection of the Estimated Set, Showing the Mean of α_i; Var(β_i) set to 10



on a coarse grid of values of $(\bar{\alpha}, \bar{\beta}, \sigma_{\alpha\beta})$ for each of the three values of the variance. Since the initial grid-based estimates suggested that the set is continuous[14], the next step of the estimation was to perform, for each level of the variance, a uniformly random draw of 2×10^5 (or more[15]) points within bounds based on the coarse-grid estimates[16]. The points were then tested against inequality (2.7) and those which satisfied it were included in the presented sets.

For variances of 10 and 100, for which the estimated set appears to be convex, for points which satisfied the inequality an approximation to the set was used by obtaining an α-shape (α-hull) based on the refined set of points and is shown in Figures 2.2 - 2.4. The figures were obtained using the `alphaShape` command in Matlab, with the α parameter set to 5.

Figure 2.4: Estimated Set - Var(β_i) set to 100



The results show that, even under the weakest assumptions: positing only the outcome equation, the exogeneity of the instruments and the family of distributions to which the random coefficients belong; and even in the simplest model without exogenous included variables and with a binary instrument and a binary measure of education, the model still identifies a narrow interval for the mean of the random intercept. Moreover, the estimated interval for $\bar{\alpha}$ remains unchanged[17] as the variance σ^2_β changes, which is detailed in Table 2.8 below. This finding is in line with the high precision of estimates of the intercept in the bivariate probit model.

[14]Further calculations revealed that, in fact, in the case of $\sigma^2_\beta = 1000$ and a binary instrument (only), the set actually has a discontinuity.

[15]The number of points drawn in the case of $\sigma^2_\beta = 10$ was increasedd by 1/3 due to the initial refinement reaching the lower limit of the range based on which the refinement was calculated.

[16]In fact, enlargements of the bounds were used to account for the possibility of points falling between points on the coarse grid.

[17]Up to the 6th decimal place, with changes of this magnitude likely attributable to computer precision and in particular the number of simulated points and the precision of estimating integrals of the multivariate normal distribution.

Following equation (2.15) and noting that if $E_i = 0$, then $Y_i = \alpha_i$, one may wish to estimate the mean $\bar{\alpha}$ by estimating the proportion of those working among those without university education. However, this approach is incorrect if the conditional means $E[\alpha_i|E_i = 0]$ and $E[\alpha_i|E_i = 1]$ are not equal, which will be the case if E_i is correlated with α_i. Obtaining only interval estimates of $\bar{\alpha}$ rather than point estimates is the price paid for allowing for arbitrary correlation.

As illustrated in Figures 2.2 and 2.4, the model imposes no significant overall constraints on the covariance $\sigma_{\alpha\beta}$ between the random coefficients. However, certain values of the covariance are rejected for particular combinations of the means of random coefficients, which can be seen in the graphs of the identified sets. These excluded values occur close to the bounds of the estimated intervals for $\bar{\alpha}$ and $\bar{\beta}$. This finding is corroborated with further estimates based on stronger instruments in the next section.

Table 2.8: Extreme Points of the Estimated Sets - Binary Instrument

	$\bar{\alpha}$		$\bar{\beta}$	
σ^2_β	min	max	min	max
10	0.385	0.526	-4.44	5.49
100	0.385	0.526	-12.88	16.22
1000	0.385	0.526	-29.29	49.78

The model does not impose strong restrictions on the mean of the slope coefficient, $\bar{\beta}$. It also admits means on both sides of zero, although the half-point of the interval is positive. Importantly, as shown in Table 2.8, the interval[18] of estimated values of $\bar{\beta}$ increases as the variance of β_i increases. As will be discussed in Section 2.7.4, the model does not seem to bound the variance of β_i, thus not providing a bound on the mean, $\bar{\beta}$. The following sections investigate the problem in further detail.

2.7.2.2 Triangular Distribution

This section estimates the same model, Model A, under the assumption that the random coefficients are independently distributed, each with a symmetric triangular distribution. As in the previous section, only an indicator of early smoking is used as an instrument.

The motivation for this choice is to investigate whether the lack of bounds on the mean

[18]There is a discontinuity in the estimated set for $\bar{\beta}$ in the case of $\sigma^2_\beta = 1000$ and a binary instrument. The extreme point refers to the furthest point belonging to the set.

and variance of β_i can be overcome if the distribution of coefficients has a finite support. The triangular distribution has been selected as a symmetric distribution with a somewhat similar the shape of the normal, yet with bounded support. The assumption of independence is made for ease of computation, especially in view of excessive demands on computational time, since the aim of these additional estimates is only to investigate the source of the lack of bounds rather than estimating the full model across all parameters. The lack of constraints imposed on the correlation $\sigma_{\alpha\beta}$ in the above estimation under the assumption of normality is unlikely to be the source of the apparent lack of a bound on σ_β^2 and, therefore, $\bar{\beta}$. The reason is that the size of the estimated range of values of $\bar{\beta}$ appears to increase as σ_β^2 increases, for all values of the correlation $\sigma_{\alpha\beta}$. In particular, Figures 2.2 - 2.4 do not suggest that the range of values of $\bar{\beta}$ appears to be stable in the vicinity of $\sigma_{\alpha\beta}$ as σ_β^2 increases.

Analogously to the case of the normal distribution, the estimates are obtained by first testing a coarse grid of parameters $(\bar{\alpha}, \bar{\beta}, s_\beta)$, where s_β is the spread of the distribution of β_i, for whether they satisfy inequality (2.7). These initial estimates suggest that, for a given level of the variance, the estimated intervals for $\bar{\alpha}$ and $\bar{\beta}$ are continuous, bounded subsets of the real line. Results are presented in Table 2.9. The results present an enlargement of the set by 1 grid step in order to account for the points, belonging to the identified set, which lie between the last point on the grid which falls into the set and the first point which does not. Since the grid steps were 0.1 for $\bar{\alpha}$ and 0.05 for $\bar{\beta}$, and with an additional refinement done for the estimates of $\bar{\alpha}$ after the initial grid search, the results presented in Table 2.9 may over-estimate each bound of each interval by up to 0.05. However, the aim of these estimates is to present certain regularities, in particular regarding the behaviour of the estimates of $\bar{\beta}$ and s_β, rather than estimating the set precisely. Moreover, the estimates were refined for illustration for a particular level of the variance (see below).

The estimates are based on values of the spread rather than the variance itself, resulting in non-integral values. In the case of a symmetric triangular distribution the variables are related through the following equation:

$$\sigma_\beta^2 = \frac{1}{6} s_\beta^2. \tag{2.22}$$

Results in Table 2.9 show that for very small values of the variance of β_i, the estimated interval contains only positive values. Moreover, for the smallest value of spread on the grid,

$s_\beta = 0.1$, which corresponds to a variance of 0.002, the distribution of β_i itself has support fully contained in the positive axis since the spread is less than the (conservative) lowest bound on the mean. However, the estimated interval for $\bar{\beta}$ clearly increases with an increase in the variance of β_i - a result in line with estimates based on the normal model. This suggests that the empirical regularity of the estimated interval for $\bar{\beta}$ increasing, potentially without bound, with an increase in σ^2_β is not confined to the normal distribution and is, therefore, not driven by the assumption of infinite support implied in assuming normality.

As in the case of normality, the set estimates of the mean of the random intercept, $\bar{\alpha}$, are stable with an increase in the variance of β_i.

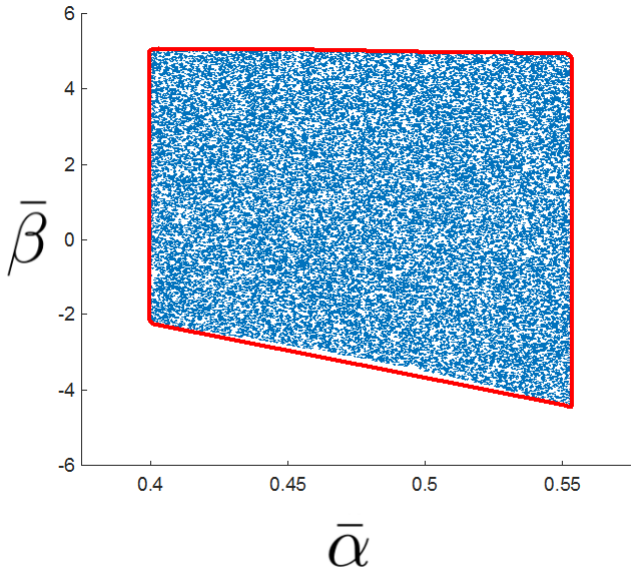
Table 2.9: Extreme Points of the (Enlargement of the) Estimated Intervals — Triangular Distribution

	$\bar{\alpha}$		$\bar{\beta}$	
σ^2_β	min	max	min	max
0.002	0.35	0.55	0.10	1.25
0.11	0.35	0.60	-0.20	1.35
0.14	0.35	0.60	-0.25	1.40
2.67	0.35	0.60	-2.05	2.80
10.67	0.35	0.60	-4.60	5.30

The estimates are further refined for a particular level of the spread s_β, which is chosen so that the variance of β_i is equal to 10 and is comparable with the results based on the normal distribution above. The refinement is performed by randomly uniformly drawing a set of 81,102 points from within the enlargement of the bounds based on grid search and testing whether inequality (2.7) holds for those. Figure 2.5 presents the estimated set as well as its convex hull. Note that there is a minor departure from convexity at the bottom of the figure, which corresponds to the lower bound of $\bar{\beta}$. Evidence of the presence of nonconvexities in the estimated set is important because it indicates that the implementation of the method can not, in general, be simplified by focusing only on finding the furthest-lying points of the set since points inside those may not belong to the set.

Results in Figure 2.9 show that the estimates of the means $\bar{\alpha}$ and $\bar{\beta}$, both in terms of location and the width of the interval are close to those obtained under normality for the same level of the variance ($\sigma^2_\beta = 10$) and an arbitrary correlation, as shown in Table 2.8, suggesting that the identifying power for the means stems from the model itself rather than the particular

Figure 2.5: Estimated Set - Var(β_i) set to 10



assumption of normality. Even though α_i and β_i are assumed to be independent under the current specification, there figure is far from being a rectangle: low values of $\bar{\alpha}$ are precluded if $\bar{\beta}$ is highly negative. On the contrary, if $\bar{\beta}$ is above approximately -2, all values in the range [0.400, 0.533] are included in the estimated set, except for when $\bar{\beta}$ is very close to its upper range, as indicated by the slight slope in the upper boundary of the set.

2.7.3 Set-identifying Method: Adding More Instruments (Model B)

This section shows the impact of increasing the support of the vector of instruments on the shape and volume of the identified set. Estimates are obtained for Model B from Section 2.4.2, which includes only one endogenous variable (education) in the outcome equation, and under the assumption of joint normality of the random coefficients with an arbitrary covariance. Two sets of instruments are used: a binary indicator of smoking when young together with the mother's further education status; and the smoking indicator together with the father's further education status. Results are obtained for the same levels of the variance of β_i as in the case of Model A: with values of 10, 100 and 1000. Numerically, estimates were based on the same draws of 2×10^5 points used in the case of a binary instrument in Section 2.7.2.1.

Recall that increasing the cardinality of support of the vector of instruments is expected to increase the identifying power of the model by increasing the number of inequalities that each

prospective vector of parameters must satisfy.

Figure 2.6, which is obtained for the combination of instruments of early smoking and mother's further education and setting $\sigma^2_\beta = 10$, is qualitatively representative of the changes arising due to an introduction of the second instrument. Firstly, more combinations of the parameters $(\bar{\alpha}, \bar{\beta}, \sigma_{\alpha\beta})$ are rejected by the model, which is captured by the more acute angles present in the shape of the estimated set. Secondly, the interval estimates for both means are narrower than in the case of the binary instrument. The impact on the width of the estimated intervals is particularly vivid when comparing the results on their extreme points, presented in Tables 2.10 and 2.11, with those in Table 2.8 for the case of a binary instrument. At each level of the variance the intervals are narrower. Moreover, in the case of mother's further education used as an in instrument, the reduction in the estimated bounds on $\bar{\beta}$ occurs disproportionately for negative values which moves the midpoints of the estimated intervals further into positive values.

Table 2.10: Extreme Points of the Estimated Sets - Smoking + Father's FE

	$\bar{\alpha}$		$\bar{\beta}$	
σ^2_β	min	max	min	max
10	0.398	0.490	-3.24	5.06
100	0.398	0.490	-9.46	14.98
1000	0.398	0.490	-20.09	37.80

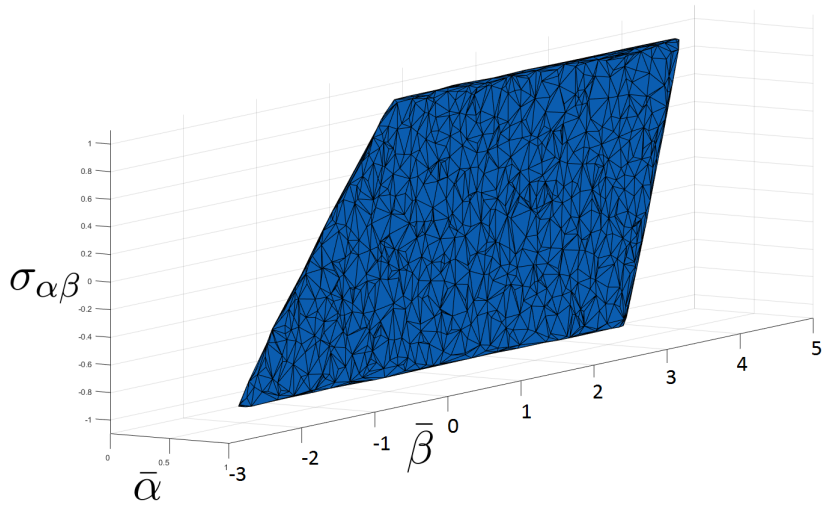
Table 2.11: Extreme Points of the Estimated Sets - Smoking + Mother's FE

	$\bar{\alpha}$		$\bar{\beta}$	
σ^2_β	min	max	min	max
10	0.418	0.473	-2.01	4.05
100	0.418	0.473	-5.54	12.24
1000	0.418	0.473	-16.01	29.29

In line with the results presented in Section 2.7.2.1, the mean of the random intercept is well-identified and the estimates of the extreme points are stable as the variance increases. There are no overall bounds on the covariance between the random coefficients, however, some values of the covariance are excluded for particular combinations of the random coefficients, which can be clearly seen in Figure 2.6.

Crucially, the reduction in the width of the estimated intervals is significantly greater when mother's further education is used as the additional instrument, rather than father's further

Figure 2.6: Estimated Set - Var(β_i) set to 10



education. This is clear both from the estimates of the extreme points of the intervals in Tables 2.10 and 2.11 and from Table 2.12, which, for each level of the variance, shows the volume of the 3-dimensional set of parameters estimated using different sets of instruments. Adding the father's further education indicator as an instrument approximately halves the volume of the estimated set. Adding the mother's further education indicator reduces its volume 4 times compared to estimates obtained using an indicator of early smoking alone. Following the findings of Section 2.4.1, this can be explained by noting that mother's further education indicator is a stronger predictor of attending university. This strength of prediction translates into higher identifying power.

Table 2.12: Strengthening the Instruments and the Volume of the Identified Set

σ^2_β	Smoking	Smoking + Dad	Smoking + Mum
10	2.24	1.18	0.50
100	7.50	4.03	1.78
1000	19.95	10.55	4.78

Notably, the current approach recovers the entire distribution of random parameters for the whole sample at hand, unlike point-identifying approaches imposing constancy of coefficients which only estimate the Local Average Treatment Effect. Thus, the narrowing of the estimated intervals and changes in the shapes of the 3-dimensional estimated sets are due to stronger identifying power of the instruments and not due to estimating the mean of random coefficients for a different sub-sample.

Finally, estimation using the full set of three instruments was performed. However, even in the case of drawing 2×10^5 random points, identical to those used for estimates above, there was not a single point that satisfied inequality (2.7). A possible explanation is due to small cell sizes: as further instruments are added, there are fewer and fewer observations with each particular combination of variables. For example, as shown in Table 2.1, several combinations of instruments occur for less than 5% of the sample, which corresponds to approximately 30 individuals. These numbers are further broken down into four categories based on combinations of the outcome and endogenous variables. If, due to sampling variation, a particular combination occurs with an unusually high or an unusually low probability, the estimated set could, in principle, be empty, in particular if the true identified set is small.

2.7.4 Models A and B: The Issue of an Increasing Variance

This section discusses the problem of the lack of a bound on the variance of β_i, matched by a lack of a bound on its mean.

To find out whether the variance is, indeed, increasing without bound, additional estimates of Model A in the case of normality and a binary smoking instrument were performed using the projection method: for each level of σ_β^2 on a coarse grid, the approach checked whether there is any combination of the remaining parameters: $(\bar{\alpha}, \bar{\beta}, \sigma_{\alpha\beta})$ for which inequality (2.7) holds.

The grid contained values from 50 to 5000, in increasing steps. Each of those values satisfied the requirements of the model as stated in inequality (2.7), suggesting that the variance is either actually unbounded, or it is unbounded for all practical purposes. Since the parameters of the model are scaled by setting $\sigma_\alpha^2 = 1$, the model admits a distribution for which the variance of the slope coefficient is 5000 times larger than that of the random intercept.

There are several possible explanations for the lack of a bound on the variance of β_i. Firstly, the assumption of the random coefficients having a jointly normal distribution, with its unbounded support, may be contributing to the unbounded variance. Secondly, the modelling assumptions about the random coefficients α_i and β_i may be too weak due to permitting an arbitrary degree of correlation between them. Thirdly, the data themselves could be contributing to the problem if the instrument Z_i does not substantially change the conditional probabilities on the right hand side of the identifying inequality (2.7); or if data quality is an issue more generally. Finally, the structure imposed by the model may be too weak due to the small number

of core-determining sets, arising from using a single binary explanatory variable.

The first two considerations are analysed by estimating Model A under the assumption of a triangular distribution, which has a bounded support, and with the assumption that the random coefficients α_i and β_i are independent. As shown in Section 2.7.2.2, the problem of an unbounded variance persists. Thus, the assumption of normality and the lack of restrictions on the relationship between the random coefficients cannot explain the problem of an unbounded variance.

Turning to the potential limitations of the data, regressions of the endogenous variable on the instruments, shown in Table 2.3, suggest that the instruments are strong. Furthermore, Section 2.7.3 shows that, conditional on a given value of the variance of β_i, the volume of the identified set is reduced significantly when more instruments are added: combining the early smoking indicator and father's further education indicator reduces the volume of the identified set by a factor of approximately 2 relative to using smoking as the only instrument; while combining mother's further education indicator with early smoking indicator reduces the volume of the identified set by a factor of approximately 4, compared to using the smoking indicator alone, which shows that strengthening the instruments increases identifying power. Nonetheless, neither of the two pairs of instruments were sufficient to impose a bound on the variance of β_i.

However, the apparently empty sets obtained when estimating Model B using a combination of all three instruments (Section 2.7.3), as well as when estimating Model C (Section 2.7.5.1), point to possible limitations of the data due to small cell sizes, resulting in unreliable estimates of the conditional probabilities on the right-hand side of Artstein's inequality (2.7). A detailed discussion of the issue of cell sizes is deferred to Section 2.7.5.1.

While the problem of an unbounded variance of β_i persists in the presence of stronger instruments, which increase the number of inequalities that need to be satisfied due to the right-hand side of inequality (2.7), it is important to note that added instruments do not increase the number, nor change the shape, of the core-determining sets S on the left-hand side of inequality (2.7). This suggests that it may be fruitful to focus on model specification changes which alter the collection of core-determining sets.

In particular, in their original paper, Chesher and Rosen (2014) provide a simulated example based on a model similar to Model B: with no included exogenous variables; an instrument with 4 points of support; and the assumption of a bivariate normal distribution imposed on

the random coefficients, which are not assumed to be independent. The key difference between their simulated example and Model B estimated in this chapter lies in the cardinality of the endogenous variable E_i, which is binary in Model B, but has 4 points of support in the simulation in Chesher and Rosen (2014). Since their simulation does not exhibit the problem of an unbounded variance, this suggests that the structure of Models A and B, which only include one binary explanatory endogenous variable in the outcome equation, may be too weak to provide a bound on the variance of β_i. This suggests an interesting question for future research, in the form of determining, based on simulated data, the minimal cardinality of the endogenous variable required in order for all estimated parameters to be bounded. However, this is beyond the scope of the current chapter.

Notably, the number of core-determining sets increases not only with the cardinality of the endogenous variable, but also with the inclusion of exogenous explanatory variables. Thus, an alternative solution to the problem of an unbounded variance of β_i is to include exogenous variables which introduce new, non-trivial, partitions of the parameter space, thus increasing and changing the shape of core-determining sets. However, this comes at the cost of significantly increasing computational time, as detailed in Section 2.4.2 and requires a larger dataset to ensure sufficient cell sizes for reliable estimation, as detailed in Section 2.7.5.1.

2.7.5 Model C: Adding Exogenous Variables

This section discusses results of estimates of the final specification, Model C. Under the assumption of normality, the model includes a coarse measure of age which enters the outcome equation linearly with a constant coefficient, as detailed in equation (2.16). As detailed in Section 2.4.2, this choice is driven by the limitations of data and computational power. Only one instrument is used: a binary indicator of smoking when young.

The age variable includes four categories: ages 25-30, 31-35, 36-40 and 41-45. In estimates below the age groupings are used as a single variable taking on 4 values, consisting of cell means of the respective age groups.

2.7.5.1 Point-identifying Model - Results

Table 2.13 shows estimates of Model C obtained using bivariate probit, which imposes constant coefficients.

Table 2.13: Point-identifying Model - Results with Exogenous Variables

	(1)
	Working (0/1)
University Education (0/1)	-0.176
	(-0.19)
Age category	0.233***
	(3.71)
Intercept	0.00389
	(0.03)
	University Education (0/1)
Smoking when young	-0.650***
	(-4.21)
Age category	0.210**
	(3.15)
Intercept	-1.540***
	(-7.91)
Athrho	
Constant	0.0298
	(0.06)
N	672

NB: t statistics, rather than standard errors, in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As in the case of Models A and B, the coefficients in the second equation, relating university attendance to the instrument and the exogenous variable, are closely in line with probit estimates in Table 2.3 showing the strength of the instrument. The instrument remains strong and highly statistically significant with a coefficient the same up to the second decimal place.

Interestingly, the results suggest that both the probability of working and the probability of attending university significantly increase with age. Even more surprisingly, after the inclusion of the age variable, education no longer has a statistically significant effect on the probability of working, and the point estimate is negative. The effect could arise, for example, due to women who went to university having children later, in which case, when observed in the sample (ages 25-45) they would be out of the labour force while those who did not attend university would be back in the labour force after giving birth. However, the coefficient is very noisily estimated with a t-statistic of -0.19. Thus, the model is highly uninformative about the mean of β_i.

2.7.5.2 Partially-identifying Model

Model C was subsequently estimated using the partial identification approach. As explained in Section 2.4.2, coefficient γ on the exogenous variable is assumed to be constant and thus there are two random coefficients in this model, α_i and β_i, whose joint distribution is assumed to be normal.

In estimation the projection approach was used to test a grid of values of three parameters: $(\bar{\alpha}, \bar{\beta}, \gamma)$ while allowing $\sigma_{\alpha\beta}$ and σ^2_β to vary arbitrarily.

The grid contained 7932 points, with 14 values of $\bar{\alpha}$ (from -0.1 to 1.2 in steps of 0.1), 44 values of $\bar{\beta}$ (from -10 to 10, in steps of just under 0.5) and 12 values of γ (from -0.1 to 0.45 in steps of 0.05).

None of the tested points satisfied inequality (2.7) despite allowing parameters $\sigma_{\alpha\beta}$ and σ^2_β to vary arbitrarily.

There are two possible explanations. Firstly, the inclusion of exogenous variables may have reduced the size of the estimated set so that it falls between points on the grid. In such case the suggested solution is estimation with more grid points. The feasibility of this solution is reduced by computational time constraints.

Table 2.14: Cell Sizes — Combinations of Data

	Smoking when young	Not smoking
Working ($Y_i = 1$); univ. educ. ($E_i = 1$)	10	46
Working ($Y_i = 1$); no univ. educ. ($E_i = 0$)	167	252
Not working ($Y_i = 0$); univ. educ. ($E_i = 1$)	3	21
Not working ($Y_i = 0$); no univ. educ. ($E_i = 0$)	77	96

An alternative explanation is that the identified set in this case is, in fact, empty. This would be the case if the model was rejected by the data. Two types of reasons could lead to such a situation. Firstly, the model itself may be misspecified and rejected by the data. However, the existence of bivariate probit estimates casts doubt on this explanation, since the bivariate probit is nested in the current specification. Alternatively, it could be the case that the model is correctly specified but the data do not allow for precise estimation. In this particular application this explanation seems highly plausible, especially in view of the cell sizes: the overall size of the sample is 672 observations. With 2 values of the outcome variable, 2 values of the endogenous variable, 2 values of the instrument and 4 of the exogenous variable, this gives $672 \div 32 = 21$

observations per cell on average. However, as shown in Table 2.14 in reality there distribution of observations across cells is much more varied, with some combinations observed with near-zero probability and leading to inference based on unreliably small numbers of observations. Such small cell sizes affect the reliability of estimates of the distribution of observables, F_{obs}, which is used on the right-hand side of inequality (2.7) which, in turn, affects which points are included or rejected from the estimated set.

Distinguishing between the two alternative explanations, i.e. whether the set is very small or empty, would require additional estimates, on a fine grid and, ideally, with a larger dataset. The feasibility of such further estimates is limited by computational time constraints.

Since the focus of the chapter is on analysing computational issues of implementing partial identification in practice, the finding of a small, and possibly empty, set is in itself valuable as it highlights the importance of not only sufficiently large sample sizes, but also sample sizes which achieve a sufficient coverage of all cells to make the estimation of the distribution of observables reliable, even when variables with non-trivial cardinality are included.

2.8 Conclusion and Discussion

This chapter discusses practical and computational aspects of implementing the partial-identifying method of Chesher and Rosen (2014) to problems involving estimation of random-coefficients binary-outcome models with endogeneity.

After setting out the model, the chapter compares methods which can be used to obtain point estimates under more restrictive assumptions. It presents the set-identifying method of Chesher and Rosen (2014), including key results, explaining the underlying intuition, and illustrating them graphically.

The chapter discusses two approaches to implementing the method: the full estimation approach and the projection approach, including outlining their strengths and weaknesses. In order to facilitate the implementation of the method of Chesher and Rosen (2014), the Appendix to this chapter includes pseudo-code as well as example Matlab code which translate the key identification results into a usable program and show the key steps which need to be taken in estimation.

The chapter discusses computational time and, in particular, the impact of the cardinality

of support of different variables (instruments, endogenous variables, exogenous variables) on the computational complexity of the model. It also discusses issues pertaining to adding more endogenous variables without increasing the number of instruments, as well as those relating to continuous variables.

The method is applied to a model of female labour force participation based on BHPS data. In order to illustrate the type of results that a user can expect, a baseline model with one binary endogenous variable and one binary instrument is estimated. Since the results suggest that the model does not impose bounds on the mean and variance of the random slope coefficient, the model is re-estimated with an alternative, triangular, distribution and the problem of unbounded variance discussed in detail.

The method is applied to two extensions of the baseline model. The first adds more instruments, the latter adds an exogenous variable. The key identifying inequality of the model of Chesher and Rosen (2014) suggests that adding more instruments or more exogenous variables weakly increases the identification strength of the model. This is corroborated by results based on including additional instruments: depending on the particular variable used, adding another binary instrument reduces the volume of the estimated set 2 or 4 times.

In the case of exogenous variables initial estimates suggest that the set is empty. This could be either because of the set being very narrow, due to the model being misspecified; or, more likely, because of the limitations of the data and, in particular, very small cell sizes affecting the reliability of the estimates of the distribution of observables which underpin the calculations of the estimated set.

Overall, the analysis presented in this chapter as well as the empirical application suggest that while the partially-identifying method has a number of advantages, such as applicability to settings in which point identification is not attainable otherwise and allowing to dispose of implausible assumptions, its practicability is still somewhat limited due to the computational time required in estimation as well as data limitations.

This chapter aims to increase the practicability of the method by familiarising the reader with key identification results and their intuition, discussing computational aspects of estimation and presenting step-by-step how the method can be implemented algorithmically. With ongoing advances in computational power as well as gains in the sizes and quality of datasets, the chapter hopes to move the method one step closer to being widely used in practice.

2.A Appendix to Computational Aspects of Partial Identification in Practice

This Appendix presents and discusses in detail a Matlab implementation of the method of Chesher and Rosen (2014), as outlined in pseudeo-code in Section 2.5.2. All comments refer to the code below.

The first section of the code reads in the data and determines the cardinality of the instrument vector. Note that the endogenous variable is defined as `'X'` and is assumed to be binary.

The second section of the code creates a grid of values of $(\bar{\alpha}, \bar{\beta})$ to be tested. The `linspace` function creates a vector of equally spaced elements, with minimum and maximum values given by the first two arguments and the number of points specified in the third argument. Example grid values are provided. The function `cartprod` returns the Cartesian product of the two vectors. It is a user-written command which can be found online.

The next section of the code calculates empirical probabilities of observables for the right hand side of inequality (2.7). Probabilities `P_raw` correspond to particular combinations of observables, while probabilities in the matrix `P` refer to the empirical probabilities corresponding to core-determining sets S. Using sets depicted in Figure 2.1, these correspond to, respectively, sets $A, B, C, D, A \cup B, A \cup D, B \cup C, C \cup D$. While the code applies to instruments of arbitrary cardinality, if the instrument is binary, the counter in the for-loop in line 18 should start at 0 and end at 1.

Importantly, in the absence of exogenous variables in the outcome equation, the left hand side of inequality (2.7) is the same for each value of the instrument Z. Therefore, the number of calculations involved in estimation can be reduced by first calculating the maximal, across values of Z, probability corresponding to the sets S and then checking inequality (2.7) only for that maximal probability — since, if it holds for the maximal probability, it holds for the other values as well. The maximal probabilities corresponding to the 8 sets S are collected in the matrix `P_max`, which is used as an argument in the objective function.

The remaining section of the main code is the optimisation stage. Using the syntax of Matlab's in-built `fmincon` command, the objective function `S_min` is minimised over $(\sigma_{\alpha\beta}, \sigma^2_{\beta})$, subject to the following constraints. The covariance, $\sigma_{\alpha\beta}$ is set to be smaller or equal than 1, which is implemented with the matrices `A` and `b`. The (empty) options `Aeq` and `beq` indicate that there

are no equality constraints on the parameters. The variance σ_β^2 is assumed to be strictly positive (bound from below by 0.001 for computational purposes) in order to avoid degeneracy. The covariance $\sigma_{\alpha\beta}$ is bounded from below by -1. Furthermore, the nonlinear constraint specified by the function `determinant_nonzero`, the code for which is provided below, bounds the determinant of the variance-covariance matrix away from zero[19].

Tests of whether each of the vectors of means on the test grid satisfy inequality (2.7) are performed in the for-loop between lines 54 and 66. For each row of the matrix containing the grid of means, the means as well as the maximal probabilities are passed onto the objective function `S_min` as parameters. The vector `x0` is chosen as an arbitrary starting point which satisfies the constraints of the optimisation problem. The function `S_min` is then minimised with respect to parameters $(\sigma_{\alpha\beta}, \sigma_\beta^2)$. The resulting minimised value is saved in the vector `optimised_value`. Recall that the function `S_min`, to be defined below, takes a negative value if and only if for the given vector of parameters the identifying inequality (2.7) holds for all core determining sets. Thus, the vector `points_in` collects information on whether the corresponding row of means belongs to the estimated set. The respective means are collected in the matrix `identified_points`, defined in the last line of the code. The points can then be used to obtain a refinement of the coarse grid, for example by randomly drawing test vectors of means locally to the points which belong to the estimated set. (Step not shown as implementation is analogous to the current code).

Note that the code can be easily parallelised, by replacing the `for` loop with a `parfor` loop with an appropriate specification of a parallel pool of workers. Although this is how the code producing results in Section 6 was implemented, the details are dropped from the current example, for the ease of exposition.

```
1 %%% Main Code
2 %% Loading the data
3 load('BHPS_Matlab_Z3.mat');
4 Y = double(st_dataset(:,1));
5 X = double(st_dataset(:,2));
6 Z = double(st_dataset(:,3));
```

[19] More precisely, it specifies that the determinant be greater than 0.1, in order to avoid near-singularity.

```
dimZ = size(unique(Z),1); % Cardinality of Z

%% Creating a grid of values of means of the random parameters
Alpha = linspace(0,1,101);
Beta = linspace(-5,5,101);
MU_combs = cartprod(Alpha, Beta);

%% Evaluating empirical probabilities of each of the combinations of
    observables
P_raw = NaN(7, dimZ);
P = NaN(8, dimZ);

for i = 1:dimZ
P_raw(1,i) = mean(Y.*(1*(Z==i)));
P_raw(2,i) = mean((1*Y==0).*(1*(Z==i)));
P_raw(3,i) = mean(Y.*X.*(1*(Z==i)));
P_raw(4,i) = mean(Y.*(1*(X==0)).*(1*(Z==i)));
P_raw(5,i) = mean((1*(Y==0)).*X.*(1*(Z==i)));
P_raw(6,i) = mean((1*(Y==0)).*(1*(X==0)).*(1*(Z==i)));
P_raw(7,i) = mean(1*(Z==i));
end

for i = 1:dimZ
P(1,i) = P_raw(3,i) ./ P_raw(7,i);
P(2,i) = P_raw(4,i) ./ P_raw(7,i);
P(3,i) = P_raw(5,i) ./ P_raw(7,i);
P(4,i) = P_raw(6,i) ./ P_raw(7,i);
P(5,i) = P_raw(1,i) ./ P_raw(7,i);
P(6,i) = (P_raw(3,i) + P_raw(6,i)) ./ P_raw(7,i);
P(7,i) = (P_raw(4,i) + P_raw(5,i)) ./ P_raw(7,i);
P(8,i) = P_raw(2,i) ./ P_raw(7,i);
end

P_max = transpose(max(P,[],2));

%% Optimisation

% Constraints
A = [1,0];
```

```
b = [1];
Aeq = [];
beq = [];
LB = [-1.000, 0.001];
UB = [];
nonlcon = @determinant_nonzero; % nonlinear constraint
points_in = NaN(size(MU_combs,1),1);
optimised_value = NaN(size(MU_combs,1),1);

for i = 1: size(MU_combs,1)
obj_fun = @(x) S_min(x(1),x(2),i,MU_combs,P_max);
x0 = [0.5,6]; %an arbitrary starting vector for the search,
    satisfying the constraints
     try
     [val fval] = fmincon(@(x)obj_fun(x), x0, A ,b, Aeq, beq, LB, UB,
         nonlcon);

     optimised_value(i,1) = fval;
     points_in(i,1) = 1*(fval<=0);
     catch
     error_num = ['Code broke at iteration: ',num2str(i)];
     disp(error_num);
     end
end

identified_points = MU_combs(find(points_in==1),:);
```

```
%% Nonlinear constraint
function [c, ceq] = determinant_nonzero(x)
c = -1 * (x(2) - (x(1))^2 ) + 0.1;
ceq = [];
end
```

The Matlab code below defines the objective function to be minimised. The function takes 5 arguments: `g` — the correlation coefficient $\sigma_{\alpha\beta}$, `sigma2` — the variance σ^2_β, `p` — the counter

of the `for`-loop, `MU` — the full matrix with the grid of means, `prob` — a matrix of empirical probabilities against which inequality (2.7) will be evaluated. As detailed above, in the absence of exogenous variables in the outcome equation this last matrix should contain the maximal probabilities, with the maximum taken over the values of the instrument vector Z.

In the first section of the function definition the vector of means as well as the covariance and variance are used in order to define a multivariate normal function, which, in what follows, will be integrated over the core-determining test sets S, in line with inequality (2.7). The function `lim_fun` is used to define the limits of integration for sets whose boundaries are not parallel to the axes, such as sets A, C, and their unions with other sets, shown in Figure 2.1. The limits of integration are further restricted to within 10 standard deviations from each mean. This is done in order to ease the computational burden of multiple integrations of a multivariate normal distribution at each iteration of the optimisation algorithm.

Lines 19-26 calculate the main components of the objective function, under the following reasoning. Inequality (2.7) must hold for each set of exogenous variables and for each core-determining set S. As explained above, since the left hand side of the inequality does not change with the values of the instrument, in the current setting it is sufficient to compare the left hand side with the maximal empirical probability, contained in the matrix `P_max`. Lines 19-26 calculate the difference between the maximal empirical probability corresponding to a given core-determining set S and the probability assigned to the same set S by a multivariate normal distribution with the vector of parameters consisting of the vector of means `MU(p,:)` and the covariance and variance at the current iteration of the optimisation algorithm. If inequality (2.7) is satisfied for a given test set S, the difference is nonpositive. Since the inequality must hold for each core-determining set S, each of the differences must be nonpositive. Thus, the last line of the code selects the largest of those differences — if it is nonpositive, then all the other differences are also nonpositive and the vector of parameters belongs to the identified set.

```
1 %% Objective function
2 function f = S_min(g, s2, p, MU, prob)
3     % Note on the arguments: g and s2 are the parameters with respect
          to which we are optimising
4     % p is used as a counter in the for-loop
```

```
5       % MU is a matrix of means
6       % prob is a matrix of probabilities
7
8       mu_p = [MU(p,2), MU(p,1)];
9       sigma_p = @(gamma, sigma2) [ sigma2, gamma; gamma, 1];
10      fun_norm= @(x,y) mvnpdf([x y], mu_p, sigma_p(g, s2));
11      lim_fun = @(x)  -1 .* x;
12      beta0 = MU(p,1);
13      beta1 = MU(p,2);
14      st_dev0 = 10*sqrt(1); % due to the normalisation to 1
15      st_dev1 = 10*sqrt(s2);
16
17      sum_temp = NaN(8,1);
18
19  sum_temp(1,1) = prob(1,1) - integral2(@(x,y)arrayfun(fun_norm,x,y),(
        beta1 -st_dev1), (beta1 + st_dev1) ,lim_fun, (beta0 + st_dev0));
20  sum_temp(2,1) = prob(1,2) - integral2(@(x,y)arrayfun(fun_norm,x,y),(
        beta1 -st_dev1), (beta1 + st_dev1),0,(beta0 + st_dev0));
21  sum_temp(3,1) = prob(1,3) - integral2(@(x,y)arrayfun(fun_norm,x,y),(
        beta1 -st_dev1), (beta1 + st_dev1),(beta0 -st_dev0),lim_fun);
22  sum_temp(4,1) = prob(1,4) - integral2(@(x,y)arrayfun(fun_norm,x,y),(
        beta1 -st_dev1), (beta1 + st_dev1),(beta0 -st_dev0),0);
23  sum_temp(5,1) = prob(1,5) - (integral2(@(x,y)arrayfun(fun_norm,x,y),(
        beta1 -st_dev1), 0,0,(beta0 + st_dev0)) + integral2(@(x,y)arrayfun
        (fun_norm,x,y),0, (beta1 + st_dev1),lim_fun, (beta0 + st_dev0)) );
         % ab
24  sum_temp(6,1) = prob(1,6) - ( 1 - integral2(@(x,y)arrayfun(fun_norm,x
        ,y),(beta1 - st_dev1), 0,0,lim_fun) ); % ad
25  sum_temp(7,1) = prob(1,7) - ( 1 - integral2(@(x,y)arrayfun(fun_norm,x
        ,y),0, (beta1+ st_dev1), lim_fun,0) ); % bc
26  sum_temp(8,1) = prob(1,8) - (integral2(@(x,y)arrayfun(fun_norm,x,y),(
        beta1 -st_dev1), 0, (beta0 -st_dev0),lim_fun) + integral2(@(x,y)
        arrayfun(fun_norm,x,y),0, (beta1 + st_dev1), (beta0 - st_dev0),0)
        ); %cd
27
28  f = max(sum_temp);
29  end
```

Chapter 3

Marked and Weighted Empirical Processes of Residuals from Non-Linear Regressions

Chapter Abstract

This chapter analyses the asymptotic behaviour of marked and weighted empirical processes of residuals from non-linear regressions in which the final set of estimates of the parameters of the non-linear function is obtained based on a subset of all observations, after deselecting observations classified as outliers based on the size of their residuals. I develop asymptotic theory for the marked and weighted empirical distribution of residuals, its conditional expectation, and the empirical process. I show that, under a number of conditions imposed on the regressors, marks, weights, the non-linear function, and the distribution of the errors, the existing asymptotic theory developed for marked and weighted functions of residuals from linear models with outlier detection can be extended to non-linear models with outlier detection. My results do not depend on unverifiable assumptions. To illustrate this, I show through an example how each of the assumptions can be checked. I suggest several applications of the theory developed in this chapter.

3.1 Introduction

The importance of non-linear modelling in Economics can be illustrated by the range of topics covered in the December 1992 special issue of the Journal of Applied Econometrics, which is entirely focused on nonlinear dynamics. Examples of applications include modelling aggregate behaviour of mergers and acquisitions by Town (1992); non-linear dynamics in employment mod-

elled by Burgess (1992); modelling non-linearities in business cycles by Terasvirta and Anderson (1992); and LeBaron's (1992) improved methodology for forecasting stock returns and exchange rates which exploits non-linearities in the time series.

The increasing availability of statistical methods for non-linear models with outlier detection has led to applications of such models to a variety of areas such as Chemistry (Vankeerberghen et. al., 1995); Biology (Motulsky and Brown, 2006) and Engineering (Marciukaitis et. al., 2017).

This chapter contributes to the literature on non-linear modelling with outlier detection by deriving asymptotic theory for marked and weighted empirical processes of residuals from non-linear regressions in which outlier detection is used. The theory can be used to develop specification tests such as normality and heteroskedasticity tests in non-linear robust regressions.

The chapter is organised as follows. Section 3.2 reviews existing literature in this area and discusses the contributions of the chapter. Section 3.3 introduces the modelling framework and key notation. Section 3.4 presents the key theorems proven in this chapter. Section 3.5 shows how each of the assumptions made in the theorems can be verified. Finally, Section 3.6 suggests applications of the theory derived in the chapter. Section 3.7 concludes.

3.2 Related Literature and Contributions

This chapter builds on two strands of theoretical Econometrics literature. Firstly, I contribute to existing research on asymptotic results and testing in regression models with outlier detection. Secondly, by focusing exclusively on non-linear regressions, I contribute to research on empirical processes of residuals from non-linear regressions.

The search for alternatives to least squares which are robust to the presence of outliers can be dated back to the 18th century and the attempts to estimate the masses of celestial bodies. Building on this literature, Edgeworth (1887) proposed an early version of what is currently known as the least absolute deviations estimator, as a solution to the shortcomings of the method of least squares. Further proposals include Huber's (1964, 1973) estimator which replaces the least squares criterion with a function $\rho(\cdot)$; Mallows' (1975) estimator, which uses a weighting function to reduce the influence of outlying observations; and Rousseeuw's (1984) least median of squares estimator, which minimises the median of squared residuals, rather than their sum.

Anscombe and Tukey (1963) propose an alternative solution to the problem of outliers,

which consists of examining the residuals from a least-squares regression to classify observations as outliers, and using weights to modify their influence or, indeed, remove outliers altogether. Literature on outlier detection, and estimation after deselection of outlying observations, has developed significantly since. Early work by Srikantan (1961) proposes a test to detect a single outlier in a linear regression. Subsequent literature offers a variety of methods and criteria for detecting multiple outliers. These include Hoaglin and Welsch's (1978) analysis of the 'hat matrix' to detect high-leverage observations; Cook and Weisberg's (1982) empirical influence curves, which detect influential observations in linear regressions and can also be applied to non-linear regressions; and Hadi's (1992) measure of potential which detects outliers both in the x- and y-spaces. Other measures include the 'difference in fits' proposed by Belsley et. al. (1980) and Atkinson's distance proposed by Atkinson (1981). Once observations are classified as outliers, they can be deselected and excluded from calculating least squares estimates, resulting in estimators such as the Least Trimmed Squares of Rousseeuw (1984), Huber-skip (Huber, 1964), or forward search of Hadi and Simonoff (1933) and Atkinson (1994).

The current chapter builds on this literature by assuming that outlier detection has been applied, and that the final set of estimates is obtained after deselecting observations with large residuals. However, unlike the above, the focus of this chapter is not on estimation methods, but on the large-sample properties of marked and weighted functions (empirical processes) of residuals obtained from the final model, after outlier deselection. In this context, weights refer to functions of regressors, and marks refer to functions of the residuals from the model estimated on the final sample of non-outlying observations. Understanding of the behaviour of marked and weighted empirical processes of residuals is essential for developing asymptotic theory for tests based on residuals such as normality or heteroskedasticity tests. Furthermore, our theory an be used to analyse the behaviour of the Huber-skip model when applied to non-linear least squares which, to the best of my knowledge, has not been done in this context before.

Asymptotic theory for empirical processes of residuals from linear regression models with outlier deselection is well-developed. Koul and Ossiander (1994) analyse the asymptotic properties of weighted empirical processes of residuals from linear regressions, though unlike the current chapter, they do not allow for marks. Koul and Levental (1989) analyse weak convergence of the residual empirical process without marks or weights, in the context of linear explosive autoregressive models, and they apply their results to the Kolmogorov-Smirnov test. Johansen

and Nielsen (2016) analyse empirical processes in the context of the forward search algorithm applied to linear regressions. Jiao and Nielsen (2017) analyse the properties of the gauge in Huber-skip M-estimators which involves analysing marked and weighted empirical processes of residuals in which the marks are in the form ε_i^p, where ε_i is the error term in the regression equation. Similarly, Berenguer-Rico et. al. (2019a) and (2019b) develop asymptotic theory for a class of marked and weighted empirical processes of residuals from linear regressions, where the marks are also in the form of ε_i^p. Berenguer-Rico and Nielsen (2017) develop a more general theory, allowing for both marks and weights, and where the marks need only to satisfy specific smoothness and boundedness conditions. Furthermore, in the context of the linear model with outlier detection, Berenguer-Rico and Nielsen (2017) analyse normality testing; while Kalina (2012), Alih and Ong (2020), and Berenguer-Rico and Wilms (2021) analyse heteroskedasticity testing.

The theory presented in this chapter extends the results of Berenguer-Rico and Nielsen (2017) from linear to non-linear regressions. I derive asymptotic results for a general class of marked and weighted empirical processes of residuals from non-linear regressions in which outlier detection is used. My focus is not on estimation but rather on analysing the asymptotic properties of the empirical processes of residuals. Details of estimation in non-linear regression models can be found in Amemiya (1983). See also Riazoshams et. al. (2018) for details of estimating linear and non-linear models with outlier detection.

Existing approaches in the literature on empirical processes of residuals from non-linear models are related to, but less general than, my work. Su and Wei (1991) analyse goodness-of-fit tests in models in which the distribution of the dependent variable belongs to an exponential family of distributions and in which there exists a link function which relates the mean of the dependent variable to a linear function of the regressors. While their test statistic is based on an empirical process, my work is more general as it applies to a wide class of non-linear regression models and it does not impose specific functional form assumptions on the distribution of the dependent variable. Koul (1996) derives asymptotic expansions of empirical processes of residuals from non-linear regressions, while Koul (2002) investigates weighted empirical processes in dynamic non-linear models, including analysing goodness-of-fit tests such as supremum distance tests. However, unlike the theory presented in the current chapter, they do not allow for marks in addition to weights in their empirical processes. Koul and Stute (1999) analyse the asymp-

totic properties of marked empirical processes of residuals from a general class of autoregressive models, which take the linear and non-linear regression models as special cases; and they apply the results to goodness-of-fit tests. However, their theory does not account for weights in the empirical processes. Wang and Phillips (2012) develop asymptotic theory for specification tests in non-linear nonstationary models, such as non-linear cointegrating relationships, in which the test is based on a kernel-smoothed test statistic. However, they do not account for outlier detection. Similarly, Escanciano (2007) considers marked and weighted empirical processes of residuals from a general class of possibly non-linear models of weakly stationary and ergodic processes, but without accounting for outlier detection.

Other relevant contributions consider applications of specific tests to non-linear models, possibly with non-stationary regressors. Berenguer-Rico and Nielsen (2020) develop the asymptotic distribution of the cumulated sum of squares test statistic in non-linear regression models, and allow for a general class of regressors which includes non-stationary regressors. Their approach is substantively different to my work since their test statistic belongs to a different class of models, and they do not allow for outlier detection. Khmaladze and Koul (2004) analyse the properties of the Kolmogorov-Smirnov test in non-linear regressions. My theory is more general and takes the Kolmorogov-Smirnov test as a special case in which both the weights and marks are set to 1. Furthermore, unlike Khmaladze and Koul (2004) I do not require that the estimator $\hat{\beta}$ be $n^{1/2}$-consistent. Instead, the scaled errors are allowed to diverge slightly.

In addition to the above distinctions between my approach and the existing literature, the approach presented in this chapter has two further advantages. Firstly, the theorems encompass a wide range of regressors, including both stationary and non-stationary regressors. The latter may include deterministic trends and stochastic trends, subject to the need to satisfy all assumptions detailed in Section 3.4. Secondly, the results do not rely on *ad hoc* assumptions which cannot be verified. Indeed, I show through an example how each of the assumptions underpinning our key theorems can be checked.

Since the results of this chapter generalise those of Berenguer-Rico and Nielsen (2017), my work builds on the framework presented there. Additionally, selected lemmas in the appendix to this chapter extend lemmas of Berenguer-Rico et. al. (2019a) from the linear case to non-linear models. Furthermore, certain proofs in this chapter use the iterated exponential martingale inequality of Johansen and Nielsen (2016a) which itself builds on Bercu and Touati (2008).

Martingale inequalities which allow for unbounded martingales allow me to incorporate both weights and marks into the empirical processes, which would not have been possible if one relied on the Freedman (1975) inequality, which is commonly used in the literature.

3.3 Marked and Weighted Empirical Processes of Residuals

This section introduces the underlying model as well as notation used throughout the chapter.

The aim of the chapter is to analyse marked and weighted empirical processes of residuals from the following non-linear regression, where the functional form of $g(x_i, \beta_0)$ is known

$$y_i = g(x_i, \beta_0) + \varepsilon_i. \tag{3.1}$$

The vector of explanatory variables, x_i, and the vector of parameters, β, are both k-vectors with β_0 being the true, unknown, value of the parameter vector. The errors ε_i are assumed to be identically and independently distributed with some unknown distribution function $F(c) = P(\varepsilon_i \leq \sigma c)$, where the scale σ is unknown. The theory developed in this chapter applies weights w_{in}, which will typically be a function of the regressors x_i. It is assumed that ε_i/σ is independent of the filtration $\mathcal{F}_{i-1}$ and that ε_{i-1}, x_i, and w_{in} are $\mathcal{F}_{i-1}$-measurable. I do not specify whether x_i is stationary as my framework is general enough to allow for both stationary and non-stationary regressions.

The chapter focuses on regressions which are robustified using outlier detection algorithms such as the 1-step Huber-skip estimator or the forward search algorithm. I model outlier deselection by assuming that the researcher obtains initial estimates $(\tilde{\beta}, \tilde{\sigma})$ of the parameters of interest (β, σ); and initial estimates of the residuals $\tilde{\varepsilon}_i = y_i - g(x_i, \tilde{\beta})$. Observations with large residuals, i.e. those for which $\tilde{\varepsilon}_i \geq \tilde{\sigma} c$ for a specified cut-off c, are deselected and the final estimates $\hat{\beta}, \hat{\sigma}$ are obtained based on the remaining observations. The final residuals $\hat{\varepsilon}_i = y_i - g(x_i, \hat{\beta})$ are calculated on the basis of the final set of estimates.

I analyse the asymptotic behaviour of truncated marked and weighted functions of the residuals, defined as

$$\hat{\mathsf{F}}_n^{w,m,g}(c) := n^{-1} \sum_{i=1}^{n} w_{in} m\left(\frac{\hat{\varepsilon}_i}{\hat{\sigma}}\right) 1_{(\tilde{\varepsilon}_i \leq \tilde{\sigma} c)}, \tag{3.2}$$

where $m(\cdot)$ is the mark function and w_{in} are the weights, as defined above and usually chosen in

such a way that $n^{-1}\sum_{i=1}^{n} w_{in}$ has a limit. As an example, setting both weights w_{in} and marks $m\left(\frac{\hat{\varepsilon}_i}{\hat{\sigma}}\right)$ to be equal to 1 for all observations gives us the empirical distribution of the scaled residuals.

As noted by Johansen and Nielsen (2016a), there is an important distinction between outlier deselection in cross-sectional and time series models. While in the former an outlying observation can simply be removed from the sample, in the case of time series data the same outlying observation may feature as a regressor in multiple time periods. I follow the approach of Johansen and Nielsen (2016a) in applying weights of zero to those observations whose residuals (based on a preliminary estimator) are larger than a cut-off value; while keeping them in the sample so that they can be used as regressors in other time periods. As detailed in Johansen and Nielsen (2016, p. 322) this corresponds to the assumption of the observations being correctly measured but possibly generated by a different model.

3.3.1 Linearisation of the Residuals

The aim of the chapter is to show that asymptotic results for empirical processes of residuals from linear models can be extended to non-linear models, if certain conditions are met. To this aim, throughout this chapter, I work with linearised versions of residuals $\tilde{\varepsilon}_i, \hat{\varepsilon}_i$. The advantage of doing so is that the linearised version emphasises which aspects are shared between residuals based on linear an non-linear models, and which are different. This gives rise to restrictions which the function $g(\cdot,\cdot)$ must satisfy for the results derived for linear models to be applicable to non-linear models.

Let β_0 be the true value of the vector of parameters β. In line with Berenguer-Rico and Nielsen (2017), in this chapter I apply normalisations by a matrix N. The matrix N is chosen so that $\sum_{i=1}^{n} \dot{g}_{in}^{0\prime}\dot{g}_{in}^{0}$, where $\dot{g}_{in}^{0} := N'\dot{g}(x_i,\beta_0)$, has a positive limit. For example, in a linear case with stationary regressors, $N = n^{-1/2}I_k$. The estimation errors are normalised as $\tilde{b}_1 = N^{-1}(\tilde{\beta}-\beta_0)$ and $\hat{b}_m = N^{-1}(\hat{\beta}-\beta_0)$ for the parameters of the non-linear function and as $\tilde{a}_1 = n^{1/2}(\tilde{\sigma}-\sigma)$ and $\hat{a}_m = n^{1/2}(\hat{\sigma}-\sigma)$, for the initial and final estimation errors in the scale.

To linearise the residuals I apply a second-order Taylor series expansion, for an intermediate point β_{1i}^{*} such that $|\beta_{1i}^{*}-\beta_0| \leq |\tilde{\beta}-\beta_0|$, which allows me to express the non-linear function as

$$g(x_i,\tilde{\beta}) = g(x_i,\beta_0) + \left(\tilde{\beta}-\beta_0\right)'\dot{g}(x_i,\beta_0) + \frac{1}{2}\left(\tilde{\beta}-\beta_0\right)'\ddot{g}(x_i,\beta_{1i}^{*})\left(\tilde{\beta}-\beta_0\right), \tag{3.3}$$

or equivalently:

$$g(x_i, \tilde{\beta}) - g(x_i, \beta_0) = \tilde{b}_1' N' \dot{g}(x_i, \beta_0) + \frac{1}{2} \tilde{b}_1' N' \ddot{g}(x_i, \beta_{1i}^{1*}) N \tilde{b}_1. \tag{3.4}$$

Using the above expansions, the residual $\tilde{\varepsilon}_i$ can be linearised as follows.

$$\begin{aligned} \tilde{\varepsilon}_i &= y_i - \tilde{y}_i \\ &= g(x_i, \beta_0) + \varepsilon_i - g(x_i, \tilde{\beta}) \\ &= \varepsilon_i - \left[\tilde{b}_1' N' \dot{g}(x_i, \beta_0) + \frac{1}{2} \tilde{b}_1' N' \ddot{g}(x_i, \beta_{1i}^*) N \tilde{b}_1 \right] \\ &= \varepsilon_i - \tilde{b}_1' \dot{g}_{in}^0 - \frac{1}{2} \tilde{b}_1' \ddot{g}_{in}^{1*} \tilde{b}_1, \end{aligned} \tag{3.5}$$

where $\dot{g}_{in}^0$ was defined above as $\dot{g}_{in}^0 := N' \dot{g}(x_i, \beta_0)$ and $\ddot{g}_{in}^{1*}$ is further defined as $\ddot{g}_{in}^{1*} = N' \ddot{g}(x_i, \beta_{1i}^*) N$, where the derivatives are taken with respect to β.

The residual $\hat{\varepsilon}_i$ can be linearised analogously as

$$\begin{aligned} \hat{\varepsilon}_i &= \varepsilon_i - \left[\hat{b}_m' N' \dot{g}(x_i, \beta_0) + \frac{1}{2} \hat{b}_m' N' \ddot{g}(x_i, \beta_{mi}^*) N \hat{b}_m \right] \\ &= \varepsilon_i - \hat{b}_m' \dot{g}_{in}^0 - \frac{1}{2} \hat{b}_m' \ddot{g}_{in}^{m*} \hat{b}_m. \end{aligned} \tag{3.6}$$

Further, accounting for the estimation of the scale σ and denoting the rescaled error by $\hat{a}_m = n^{1/2}(\hat{\sigma} - \sigma)$, we can express the normalised residual $\hat{\varepsilon}_i / \hat{\sigma}$ as

$$\frac{\hat{\varepsilon}_i}{\hat{\sigma}} = \frac{\varepsilon_i - \hat{b}_1' \dot{g}_{in}^0 - \frac{1}{2} \hat{b}_m' \ddot{g}_{in}^* \hat{b}_m}{\sigma + n^{-1/2} \hat{a}_m}. \tag{3.7}$$

3.3.2 Replacing Stochastic Errors with Non-Stochastic Terms

Recall that the initial estimation errors are scaled as: $\tilde{b}_1 = N^{-1}(\tilde{\beta} - \beta)$ and $\tilde{a}_1 = n^{1/2}(\tilde{\sigma} - \sigma)$. The errors in the updated estimates are scaled analogously: $\hat{b}_m = N^{-1}(\hat{\beta} - \beta)$ and $\hat{a}_m = n^{1/2}(\hat{\sigma} - \sigma)$. Although in standard situations the estimation errors can be expected to converge in distribution, in the following the rescaled errors are allowed to diverge at a rate $n^{1/4-\eta}$, where $0 < \eta \leq 1/4$. Thus, each term of $\hat{\theta} = (\hat{a}_m, \hat{b}_m, \tilde{a}_1, \tilde{b}_1)$ is $o_p(n^{1/4-\eta})$.

Analogously to Berenguer-Rico and Nielsen (2017), I analyse the behaviour of the quantities of interest by focusing on their suprema over a compact set, uniformly in that set. Specifically,

the aim is to show that a given quantity, which is a function of the estimation errors $\hat{\theta} = (\hat{a}_m, \hat{b}_m, \tilde{a}_1, \tilde{b}_1)$ and the cut-of c, is $o_p(1)$. Let $Q_n(\hat{\theta}, c)$ denote a generic example of such a quantity. My aim is to show that $Q_n(\hat{\theta}, c) = o_p(1)$. Following Lemma A.1 of Berenguer-Rico and Nielsen (2017), if a compact set Θ exists such that $\lim_{n\to\infty} P(\hat{\theta} \in \Theta) > 1 - \epsilon$, then $P\{|Q_n(\hat{\theta}, c)| > \epsilon\} \leq P(\sup_{\theta\in\Theta} |Q_n(\theta, c)| > \epsilon) + \epsilon$. Thus, it is sufficient to study the behaviour of the quantity $Q_n(\theta, c)$ uniformly in $\theta \in \Theta$. In the context of the model analysed in this chapter, $\Theta := \{\theta : |\theta| \leq Bn^{1/4-\eta}\}$ for some $B > 0$ since $|\hat{\theta}| = o_p(n^{1/4-\eta})$.

As a result, the stochastic estimation errors $\hat{\theta}$ can be replaced with their non-stochastic counterparts $\theta = (a_m, b_m, a_1, b_1)$, which vary within the compact set $\{\theta : |\theta| \leq Bn^{1/4-\eta}\}$ for some $B > 0$. For example, the term in equation (3.7) is replaced with

$$\frac{\varepsilon_i - b_1' \dot{g}_{in}^0 - \frac{1}{2} b_m' \ddot{g}_{in}^* b_m}{\sigma + n^{-1/2} a_m}. \tag{3.8}$$

Accordingly, all results in this chapter which involve estimation errors are derived uniformly in $\{\theta : |\theta| \leq Bn^{1/4-\eta}\}$.

3.3.3 Other Notation

Throughout this chapter, the following shorthand will be used, to simplify notation:

$$\begin{aligned} \sup_{\theta} &:= \sup_{\theta:|\theta|\leq Bn^{1/4-\eta}}, \\ \sup_{\beta_i} &:= \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}}, \\ \sup_{c} &:= \sup_{c\in\mathbb{R}}. \end{aligned} \tag{3.9}$$

3.4 Key Results

This chapter extends the results of Berenguer-Rico and Nielsen (2017), who focus on linear regressions, to empirical processes of residuals based on non-linear regressions. I maintain notation consistent with their work. Berenguer-Rico and Nielsen (2017) define three key quantities in the linear model: the marked and weighted empirical distribution, the compensator, and the empirical process. In the following equation, I define the marked and weighted empirical distribution applicable to the linear case. However, to make comparisons with the non-linear

case more explicit, the normalised vector of regressors x_{in} used by Berenguer-Rico and Nielsen (2017) in the linear model, is here replaced with the normalised first derivative of the non-linear function $g(\cdot,\cdot)$. This reflects the fact that if $g(x_i,\beta)=\beta' x_i$ then $\dot{g}(x_i,\beta_0)=x_i$. Analogously to Sections 3.3.1 and 3.3.2, the errors have been linearised and replaced with their non-stochastic equivalents. Thus, the marked and weighted empirical distribution function is defined as

$$F_n^{w,m}(\theta,c):=n^{-1}\sum_{i=1}^{n} w_{in} m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)1_{\left(\varepsilon_i\leq \sigma c+n^{-1/2}a_1 c+b_1'\dot{g}_{in}^0\right)}. \tag{3.10}$$

The corresponding compensator is the weighted sum of conditional expectations, which is required in order to create martingales and apply martingale theory

$$\bar{F}_n^{w,m}(\theta,c):=n^{-1}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)1_{\left(\varepsilon_i\leq \sigma c+n^{-1/2}a_1 c+b_1'\dot{g}_{in}^0\right)}\right\}. \tag{3.11}$$

The empirical process combines the empirical distribution and the compensator into a martingale, and applies scaling by $n^{1/2}$:

$$\mathbb{F}_n^{w,m}(\theta,c):=n^{1/2}\left\{F_n^{w,m}(\theta,c)-\bar{F}_n^{w,m}(\theta,c)\right\}. \tag{3.12}$$

For the non-linear model studied in the current chapter, three analogous terms are defined. The terms, based on residuals from non-linear regressions, are distinguished by the addition of the non-linear function g in the superscript. Given the first-order Taylor series expansions derived in Section 3.3.1, and the method of replacing stochastic errors with non-stochastic terms, as detailed in Section 3.3.2, the linearised equivalent of the marked and weighted empirical distribution in (3.2) can be expressed as

$$F_n^{w,m,g}(\theta,c):=n^{-1}\sum_{i=1}^{n} w_{in} m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0-\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)1_{\left(\varepsilon_i\leq \sigma c+n^{-1/2}a_1 c+b_1'\dot{g}_{in}^0+\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1\right)}. \tag{3.13}$$

The compensator in the non-linear model is defined accordingly as the following weighted sum of conditional expectations:

$$\bar{F}_n^{w,m,g}(\theta,c):=n^{-1}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0-\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)1_{\left(\varepsilon_i\leq \sigma c+n^{-1/2}a_1 c+b_1'\dot{g}_{in}^0+\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1\right)}\right\}. \tag{3.14}$$

The above combined give the definition of the empirical process in the non-linear model:

$$\mathbb{F}_n^{w,m,g}(\theta, c) := n^{1/2} \left\{ F_n^{w,m,g}(\theta, c) - \bar{F}_n^{w,m,g}(\theta, c) \right\}. \tag{3.15}$$

Berenguer-Rico and Nielsen (2017) further define terms $\mathbb{F}_n^{w,m}(0, c)$ and $\bar{F}_n^{w,m}(0, c)$, which are special cases of $\mathbb{F}_n^{w,m}(\theta, c)$ and $\bar{F}_n^{w,m}(\theta, c)$ with no estimation errors. Theorem 2.1 of their paper shows that the term $\mathbb{F}_n^{w,m}(\theta, c)$ is asymptotically equivalent to $\mathbb{F}_n^{w,m}(0, c)$, uniformly in θ and c. Mathematically, their result can be stated as

$$\sup_{\theta} \sup_{c} |\mathbb{F}_n^{w,m}(\theta, c) - \mathbb{F}_n^{w,m}(0, c)| = o_p(1). \tag{3.16}$$

Berenguer-Rico and Nielsen (2017) also provide an asymptotic uniform linearisation of the compensator $\bar{F}_n^{w,m}(\theta, c)$, which is asymptotically equivalent, uniformly in θ and c, to the sum of $\bar{F}_n^{w,m}(0, c)$ and a bias term $\mathcal{B}_n^{w,m}$. Following Theorem 2.1 of Berenguer-Rico and Nielsen, the linearisation can be stated as follows, where the linear vector of normalised regressors x_{in} used in their work has been replaced with the normalised vector of first derivatives $\dot{g}_{in}^0$,

$$\sup_{\theta} \sup_{c} \left| n^{1/2} \{ \bar{F}_n^{w,m}(\theta, c) - \bar{F}_n^{w,m}(0, c) \} - \mathcal{B}_n^{w,m}(\theta, c) \right| = o_p(1), \tag{3.17}$$

where $\mathcal{B}_n^{w,m}(\theta, c) = \mathcal{B}_{1,n}^{w,m}(\theta_1, c) - \mathcal{B}_{m,n}^{w,m}(\theta_m, c)$ and the terms $\mathcal{B}_{1,n}^{w,m}(\theta_1, c)$ and $\mathcal{B}_{m,n}^{w,m}(\theta_m, c)$ are defined as

$$\begin{aligned}
\mathcal{B}_{1,n}^{w,m}(\theta_1, c) &:= \sigma^{-1} m(c) f(c) n^{-1/2} \sum_{i=1}^{n} w_{in} (n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0), \\
\mathcal{B}_{m,n}^{w,m}(\theta_m, c) &:= \sigma^{-1} n^{-1/2} \sum_{i=1}^{n} w_{in} \left\{ n^{-1/2} a_m \mathsf{E} \frac{\varepsilon_i}{\sigma} \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) 1_{(\varepsilon_i \leq \sigma c)} + b_m' \dot{g}_{in}^0 \mathsf{E} \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) 1_{(\varepsilon_i \leq \sigma c)} \right\}.
\end{aligned} \tag{3.18}$$

The aim of this chapter is to show that the asymptotic results of Berenguer-Rico and Nielsen (2017) can be extended to encompass empirical processes of residuals from non-linear regressions. This is achieved through a combination of the following three theorems. Firstly, in Theorem 3.4.1 I show that the compensators in the linear and non-linear models are asymptotically equivalent, uniformly in θ and c.

Assumption 3.4.1. Assume the following, where $0 < \eta \leq 1/4$ was defined in Section 3.3.2.

1. *The regressors and the function $g(\cdot)$:*

   (a) $\max_{1\leq i\leq n} \|N'\dot{g}(x_i;\beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$ for some $\kappa > 0$

   (b) $\max_{1\leq i\leq n} \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\| = O_p(n^{-1/2+2\eta-\kappa})$ for some $\kappa > 0$

   (c) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{2\eta})$

   (d) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\|^2 = o_p(n^{-1/2+4\eta})$

   (e) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\dot{g}(x_i;\beta_0)\|\|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{-1/4+3\eta})$

   (f) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\dot{g}(x_i;\beta_0)\|^2\|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{-1/2+4\eta})$

2. *The mark function $m(u)$*:

   (a) $\sup_{u\in\mathbb{R}} |m(u)|\, f(u) < \infty$

   (b) $\sup_{u\in\mathbb{R}} (1+|u|)\, |\dot{m}(u)|\, f(u) < \infty$

   (c) $\int_{-\infty}^{\infty} |\dot{m}(u)|\, f(u)du < \infty$

   (d) (local Lipschitz) $\exists\omega > 0$ and a function $\mathring{m}(u) \geq 0$ such that if $\forall u^*, u$ such that $|u^* - u| \leq (1+|u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u|\mathring{m}(u)$ and where:

      - $\sup_{u\in\mathbb{R}} (1+|u|+u^2)\mathring{m}(u)f(u) < \infty$

Remark 3.4.1. *Assumption 3.4.1 imposes restrictions on the non-linear function $g(x_i,\beta)$, the regressors x_i, the mark function $m(u)$, and the density of the errors $f(\varepsilon_i/\sigma)$. Section 3.5, which illustrates how the assumptions can be verified, shows that Assumption 3.4.1/1 is satisfied, for example, for a version of the Wu (1981) model. Assumption 3.4.1/2 is satisfied, for example, when the marks are polynomial and the errors ε_i/σ are normally distributed. Section 3.5 shows this for the special cases of $m(u) = u^3$ and $m(u) = u^4$ with errors ε_i/σ distributed standard normal.*

Theorem 3.4.1. Under Assumption 3.4.1, the following holds as $n \to \infty$,

$$\sup_{\theta} \sup_{c} \left| n^{1/2}\{\bar{F}_n^{w,m,g}(\theta,c) - \bar{F}_n^{w,m}(\theta,c)\} \right| = o_p(1). \tag{3.19}$$

Theorem 3.4.1 shows that the compensator in the non-linear model is asymptotically equivalent to the compensator in the linear model, where $N'\dot{g}(x_i, \beta_0)$ is used as a regressor, in place of $x_{in} = N'x_i$. With the key result established for the compensator, the following theorem derives an analogous asymptotic equivalence result for the marked and weighted empirical distribution.

Definition 3.4.1. Let $h(c) \geq 0$ have support $\mathcal{S} =]\underline{c}, \bar{c}[$ such that $-\infty \leq \underline{c} < \bar{c} \leq \infty$.

i) $h(\cdot)$ is Lipschitz if $\exists C_L > 0 : \forall c, c^\dagger \in \mathcal{S} : |h(c) - h(c^\dagger)| \leq C_L|c - c^\dagger|$;

ii) $h(\cdot)$ has weakly unimodal bounds if a constant $C_u \geq 1$ and a function u exist such that $\forall c \in \mathbb{R} : u(c) \leq h(c) \leq C_u u(c)$, where u has a finite mode at $c_{mode} \in \mathcal{S}$ and $u(c)$ is non-increasing for $c > c_{mode}$ and non-decreasing for $c < c_{mode}$.

Assumption 3.4.2. Let $\tilde{v}$ represent each of w_{in}, $w_{in}n^{-1/4-\eta}$ and $w_{in}n^{1/4-\eta}|N'\dot{g}(x_i;\beta_0)|$. Let $\tilde{h}(u)$ represent each of $m(u)$, $\dot{m}(u)$ and $u\dot{m}(u)$. Let r be such that $2^{r-1} > 1 + (1/4 - \eta)(\dim b_1)$ and q such that $1 \leq q \leq r$. Further, define:

$$\tilde{H}_r(c) = \int_{-\infty}^{c} \left(1 + \{\tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right) f\left(\frac{\varepsilon_i}{\sigma}\right) \mathrm{d}\left(\frac{\varepsilon_i}{\sigma}\right) = \mathsf{E}\left(1 + \{\tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right) 1_{(\varepsilon_i/\sigma \leq c)}. \tag{3.20}$$

Assume the following, where $0 < \eta \leq 1/4$ was defined in Section 3.3.2.

1. *The weights* $\tilde{v}_{in}$:

   (a) $\mathsf{E}\sum_{i=1}^{n}\left(1 + |\tilde{v}_{in}|^{2^r}\right) = O(n)$

2. *The regressors and the function* $g(\cdot)$*, for all* $1 \leq q \leq r$*:*

   (a) $\max\limits_{1\leq i\leq n} \|N'\dot{g}(x_i;\beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$ for some $\kappa > 0$

   (b) $\max\limits_{1\leq i\leq n} \sup\limits_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\| = O_p(n^{-1/2+2\eta-\kappa})$ for some $\kappa > 0$

   (c) $\sum_{i=1}^{n} |w_{in}|^2 \sup\limits_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\|^2 = o_p(n^{4\eta-1})$

   (d) $\mathsf{E}\sum_{i=1}^{n} |w_{in}| \sup\limits_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\|\|N'\dot{g}(x_i;\beta_0)\| = O(n^{-1/4+2\eta})$

   (e) $\mathsf{E}\sum_{i=1}^{n} |w_{in}| \sup\limits_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\|^2 = O(n^{-1/2+3\eta})$

   (f) $\mathsf{E}\sum_{i=1}^{n}\left(1 + |\tilde{v}_{in}|^{2^r}\right) \sup\limits_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\| = O(n^{2\eta-\kappa})$ for some $\kappa > 0$

   (g) $\mathsf{E}\sum_{i=1}^{n}\left(1 + |\tilde{v}_{in}|^{2^r}\right)|N'\dot{g}(x_i;\beta_0)| = O(n^{1/2})$

(h) $\mathsf{E}\sum_{i=1}^{n}\left(1+|\tilde{v}_{in}|^{2^r}\right)|N'\dot{g}(x_i;\beta_0)|^2 = O(n^{\eta})$

3. *The mark function* $m(u)$:

   (a) $\mathsf{E}\,|\dot{m}(u)|^2 < \infty$

   (b) (local Lipschitz) $\exists\omega > 0$ and a function $\mathring{m}(u) \geq 0$ such that if $\forall u^*, u$ such that $|u^* - u| \leq (1+|u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u|\mathring{m}(u)$ and where:

   - $\sup_{u\in\mathbb{R}}(1+u)\mathring{m}(u)f(u) < \infty$
   - $\int_{-\infty}^{+\infty}(1+|u|+u^2)\mathring{m}(u)f(u)du < \infty$.

4. *The mark function* $\tilde{h}(\cdot)$*:*

   (a) $\mathsf{E}\left|\tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right)\right|^{2^r} < \infty$

5. *The distance function* $\tilde{H}_r(c)$:

   (a) the support of $\dot{\tilde{H}}_r$, that is $\mathcal{C} = \{c : \dot{\tilde{H}}_r(c) > 0\}$, is an interval with endpoints $\underline{c}, \bar{c}$ such that $-\infty \leq \underline{c} < \bar{c} \leq \infty$;

   (b) $\sup_{x\in\mathcal{S}} 0 \leq \dot{\tilde{H}}_r(x) < \infty$;

   (c) $\tilde{H}_r = \tilde{H}_r(\infty) = \int_{\mathbb{R}} \dot{\tilde{H}}_r(c)dc < \infty$;

   (d) $\dot{\tilde{H}}_r(c)$ is Lipschitz, as defined in Definition 3.4.1;

   (e) $\dot{\tilde{H}}_r(c)$ has weakly unimodal bounds as defined in Definition 3.4.1.

Remark 3.4.2. *Analogously to Assumption 3.4.1, Assumption 3.4.2 imposes restrictions on the non-linear function* $g(x_i, \beta)$*, the regressors* x_i*, the mark function* $m(u)$*, and the density of the errors* $f(\varepsilon_i/\sigma)$*. As in the case of Assumption 3.4.1, Section 3.5 shows how the assumptions can be verified. The two new sets of assumptions are those on the mark function* $\tilde{h}(u)$ *and the distance function* $\tilde{H}_r(c)$*. The two new Assumptions: 3.4.1/4 and 3.4.1/5 are satisfied, for example, when the mark function* $m(u)$ *is polynomial and the errors* ε_i/σ *are normally distributed. An example of how the assumptions can be verified for the special case where* $m(u) = u^3$ *or* $m(u) = u^4$ *with errors* ε_i/σ *distributed standard normal, is presented in Section 3.5.*

Theorem 3.4.2. Under Assumption 3.4.2, the following holds as $n \to \infty$,

$$\sup_{\theta}\sup_{c}\left|n^{1/2}\{F_n^{w,m,g}(\theta,c) - F_n^{w,m}(\theta,c)\}\right| = o_p(1). \tag{3.21}$$

Theorem 3.4.2 shows that the marked and weighted empirical distribution in the non-linear model is asymptotically equivalent to the marked and weighted empirical distribution in the linear model with $N'\dot{g}(x_i, \beta_0)$ used as a regressor in place of x_{in}. The result applies uniformly in θ and c. Given the definition of $\mathbb{F}_n^{w,m,g}(\theta, c)$, Theorems 3.4.1 and 3.4.2 imply the following result.

Theorem 3.4.3. Under Assumptions 3.4.1 and 3.4.2, the following holds as $n \to \infty$,

$$\sup_{\theta} \sup_{c} |\mathbb{F}_n^{w,m,g}(\theta, c) - \mathbb{F}_n^{w,m}(\theta, c)| = o_p(1). \tag{3.22}$$

Theorem 3.4.3 shows that the empirical process in the non-linear model, $\mathbb{F}_n^{w,m,g}(\theta, c)$, is asymptotically equivalent, uniformly in θ and c, to the empirical process analysed by Berenguer-Rico and Nielsen (2017) in the linear model, with $N'\dot{g}(x_i, \beta_0)$ used as a regressor, in place of x_{in}. This key result allows one to apply asymptotic theory derived for linear models, to empirical processes of residuals from non-linear models. To illustrate this, below I cite existing results derived by Berenguer-Rico and Nielsen (2017) and show the implications of combining the theorems derived in this chapter with existing asymptotic results.

The following assumption restates Assumptions 2.1 and 2.2 of Berenguer-Rico and Nielsen (2017), which are adapted to account for the normalised vector of regressors in the linear model, x_{in}, being replaced with the normalised vector of first derivatives, $\dot{g}_{in}^0$. This additional assumption is required to combine the results of this chapter with those of Berenguer-Rico and Nielsen (2017).

Assumption 3.4.3 (Assumptions 2.1 and 2.2 of Berenguer-Rico and Nielsen (2017))**.**
Let Assumptions 2.1 and 2.2 of Berenguer-Rico and Nielsen (2017) hold for $N'x_i = N'\dot{g}(x_i, \beta_0)$. This requires that for $0 \leq \zeta < \eta < 1/4$ the following holds, where $\tilde{m}(u)$ represents each of $m(u)$, $u\dot{m}(u)$, $\dot{m}(u)$; and $\tilde{h}(u) = \{1 + m^4(u)\}f(u)$.

1. *The filtration:*

   (a) Let $\mathcal{F}_{i-1}$ be a filtration such that ε_{i-1}, $x_i = \dot{g}(x_i, \beta_0)$, and w_{in} are $\mathcal{F}_{i-1}$-measurable and that ε_i/σ is independent of $\mathcal{F}_{i-1}$ with distribution function F and positive density f on $\mathbb{R}$ with derivative $\dot{\mathsf{f}}$.

2. *The density and marks:*

   (a) $\int_{\infty}^{\infty} \tilde{m}^4(u) f(u) du < \infty$

(b) $\sup_{u\in\mathbb{R}} |u|\{1+m^4(u)\}f(u) < \infty$

(c) $\sup_{u\in\mathbb{R}} |\frac{\partial}{\partial u}\left[\{1+\tilde{m}^4(u)\}f(u)\right] < \infty$

(d) $\sup_{u\in\mathbb{R}}(1+u^2)|\dot{m}(u)f(u)| < \infty$

(e) $\sup_{c>0} \frac{\sup_{u\geq c}\tilde{h}(u)}{\inf_{0\leq u\leq c}\tilde{h}(u)} < \infty$

(f) $\sup_{c>0} \frac{\sup_{u\leq -c}\tilde{h}(u)}{\inf_{-c\leq u\leq 0}\tilde{h}(u)} < \infty$

(g) (local Lipschitz) $\exists\omega > 0$ and a function $\mathring{m}(u) \geq 0$ such that if $\forall u^*, u$ such that $|u^* - u| \leq (1+|u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u|\mathring{m}(u)$ and where $\int_{-\infty}^{+\infty}(1+|u|^2)\mathring{m}(u)f(u)du < \infty$

3. *Regressors:*

(a) $\max_{1\leq i\leq n} \left|n^{1/2-\zeta}N'\dot{g}(x_i,\beta_0)\right| = O_p(1)$

4. *Weights:*

(a) $\mathsf{E}\sum_{i=1}^n |w_{in}|^{2+\xi} = O(n)$ for some $\xi > 0$

Remark 3.4.3. *There is a notable overlap between Assumption 3.4.3, which restates Assumptions 2.1 and 2.2. of Berenguer-Rico and Nielsen (2017) for the special case of $N'x_i = N'\dot{g}(x_i,\beta_0)$; and Assumptions 3.4.1 and 3.4.2. Specifically, Assumption 3.4.3/1a is satisfied throughout this chapter; Assumption 3.4.3/2d is a strengthened version of Assumption 3.4.1/2b; Assumption 3.4.3/2g is implied by Assumption 3.4.1/2d; Assumption 3.4.3/3a is a stronger version of Assumption 3.4.1/1a; and Assumption 3.4.3/4a is related to Assumption 3.4.2/1a — whether it is stronger or weaker depends on the value of r in Assumption 3.4.2. Assumptions 3.4.3/2a, 2b, 2c, 2e, and 2f are new relative to this chapter.*

Combining Theorem 3.4.3 with Theorem 2.1 of Berenguer-Rico and Nielsen (2017), as summarised in equation (3.16), implies the following result.

Theorem 3.4.4. Suppose that Assumptions 3.4.1, 3.4.2, and 3.4.3 hold. Then, the following holds as $n \to \infty$,

$$\sup_{\theta}\sup_{c} |\mathbb{F}_n^{w,m,g}(\theta,c) - \mathbb{F}_n^{w,m}(0,c)| = o_p(1). \tag{3.23}$$

Theorem 3.4.4 shows that the empirical process based on residuals from the non-linear model is asymptotically equivalent, uniformly in θ and c, to the empirical process based on the linear

model without estimation errors. This has two profound implications. Firstly, the non-linearity of the model, as captured by a non-zero second derivative $\ddot{g}(x_i, \beta)$, does not affect the asymptotic behaviour of the empirical process. Secondly, analogously to the result proven for the linear case by Berenguer-Rico and Nielsen (2017), the estimation errors have no impact on the asymptotic behaviour of the empirical process.

Combining Theorem 3.4.1 with Theorem 2.1 of Berenguer-Rico and Nielsen (2017) gives an analogous result for the compensator $\bar{F}_n^{w,m,g}(\theta, c)$:

Theorem 3.4.5. Suppose that Assumptions 3.4.1 and 3.4.3 hold. Then, the following holds as $n \to \infty$,

$$\sup_{\theta} \sup_{c} \left| n^{1/2} \{ \bar{F}_n^{w,m,g}(\theta, c) - \bar{F}_n^{w,m}(0, c) \} - \mathcal{B}_n^{w,m}(\theta, c) \right| = o_p(1), \tag{3.24}$$

where $\mathcal{B}_n^{w,m,g}(\theta, c)$ is defined in equation (3.18).

Theorem 3.4.5 shows that the linearisation of the compensator term derived by Berenguer-Rico and Nielsen (2017) in the linear case can also be applied to the compensator term based on residuals from the non-linear model. The non-linearity of the model, as captured by $\ddot{g}(x_i, \beta)$, does not affect the compensator asymptotically.

Importantly, terms $\mathbb{F}_n^{w,m}(0, c)$ and $\bar{F}_n^{w,m}(0, c)$ have no estimation errors and are, therefore, based on the true innovations ε_i/σ:

$$F_n^{w,m}(0, c) = n^{-1} \sum_{i=1}^{n} w_{in} m \left(\frac{\varepsilon_i}{\sigma} \right) 1_{(\varepsilon_i \leq \sigma c)} \tag{3.25}$$

and

$$\bar{F}_n^{w,m}(0, c) = n^{-1} \sum_{i=1}^{n} w_{in} \mathsf{E}_{i-1} \left\{ m \left(\frac{\varepsilon_i}{\sigma} \right) 1_{(\varepsilon_i \leq \sigma c)} \right\}, \tag{3.26}$$

which results in the following empirical process:

$$\mathbb{F}_n^{w,m}(0, c) = n^{1/2} \left\{ F_n^{w,m}(0, c) - \bar{F}_n^{w,m}(0, c) \right\}. \tag{3.27}$$

Since they are based on i.i.d. innovations, the properties of $F_n^{w,m}(0, c)$, $\bar{F}_n^{w,m}(0, c)$ and $\mathbb{F}_n^{w,m}(0, c)$ can be analysed using existing theory.

3.5 Verification of Assumptions: Example

As detailed in Section 3.2, all assumptions of Theorems 3.4.1, 3.4.2 and 3.4.3 can be verified. In this section I illustrate the process of verification through an example. I exclude Assumption 3.4.3 from this example because this assumption merely re-states Assumptions 2.1 and 2.2 of Berenguer-Rico and Nielsen (2017), which are discussed at length therein, including examples of models to which they apply.

3.5.1 Example: Wu (1981) model

This section focuses on the Wu (1981) power curve model. The model is used in Sections 3.5.1.1 and 3.5.1.2, to illustrate through an example how the assumptions underpinning Theorems 3.4.1, 3.4.2 and 3.4.3 can be verified.

The model can be stated as

$$y_i = (x_i + \beta)^d + \varepsilon_i. \tag{3.28}$$

In what follows I assume that $d = 3$, which is the simplest case which allows for a non-trivial second derivative.

For simplicity, I assume that x_i is stationary, whereas Wu (1981) analyses $x_i = t$ where t is a linear time trend. Additionally, I assume that $\mathsf{E}(x_i^2)^{2^r+2} < \infty$, where r is as defined in Assumption 3.4.2. In particular, r must satisfy the inequality: $2^{r-1} > 1 + (1/4 - \eta)(\dim b_1)$. In the example considered in this section, the parameter β is a scalar and, therefore, $\dim b_1 = 1$. Thus, in this example it is sufficient that $r = 2$.

For the purposes of this example, I also set $w_{in} = 1$ for all i, which corresponds to all observations being equally weighted. I assume that $m(u) = u^p$ for $p = 3, 4$, which is in line with components of tests for normality, which use estimates of the third and fourth moments. The density of errors, $f(\varepsilon_i/\sigma)$, is assumed to be a standard normal density.

Since x_i is stationary, $N = n^{-1/2}$. This gives the following normalised first derivative, where β_0 is the true value of the parameter vector β,

$$N'\dot{g}(x_i, \beta_0) = 3n^{-1/2}(x_i + \beta_0)^2. \tag{3.29}$$

For any $y, z \in \mathbb{R}$, the following inequality holds: $(y + z)^2 \leq 2y^2 + 2z^2$. Therefore, the normalised

first derivative can be bounded as follows, where $C < \infty$ is a finite constant,

$$\|N'\dot{g}(x_i, \beta_0)\| \leq Cn^{-1/2}(x_i^2 + \beta_0^2). \tag{3.30}$$

Assumptions on the normalised second derivative are imposed on its supremum over β_i such that $\|N^{-1}(\beta_i - \beta_0)\| \leq Bn^{1/4-\eta}$. Given the normalisation by $N = n^{-1/2}$ in the stationary case, the normalised second derivative can be expressed as:

$$N'\ddot{g}(x_i, \beta_i)N = n^{-1}6(x_i + \beta_i). \tag{3.31}$$

In order to account for the supremum, one can further decompose the normalised derivative into a function of the constant β_0 and the deviation of β_i from β_0, which is bounded. Further, use the triangle inequality to bound the absolute value of a sum with a sum of absolute values, which results in the bound

$$\sup_{\beta_i}\|N'\ddot{g}(x_i, \beta_i)N\| \leq Dn^{-1}|x_i| + D\sup_{\beta_i} n^{-1}NN^{-1}\|\beta_i - \beta_0\| + Dn^{-1}NN^{-1}|\beta_0|, \tag{3.32}$$

where $D < \infty$ is a finite constant. Note that the supremum with respect to β_i is taken over the set $\{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq Bn^{1/4-\eta}\}$. Further, as noted above, $N = n^{-1/2}$. Thus, incorporating B into the constant D gives the following bound on the normalised second derivative:

$$\sup_{\beta_i}\|N'\ddot{g}(x_i, \beta_i)N\| \leq Dn^{-1}|x_i| + Dn^{-5/4-\eta} + Dn^{-1}|\beta_0|. \tag{3.33}$$

3.5.1.1 Verifying Assumption 3.4.1

Assumption 1a.

The aim is to show that $\max_i\|N'\dot{g}(x_i; \beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$. Following equation (3.30), $\|N'\dot{g}(x_i, \beta_0)\| \leq Cn^{-1/2}(x_i^2 + \beta_0^2)$. Since x_i has a finite second moment, it follows from Markov's inequality that, for any ϵ one can find an M such that:

$$P\left(\frac{1}{n^{-1/4+\eta-\kappa}} \cdot \frac{x_i^2}{n^{1/2}} > M\right) \leq \frac{\mathsf{E}(x_i^2)}{Mn^{1/4+\eta-\kappa}} < \epsilon, \tag{3.34}$$

since $\mathsf{E}(x_i^2) < \infty$ and κ can be any positive number, hence one can choose $\kappa < \eta$. Thus, the

assumption is satisfied.

Assumption 1b.

The aim is to show that $\max_i \sup_{\beta_i} \|N'\ddot{g}(x_i;\beta_i)N\| = O_p(n^{-1/2+2\eta-\kappa})$. Following equation (3.33), $\sup_{\beta_i} \|N'\ddot{g}(x_i,\beta_i)N\| \leq Dn^{-1}|x_i| + Dn^{-5/4-\eta} + Dn^{-1}|\beta_0|$. Analogously to the proof for Assumption 1a, the premise follows from the Markov inequality since $\mathsf{E}|x_i| < \infty$.

Assumption 1c.

The aim is to show that: $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{2\eta})$. With $w_{in} = 1$ for all i and given inequality (3.33), the term can be expressed as

$$\begin{aligned}\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N'\ddot{g}(x_i;\beta_i)N\| &\leq \sum_{i=1}^n \left(Dn^{-1}|x_i| + Dn^{-5/4-\eta} + Dn^{-1}|\beta_0|\right) \\ &= Dn^{-1}\sum_{i=1}^n |x_i| + Dn^{-1/4-\eta} + D|\beta_0|.\end{aligned} \tag{3.35}$$

Since x_i is assumed to be stationary, the requirement that the sum be $o_p(n^{2\eta})$ is satisfied.

Assumption 1d.

The aim is to show that $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N'\ddot{g}(x_i;\beta_i)N\|^2 = o_p(n^{-1/2+4\eta})$. Given $w_{in} = 1$ and due to inequality (3.33), one can bound the term as follows:

$$\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N'\ddot{g}(x_i;\beta_i)N\|^2 \leq \sum_{i=1}^n \left(Dn^{-1}|x_i| + Dn^{-5/4-\eta} + Dn^{-1}|\beta_0|\right)^2. \tag{3.36}$$

For any $x, y, z \in \mathbb{R}$ the following inequality holds: $(x+y+z)^2 \leq 4(x^2+y^2+z^2)$. Incorporating the constant 4 into D^2 gives the bound:

$$\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N'\ddot{g}(x_i;\beta_i)N\|^2 \leq n^{-2}D^2 \sum_{i=1} x_i^2 + D^2 n^{-3/2-2\eta} + D^2 n^{-1}\beta_0^2. \tag{3.37}$$

Since x_i is assumed to be stationary, all terms are at most $O_p(n^{-1})$ and, therefore, the assumption is satisfied.

Assumption 1e.

The aim is to show that $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N'\dot{g}(x_i;\beta_0)\| \|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{-1/4+3\eta})$. Since

$w_{in} = 1$ then given inequalities (3.30) and (3.33), the term can be bounded as follows:

$$\sum_{i=1}^{n} |w_{in}| \sup_{\beta_i} \|N'\dot{g}(x_i;\beta_0)\| \|N'\ddot{g}(x_i;\beta_i)N\| \leq \sum_{i=1}^{n} Cn^{-1/2}(x_i^2 + \beta_0^2) \left(Dn^{-1}|x_i| + Dn^{-5/4-\eta} + Dn^{-1}|\beta_0| \right) \tag{3.38}$$

Multiplying out the parentheses gives the following sum, where D has been redefined to incorporate the constants C and β_0, i.e. $D = \max(CD, CD\beta_0^2, CD\beta_0^3)$,

$$Dn^{-3/2} \sum_{i=1}^{n} |x_i|^3 + Dn^{-7/4-\eta} \sum_{i=1}^{n} x_i^2 + Dn^{-3/2} \sum_{i=1}^{n} x_i^2 + Dn^{-3/2} \sum_{i=1}^{n} |x_i| + Dn^{-3/4-\eta} + Dn^{-1/2}. \tag{3.39}$$

Since $|x_i|$ has a finite third moment, the terms are at most $O_p(n^{-1/2})$ and the assumption is satisfied.

Assumption 1f.

The aim is to show that $\sum_{i=1}^{n} |w_{in}| \sup_{\beta_i} \|N'\dot{g}(x_i;\beta_0)\|^2 \|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{-1/2+4\eta})$. Since $w_{in} = 1$ then given inequalities (3.30) and (3.33), the expression is bounded by

$$\sum_{i=1}^{n} |w_{in}| \sup_{\beta_i} \|N'\dot{g}(x_i;\beta_0)\|^2 \|N'\ddot{g}(x_i;\beta_i)N\| \leq \sum_{i=1}^{n} \left(Cn^{-1/2}(x_i^2 + \beta_0^2) \right)^2 \left(Dn^{-1}|x_i| + Dn^{-5/4-\eta} + Dn^{-1}|\beta_0| \right). \tag{3.40}$$

In line with Section 3.5.1, note that for any $y, z \in \mathbb{R}$, the following inequality holds: $(y+z)^2 \leq 2y^2 + 2z^2$. Incorporating $2, C$ and β_0 into the constant D gives the bound:

$$Dn^{-2} \sum_{i=1}^{n} |x_i|^5 + Dn^{-9/4-\eta} \sum_{i=1}^{n} x_i^4 + Dn^{-2} \sum_{i=1}^{n} x_i^4 + Dn^{-2} \sum_{i=1}^{n} |x_i| + Dn^{-5/4-\eta} + Dn^{-1}. \tag{3.41}$$

Since the 1st, 4th and 5th moments of $|x_i|$ exist, all terms are at most $O_p(n^{-1})$. Therefore, the assumption is satisfied.

Assumption 2.

Assumption 2 is satisfied in the example analysed here in which the errors follow a standard normal distribution and the mark function is a power function. In the case of Assumptions 2a and 2b this follows from the fact that $m(u) = u^p$ increases to infinity at a polynomial rate while $f(u)$ decreases to zero at an exponential rate, which is faster. Assumption 2c is satisfied since all moments of the standard normal distribution are finite. The proof for the Lipschitz condition (Assumption 2d) is as follows.

Note that for any $x, y \in \mathbb{R}$, the following two results hold. Firstly, use the formula for the

difference in powers and apply the triangle inequality to obtain

$$\left|x^k - y^k\right| = \left|(x-y)\sum_{j=0}^{k-1} x^{k-j-1}y^j\right| \leq |x-y|\sum_{j=0}^{k-1}|x|^{k-j-1}|y|^j. \tag{3.42}$$

Secondly, note that for any $x, y \in \mathbb{R}$, the following holds:

$$|x| \leq |y| + |x-y|\,. \tag{3.43}$$

In order to prove Assumption 2d one needs to show that if $|u^* - u| \leq (1+|u|)\omega$ for some known ω, then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u|\mathring{m}(u)$, where $\mathring{m}(u)$ is a nonnegative function which satisfies the condition: $\sup_u(1+|u|+u^2)\mathring{m}(u)f(u) < \infty$.

The derivative of $m(u) = u^p$ is $\dot{m}(u) = pu^{p-1}$. To use results (3.42) and (3.43), set $x = u^*$, $y = u$ and $k = p-1$, which gives the bound

$$|\dot{m}(u^*) - \dot{m}(u)| = p\left|(u^*)^{p-1} - u^{p-1}\right| \leq p\,|u^* - u|\sum_{j=0}^{p-2}|u^*|^{p-2-j}|u|^j. \tag{3.44}$$

Given the premise of the Lipshitz condition, bound $|u^*-u|$ by $(1+|u|)\omega$ and use equation (3.43) to bound the term $|u^*|$ as follows: $|u^*| \leq |u| + |u^* - u| \leq \omega + (1+\omega)|u|$. This leads to the bound

$$p\,|u^* - u|\sum_{j=0}^{p-2}|u^*|^{p-2-j}|u|^j \leq p(1+|u|)\omega\sum_{j=0}^{p-2}[\omega + (1+\omega)|u|]^{p-2-j}|u|^j =: \mathring{m}(u). \tag{3.45}$$

Thus, $|\dot{m}(u^*)-\dot{m}(u)| \leq |u^*-u|\mathring{m}(u)$. The function $\mathring{m}(u)$ is nonnegative since it is based on powers of the absolute value of u. It also satisfies the Lipschitz condition $\sup_u(1+|u|+u^2)\mathring{m}(u)f(u) < \infty$ since $\mathring{m}(u)$ increases polynomially in u while $f(u)$ decreases to zero at an exponential rate.

Summary

This example shows that all parts of Assumption 3.4.1 can be verified.

3.5.1.2 Verifying Assumption 3.4.2

Analogously to Section 3.5.1.1, in the current section I verify Assumption 3.4.2 of Theorem 3.4.2. Note that Assumptions 3.4.2/2a, 2b and 3 are verified in Section 3.5.1.1 since they overlap with Assumption 3.4.1. Assumptions 2f, 2g, 2h must hold for $\tilde{v}_{in}$ which represents each of:

$w_{in}, w_{in}n^{-1/4-\eta}, w_{in}n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|$. However, since all three assumptions are stated in the form of an asymptotic upper bound, if the assumptions hold for $\tilde{v}_{in} = w_{in}$ then this implies that they also hold for $\tilde{v}_{in} = w_{in}n^{-1/4-\eta}$ and, therefore, assumptions on the latter need not be verified.

Assumption 1a

The aim is to show that $\mathsf{E}\sum_{i=1}^{n}\left(1+|\tilde{v}_{in}|^{2^r}\right) = O(n)$, where $\tilde{v}_{in}$ represents each of: w_{in} and $w_{in}n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|$.

For $\tilde{v}_{in} = w_{in}$, since $w_{in} = 1$ for the purposes of this derivation, it immediately follows that $\mathsf{E}\sum_{i=1}^{n}\left(1+|1|^{2^r}\right) = 2n = O(n)$, as required.

For $\tilde{v}_{in} = w_{in}n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|\|$, due to equation (3.30), the expectation can be expressed as

$$\mathsf{E}\sum_{i=1}^{n}\left(1+|\tilde{v}_{in}|^{2^r}\right) \leq \mathsf{E}\sum_{i=1}^{n}\left(1+\left(n^{1/4-\eta}Cn^{-1/2}(x_i^2+\beta_0^2)\right)^{2^r}\right). \tag{3.46}$$

The above can be simplified resulting in the bound

$$\mathsf{E}\sum_{i=1}^{n}\left(1+|\tilde{v}_{in}|^{2^r}\right) \leq n+\mathsf{E}\sum_{i=1}^{n}\left(Cn^{-1/4-\eta}(x_i^2+\beta_0^2)\right)^{2^r}. \tag{3.47}$$

The first term is trivially $O(n)$. In the second term, analogously to the arguments in Section 3.5.1.1, the dominant terms in the power of the sum $(x_i^2+\beta_0^2)^{2^r}$ are $x_i^{2^{r+1}}$ and $\beta_0^{2^{r+1}}$. Incorporate the lower-order terms into the constant C. Thus:

$$\mathsf{E}\sum_{i=1}^{n}\left(Cn^{-1/4-\eta}(x_i^2+\beta_0^2)\right)^{2^r} \leq Cn^{(-1/4-\eta)2^r}\mathsf{E}\sum_{i=1}^{n}x_i^{2^{r+1}} + Cn^{(-1/4-\eta)2^r}\mathsf{E}\sum_{i=1}^{n}\beta_0^{2^{r+1}}. \tag{3.48}$$

If x_i has a finite 2^{r+1}'th moment, then both quantities are $O(n^{1-(1/4+\eta)2^r})$. Thus, Assumption 1a is satisfied.

Assumption 2c

Under the restriction that $w_{in} = 1$, this assumption is equivalent to the following, since $1^2 = 1$:

$$\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|^2 = o_p(n^{4\eta-1}). \tag{3.49}$$

In Section 3.5.1.1 it was shown that:

$$\sum_{i=1}^{n} |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\|^2 \leq n^{-2} D^2 \sum_{i=1} x_i^2 + D^2 n^{-3/2-2\eta} + D^2 n^{-1} \beta_0^2. \tag{3.50}$$

Since x_i is assumed to be stationary, all terms are at most $O_p(n^{-1})$ and, therefore, Assumption 3.4.2/2c is satisfied.

Assumption 2d

The aim is to verify that

$$\mathsf{E} \sum_{i=1}^{n} |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| \|N' \dot{g}(x_i; \beta_0)\| = O(n^{-1/4+2\eta}). \tag{3.51}$$

Result (3.38) implies the following bound, where the constant D has been incorporated into the constant C:

$$\mathsf{E} \sum_{i=1}^{n} |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| \|N' \dot{g}(x_i; \beta_0)\| \leq \mathsf{E} \sum_{i=1}^{n} C n^{-1/2} (x_i^2 + \beta_0^2) \left(n^{-1} |x_i| + n^{-5/4-\eta} + n^{-1} |\beta_0| \right). \tag{3.52}$$

After multiplying out the terms, the bound can be expressed as

$$\begin{aligned} &\mathsf{E} \sum_{i=1}^{n} C n^{-3/2} |x_i|^3 + \mathsf{E} \sum_{i=1}^{n} C n^{-7/4-\eta} x_i^2 + \mathsf{E} \sum_{i=1}^{n} C n^{-3/2} x_i^2 |\beta_0| + \\ &+ \mathsf{E} \sum_{i=1}^{n} C n^{-3/2} \beta_0^2 |x_i| + \mathsf{E} \sum_{i=1}^{n} C n^{-7/4-\eta} \beta_0^2 + \mathsf{E} \sum_{i=1}^{n} C n^{-3/2} |\beta_0|^3. \end{aligned} \tag{3.53}$$

Under the assumption that $\mathsf{E}|x_i|^3 < \infty$, the first, third, fourth and sixth terms are all $O(n^{-1/2})$, while the second and fifth terms are $O(n^{-3/4-\eta})$. Since the assumption only requires that the terms be $O(n^{-1/4+2\eta})$, the assumption is satisfied.

Assumption 2e

The aim is to verify that

$$\mathsf{E} \sum_{i=1}^{n} |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\|^2 = O(n^{-1/2+3\eta}). \tag{3.54}$$

Due to result (3.37) the expression can be bounded by

$$\mathsf{E}\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|^2 \leq n^{-2}D^2\mathsf{E}\sum_{i=1}x_i^2 + D^2n^{-3/2-2\eta} + D^2n^{-1}\beta_0^2 \tag{3.55}$$

Under the assumption that $\mathsf{E}x_i^2 < \infty$, the first term is $O(n^{-1})$, the second term is $O(n^{-3/2-\eta})$ and the last term is $O(n^{-1})$. Thus the requirement that all terms be $O(n^{-1/2+3\eta})$ is satisfied.

Assumption 2f

Let $\tilde{v}_{in} = w_{in}$. Furthermore, in this example $w_{in} = 1$. Thus, the assumption to be verified can be stated as

$$2\mathsf{E}\sum_{i=1}^{n}\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\| = O(n^{2\eta-\kappa}). \tag{3.56}$$

Substituting equation (3.33) and incorporating the constant 2 into the constant D gives

$$2\mathsf{E}\sum_{i=1}^{n}\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\| \leq D\mathsf{E}\sum_{i=1}^{n}n^{-1}|x_i| + D\mathsf{E}\sum_{i=1}^{n}n^{-5/4-\eta} + D\mathsf{E}\sum_{i=1}^{n}n^{-1}|\beta_0|. \tag{3.57}$$

Under the assumption that $\mathsf{E}|x_i| < \infty$, the first term is $O(1)$. The second term is $O(n^{-1/4-\eta})$. The final term is $O(1)$. The assumption requires only that the terms be $O(n^{2\eta-\kappa})$. Thus, the assumption is satisfied.

For the second part of the proof, let $\tilde{v}_{in} = w_{in}n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|$ which leads to assumption 2f being stated as

$$\mathsf{E}\sum_{i=1}^{n}\left(1+\left(n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|\right)^{2^r}\right)\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\| = O(n^{2\eta-\kappa}), \tag{3.58}$$

for some $\kappa > 0$. Given the derivations in equations (3.56) and (3.57), it suffices to consider the following term, where equation (3.33) has been used to bound $\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|$ and equation (3.30) has been used to bound $\|N'\dot{g}(x_i;\beta_0)\|$,

$$\mathsf{E}\sum_{i=1}^{n}|\tilde{v}_{in}|^{2^r}\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\| \leq \mathsf{E}\sum_{i=1}^{n}\left(n^{1/4-\eta}Cn^{-1/2}(x_i^2+\beta_0^2)\right)^{2^r}\left(Dn^{-1}|x_i| + Dn^{-5/4-\eta} + Dn^{-1}|\beta_0|\right). \tag{3.59}$$

Analogously to the steps involved in verifying Assumption 1a, note that in the expression $(x_i^2+\beta_0^2)^{2^r}$, the dominant terms are $x_i^{2^{r+1}}$ and $\beta_0^{2^{r+1}}$. The constant C as well as the remaining

terms are incorporated into the constant D. Collecting the terms gives the following sum.

$$Dn^{(-1/4-\eta)2^r}\mathsf{E}\sum_{i=1}^{n}x_i^{2^{r+1}}n^{-1}|x_i|+n^{(-1/4-\eta)2^r}\mathsf{E}\sum_{i=1}^{n}x_i^{2^{r+1}}n^{-5/4-\eta}+n^{(-1/4-\eta)2^r}\mathsf{E}\sum_{i=1}^{n}x_i^{2^{r+1}}n^{-1}|\beta_0|+$$
$$+Dn^{(-1/4-\eta)2^r}\mathsf{E}\sum_{i=1}^{n}\beta_0^{2^{r+1}}n^{-1}|x_i|+n^{(-1/4-\eta)2^r}\mathsf{E}\sum_{i=1}^{n}\beta_0^{2^{r+1}}n^{-5/4-\eta}+n^{(-1/4-\eta)2^r}\mathsf{E}\sum_{i=1}^{n}\beta_0^{2^{r+1}}n^{-1}|\beta_0| \tag{3.60}$$

Under the assumption that $\mathsf{E}|x_i|^{2^{r+1}+1}<\infty$, the first, third, fourth and sixth terms are $O(n^{(-1/4-\eta)2^r})$; while the second and fifth terms are $O(n^{(-1/4-\eta)(2^r+1)})$. Thus, the requirement of the expression being $O(n^{2\eta-\kappa})$ is satisfied for example for $0<\kappa<\eta$.

Assumption 2g

Let $\tilde{v}_{in}=w_{in}=1$. The aim is to prove that

$$\mathsf{E}\sum_{i=1}^{n}\left(1+|\tilde{v}_{in}|^{2^r}\right)\left|N'\dot{g}(x_i;\beta_0)\right|=O(n^{1/2}). \tag{3.61}$$

Substitute $\tilde{v}_{in}=w_{in}=1$ and substitute result (3.30). Incorporating the constant 2 into the constant C gives the following bound:

$$\mathsf{E}\sum_{i=1}^{n}\left(1+|1|^{2^r}\right)\left|N'\dot{g}(x_i;\beta_0)\right|\leq Cn^{-1/2}\mathsf{E}\sum_{i=1}^{n}(x_i^2+\beta_0^2). \tag{3.62}$$

Under the assumption that $\mathsf{E}x_i^2<\infty$, the above term is $O(n^{1/2})$, which satisfies the assumption.

Now let $\tilde{v}_{in}=w_{in}n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|$. Thus, the aim is to show the following:

$$\mathsf{E}\sum_{i=1}^{n}\left[1+(n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|)^{2^r}\right]\|N'\dot{g}(x_i;\beta_0)\|=O(n^{1/2}). \tag{3.63}$$

In view of result (3.62) it suffices to show that

$$\mathsf{E}\sum_{i=1}^{n}(n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|)^{2^r}\|N'\dot{g}(x_i;\beta_0)\|=O(n^{1/2}). \tag{3.64}$$

Substituting result (3.30) gives the following bound, where the constant C incorporates C^{2^r}:

$$\mathsf{E}\sum_{i=1}^{n}(n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|)^{2^r}\|N'\dot{g}(x_i;\beta_0)\|\leq C\mathsf{E}\sum_{i=1}^{n}(n^{1/4-\eta}n^{-1/2}(x_i^2+\beta_0^2))^{2^r}n^{-1/2}(x_i^2+\beta_0^2) \tag{3.65}$$

The bound can be simplified as follows.

$$\mathsf{E}\sum_{i=1}^{n}(n^{1/4-\eta}|N'\dot{g}(x_i;\beta_0)|)^{2^r}\left|N'\dot{g}(x_i;\beta_0)\right| \leq Cn^{(-1/4-\eta)2^r-1/2}\mathsf{E}\sum_{i=1}^{n}(x_i^2+\beta_0^2)^{2^r+1} \tag{3.66}$$

As above, note that the dominant terms in the expression $(x_i^2+\beta_0^2)^{2^r+1}$ are $(x_i^2)^{2^r+1}$ and $(\beta_0^2)^{2^r+1}$, the latter of which is a constant. Thus, provided that $\mathsf{E}(x_i^2)^{2^r+1} < \infty$, the expression in (3.66) is $O(n^{1/2-(1/4+\eta)2^r})$, which satisfies the requirement of it being $O(n^{1/2})$.

Assumption 2h

Let $\tilde{v}_{in} = w_{in} = 1$. The aim is to show that

$$\mathsf{E}\sum_{i=1}^{n}\left(1+|\tilde{v}_{in}|^{2^r}\right)\left|N'\dot{g}(x_i;\beta_0)\right|^2 = O(n^{\eta}). \tag{3.67}$$

Substitute $\tilde{v}_{in} = w_{in} = 1$ and substitute result (3.30). Incorporating the constants 2 and C^2 into the constant C gives

$$\mathsf{E}\sum_{i=1}^{n}\left(1+|1|^{2^r}\right)\left|N'\dot{g}(x_i;\beta_0)\right|^2 \leq C\mathsf{E}\sum_{i=1}^{n}n^{-1}(x_i^2+\beta_0^2)^2. \tag{3.68}$$

The dominant terms in expression $(x_i^2+\beta_0^2)^2$ are x_i^4 and β_0^4, the latter of which is a constant. Thus, since $\mathsf{E}x_i^4 < \infty$, the expression in (3.68) is $O(1)$, which satisfies the requirement that it be $O(n^{\eta})$.

Now let $\tilde{v}_{in} = w_{in}n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|$. The aim is to show the following:

$$\mathsf{E}\sum_{i=1}^{n}\left(1+(n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|)^{2^r}\right)\|N'\dot{g}(x_i;\beta_0)\|^2 = O(n^{\eta}). \tag{3.69}$$

In view of result (3.68), it suffices to show that

$$\mathsf{E}\sum_{i=1}^{n}(n^{1/4-\eta}|N'\dot{g}(x_i;\beta_0)|)^{2^r}\left|N'\dot{g}(x_i;\beta_0)\right|^2 = O(n^{\eta}). \tag{3.70}$$

Substitute result (3.30) and incorporate C^{2^r+1} into the constant C to obtain

$$\mathsf{E}\sum_{i=1}^{n}(n^{1/4-\eta}\|N'\dot{g}(x_i;\beta_0)\|)^{2^r}\|N'\dot{g}(x_i;\beta_0)\|^2 \leq Cn^{-1}\mathsf{E}\sum_{i=1}^{n}n^{(-1/4-\eta)2^r}(x_i^2+\beta_0^2)^{2^r+2} \tag{3.71}$$

The dominant terms in the expansion of $(x_i^2 + \beta_0^2)^{2^r+2}$ are $(x_i^2)^{2^r+2}$ and $(\beta_0)^{2^r+2}$, the latter of which is a finite constant. Thus, under the assumption that $\mathsf{E}(x_i^2)^{2^r+2} < \infty$, the bound in equation (3.71) is $O(n^{(-1/4-\eta)2^r})$, which satisfies the assumption of it being $O(n^\eta)$.

Assumption 4

To verify Assumption 4, one needs to show the following, where $\tilde{h}(u)$ stands for each of: $m(u)$, $\dot{m}(u)$ and $u\dot{m}(u)$,

$$\mathsf{E}\left|\tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right)\right|^{2^r} < \infty. \tag{3.72}$$

The assumption is satisfied for both $m(u) = u^3$ and $m(u) = u^4$ since in this example ε_i/σ is distributed standard normal, all of which moments are finite.

Assumption 5

All parts of Assumption 5 are satisfied due to the assumption that u is distributed standard normal and $\tilde{h}(u)$ represents the second, third and fourth powers of u. Assumption 5a is satisfied due to the support of the standard normal distribution being the real line. Assumptions 5b and 5d are satisfied due to the fact that $\tilde{h}(u)$ increases to infinity at a polynomial rate, due to $m(u)$ being a power of u, while $f(u)$ decreases to zero at an exponential rate. Assumption 5c follows from the finiteness of the moments of the normal distribution. To verify assumption 5c, note that $\dot{\tilde{H}}_r(c) = (1 + h(c)^{2^r})f(c)$, which is bounded from below by the unimodal standard normal density $f(\cdot)$ and is bounded from above by $C_u f(c)$ where the constant C_u depends on the moment since $C_u = 1 + h(c^\dagger)^{2^r}$, where $c^\dagger := \arg\max h(c)^{2^r} f(c)$.

Summary

This example shows that all parts of Assumption 3.4.2 can be verified.

3.5.1.3 Wu (1981) Example: Summary

Sections 3.5.1.1 and 3.5.1.2 show how assumptions underlying Theorems 3.4.1, 3.4.2 and 3.4.3 can be verified, thus proving that the theorems do not rely on unverifiable assumptions. The section also shows that the assumptions are satisfied for the example non-linear model of Wu (1981) with stationary regressors which have 12 finite moments, normally distributed errors, and where the mark function of interest is a power function, which can be used e.g. in moments-based test for normality.

The verification of assumptions in Sections 3.5.1.1 and 3.5.1.2 assumed that the single regres-

sor was stationary. Therefore, the normalisation matrix simplified to the constant $N = n^{-1/2}$. In the more general case of k stationary regressors, the normalisation matrix would be a diagonal matrix $N = n^{-1/2}I_k$. In the case of non-stationary regressors, the normalisation matrix N would need to be constructed so as to ensure that $\sum_{i=1}^{n} \dot{g}_{in}^{0\prime}\dot{g}_{in}^{0}$, where $\dot{g}_{in}^{0} := N'\dot{g}(x_i, \beta_0)$, has a positive limit. For example, in a linear regression model with random walk regressors the normalisation matrix would be specified as $N = n^{-1}I_k$. In a linear regression model with a constant and a linear time trend, the normalisation matrix would be $N = diag(n^{-1/2}, n^{-3/2})$. Thus, in order to apply Theorems 3.4.1, 3.4.2 and 3.4.3 in the presence of non-stationary regressors, assumptions underpinning the theorems need to be verified analogously to Sections 3.5.1.1 and 3.5.1.2, for the appropriate definition of the normalisation matrix N.

3.6 Applications

Since Theorems 3.4.1, 3.4.2 and 3.4.3 focus on the asymptotic properties of the empirical distribution, compensator, and empirical process of marked and weighted residuals, the results have a variety of applications. As detailed above, the results have the advantage of not requiring that the regressors be stationary, and they are based on assumptions which can be verified.

The Kolmogorov-Smirnov (K-S) test can be implemented within this framework by setting both the weights w_{in} and the marks $m(\cdot)$ equal to 1, which allows for an application of the K-S test to residuals from non-linear regressions. Unlike Khmaladze and Koul (2004), who also analyse the K-S test in non-linear regressions, I do not require that the estimator $\hat{\beta}$ be $n^{1/2}$-consistent but instead I allow the scaled errors $\tilde{b} = N^{-1}(\tilde{\beta} - \beta)$ to diverge slightly since $|\tilde{b}| = o_p(n^{1/4-\eta})$.

Another application of the theory presented in this chapter is to moments-based tests for normality such as the Jarque-Bera (1987) test based on skewness and kurtosis, whose application to robust regressions was analysed by Onder and Zaman (2003) in the context of linear regressions. Our theory can be used to derive the properties of the test when applied to residuals from non-linear regressions with outlier detection.

Furthermore, Berenguer-Rico and Wilms (2021) derive the asymptotic distribution of the White (1980) test for heteroskedasticity when applied to linear regressions with outlier detection, by applying the results of Berenguer-Rico et. al. (2019a, 2019b), who analyse marked and

weighted empirical processes of residuals from linear regressions with outlier detection in which the mark function takes the form ε_i^p. Results presented in this chapter can similarly be used to derive the asymptotic properties of the White (1980) test of heteroskedasticity in non-linear regressions with outlier detection.

Finally, Johansen and Nielsen (2013, 2016b), Jiao and Nielsen (2017) and Berenguer-Rico et.al. (2019b) employ marked and weighted empirical processes of residuals to analyse the properties of the Huber-skip estimator in the linear model. Results presented here can be used to analyse the properties of the Huber-skip estimator when applied to non-linear least squares.

3.7 Conclusion

In this chapter I develop asymptotic theory for marked and weighted empirical processes of residuals from non-linear regressions with outlier detection. I show that the assumptions of the theorems do not rely on unverifiable assumptions but rather, that the assumptions can be checked, which I illustrate with an example. In proving the results, I also prove a number of supplementary lemmas, which are detailed in the following appendix, including new lemmas as well as extending to non-linear models selected lemmas applicable to linear models. I suggest applications of the results to the Kolmogorov-Smirnov test, the Jarque-Bera test, the White test for heteroskedasticity, and analysing the properties of the Huber-skip in non-linear least squares; although their implementations are left for future research.

3.A Appendix 3.A: Details and Proofs of Lemmas

3.A.1 Structure of the Appendix

Appendix 3.A includes the details of all lemmas used in proving theorems presented in this chapter. Subsection 3.A.2 details the shorthand notation used throughout both Appendices 3.A and 3.B. Subsection 3.A.3 presents several smaller results including rescaling and simpler lemmas which are used throughout the derivations. Subsection 3.A.4 outlines each of the results proven in the lemmas in Appendix 3.A. Finally, subsection 3.A.5 includes the assumptions and details of each lemma, as well as their proofs.

3.A.2 Notation and Definitions

Throughout the appendix, the following shorthand will be used, to simplify notation:

$$\begin{aligned}
\max_i &:= \max_{1\le i\le n}, \\
\max_p &:= \max_{1\le p\le P}, \\
\max_k &:= \max_{1\le k\le K}, \\
\sup_u &:= \sup_{u\in\mathbb{R}}, \\
a_n^s &:= n^{-1/2}a_m/\sigma, \\
c_{in}^0 &:= \frac{\sigma c + n^{-1/2}a_1 c + b_1'\dot{g}_{in}^0}{\sigma}, \\
d_{in}^* &:= \frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}.
\end{aligned} \tag{3.73}$$

The scale s is defined in Definition 3.A.1.

Definition 3.A.1. Scale s is defined as $s := 1 + n^{-1/2}a_m/\sigma$. In particular, the following quantities, which are used throughout the derivations, can be expressed using s:

$$\begin{aligned}
&\frac{a_m}{1+n^{-1/2}a_m/\sigma} = \frac{1}{s}a_m, \\
&\frac{\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{1+n^{-1/2}a_m/\sigma} = \frac{1}{s}\frac{\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma}, \\
&\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m} = \frac{\varepsilon_i}{\sigma}\left(1 - n^{-1/2}\frac{1}{s}a_m/\sigma\right) - \frac{1}{s}\frac{b_m'\dot{g}_{in}^0 + \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma}.
\end{aligned} \tag{3.74}$$

3.A.3 Preliminary Lemmas

Assumption 3.A.1. The error a_m is $O(n^{1/4-\eta})$.

Lemma 3.A.1. Under Assumption 3.A.1, for large n:

$$\sup_{\theta} \left| \frac{1}{s} \right| < 2. \tag{3.75}$$

Proof. Since $|a_m| = O(n^{1/4-\eta})$ then, $n^{-1/2}a_m/\sigma = O(n^{-1/4-\eta})$ and, therefore, there exists $\bar{N}$ such that for all $n \geq \bar{N}$ it holds that $n^{-1/2}a_m/\sigma \in (-\frac{1}{2}, \frac{1}{2})$. Consequently, $s = 1 + n^{-1/2}a_m/\sigma \in (\frac{1}{2}, \frac{3}{2})$ and therefore $0 < \frac{1}{s} < 2$. ■

Remark 3.A.1. In particular, Lemma 3.A.1 leads to the following bounds:

$$\begin{aligned} &\left| \frac{1}{s} a_m \right| < 2 \left| a_m \right|, \\ &\left| \frac{1}{s} b_m' \dot{g}_{in}^0 \right| < 2 \left| b_m' \dot{g}_{in}^0 \right|, \\ &\left| \frac{1}{s} \frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m \right| < 2 \left| \frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m \right|. \end{aligned} \tag{3.76}$$

Assumption 3.A.2. Let:

1. $\max_{1 \leq i \leq n} \|N' \dot{g}(x_i; \beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$,
2. $\max_{1 \leq i \leq n} \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = O_p(n^{-1/2+2\eta-\kappa})$,

for some $\kappa > 0$.

Definition 3.A.2. The following sets are defined, for $0 < \eta < 1/4$ and $\kappa > 0$ and for a $B > 0$:

$$\begin{aligned} S_{1,n} &:= (\max_i \|N' \dot{g}(x_i, \beta_0)\| \leq B n^{-1/4+\eta-\kappa}), \\ S_{2,n} &:= (\max_i \sup_{\beta_i} \|N' \ddot{g}(x_i, \beta_i) N\| \leq B n^{-1/2+2\eta-\kappa}), \\ S_n &:= S_{1,n} \cap S_{2,n}. \end{aligned} \tag{3.77}$$

Remark 3.A.2.

1. Under Assumption 3.A.2/1, as $n \to \infty$, then for any $\epsilon_1 > 0$, $P(S_{1,n}^c) < \epsilon_1$.
2. Under Assumption 3.A.2/2, as $n \to \infty$, then for any $\epsilon_2 > 0$, $P(S_{2,n}^c) < \epsilon_2$.
3. Under Assumption 3.A.2, as $n \to \infty$, then for any $\epsilon_1, \epsilon_2 > 0$: $P(S_{1,n}^c \cup S_{2,n}^c) < \epsilon_1 + \epsilon_2$.

This follows from the definition of boundedness in probability. For example, in the case of set $S_{1,n}$, given Assumption 3.A.2/1, for any ε_1 one can choose $B > 0$ such that the inequality is satisfied. The remaining two inequalities above follow analogously.

Thus, in what follows it suffices to consider the properties of quantities of interest on the sets $S_{1,n}$, $S_{2,n}$, or $S_n := S_{1,n} \cap S_{2,n}$.

3.A.4 Summary of Key Lemmas

The purpose of this section is to present a brief outline of the twelve lemmas proven in this Appendix which underpin the results derived in this chapter.

Our over-arching aim is to prove results (3.19) and (3.21). The overall strategy for the proof involves bounding the terms of interest by sums of simpler terms, whose bounds are derived in individual theorems and lemmas. Section 3.B.3 shows that (3.19) can be bounded by the following sum of two terms, which allow us to focus separately on the difference in the mark functions and the difference in the indicator functions.

$$\begin{aligned}
&\sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\mathsf{E}_{i-1}\left|m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\left\{1_{\left(\varepsilon_i/\sigma\le c_{in}^0+d_{in}^*\right)}-1_{\left(\varepsilon_i/\sigma\le c_{in}^0\right)}\right\}\right| + \\
&+\sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\mathsf{E}_{i-1}\left|\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)-m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right\}\right|1_{\left(\varepsilon_i/\sigma\le c_{in}^0\right)}
\end{aligned} \tag{3.78}$$

The proof that the first term is $o_p(1)$ is provided in Theorem 3.B.1. The proof that the second term is $o_p(1)$ is provided in Theorem 3.B.2.

Further, Section 3.B.4 shows that (3.21) can be bounded by the following sum. This follows a decomposition into a difference of marks and a difference of indicators analogous to equation (3.78) but with a further step to incorporate a value of the mark function evaluated without estimation errors.

$$\begin{aligned}
&\sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)-m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right\}1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)}\right| + \\
&+\sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)-m\left(\frac{\varepsilon_i}{\sigma}\right)\right\}\left\{1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0\right)}\right\}\right| + \\
&+\sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in}m\left(\frac{\varepsilon_i}{\sigma}\right)\left\{1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0\right)}\right\}\right|
\end{aligned} \tag{3.79}$$

The proof that the first term is $o_p(1)$ is provided in Theorem 3.B.3. The proof that the second term is $o_p(1)$ is given in Theorem 3.B.4. The proof that the third term is $o_p(1)$ is provided in Theorem 3.B.5.

The proofs of the theoreams listed above rely on a number of lemmas. The proof of Theorem 3.B.1 decomposes the first term in equation (3.78) further, and builds on Lemmas 3.A.4 and 3.A.5. Both lemmas derive bounds on weighted sums of the mark function, weighted by the weights w_{in} and the density $f(u)$. Lemma 3.A.4 focuses on the mark function itself, with the

following bound.

$$\sup_{u\in\mathbb{R}}\sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}\right||m(u)|\,f(u)=o_p(1) \tag{3.80}$$

Lemma 3.A.5 derives the following bound on a weighted sum of the first derivative of the mark function.

$$\sup_{u\in\mathbb{R}}\sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}\right|\left|u\left(n^{-1/2}\frac{1}{s}\frac{1}{\sigma}a_m\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^{0}\right||\dot{m}(u)|\,f(u)=o_p(1) \tag{3.81}$$

The proof of Theorem 3.B.2, which focuses on the second term in equation (3.78), relies on Lemmas 3.A.6 and 3.A.7. In contrast to Lemmas 3.A.4 and 3.A.5, the focus is on the expectations of weighted sums of the mark function. Lemma 3.A.6 derives the following bound on the expectation of a weighted sum of the first derivative of the mark function.

$$\sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\int_{-\infty}^{\infty}\frac{|b_m'\ddot{g}_{in}^{m*}b_m|}{\sigma}\left|\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|f\left(\frac{\varepsilon_i}{\sigma}\right)d\left(\frac{\varepsilon_i}{\sigma}\right)=o_p(1) \tag{3.82}$$

Lemma 3.A.7 uses a Lipschitz condition to derive the following bound on the expectation of a weighted sum of the difference between the values of the first derivative of the mark function evaluated at two different points.

$$\sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\int_{-\infty}^{\infty}\left|\left(\frac{\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)\left\{\dot{m}\left(\frac{\varepsilon_i-b_m'\dot{g}_{in}^{0}-\frac{1}{2}b_m'\bar{g}_m b_m}{\sigma+n^{-1/2}a_m}\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right\}\right|f\left(\frac{\varepsilon_i}{\sigma}\right)d\left(\frac{\varepsilon_i}{\sigma}\right)=o_p(1) \tag{3.83}$$

In the proof of result (3.21), Theorem 3.B.3 relies on further decomposing the first term of equation (3.79) and using Lemmas 3.A.8 and 3.A.9 on the resulting terms. Lemma 3.A.8 derives a bound on a weighted sum of the first derivative of the mark function, $\dot{m}(u)$. In contrast to Lemma 3.A.4, the term of interest is not weighted by the density $f(u)$ and the focus is on the first derivative, rather than the mark function itself, as shown in the following equation.

$$\sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|=o_p(1) \tag{3.84}$$

Lemma 3.A.9 uses the Lipschitz condition to derive a bound on the weighted sum of the difference between the values of the first derivative of the mark function, $\dot{m}(u)$, evaluated at two different points. In contrast to Lemma 3.A.7, the bound is on the weighted sum, rather than its expectation, as follows.

$$\sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\left\{\dot{m}\left(\frac{\varepsilon_i-b_m'\dot{g}_{in}^{0}-\frac{1}{2}b_m'\bar{g}_m b_m}{\sigma+n^{-1/2}a_m}\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right\}\right|=o_p(1) \tag{3.85}$$

Theorem 3.B.4, which focuses on the second term in equation (3.79), uses Lemmas 3.A.10 and 3.A.11. Both lemmas derive bounds on weighted sums of the first derivative of the mark function, $\dot{m}(u)$. Lemma 3.A.10 focuses on the derivative itself, as detailed below.

$$\sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right|\left|\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|\cdot \\ \cdot\left|1_{\left(\frac{\varepsilon_i}{\sigma}\le c+\frac{b'_1\dot{g}^0_{in}}{\sigma}+d^*_{in}\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\le c+\frac{b'_1\dot{g}^0_{in}}{\sigma}\right)}\right|=o_p(1) \tag{3.86}$$

Lemma 3.A.11 relies on the Lipschitz condition to derive a bound on a weighted sum of the values of the first derivative of the mark function, $\dot{m}(u)$ evaluated at the following two points.

$$\sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right|\left|\dot{m}\left(u^*\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|\cdot \\ \cdot\left|1_{\left(\frac{\varepsilon_i}{\sigma}\le c+\frac{b'_1\dot{g}^0_{in}}{\sigma}+d^*_{in}\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\le c+\frac{b'_1\dot{g}^0_{in}}{\sigma}\right)}\right|=o_p(1) \tag{3.87}$$

Furthermore, a number of proofs use Lemmas 3.A.2 and 3.A.3. Lemma 3.A.2 introduces the concept of a distance function, which is based on a general function $h(\cdot)$, and derives bounds on the distance function. Lemma 3.A.3 applies Lemma 3.A.2 to derive a bound on the following process, based on a general function $h(\cdot)$ and a general set of weights v_{in}.

$$\sup_{c}\sup_{\theta} n^{-1/2}\sum_{i=1}^{n}\left|v_{in}h\left(\frac{\varepsilon_i}{\sigma}\right)\left\{1_{\left(\varepsilon_i\le\sigma c+b'_1\dot{g}^0_{in}+\frac{1}{2}b'_1\ddot{g}^{1*}_{in}b_1\right)}-1_{\left(\varepsilon_i\le\sigma c+b'_1\dot{g}^0_{in}\right)}\right\}\right| \tag{3.88}$$

Lemmas 3.A.3 and 3.A.2 are used to prove Lemma 3.A.10 and Theorems 3.B.4 and 3.B.5, with the general function $h(\cdot)$ replaced by functions of the mark function $m(\cdot)$ and the non-linear function $g(\cdot)$. The weights v_{in} are replaced by specified functions of the weights w_{in}. The weights v_{in} and the mark function $h(\cdot)$ are defined separately for each application.

3.A.5 Main Lemmas

Assumption 3.A.3. Define:

$$H_r(c)=\int_{-\infty}^{c}\left(1+\{h\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right)f\left(\frac{\varepsilon_i}{\sigma}\right)\mathrm{d}\left(\frac{\varepsilon_i}{\sigma}\right)=\mathsf{E}\left(1+\{h\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right)1_{(\varepsilon_i/\sigma\le c)} \tag{3.89}$$

1. *Assumptions on the distance function $H_r(c)$.*
   $H_r(\cdot)$ is a distance function where $H_r(c)=\int_{-\infty}^{c}\dot{H}_r(x)dx$. Assume:

   (a) the support of $\dot{H}_r$, that is $\mathcal{S}=\{c:\dot{H}_r(c)>0\}$, is an interval with endpoints $\underline{c},\bar{c}$ such that $-\infty\le\underline{c}<\bar{c}\le\infty$;

(b) $\sup_{x \in \mathcal{S}} 0 \leq \dot{H}_r(x) < \infty$;

(c) $H_r = H_r(\infty) = \int_{\mathbb{R}} \dot{H}_r(c)dc < \infty$;

(d) $\dot{H}_r(c)$ is Lipschitz, as defined in Definition 3.4.1;

(e) $\dot{H}_r(c)$ has weakly unimodal bounds as defined in Definition 3.4.1.

Lemma 3.A.2. Let $c \in \mathbb{R}$ and $b \in |b| \leq Bn^{1/4-\eta}$, where $0 < B < \infty$ and $\eta < 1/4$. Further, let $c_{k-1}, c_k \in \mathbb{R}$ be $K \in \mathcal{N}$ grid points such that $c_{k-1} < c \leq c_k$. Let $b_p \in \mathbb{R}^{dimb}$ be $P \in \mathcal{N}$ ball centres such that $|b_p - b| \leq \delta$, where $0 < \delta < \infty$. Crucially, δ does not depend on n and it can be chosen for a particular application. Define:

$$\ddot{g}^s_{in} = \sup_{\beta_i} \| N' \ddot{g}(x_i, \beta_i) N \| \frac{1}{2\sigma}. \tag{3.90}$$

Further, define the following values:

$$\underline{c}_{kpi} = c_{k-1} + b_p' \dot{g}^0_{in}/\sigma + b_p' \ddot{g}^s_{in} b_p - \delta \|\dot{g}^0_{in}\|/\sigma - \delta^2 \ddot{g}^s_{in} - 2\delta \ddot{g}^s_{in} |b_p| \tag{3.91}$$

and:

$$\bar{c}_{kpi} = c_k + b_p' \dot{g}^0_{in}/\sigma + b_p' \ddot{g}^s_{in} b_p + \delta \|\dot{g}^0_{in}\|/\sigma + \delta^2 \ddot{g}^s_{in} + 2\delta \ddot{g}^s_{in} |b_p|. \tag{3.92}$$

Under Assumption 3.A.3, the following holds:

$$\begin{aligned} &\max_k \max_p \mathcal{H}_{kpi} = \max_k \max_p \left| H_r(\bar{c}_{kpi}) - H_r(\underline{c}_{kpi}) \right| \\ &\leq C(K^{-1} + \|\dot{g}^0_{in}\| K^{-1/2} n^{1/4-\eta} + \|\dot{g}^0_{in}\| \delta + \|\dot{g}^0_{in}\|^2 n^{1/2-2\eta} + n^{1/2-2\eta} \ddot{g}^s_{in} + \delta^2 \ddot{g}^s_{in} + n^{1/4-\eta} 2\delta \ddot{g}^s_{in}), \end{aligned} \tag{3.93}$$

where $\delta, C < \infty$ are finite constants which do not depend on n or i.

Proof. By a first-order Taylor series expansion of $H_r(\bar{c}_{kpi})$ around $H_r\left\{c_k + b_p' \dot{g}^0_{in}/\sigma + \delta \|\dot{g}^0_{in}\|/\sigma\right\}$, there exists an intermediate point c^*, such that:

$$\begin{aligned} H_r(\bar{c}_{kpi}) &= H_r(c_k + b_p' \dot{g}^0_{in}/\sigma + b_p' \ddot{g}^s_{in} b_p + \delta \|\dot{g}^0_{in}\|/\sigma + \delta^2 \ddot{g}^s_{in} + 2\delta \ddot{g}^s_{in} |b_p|) = \\ &= H_r\left\{c_k + b_p' \dot{g}^0_{in}/\sigma + \delta \|\dot{g}^0_{in}\|/\sigma\right\} + \left\{b_p' \ddot{g}^s_{in} b_p + \delta^2 \ddot{g}^s_{in} + 2\delta \ddot{g}^s_{in} b_p\right\} \dot{H}_r(c^*) \end{aligned} \tag{3.94}$$

and:

$$\left| c^* - \left(c_k + b_p' \dot{g}^0_{in}/\sigma + \delta \|\dot{g}^0_{in}\|/\sigma\right) \right| \leq \left| \bar{c}_{kpi} - \left(c_k + b_p' \dot{g}^0_{in}/\sigma + \delta \|\dot{g}^0_{in}\|/\sigma\right) \right|. \tag{3.95}$$

Analogously, $H_r(\underline{c}_{kpi})$ can be expanded around $H_r\left\{c_{k-1} + b_p' \dot{g}^0_{in}/\sigma - \delta \|\dot{g}^0_{in}\|/\sigma\right\}$ for an inter-

mediate point c_* such that:

$$\begin{aligned} H_r(\underline{c}_{kpi}) &= H_r(c_{k-1} + b_p'\dot{g}_{in}^0/\sigma + b_p'\ddot{g}_{in}^s b_p - \delta\|\dot{g}_{in}^0\|/\sigma - \delta^2\ddot{g}_{in}^s - 2\delta\ddot{g}_{in}^s b_p) = \\ &= H_r\left\{c_{k-1} + b_p'\dot{g}_{in}^0/\sigma - \delta\|\dot{g}_{in}^0\|/\sigma\right\} + \left\{b_p'\ddot{g}_{in}^s b_p - \delta^2\ddot{g}_{in}^s - 2\delta\ddot{g}_{in}^s b_p\right\}\dot{H}_r(c_*) \end{aligned} \tag{3.96}$$

and:

$$\left|c_{k-1} + b_p'\dot{g}_{in}^0/\sigma - \delta\|\dot{g}_{in}^0\|/\sigma - c_*\right| \leq \left|c_{k-1} + b_p'\dot{g}_{in}^0/\sigma - \delta\|\dot{g}_{in}^0\|/\sigma - \underline{c}_{kpi}\right|. \tag{3.97}$$

Thus:

$$\begin{aligned} &\mathcal{H}_{kpi} = H_r(\bar{c}_{kpi}) - H_r(\underline{c}_{kpi}) \\ &= H_r(c_k + b_p'\dot{g}_{in}^0/\sigma + \delta\|\dot{g}_{in}^0\|/\sigma) - H_r(c_{k-1} + b_p'\dot{g}_{in}^0/\sigma - \delta\|\dot{g}_{in}^0\|/\sigma) + \\ &+ (b_p'\ddot{g}_{in}^s b_p + \delta^2\ddot{g}_{in}^s + 2\delta\ddot{g}_{in}^s b_p)\dot{H}_r(c^*) - (b_p'\ddot{g}_{in}^s b_p - \delta^2\ddot{g}_{in}^s - 2\delta\ddot{g}_{in}^s b_p)\dot{H}_r(c_*). \end{aligned} \tag{3.98}$$

The difference of the first two terms in equation (3.98) is bounded uniformly in k by Lemma A.9. of Berenguer-Rico et. al. (2019a) under Assumption 3.A.3, which subsumes Assumption A.1 of Berenguer-Rico et. al. (2019a). In particular, note that Berenguer-Rico et. al. (2019a) scale the non-stochastic estimation errors (b_m) by $1/\sigma$ whereas in our derivation the scaling by $1/\sigma$ is explicit in the distance function as b_1 and b_m are not scaled by $1/\sigma$. Applying Lemma A.9 of Berenguer-Rico et.al. (2019a) gives:

$$\begin{aligned} &\max_k \left|H_r(c_k + b_p'\dot{g}_{in}^0/\sigma + \delta\|\dot{g}_{in}^0\|/\sigma) - H_r(c_{k-1} + b_p'\dot{g}_{in}^0/\sigma - \delta\|\dot{g}_{in}^0\|/\sigma)\right| \leq \\ &\leq C_1(K^{-1} + |b_p'|\|\dot{g}_{in}^0\|K^{-1/2} + \|\dot{g}_{in}^0\|\delta + |b_p|^2\|\dot{g}_{in}^0\|^2), \end{aligned} \tag{3.99}$$

where $C_1 < \infty$ is a constant which does not depend on n. Note that the scaling by $1/\sigma$ and $1/\sigma^2$ (where relevant) is contained in the constant C_1. The remaining terms in equation (3.98) can be written as:

$$(b_p'\ddot{g}_{in}^s b_p)(\dot{H}_r(c^*) - \dot{H}_r(c_*)) + \delta^2\ddot{g}_{in}^s(\dot{H}_r(c^*) + \dot{H}_r(c_*)) + 2\delta\ddot{g}_{in}^s b_p(\dot{H}_r(c^*) + \dot{H}_r(c_*)) \tag{3.100}$$

Since $|\dot{H}_r(c)| < \infty$ for all $c \in \mathbb{R}$, then the overall bound can be stated as follows, for $(\dot{H}_r(c^*) - \dot{H}_r(c_*)) = C_2 < \infty$ and $(\dot{H}_r(c^*) + \dot{H}_r(c_*)) = C_3 < \infty$, where C_2 and C_3 are constants which do not depend on n:

$$\max_k \mathcal{H}_{kpi} \leq C_1(K^{-1} + |b_p'|\|\dot{g}_{in}^0\|K^{-1/2} + \|\dot{g}_{in}^0\|\delta + |b_p|^2\|\dot{g}_{in}^0\|^2) + C_2 b_p'\ddot{g}_{in}^s b_p + C_3(\delta^2\ddot{g}_{in}^s + 2\delta\ddot{g}_{in}^s b_p) \tag{3.101}$$

By construction, $|b_p| \leq Bn^{1/4-\eta}$. Hence:

$$\begin{aligned}\max_k \mathcal{H}_{kpi} \leq C_1(K^{-1} + \|\dot{g}_{in}^0\|K^{-1/2}Bn^{1/4-\eta} + \|\dot{g}_{in}^0\|\delta + \|\dot{g}_{in}^0\|^2B^2n^{1/2-2\eta}) + \\ + C_2B^2n^{1/2-2\eta}\ddot{g}_{in}^s + C_3\delta^2\ddot{g}_{in}^s + C_3Bn^{1/4-\eta}2\delta\ddot{g}_{in}^s\end{aligned} \tag{3.102}$$

Taking $C := \max\{C_1, C_1B, c_1B^2, C_2B^2, C_3, C_3B\}$ gives the following, as required.

$$\begin{aligned}\max_k \mathcal{H}_{kpi} \leq & C(K^{-1} + \|\dot{g}_{in}^0\|K^{-1/2}n^{1/4-\eta} + \|\dot{g}_{in}^0\|\delta + \|\dot{g}_{in}^0\|^2n^{1/2-2\eta} + \\ & + n^{1/2-2\eta}\ddot{g}_{in}^s + \delta^2\ddot{g}_{in}^s + n^{1/4-\eta}2\delta\ddot{g}_{in}^s)\end{aligned} \tag{3.103}$$

■

Assumption 3.A.4. Let v_{in} be a general set of weights and let $h(\frac{\varepsilon_i}{\sigma})$ be a general mark function. Let r be such that $2^{r-1} > 1 + (1/4 - \eta)(\dim b_1)$ and q such that $1 \leq q \leq r$. Further, define $H_r(c)$ analogously to Assumption 3.A.3:

$$H_r(c) = \int_{-\infty}^{c} \left(1 + \{h\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right) f\left(\frac{\varepsilon_i}{\sigma}\right) \mathrm{d}\left(\frac{\varepsilon_i}{\sigma}\right) = \mathsf{E}\left(1 + \{h\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right) 1_{(\varepsilon_i/\sigma \leq c)}. \tag{3.104}$$

Assume the following:

1. *The weights v_{in}*:
   (a) The weights v_{in} are $\mathcal{F}_{i-1}$-measurable, noting as above that ε_i is independent of $\mathcal{F}_{i-1,n}$.
   (b) $\mathsf{E}|v_{in}|^{2^r} < \infty$
   (c) For all $1 \leq q \leq r$: $\mathsf{E}\sum_{i=1}^n |v_{in}|^{2^q} = O(n)$
   (d) $\sum_{i=1}^n |v_{in}| = O_p(n)$
2. *The regressors and the function $g(\cdot)$, for all $1 \leq q \leq r$:*
   (a) $\mathsf{E}\sum_{i=1}^n |v_{in}|^{2^q}\|N'\dot{g}(x_i;\beta_0)\| = O(n^{3/4-\eta})$
   (b) $\mathsf{E}\sum_{i=1}^n |v_{in}|^{2^q}\|N'\dot{g}(x_i;\beta_0)\|^2 = O(n^{1/4+\eta})$
   (c) $\mathsf{E}\sum_{i=1}^n |v_{in}|^{2^q} \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\| = O(n^{2\eta})$
   (d) $\sum_{i=1}^n |v_{in}|\|N'\dot{g}(x_i;\beta_0)\| = O_p(n^{1/2})$
   (e) $\sum_{i=1}^n |v_{in}|\|N'\dot{g}(x_i;\beta_0)\|^2 = o_p(n^{2\eta})$
   (f) $\sum_{i=1}^n |v_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{2\eta})$
3. *The mark function $h(\cdot)$*:

(a) $\mathsf{E}\left|h\left(\frac{\varepsilon_i}{\sigma}\right)\right|^{2^r} < \infty$

4. *The distance function* $H_r(c)$:

   (a) the support of $\dot{H}_r$, that is $\mathcal{S} = \{c : \dot{H}_r(c) > 0\}$, is an interval with endpoints $\underline{c}, \bar{c}$ such that $-\infty \leq \underline{c} < \bar{c} \leq \infty$;

   (b) $\sup_{x \in \mathcal{S}} 0 \leq \dot{H}_r(x) < \infty$;

   (c) $H_r = H_r(\infty) = \int_{\mathbb{R}} \dot{H}_r(c)dc < \infty$;

   (d) $\dot{H}_r(c)$ is Lipschitz, as defined in Definition 3.4.1;

   (e) $\dot{H}_r(c)$ has weakly unimodal bounds as defined in Definition 3.4.1.

Lemma 3.A.3. Let Assumption 3.A.4 be satisfied. Then for a general set of weights v_{in} and a general mark function $h(\frac{\varepsilon_i}{\sigma})$, the following holds:

$$\sup_c \sup_\theta n^{-1/2} \sum_{i=1}^n \left| v_{in} h\left(\frac{\varepsilon_i}{\sigma}\right) \{1_{\left(\varepsilon_i \leq \sigma c + b_1' \dot{g}_{in}^0 + \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1\right)} - 1_{\left(\varepsilon_i \leq \sigma c + b_1' \dot{g}_{in}^0\right)}\} \right| = o_p(1). \tag{3.105}$$

Proof.

1. *Defining the metric and cover.*

Define $H_r = \int_{-\infty}^c (1 + h(u)^{2^r}) f(u)d(u) = \mathsf{E}(1 + h(\frac{\varepsilon_i}{\sigma})^{2^r}) 1_{(\varepsilon_i \leq \sigma c)}$ with derivative $\dot{H}_r(c) = (1 + h(c)^{2^r}) f(c)$. The function $H_r(\cdot)$ is increasing in c by construction and is bounded by Assumptions 3.A.4/4b and 4c.

Define further:

$$H_r := H_r(\infty) = \int_{-\infty}^{\infty} (1 + |h(u)|^{2^r}) f(u)du = \mathsf{E}(1 + |h(\varepsilon/\sigma)|^{2^r}) < \infty. \tag{3.106}$$

For $c^\dagger \geq c$ we define the H_r-distance between c and $c^\dagger$ as:

$$H_r(c^\dagger) - H_r(c) = \mathsf{E}\left(1 + h(\varepsilon_i/\sigma)^{2^r}\right) 1_{(\sigma c < \varepsilon_i \leq \sigma c^\dagger)}. \tag{3.107}$$

2. *Constructing c-balls.*

The support of $H_r(\cdot)$ is $\mathbb{R}$. We partition the $\mathbb{R}$ axis on the basis of the range of $H_r(c)$. Since $\forall c \in \mathbb{R} : 0 \leq H_r(c) \leq H_r < \infty$, we partition the range of $H_r(c)$, i.e. the interval $[0, H_r]$ into K intervals of *equal* size H_r/K. This gives a partition of the support into K intervals defined by the grid points:

$$-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = \infty. \tag{3.108}$$

Consequently, for $1 \leq k \leq K$:

$$H_r(c_k) - H_r(c_{k-1}) = H_r/K. \tag{3.109}$$

Analogously to Berenguer-Rico et. al. (2019a) we choose $K = \mathsf{int}(H_r n^{1/2}/\delta)$ where $0 < \delta < \infty$ is a constant which does not depend on n and will be specified in part 7. of this proof. Consequently, the length of each interval in the range of $H_r(c)$ is $H_r(c_k) - H_r(c_{k-1}) = H_r/K \sim \delta n^{-1/2}$.

3. *Constructing b-balls.*

Cover the set $|b| \leq n^{1/4-\eta}$ with balls of radius δ (as in 2. above), and centres b_p, where $p = 1, ..., P$ and $P \sim (n^{1/4-\eta}B/\delta)^{\dim b_1} \sim n^{(1/4-\eta)\dim b_1}/\delta^{\dim b_1}$. Consequently, for any point b there exists a centre b_p such that $|b - b_p| \leq \delta$.

4. *Applying chaining.*

Recall that we wish to bound the following quantity:

$$\sup_c \sup_{b_1} n^{-1/2} \sum_{i=1}^{n} \left| v_{in} h\left(\frac{\varepsilon_i}{\sigma}\right) \left\{ 1_{\left(\varepsilon_i/\sigma \leq c + b_1' \dot{g}_{in}^0/\sigma + \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1/\sigma\right)} - 1_{\left(\varepsilon_i/\sigma \leq c + b_1' \dot{g}_{in}^0/\sigma\right)} \right\} \right|. \tag{3.110}$$

We start by focusing on the terms in the indicator function, which we generalise from a specific b_1 to a general $b \in \left\{|b| \leq Bn^{1/4-\eta}\right\}$, where $0 < B < \infty$:

$$c + b' \dot{g}_{in}^0/\sigma + b' \ddot{g}_{in}^{1*} b/2\sigma. \tag{3.111}$$

By construction, given the grid points in 2. and b-balls in 3., for any $c \in \mathbb{R}$ and $b \in \left\{|b| \leq Bn^{1/4-\eta}\right\}$, where $0 < B < \infty$, the term can be expressed as a function of the deviations of point b from the nearest ball centre, b_p:

$$\begin{aligned} &c + b' \dot{g}_{in}^0/\sigma + b' \ddot{g}_{in}^{1*} b/2\sigma = \\ &= c + (b - b_p)' \dot{g}_{in}^0/\sigma + b_p' \dot{g}_{in}^0/\sigma + (b - b_p)' \ddot{g}_{in}^{1*}(b - b_p)/2\sigma + 2(b - b_p)' \ddot{g}_{in}^{1*} b_p/2\sigma + b_p' \ddot{g}_{in}^{1*} b_p/2\sigma. \end{aligned} \tag{3.112}$$

Hence, given the construction of c-balls and b-balls, or any $b \in |b| \leq Bn^{1/4-\eta}$ and $c \in \mathbb{R}$ we can find a ball centre b_p and grid points c_k, c_{k-1} such that $c_{k-1} < c \leq c_k$ and $|b_p - b| \leq \delta$. Consequently, for any $b \in |b| \leq Bn^{1/4-\eta}$ and $c \in \mathbb{R}$ we can find c_k and c_{k-1} such that:

$$\mathring{\underline{c}}_{kpi} < c + b' \dot{g}_{in}^0/\sigma + b' \ddot{g}_{in}^{1*} b/2\sigma \leq \mathring{\bar{c}}_{kpi}, \tag{3.113}$$

where:

$$\mathring{\underline{c}}_{kpi} = c_{k-1} + b_p' \dot{g}_{in}^0/\sigma + b_p' \ddot{g}_{in}^{1*} b_p/2\sigma - \delta \|\dot{g}_{in}^0\|/\sigma - \delta^2 \|\ddot{g}_{in}^{1*}\|/2\sigma - 2\delta \|\ddot{g}_{in}^{1*}\| |b_p|/2\sigma \tag{3.114}$$

and:

$$\mathring{\bar{c}}_{kpi} = c_k + b_p' \dot{g}_{in}^0/\sigma + b_p' \ddot{g}_{in}^{1*} b_p/2\sigma + \delta \|\dot{g}_{in}^0\|/\sigma + \delta^2 \|\ddot{g}_{in}^{1*}\|/2\sigma + 2\delta \|\ddot{g}_{in}^{1*}\| |b_p|/2\sigma. \tag{3.115}$$

The terms are further bounded by

$$\underline{c}_{kpi} \leq \mathring{\underline{c}}_{kpi} < c + b' \dot{g}_{in}^0/\sigma + b' \ddot{g}_{in}^{1*} b/2\sigma \leq \bar{c}_{kpi}, \tag{3.116}$$

where $\underline{c}_{kpi}$ and $\bar{c}_{kpi}$ are definedin equations (3.91) and (3.92). Note that the scaling by $1/2\sigma$ is included in the definition of $\ddot{g}^s_{in}$. Thus, the term of interest can be bounded as follows

$$\begin{aligned}&\sup_c \sup_{b_1} n^{-1/2} \sum_{i=1}^n |v_{in}| \left| h\left(\frac{\varepsilon_i}{\sigma}\right) \right| \left| \{1_{\left(\varepsilon_i \leq \sigma c + b_1' \dot{g}^0_{in} + \frac{1}{2} b_1' \ddot{g}^{1*}_{in} b_1\right)} - 1_{\left(\varepsilon_i \leq \sigma c + b_1' \dot{g}^0_{in}\right)}\} \right| \\ &\leq \max_k \max_p n^{-1/2} \sum_{i=1}^n |v_{in}| \left| h\left(\frac{\varepsilon_i}{\sigma}\right) \right| \{1_{\left(\varepsilon_i/\sigma \leq \bar{c}_{kpi}\right)} - 1_{\left(\varepsilon_i/\sigma \leq \underline{c}_{kpi}\right)}\}.\end{aligned} \tag{3.117}$$

5. *Martingale and compensator.*
Denote $z_{kpi} := |v_{in}||h\left(\frac{\varepsilon_i}{\sigma}\right)|\{1_{\left(\varepsilon_i/\sigma \leq \bar{c}_{kpi}\right)} - 1_{\left(\varepsilon_i/\sigma \leq \underline{c}_{kpi}\right)}\}$. Create a martingale and a compensator term.

$$\tilde{M}_k := n^{-1/2} \sum_{i=1}^n \{z_{kpi} - \mathsf{E}_{i-1} z_{kpi}\} \tag{3.118}$$

$$\bar{M}_k := 2n^{-1/2} \sum_{i=1}^n \mathsf{E}_{i-1} z_{kpi} \tag{3.119}$$

By the triangle inequality, in order to show that the bound in equation (3.117) is $o_p(1)$, it suffices to show that: $\tilde{M}_k = o_p(1)$ and $\bar{M}_k = o_p(1)$.

6. *The martingale.*

In order to show that the martingale $\tilde{M}_k := n^{-1/2} \sum_{i=1}^n \{z_{kpi} - \mathsf{E}_{i-1} z_{kpi}\} = o_p(1)$, we use the iterated martingale inequality of Johansen and Nielsen (2016a) (also used as Lemma A.1. in Berenguer-Rico et. al. (2019a)).

Let $L = PK$. Since $K \sim n^{1/2}/\delta$; and $P \sim n^{(1/4-\eta)\dim b_1}/\delta^{\dim b_1}$, then $L = O(n^\lambda)$ with $\lambda = 1/2 + (1/4 - \eta)\dim b_1$. In order to use the iterated martingale inequality, the following must hold:

a) $\mathsf{E} z_{li}^{2^r} < \infty$ for some $r \in \mathbb{N}$.

b) Let $D_q := \max_{1 \leq l \leq L} \sum_{i=1}^n \mathsf{E}_{i-1} z_{li}^{2^q}$ for $1 \leq q \leq r$ and $\mathsf{E} D_q = O(n^\zeta)$, where $\nu > 0$ is chosen such that the following holds:

   i) $\zeta < 2\nu$, and

   ii) $\zeta + \lambda < \nu 2^r$.

Then, by the iterated martingale inequality, $\max_l |\sum_{i=1}^n (z_{li} - \mathsf{E}_{i-1} z_{li})| = o_p(n^\nu)$.

6.1. *Checking condition a).*

By the Law of Iterated Expectations:

$$\mathsf{E} z_{li}^{2^r} = \mathsf{E}\mathsf{E}_{i-1} z_{li}^{2^r}, \tag{3.120}$$

where the outer expectation is taken with respect to the filtration $\mathcal{F}_{i-1}$ and the inner expectation is taken conditional on the filtration $\mathcal{F}_{i-1}$.

By definition of z_{li} and noting that $\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}$ can be either 0 or 1 and, therefore, remains unchanged when raised to the power of 2^r, this is equivalent to:

$$\mathsf{E}z_{li}^{2^r} = \mathsf{E}\mathsf{E}_{i-1}|v_{in}|^{2^r}\left|h\left(\frac{\varepsilon_i}{\sigma}\right)\right|^{2^r}\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}. \tag{3.121}$$

Noting that $|v_{in}| \subset \mathcal{F}_{i-1}$ gives:

$$\mathsf{E}z_{li}^{2^r} = \mathsf{E}|v_{in}|^{2^r}\mathsf{E}_{i-1}\left|h\left(\frac{\varepsilon_i}{\sigma}\right)\right|^{2^r}\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}. \tag{3.122}$$

Since the term $\{1_{(\varepsilon_i\leq\bar{c}_{kpi})} - 1_{(\varepsilon_i\leq\underline{c}_{kpi})}\}$ can only be 0 or 1, it is bounded from above by 1:

$$\mathsf{E}z_{li}^{2^r} \leq \mathsf{E}|v_{in}|^{2^r}\mathsf{E}_{i-1}\left|h\left(\frac{\varepsilon_i}{\sigma}\right)\right|^{2^r}. \tag{3.123}$$

Noting that ε_i is independent of the filtration $\mathcal{F}_{i-1}$ and, therefore, the expectation conditional on $\mathcal{F}_{i-1}$ can be replaced with the unconditional expectation:

$$\mathsf{E}z_{li}^{2^r} \leq \mathsf{E}|v_{in}|^{2^r}\mathsf{E}\left|h\left(\frac{\varepsilon_i}{\sigma}\right)\right|^{2^r} \tag{3.124}$$

By Assumption 3.A.4/1b: $\mathsf{E}|v_{in}|^{2^r} < \infty$. By Assumption 3.A.4/3a: $\mathsf{E}\left|h\left(\frac{\varepsilon_i}{\sigma}\right)\right|^{2^r} < \infty$. Thus: $\mathsf{E}z_{li}^{2^r} < \infty$ and condition a) of the iterated martingale inequality is satisfied.

6.2. *Checking condition b).*

The focus is on the following quantity:

$$\begin{aligned}\mathsf{E}D_q &:= \mathsf{E}\left\{\max_{1\leq l\leq L}\sum_{i=1}^{n}\mathsf{E}_{i-1}z_{li}^{2^q}\right\} \\ &= \mathsf{E}\left\{\max_{1\leq l\leq L}\sum_{i=1}^{n}\mathsf{E}_{i-1}|v_{in}|^{2^q}|h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^q}\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}^{2^q}\right\},\end{aligned} \tag{3.125}$$

where the equality follows from the definition of D_q. Further, using the inequality $|u|^q < 1 + |u|^r$ for $0 \leq q \leq r$, the above is bounded, for all q such that $0 \leq q \leq r$:

$$\mathsf{E}D_q \leq \mathsf{E}\left\{\max_{1\leq l\leq L}\sum_{i=1}^{n}\mathsf{E}_{i-1}|v_{in}|^{2^q}\left(1 + |h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^r}\right)\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}^{2^q}\right\}. \tag{3.126}$$

Noting that $\{1_{(\varepsilon_i\leq\bar{c}_{kpi})} - 1_{(\varepsilon_i\leq\underline{c}_{kpi})}\}$ can be either 0 or 1 and, therefore, remains unchanged when raised to the power of 2^q, we can drop the power on the difference of the indicator functions,

which simplifies the bound to:

$$\mathsf{E}D_q \leq \mathsf{E}\left\{\max_{1\leq l\leq L}\sum_{i=1}^{n}\mathsf{E}_{i-1}|v_{in}|^{2^q}\left(1+|h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^r}\right)\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})}-1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}\right\}. \tag{3.127}$$

Since $v_{in}\subset\mathcal{F}_{i-1}$, it can be taken out of the conditional expectation:

$$\mathsf{E}D_q \leq \mathsf{E}\left\{\max_{1\leq l\leq L}\sum_{i=1}^{n}|v_{in}|^{2^q}\mathsf{E}_{i-1}\left(1+|h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^r}\right)\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})}-1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}\right\}. \tag{3.128}$$

Noting that the supremum of a sum is no greater than the sum of individual suprema gives:

$$\mathsf{E}D_q \leq \mathsf{E}\left\{\sum_{i=1}^{n}|v_{in}|^{2^q}\max_{1\leq l\leq L}\mathsf{E}_{i-1}\left(1+|h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^r}\right)\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})}-1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}\right\}. \tag{3.129}$$

Use the definition of $H_r(c)$ and note that ε_i is independent of the filtration $\mathcal{F}_{i-1}$ and, therefore, the expectation conditional on $\mathcal{F}_{i-1}$ can be replaced with the unconditional expectation, which leads to:

$$\begin{aligned}\mathsf{E}D_q \leq & \mathsf{E}\left\{\sum_{i=1}^{n}|v_{in}|^{2^q}\max_{1\leq l\leq L}\mathsf{E}\left(1+|h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^r}\right)\{1_{(\varepsilon_i/\sigma\leq\bar{c}_{kpi})}-1_{(\varepsilon_i/\sigma\leq\underline{c}_{kpi})}\}\right\}\\ &= \mathsf{E}\sum_{i=1}^{n}|v_{in}|^{2^q}\max_{1\leq l\leq L}\{H_r(\bar{c}_{kpi})-H_r(\underline{c}_{kpi})\}.\end{aligned} \tag{3.130}$$

By Lemma 3.A.2, the above can be further bounded as follows:

$$\begin{aligned}\mathsf{E}D_q \leq & \mathsf{E}\sum_{i=1}^{n}|v_{in}|^{2^q}C(K^{-1}+\|\dot{g}_{in}^0\|K^{-1/2}n^{1/4-\eta}+\|\dot{g}_{in}^0\|\delta+\\ &+\|\dot{g}_{in}^0\|^2n^{1/2-2\eta}+n^{1/2-2\eta}\ddot{g}_{in}^s+\delta^2\ddot{g}_{in}^s+n^{1/4-\eta}2\delta\ddot{g}_{in}^s).\end{aligned} \tag{3.131}$$

Noting further that $K\sim n^{1/2}/\delta$, simplifies the bound to:

$$\begin{aligned}\mathsf{E}D_q \leq & \mathsf{E}\sum_{i=1}^{n}|v_{in}|^{2^q}C(\delta n^{-1/2}+\delta^{1/2}\|\dot{g}_{in}^0\|n^{-\eta}+\|\dot{g}_{in}^0\|\delta+\\ &+\|\dot{g}_{in}^0\|^2n^{1/2-2\eta}+n^{1/2-2\eta}\ddot{g}_{in}^s+\delta^2\ddot{g}_{in}^s+n^{1/4-\eta}2\delta\ddot{g}_{in}^s)\end{aligned} \tag{3.132}$$

By Assumption 3.A.4/1c: fFor all $1\leq q\leq r$: $\mathsf{E}\sum_{i=1}^{n}|v_{in}|^{2^q}=O(n)$. By Assumption 3.A.4/2a: $\mathsf{E}|v_{in}|^{2^q}\|N'\dot{g}(x_i;\beta_0)\|=O(n^{3/4-\eta})$; by Assumption 3.A.4/2b: $\mathsf{E}|v_{in}|^{2^q}\|N'\dot{g}(x_i;\beta_0)\|^2=O(n^{-1/4+\eta})$;

and by Assumption 3.A.4/2c: $\mathsf{E}|v_{in}|^{2^q}\ddot{g}_{in}^s = O(n^{2\eta})$. Thus the bound becomes:

$$\mathsf{E}D_q = O(n^{1/2}) + O(n^{3/4-2\eta}) + O(n^{3/4-\eta}) + O(n^{3/4-\eta}) + O(n^{1/2}) + O(n^{2\eta}) + O(n^{1/4+\eta}) \tag{3.133}$$

Since $\eta < 1/4$, the terms are at most $O(n^{3/4-\eta})$ and $\mathsf{E}D_q = O(n^{3/4-\eta})$. Thus, $\zeta = 3/4 - \eta$. With $\nu = 1/2$ and $\lambda = 1/2 + (1/4 - \eta)\text{dim}b_1$ we check the following conditions in b):

i) $\zeta < 2\nu$, and

ii) $\zeta + \lambda < \nu 2^r$

6.2.1. Checking condition bi).

$\zeta = 3/4 - \eta < 1 = 2 * 1/2 = 2\nu$. Condition satisfied since $\eta < 1/4$.

6.2.2. Checking condition bii).

For condition bii) we require that:

$$\begin{aligned} &\zeta + \lambda = 3/4 - \eta + 1/2 + (1/4 - \eta)\text{dim}b_1 = \\ &= 1 + (1/4 - \eta)(\text{dim}b_1 + 1) < \\ &< \nu 2^r = \frac{1}{2}2^r = 2^{r-1} \end{aligned} \tag{3.134}$$

Condition bii) is satisfied since $2^{r-1} > 1 + (1/4 - \eta)(\text{dim}b_1)$ under Assumption 3.A.4.

6.3. Thus, by 6.1 and 6.2, the term $\tilde{M}_k := n^{-1/2}\sum_{i=1}^n \{z_{kpi} - \mathsf{E}_{i-1}z_{kpi}\} = o_p(1)$.

7. The compensator.

We wish to show that

$$\max_k \max_p \bar{M}_{kp} := \max_k \max_p 2n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}z_{kpi} = o_p(1). \tag{3.135}$$

The bound on $\max_k \max_p n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}z_{kpi}$ will be found by first deriving the bound on $\max_k \max_p n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}z_{kpi}^{2^q}$ and then considering the special case of $q = 0$.

Define: $V_{kpi} := n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}z_{kpi}^{2^q}$. Then, by definition of z_{kpi}:

$$\begin{aligned} \max_k \max_p V_{kpi} &= \max_k \max_p n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}z_{kpi}^{2^q} = \\ &= \max_k \max_p n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}|v_{in}|^{2^q}|h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^q}\{1_{(\varepsilon_i/\sigma \le \bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma \le \underline{c}_{kpi})}\}^{2^q}. \end{aligned} \tag{3.136}$$

Proceeding analogously to the derivations above, we note that for $0 \le q \le r$, the inequality

$|u|^q < 1 + |u|^r$, for all q such that $0 \leq q \leq r$, allows us to obtain the bound:

$$\max_k \max_p V_{kpi} < n^{-1/2} \sum_{i=1}^{n} \mathsf{E}_{i-1} \left\{ |v_{in}|^{2^q} \left(1 + |h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^r} \right) \{1_{(\varepsilon_i/\sigma \leq \bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma \leq \underline{c}_{kpi})}\}^{2^q} \right\}. \tag{3.137}$$

Since $v_{in} \subset \mathcal{F}_{i-1}$, functions of it can be passed through the conditional expectation:

$$\max_k \max_p V_{kpi} < n^{-1/2} \sum_{i=1}^{n} |v_{in}|^{2^q} \mathsf{E}_{i-1} \left(1 + |h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^r} \right) \{1_{(\varepsilon_i/\sigma \leq \bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma \leq \underline{c}_{kpi})}\}^{2^q}. \tag{3.138}$$

Noting that the term $\{1_{(\varepsilon_i/\sigma \leq \bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma \leq \underline{c}_{kpi})}\}$ can only be either 0 or 1 and is, therefore, unaffected by the power of 2^q, further simplifies the bound to

$$\max_k \max_p V_{kpi} < \max_k \max_p n^{-1/2} \sum_{i=1}^{n} |v_{in}|^{2^q} \mathsf{E}_{i-1} \left(1 + |h\left(\frac{\varepsilon_i}{\sigma}\right)|^{2^r} \right) \{1_{(\varepsilon_i/\sigma \leq \bar{c}_{kpi})} - 1_{(\varepsilon_i/\sigma \leq \underline{c}_{kpi})}\}. \tag{3.139}$$

Thus, by definition of $H_r(c)$:

$$\max_k \max_p V_{kpi} < \max_k \max_p n^{-1/2} \sum_{i=1}^{n} |v_{in}|^{2^q} \{H_r(\bar{c}_{kpi}) - H_r(\underline{c}_{kpi})\}. \tag{3.140}$$

Using the definition of V_{kpi} and noting that the maximum of a sum is no greater than the sum of maxima gives:

$$\max_k \max_p n^{-1/2} \sum_{i=1}^{n} \mathsf{E}_{i-1} z_{kpi}^{2^q} < n^{-1/2} \sum_{i=1}^{n} |v_{in}|^{2^q} \max_k \max_p \{H_r(\bar{c}_{kpi}) - H_r(\underline{c}_{kpi})\}. \tag{3.141}$$

Applying Lemma 3.A.2 to bound $\max_k \max_p \{H_r(\bar{c}_{kpi}) - H_r(\underline{c}_{kpi})\}$, gives:

$$\begin{aligned} \max_k \max_p n^{-1/2} \sum_{i=1}^{n} \mathsf{E}_{i-1} z_{kpi}^{2^q} < n^{-1/2} \sum_{i=1}^{n} |v_{in}|^{2^q} C(K^{-1} + \|\dot{g}_{in}^0\| K^{-1/2} n^{1/4-\eta} + \|\dot{g}_{in}^0\|\delta + \\ + \|\dot{g}_{in}^0\|^2 n^{1/2-2\eta} + n^{1/2-2\eta} \ddot{g}_{in}^s + \delta^2 \ddot{g}_{in}^s + n^{1/4-\eta} 2\delta \ddot{g}_{in}^s). \end{aligned} \tag{3.142}$$

In particular, for $q = 0$, and noting that $K \sim n^{1/2}/\delta$:

$$\begin{aligned} &\max_k \max_p n^{-1/2} \sum_{i=1}^{n} \mathsf{E}_{i-1} z_{kpi} \\ &< n^{-1/2} \sum_{i=1}^{n} |v_{in}| C(\delta n^{-1/2} + \delta^{1/2} \|\dot{g}_{in}^0\| n^{-\eta} + \|\dot{g}_{in}^0\|\delta + \|\dot{g}_{in}^0\|^2 n^{1/2-2\eta} + n^{1/2-2\eta} \ddot{g}_{in}^s + \delta^2 \ddot{g}_{in}^s + n^{1/4-\eta} 2\delta \ddot{g}_{in}^s) \end{aligned} \tag{3.143}$$

Under Assumption 3.A.4/1d: $\sum_{i=1}^{n}|v_{in}| = O_p(n)$. Further, under Assumption 3.A.4/2d: $\sum_{i=1}^{n}|v_{in}|\|N'\dot{g}(x_i;\beta_0)\| = O_p(n^{1/2})$; under Assumption 3.A.4/2e: $\sum_{i=1}^{n}|v_{in}|\|N'\dot{g}(x_i;\beta_0)\|^2 = o_p(n^{2\eta})$; and under Assumption 3.A.4/2f: $\sum_{i=1}^{n}|v_{in}|\ddot{g}_{in}^{s} = o_p(n^{2\eta})$. Thus:

$$\begin{aligned}
&\max_k \max_p n^{-1/2}\sum_{i=1}^{n}\mathsf{E}_{i-1}z_{kpi} = \\
&= \delta O_p(1) + \delta^{1/2}O_p(n^{-\eta}) + \delta O_p(1) + o_p(1) + o_p(1) + \delta^2 o_p(n^{-1/2+2\eta}) + \delta o_p(n^{-1/4+\eta}) \\
&= \delta O_p(1).
\end{aligned} \tag{3.144}$$

Analogously to the proof of Lemma 4.1 of Berenguer-Rico et. al. (2019a), in order to show that the term is $o_p(1)$ we exploit the fact that δ remains to be chosen, as detailed in part 2. of this proof. By definition of stochastic boundedness: $\forall\epsilon > 0, \exists M > 0 : P(\bar{M}_k \geq \delta M) \leq \epsilon$ for all $n > N_\epsilon$. For convergence in probability we require that $\forall\epsilon,\gamma > 0 : P(\bar{M}_n \geq \gamma) \leq \epsilon$ for all $n > N_{\epsilon,\gamma}$. Note that for each $\gamma > 0$ we can choose $\delta = \gamma/M$ such that $P(\bar{M}_k \geq \gamma) \leq \epsilon$. Thus, the definition of convergence in probability is satisfied and the term is $o_p(1)$.

8. *Conclusion.*

Steps 5., 6. and 7. imply the following, which completes the proof:

$$\sup_c \sup_\theta n^{-1/2}\sum_{i=1}^{n}\left|v_{in}h\left(\frac{\varepsilon_i}{\sigma}\right)\left\{1_{\left(\varepsilon_i \leq \sigma c + b_1'\dot{g}_{in}^0 + \frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1\right)} - 1_{\left(\varepsilon_i \leq \sigma c + b_1'\dot{g}_{in}^0\right)}\right\}\right| = o_p(1). \tag{3.145}$$

■

Assumption 3.A.5. Assume the following:

1. *The regressors and the function $g(\cdot)$:*
   (a) $\sum_{i=1}^{n}|w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{2\eta})$
2. *The mark function $m(u)$:*
   (a) $\sup_{u\in\mathbb{R}} |m(u)|\, f(u) < \infty$

Lemma 3.A.4. Under Assumption 3.A.5, the following holds:

$$\sup_u \sup_\theta n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}\right||m(u)|\, f(u) = o_p(1). \tag{3.146}$$

Proof. Denote the term of interest as $V_n(\theta,u)$:

$$V_n(\theta,u) := \sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}\right||m(u)|\, f(u) = o_p(1). \tag{3.147}$$

Bound $|m(u)| f(u)$ by its supremum over $u \in \mathbb{R}$:

$$\sup_u \sup_\theta V_n(\theta, u) \leq \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \left| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \right| \sup_{u \in \mathbb{R}} |m(u)| f(u). \tag{3.148}$$

By Assumption 3.A.5/2a, $\sup_{u \in \mathbb{R}} |m(u)| f(u) = C_1$ where $C_1 < \infty$ is a constant which does not depend on n. Using the definition of $\ddot{g}_{in}^{1*}$ gives:

$$\sup_u \sup_\theta V_n(\theta, u) \leq C_1 \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \left| \frac{\frac{1}{2} b_1' N' \ddot{g}(x_i; \beta_{1i}^*) N b_1}{\sigma} \right|. \tag{3.149}$$

Taking the supremum of $\|N' \ddot{g}(x_i, \beta_i) N\|$ over $\left\{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}\right\}$; and taking absolute values or norms of individual terms gives:

$$\sup_u \sup_\theta V_n(\theta, u) \leq C_1 n^{-1/2} \sum_{i=1}^n |w_{in}| \sup_{b_1} |b_1'| \frac{1}{2} \frac{1}{\sigma} \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| \sup_{b_1} |b_1|. \tag{3.150}$$

Take the supremum of b_1 over $\left\{\theta : |\theta| \leq B n^{1/4-\eta}\right\}$ to obtain

$$\sup_u \sup_\theta V_n(\theta, u) \leq C_2 n^{-2\eta} \sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\|, \tag{3.151}$$

where $C_2 := C_1 \frac{1}{2} \frac{1}{\sigma} B^2 < \infty$ is a constant which does not depend on n. Further, by Assumption 3.A.5/1a, $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{2\eta})$. Thus:

$$\sup_{u \in \mathbb{R}} \sup_\theta V_n(\theta, u) \leq o_p\left(n^{-2\eta+2\eta}\right) = o_p(1) \tag{3.152}$$

■

Assumption 3.A.6. We assume the following:

1. *The regressors and the function $g(\cdot)$:*
   (a) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{2\eta})$
   (b) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}} \|N' \dot{g}(x_i; \beta_0)\| \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{-1/4+3\eta})$
2. *The mark function $m(u)$:*
   (a) $\sup_{u \in \mathbb{R}} (1 + |u|) |\dot{m}(u)| f(u) < \infty$

Lemma 3.A.5. Under Assumption 3.A.6, the following holds:

$$\sup_u \sup_\theta n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \right| \left| u\left(n^{-1/2} \frac{1}{s} \frac{1}{\sigma} a_m \right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^{0} \right| |\dot{m}(u)| f(u) = o_p(1). \quad (3.153)$$

Proof. For the purposes of this proof, define the term of interest $V_n(\theta, u)$ as follows:

$$V_n(\theta, u) := n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \right| \left| u\left(n^{-1/2} \frac{1}{s} \frac{1}{\sigma} a_m \right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^{0} \right| |\dot{m}(u)| f(u). \quad (3.154)$$

By the triangle inequality and noting that the sum of the suprema is no smaller than the supremum of a sum:

$$\sup_u \sup_\theta V_n(\theta, u) \leq \sup_u \sup_\theta V_{1,n}(\theta, u) + \sup_u \sup_\theta V_{2,n}(\theta, u), \quad (3.155)$$

where $V_{1,n}(\theta, u)$ and $V_{2,n}(\theta, u)$ are defined as follows.

$$\begin{aligned} V_{1,n}(\theta, u) &:= n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \right| \left| u\left(n^{-1/2} \frac{1}{s} \frac{1}{\sigma} a_m \right) \right| |\dot{m}(u)| f(u) \\ V_{2,n}(\theta, u) &:= n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \right| \left| \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^{0} \right| |\dot{m}(u)| f(u) \end{aligned} \quad (3.156)$$

It suffices to show that both terms $V_{1,n}(\theta, u)$ and $V_{2,n}(\theta, u)$ are $o_p(1)$, uniformly in u and c.

Proof for $V_{1,n}(\theta, u)$.

We note that only the function $|u| |\dot{m}(u)| f(u)$ depends on u, so the supremum over u can be taken inside the sum:

$$\sup_u \sup_\theta V_{1,n}(\theta, u) = \sup_\theta n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \right| \left| n^{-1/2} \frac{1}{s} \frac{1}{\sigma} a_m \right| \sup_u |u| |\dot{m}(u)| f(u). \quad (3.157)$$

Simplify the bound using Assumption 3.A.6/2a, $\sup_u |u| |\dot{m}(u)| f(u) = C_1 < \infty$ where C_1 is a finite constant which does not depend on n, which gives:

$$\sup_u \sup_\theta V_{1,n}(\theta, u) = \sup_\theta n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \right| n^{-1/2} \frac{1}{\sigma} \left| \frac{1}{s} a_m \right| C_1. \quad (3.158)$$

Following Lemma 3.A.1 and Remark 3.A.1, for sufficiently large n: $\left|\frac{1}{s} a_m\right| < 2 |a_m|$. Hence

$$\sup_u \sup_\theta V_{1,n}(\theta, u) < \sup_\theta n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \right| n^{-1/2} \frac{1}{\sigma} |a_m| 2C_1. \quad (3.159)$$

Use the definition of $\ddot{g}_{in}^{m*}$, and take the supremum of $\|N'\ddot{g}(x_i,\beta_i)N\|$ over β_i such that $\|N^{-1}(\beta_i-\beta_0)\| \leq Bn^{1/4-\eta}$. The constant $\frac{1}{\sigma^2}$ has been incorporated into C_1.

$$\sup_u \sup_\theta V_{1,n}(\theta,u) < n^{-1/2}\sum_{i=1}^{n}|w_{in}|\sup_\theta|b_1|'\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|\sup_\theta|b_1|n^{-1/2}\sup_\theta|a_m|\,C_1. \tag{3.160}$$

Note that $|a_m| \leq Bn^{1/4-\eta}$ and $|b_1| \leq Bn^{1/4-\eta}$. Let$C_2 := 2C_1\frac{1}{2}\frac{1}{\sigma}B^3$, where $C_2 < \infty$ is a constant which does not depend on n. Hence:

$$\begin{aligned}\sup_u \sup_\theta V_{1,n}(\theta,u) &< C_2 n^{-1/2}\sum_{i=1}^{n}|w_{in}|n^{1/4-\eta}\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|n^{1/4-\eta}n^{-1/2}n^{1/4-\eta}\\ &= C_2 n^{-1/4-3\eta}\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|\end{aligned} \tag{3.161}$$

By Assumption 3.A.6/1a: $\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{2\eta})$. Hence:

$$\sup_u \sup_\theta V_{1,n}(\theta,u) = o_p\left(n^{-1/4-3\eta+2\eta}\right) = o_p\left(n^{-1/4-\eta}\right). \tag{3.162}$$

Thus, the first term of equation (3.155) is at most $o_p(1)$.

Proof for $V_{2,n}(\theta,u)$.

Note that in the term $V_{2,n}(\theta,u)$ only the function $|\dot{m}(u)|\,f(u)$ depends on u so the supremum over u can be taken inside the summation:

$$\sup_u \sup_\theta V_{2,n}(\theta,u) = \sup_\theta n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}\right|\left|\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^{0}\right|\sup_u|\dot{m}(u)|\,f(u). \tag{3.163}$$

By Assumption 3.A.6/2a, $\sup_u|\dot{m}(u)|\,f(u) = C_3$, where $C_3 < \infty$ does not depend on n. Hence:

$$\sup_u \sup_\theta V_{2,n}(\theta,u) = \sup_\theta n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}\right|\left|\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^{0}\right|C_3. \tag{3.164}$$

Following Lemma 3.A.1 and Remark 3.A.1, for sufficiently large n: $\left|\frac{1}{s}b_m'\dot{g}_{in}^{0}\right| < 2\left|b_m'\dot{g}_{in}^{0}\right|$. Thus:

$$\sup_u \sup_\theta V_{2,n}(\theta,u) < \sup_\theta n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}\right|\left|\frac{1}{\sigma}b_m'\dot{g}_{in}^{0}\right|2C_3. \tag{3.165}$$

Using the definitions of $\dot{g}_{in}^{0}$ and $\ddot{g}_{in}^{m*}$, and taking the supremum of $\|N'\ddot{g}(x_i,\beta_i)N\|$ over β_i such that $\|N^{-1}(\beta_i-\beta_0)\| \leq Bn^{1/4-\eta}$ gives the following, where constants have been collected together

in $C_4 < \infty$, which does not depend on n:

$$\sup_u \sup_\theta V_{2,n}(\theta, u) < n^{-1/2} \sum_{i=1}^n |w_{in}| \sup_\theta |b_1|' \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| \sup_\theta |b_1| \sup_\theta |b_m|' \left| N' \dot{g}(x_i; \beta_0) \right| C_4. \tag{3.166}$$

Note that $|b_1| \leq B n^{1/4-\eta}$ and $|b_m| \leq B n^{1/4-\eta}$, hence, for a constant $C < \infty$:

$$\sup_u \sup_\theta V_{2,n}(\theta, u) < C n^{1/4-3\eta} \sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| \left| N' \dot{g}(x_i; \beta_0) \right|. \tag{3.167}$$

By Assumption 3.A.6/1b: $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \dot{g}(x_i; \beta_0)\| \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{-1/4+3\eta})$. Thus:

$$\sup_u \sup_\theta V_{2,n}(\theta, u) < C n^{1/4-3\eta} \sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| \left| N' \dot{g}(x_i; \beta_0) \right| = o_p(1). \tag{3.168}$$

Thus, the second term in equation (3.155) is $o_p(1)$. This completes the proof.

■

Assumption 3.A.7. We assume the following:

1. *The regressors and the function $g(\cdot)$:*
   (a) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{2\eta})$
2. *The mark function $m(u)$:*
   (a) $\int_{-\infty}^{\infty} |\dot{m}(u)| f(u) du < \infty$

Lemma 3.A.6. Under Assumption 3.A.7, the following holds:

$$\sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} \frac{|b_m' \ddot{g}_{in}^{m*} b_m|}{\sigma} \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right) = o_p(1). \tag{3.169}$$

Proof. For the purposes of this proof, define the quantity of interest as $V_n(\theta)$:

$$V_n(\theta) := n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} \frac{|b_m' \ddot{g}_{in}^{m*} b_m|}{\sigma} \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.170}$$

Using the definition of $\ddot{g}_{in}^{m*}$ obtain:

$$\sup_\theta V_n(\theta) = \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} \frac{|b_m'| \|N' \ddot{g}(x_i, \beta_{mi}^*) N\| |b_m|}{\sigma} \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.171}$$

Taking the supremum over β_i allows us to take the term $\|N'\ddot{g}(x_i, \beta^*_{mi})N\|$ outside of the integral, as it no longer depends on ε_i/σ:

$$\sup_\theta V_n(\theta) \leq \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \frac{1}{\sigma} |b'_m| \sup_{\beta_i} \|N'\ddot{g}(x_i, \beta_i)N\| |b_m| \int_{-\infty}^{\infty} \left|\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.172}$$

Note that $\int_{-\infty}^{\infty} \left|\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right) = C < \infty$ under Assumption 3.A.7/2a. Further, incorporate $\frac{1}{\sigma}$ into the constant, to obtain the bound:

$$\sup_\theta V_n(\theta) \leq \sup_\theta C n^{-1/2} \sum_{i=1}^n |w_{in}| |b'_m| \sup_{\beta_i} \|N'\ddot{g}(x_i, \beta_i)N\| |b_m|. \tag{3.173}$$

The supremum with respect to θ is taken over $|\theta| < Bn^{1/4-\eta}$. Thus:

$$\sup_\theta V_n(\theta) \leq C n^{-1/2} n^{1/2-2\eta} \sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N'\ddot{g}(x_i, \beta_i)N\|, \tag{3.174}$$

where B^2 has been incorporated into the constant C. By Assumption 3.A.7/1a: $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N'\ddot{g}(x_i; \beta_i)N\| = o_p(n^{2\eta})$. Thus:

$$\sup_\theta V_n(\theta) = o_p(n^{-2\eta+2\eta}) = o_p(1), \tag{3.175}$$

which completes the proof.

■

Assumption 3.A.8. We assume the following:

1. *The regressors and the function* $g(\cdot)$, for $0 < \eta < 1/4$ and $\kappa > 0$:
   (a) $\max_{1\leq i\leq n} \|N'\dot{g}(x_i; \beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$
   (b) $\max_{1\leq i\leq n} \sup_{\beta_i : \|N^{-1}(\beta_i-\beta_0)\| \leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i; \beta_i)N\| = O_p(n^{-1/2+2\eta-\kappa})$
   (c) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i-\beta_0)\| \leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i; \beta_i)N\| = o_p(n^{2\eta})$
   (d) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i-\beta_0)\| \leq Bn^{1/4-\eta}} \|N'\ddot{g}(x_i; \beta_i)N\|^2 = o_p(n^{-1/2+4\eta})$
   (e) $\sum_{i=1}^n |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i-\beta_0)\| \leq Bn^{1/4-\eta}} \|N'\dot{g}(x_i; \beta_0)\| \|N'\ddot{g}(x_i; \beta_i)N\| = o_p(n^{-1/4+3\eta})$
2. *The mark function* $m(u)$:

(a) (local Lipschitz) $\exists \omega > 0$ and a function $\mathring{m}(u) \geq 0$ such that if $\forall u^*, u$ such that $|u^* - u| \leq (1 + |u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u|\mathring{m}(u)$ and where:

- $\sup_{u \in \mathbb{R}} (1 + |u|)\mathring{m}(u) f(u) < \infty$
- $\int_{-\infty}^{+\infty} (1 + |u|)\mathring{m}(u) f(u) du < \infty \iff \mathsf{E}(1 + |u|)\mathring{m}(u) < \infty$.

Lemma 3.A.7. Under Assumption 3.A.8, the following holds for an intermediate point $\bar{g}_m$ such that $|\bar{g}_m| \leq |\ddot{g}_{in}^{m*}|$.

$$\sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \left| \left(\frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \left\{ \dot{m} \left(\frac{\varepsilon_i - b'_m \dot{g}_{in}^0 - \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right\} \right| f \left(\frac{\varepsilon_i}{\sigma} \right) d \left(\frac{\varepsilon_i}{\sigma} \right) = o_p(1). \tag{3.176}$$

Proof. Define the quantity of interest as $V_n(\theta)$:

$$V_n(\theta) := n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \left| \left(\frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \left\{ \dot{m} \left(\frac{\varepsilon_i - b'_m \dot{g}_{in}^0 - \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right\} \right| f \left(\frac{\varepsilon_i}{\sigma} \right) d \left(\frac{\varepsilon_i}{\sigma} \right) \tag{3.177}$$

We wish to show that $\sup_\theta V_n(\theta) = o_p(1)$. The proof relies on the local Lipschitz condition. Let: $u_i^* := \frac{\varepsilon_i - b'_m \dot{g}_{in}^0 - \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma + n^{-1/2} a_m}$ and $u_i := \frac{\varepsilon_i}{\sigma}$. Thus:

$$|u_i^* - u_i| = \left| \frac{\varepsilon_i - b'_m \dot{g}_{in}^0 - \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} - \frac{\varepsilon_i}{\sigma} \right|. \tag{3.178}$$

As given in Definition 3.A.1, the scale s is defined as $s := 1 + n^{-1/2} a_m / \sigma$. Hence:

$$|u_i^* - u_i| = \left| \frac{\varepsilon_i}{\sigma} \left(1 - n^{-1/2} \frac{1}{s} a_m / \sigma \right) - \frac{\frac{1}{s} b'_m \dot{g}_{in}^0 + \frac{1}{s} \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma} - \frac{\varepsilon_i}{\sigma} \right| \tag{3.179}$$

By triangle inequality and Lemma 3.A.1 and Remark 3.A.1, for sufficiently large n $|\frac{1}{s}| < 2$:

$$\begin{aligned} |u_i^* - u_i| &\leq \left| \frac{\varepsilon_i}{\sigma} \right| \left| n^{-1/2} \frac{1}{s} a_m / \sigma \right| + \left| \frac{\frac{1}{s} b'_m \dot{g}_{in}^0}{\sigma} \right| + \left| \frac{\frac{1}{s} \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma} \right| \\ &< \left| \frac{\varepsilon_i}{\sigma} \right| 2 \left| n^{-1/2} a_m / \sigma \right| + 2 \left| \frac{b'_m \dot{g}_{in}^0}{\sigma} \right| + 2 \left| \frac{\frac{1}{2} b'_m \bar{g}_m b_m}{\sigma} \right|. \end{aligned} \tag{3.180}$$

Note that $\bar{g}_m$ is a point such that $|\bar{g}_m| \leq |\ddot{g}_{in}^{m*}|$, thus the above can be bounded by:

$$|u_i^* - u_i| < \left| \frac{\varepsilon_i}{\sigma} \right| 2 \left| n^{-1/2} a_m / \sigma \right| + 2 \left| \frac{b'_m \dot{g}_{in}^0}{\sigma} \right| + 2 \left| \frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \right|. \tag{3.181}$$

Use the definitions of $\dot{g}_{in}^0$ and $\ddot{g}_{in}^{m*}$:

$$|u_i^* - u_i| < \left|\frac{\varepsilon_i}{\sigma}\right| 2\left|n^{-1/2}a_m/\sigma\right| + 2\frac{|b_m'|\|N'\dot{g}(x_i,\beta_0)\|}{\sigma} + \frac{|b_m'|\|N'\ddot{g}(x_i,\beta_{mi}^*)N\||b_m|}{\sigma}. \tag{3.182}$$

Bound values by their maximum over i and supremum over $\theta : |\theta| < Bn^{1/4-\eta}$, which gives:

$$|u_i^* - u_i| < \left|\frac{\varepsilon_i}{\sigma}\right| 2\sup_\theta\left|n^{-1/2}a_m/\sigma\right| + 2\sup_\theta\max_i\frac{|b_m'|\|N'\dot{g}(x_i,\beta_0)\|}{\sigma} + \sup_\theta\max_i\frac{|b_m'|\|N'\ddot{g}(x_i,\beta_{mi}^*)N\||b_m|}{\sigma}. \tag{3.183}$$

Consider the sets $S_{1,n}$, $S_{2,n}$ and S_n, as defined in Definition 3.A.2. By Remark 3.A.2, under Assumptions 3.A.8/1a and 3.A.8/1b, which are equivalent to Assumption 3.A.2, it follows that: $P(S_n^c) < \epsilon_1 + \epsilon_2$. Following Remark 3.A.2, since $P(S_n^c) < \epsilon_1 + \epsilon_2$, it suffices to consider the quantities on the set S_n. Thus, the bound can be rewritten as:

$$\begin{aligned}
|u_i^* - u_i| &= \left|\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m} - \frac{\varepsilon_i}{\sigma}\right| \\
&< \left\{\left|\frac{\varepsilon_i}{\sigma}\right| 2\sup_\theta\left|n^{-1/2}a_m/\sigma\right| + 2\sup_\theta\max_i\frac{|b_m'|\|N'\dot{g}(x_i,\beta_0)\|}{\sigma} + \sup_\theta\max_i\frac{|b_m'|\|N'\ddot{g}(x_i,\beta_{mi}^*)N\||b_m|}{\sigma}\right\}1_{S_n} + \\
&+ \left\{\left|\frac{\varepsilon_i}{\sigma}\right| 2\sup_\theta\left|n^{-1/2}a_m/\sigma\right| + 2\sup_\theta\max_i\frac{|b_m'|\|N'\dot{g}(x_i,\beta_0)\|}{\sigma} + \sup_\theta\max_i\frac{|b_m'|\|N'\ddot{g}(x_i,\beta_{mi}^*)N\||b_m|}{\sigma}\right\}1_{S_n^c}.
\end{aligned} \tag{3.184}$$

By definition of the set S_n, the following holds deterministically on the set S_n: $\sup_\theta \max_i\|N'\dot{g}(x_i,\beta_0)\| = O(n^{-1/4+\eta-\kappa})$ and $\sup_\theta\sup_{\beta_i}\max_i\|N'\ddot{g}(x_i,\beta_{mi}^*)N\| = O(n^{-1/4+\eta-\kappa})$. The supremum with respect to θ is taken over $|\theta| \le Bn^{1/4-\eta}$. Therefore, it follows that on the set S_n: $\sup_\theta\max_i\frac{|b_m'|\|N'\dot{g}(x_i,\beta_0)\|}{\sigma} = O(n^{-\kappa})$; and $\sup_\theta\sup_{\beta_i}\max_i\frac{|b_m'|\|N'\ddot{g}(x_i,\beta_{mi}^*)N\||b_m|}{\sigma} = O(n^{-\kappa})$. Further, $\sup_\theta\left|n^{-1/2}a_m/\sigma\right| = O(n^{-1/4-\eta})$. Therefore, on the set S_n:

$$\begin{aligned}
|u_i^* - u_i| &< \left|\frac{\varepsilon_i}{\sigma}\right| 2\sup_\theta\left|n^{-1/2}a_m/\sigma\right| + 2\sup_\theta\max_i\frac{|b_m'|\|N'\dot{g}(x_i,\beta_0)\|}{\sigma} + \sup_\theta\max_i\frac{|b_m'|\|N'\ddot{g}(x_i,\beta_{mi}^*)N\||b_m|}{\sigma} = \\
&= \left|\frac{\varepsilon_i}{\sigma}\right| C_{1,n} + C_{2,n} = |u_i|\, C_{1,n} + C_{2,n},
\end{aligned} \tag{3.185}$$

where $C_{1,n} = O(n^{-1/4-\eta})$ and $C_{2,n} = O(n^{-\kappa})$. Thus, on the set S_n the local Lipschitz condition (Assumption 3.A.8/2) holds, which implies the following bound:

$$\left|\dot{m}\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m}\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| < \left|\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m} - \frac{\varepsilon_i}{\sigma}\right| \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|, \tag{3.186}$$

where $\sup_u(1+u)\mathring{m}(u)f(u) < \infty$ and $\int_{-\infty}^{+\infty}(1+|u|)\mathring{m}(u)f(u)du < \infty$. Thus, by the triangle

inequality and the Lipschitz condition, the quantity in the lemma can be bounded on the set S_n as follows:

$$\sup_\theta V_n(\theta) \leq \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} \left| \frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right| \left| \frac{\varepsilon_i - b'_m \dot{g}_{in}^0 - \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} - \frac{\varepsilon_i}{\sigma} \right| \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right) \tag{3.187}$$

Use Definition 3.A.1 of the scale as well as the notation $a_n^s := n^{-1/2} a_m / \sigma$ to simplify the bound. Further, use the triangle inequality, and the bound on the term $\left| \frac{\varepsilon_i - b'_m \dot{g}_{in}^0 - \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} - \frac{\varepsilon_i}{\sigma} \right|$ given in equation (3.180), to obtain the further bound:

$$\begin{aligned} \sup_\theta V_n(\theta) \leq & \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} \left| \frac{\frac{1}{s} \frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \right| \cdot \\ & \cdot \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| \left| \frac{1}{s} a_n^s \right| + \left| \frac{\frac{1}{s} b'_m \dot{g}_{in}^0}{\sigma} \right| + \left| \frac{\frac{1}{s} \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma} \right| \right\} \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \end{aligned} \tag{3.188}$$

Using Lemma 3.A.1 and Remark 3.A.1 the above can be further bounded by:

$$\begin{aligned} \sup_\theta V_n(\theta) < & \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} 2 \left| \frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \right| \cdot \\ & \cdot \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| 2 \left| a_n^s \right| + 2 \left| \frac{b'_m \dot{g}_{in}^0}{\sigma} \right| + 2 \left| \frac{\frac{1}{2} b'_m \bar{g}_m b_m}{\sigma} \right| \right\} \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \end{aligned} \tag{3.189}$$

Note that $\bar{g}_m$ is a point such that $|\bar{g}_m| \leq |\ddot{g}_{in}^{m*}|$, hence:

$$\begin{aligned} \sup_\theta V_n(\theta) < & \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} 4 \left| \frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \right| \cdot \\ & \cdot \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| \left| a_n^s \right| + \left| \frac{b'_m \dot{g}_{in}^0}{\sigma} \right| + \left| \frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \right| \right\} \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \end{aligned} \tag{3.190}$$

Collect the terms and note that a_n^s does not depend on i. Note that the supremum of a sum is no greater than the sum of the suprema, thus:

$$\begin{aligned} \sup_\theta V_n(\theta) < & \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| 4 \left| a_n^s \right| \int_{-\infty}^{\infty} \frac{\left| \frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m \right|}{\sigma} \left| \frac{\varepsilon_i}{\sigma} \right| \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right) + \\ & + \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| 4 \int_{-\infty}^{\infty} \frac{\left| \frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m \right|}{\sigma} \left(\frac{|b'_m \dot{g}_{in}^0|}{\sigma} + \frac{\left| \frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m \right|}{\sigma} \right) \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \end{aligned} \tag{3.191}$$

Define:

$$V_{1,n}(\theta) := \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| 4 \left|a_n^s\right| \int_{-\infty}^{\infty} \frac{|\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m|}{\sigma} \left|\frac{\varepsilon_i}{\sigma}\right| \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right)$$
$$V_{2,n}(\theta) := \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| 4 \int_{-\infty}^{\infty} \frac{|\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m|}{\sigma} \left(\frac{|b_m' \dot{g}_{in}^0|}{\sigma} + \frac{|\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m|}{\sigma}\right) \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right) \tag{3.192}$$

In order to prove the lemma, it suffices to show that both terms are $o_p(1)$. We will analyse the two terms in turn.

Term $V_{1,n}(\theta)$.

By definition of $\ddot{g}_{in}^{m*}$ and replacing $|b_m'| \|N' \ddot{g}(x_i; \beta_{mi}^*) N\| |b_m|$ with its supremum over $\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}$ gives:

$$V_{1,n}(\theta) \leq \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| 2 \left|a_n^s\right| \int_{-\infty}^{\infty} \sup_{\beta_i} \frac{|b_m'| \|N' \ddot{g}(x_i; \beta_i) N\| |b_m|}{\sigma} \left|\frac{\varepsilon_i}{\sigma}\right| \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.193}$$

The term $\|N' \ddot{g}(x_i; \beta_i) N\|$ can be taken outside of the integral since the distribution of x_i is independent of that of ε_i/σ, which leads to the following bound:

$$V_{1,n}(\theta) \leq \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| 2 \left|a_n^s\right| \sup_{\beta_i} \frac{|b_m'| \|N' \ddot{g}(x_i; \beta_i) N\| |b_m|}{\sigma} \int_{-\infty}^{\infty} \left|\frac{\varepsilon_i}{\sigma}\right| \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.194}$$

By Assumption 3.A.8/2a, $\int_{-\infty}^{\infty} \left|\frac{\varepsilon_i}{\sigma}\right| \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right) = C < \infty$. Hence, the above bound is equivalent to the following, where the constant 2 has been incorporated into C:

$$V_{1,n}(\theta) \leq C \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \left|n^{-1/2} a_m/\sigma\right| \sup_{\beta_i} \frac{|b_m'| \|N' \ddot{g}(x_i; \beta_i) N\| |b_m|}{\sigma}. \tag{3.195}$$

By Assumption 3.A.8/1c, $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{2\eta})$. Furthermore, the supremum with respect to θ is taken over θ such that $|\theta| \leq B n^{1/4-\eta}$. Thus, $V_{1,n}(\theta) = o_p(n^{-1/4-\eta})$.

Term $V_{2,n}$.

To prove the Lemma, it remains to show that the second term in equation (3.191) is $o_p(1)$:

$$\sup_\theta V_{2,n}(\theta) = \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} \frac{2|\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m|}{\sigma} \left(2\frac{|b_m' \dot{g}_{in}^0|}{\sigma} + 2\frac{|\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m|}{\sigma}\right) \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.196}$$

Bounding the term $\ddot{g}_{in}^{m*}$ by its supremum over $\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq Bn^{1/4-\eta}$, and noting that the supremum of a sum is no greater than the sum of the suprema, gives:

$$\sup_\theta V_{2,n}(\theta) \leq \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} \sup_{\beta_i} \left(\frac{|b_m'| \|N' \ddot{g}(x_i;\beta_i) N\| |b_m|}{\sigma} \right)^2 \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| d\left(\frac{\varepsilon_i}{\sigma}\right) + \\ + \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \int_{-\infty}^{\infty} 2 \left(\sup_{\beta_i} \frac{|b_m'| \|N' \ddot{g}(x_i;\beta_i) N\| |b_m|}{\sigma} \right) \frac{|b_m'| \|N' \dot{g}(x_i;\beta_0) N\|}{\sigma} \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.197}$$

The terms $\sup_{\beta_i} \|N' \ddot{g}(x_i;\beta_i) N\|$ and $|b_m'| \|N' \dot{g}(x_i;\beta_0) N\|$ can be taken outside of the integral since the distribution of x_i is independent of that of $\frac{\varepsilon_i}{\sigma}$; and b_m is a constant vector. This gives:

$$\sup_\theta V_{2,n}(\theta) \leq \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \sup_{\beta_i} \left(\frac{|b_m'| \|N' \ddot{g}(x_i;\beta_i) N\| |b_m|}{\sigma} \right)^2 \int_{-\infty}^{\infty} \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| d\left(\frac{\varepsilon_i}{\sigma}\right) + \\ + \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| 2 \left(\sup_{\beta_i} \frac{|b_m'| \|N' \ddot{g}(x_i;\beta_i) N\| |b_m|}{\sigma} \right) \frac{|b_m'| \|N' \dot{g}(x_i;\beta_0) N\|}{\sigma} \int_{-\infty}^{\infty} \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.198}$$

By Assumption 3.A.8/2a, $\int_{-\infty}^{\infty} \left|\frac{\varepsilon_i}{\sigma}\right| \left| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right) = C < \infty$. Hence, the above bound is equivalent to:

$$\sup_\theta V_{2,n}(\theta) \leq C \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \sup_{\beta_i} \left(\frac{|b_m'| \|N' \ddot{g}(x_i;\beta_i) N\| |b_m|}{\sigma} \right)^2 + \\ + 2C \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \left(\sup_{\beta_i} \frac{|b_m'| \|N' \ddot{g}(x_i;\beta_i) N\| |b_m|}{\sigma} \right) \frac{|b_m'| \|N' \dot{g}(x_i;\beta_0) N\|}{\sigma}. \tag{3.199}$$

Note that the supremum with respect to θ is taken over $|\theta| \leq Bn^{1/4-\eta}$. Hence, incorporating the constants 2 and B into the constant C, gives:

$$\sup_\theta V_{2,n}(\theta) \leq Cn^{-1/2} \sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i;\beta_i) N\|^2 n^{1-4\eta} + \\ + Cn^{-1/2} \sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i;\beta_i) N\| \|N' \dot{g}(x_i;\beta_0) N\| n^{3/4-3\eta}. \tag{3.200}$$

By Assumption 3.A.8/1d, $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i;\beta_i) N\|^2 = o_p(n^{-1/2+4\eta})$. Hence, the first term in equation (3.200) is $o_p(1)$. By Assumption 3.A.8/1e, $\sum_{i=1}^n |w_{in}| \sup_{\beta_i} \|N' \dot{g}(x_i;\beta_0)\| \|N' \ddot{g}(x_i;\beta_i) N\| = o_p(n^{-1/4+3\eta})$. Hence, the second term in equation (3.200) is $o_p(1)$. Thus, $\sup_\theta V_{2,n}(\theta) = o_p(1)$.

■

Assumption 3.A.9. Assume the following:

1. *The regressors and the function $g(\cdot)$:*

   (a) $\sum_{i=1}^{n} |w_{in}|^2 \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq Bn^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\|^2 = o_p(n^{4\eta-1})$

2. *The mark function $m(u)$:*

   (a) $\mathsf{E}\,|\dot{m}(u)|^2 < \infty$

Lemma 3.A.8. Under Assumption 3.A.9, the following holds:

$$\sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| = o_p(1). \tag{3.201}$$

Proof. For the purposes of this proof, define the term of interest:

$$V_n(\theta) := n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| = o_p(1). \tag{3.202}$$

Use Definition 3.A.1 of the scale: $\frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} = \frac{1}{s} \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma}$ to express the bound as:

$$\sup_{\theta} V_n(\theta) = \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \left(\frac{\frac{1}{s} \frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma} \right) \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right|. \tag{3.203}$$

Since the error $a_m = O(n^{1/4-\eta})$, by Lemma 3.A.1 and Remark 3.A.1 the above is bounded by:

$$\sup_{\theta} V_n(\theta) < \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| 2 \frac{\left|\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m\right|}{\sigma} \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right|. \tag{3.204}$$

By definition of $\ddot{g}_{in}^{m*}$:

$$\sup_{\theta} V_n(\theta) < \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \frac{|b_m'| \|N' \ddot{g}(x_i, \beta_{mi}^*) N\| |b_m|}{\sigma} \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right|. \tag{3.205}$$

By the Cauchy-Schwartz inequality and taking $n^{-1/2}$ inside the square root:

$$\sup_{\theta} V_n(\theta) < \sup_{\theta} \left\{ \sum_{i=1}^{n} \left(|w_{in}| \frac{|b_m'| \|N' \ddot{g}(x_i, \beta_{mi}^*) N\| |b_m|}{\sigma} \right)^2 n^{-1} \sum_{i=1}^{n} \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right|^2 \right\}^{1/2}. \tag{3.206}$$

The errors ε_i are independently and identically distributed, that $\mathsf{E}\left|\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|^2 < \infty$, hence it follows by the Law of Large Numbers (e.g. Khinchine's LLN in Greene (2008), p. 1042) that $n^{-1} \sum_{i=1}^{n} \left|\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|^2 \xrightarrow{p} \mathsf{E}\left|\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|^2 < \infty$. Further, noting that $|b_m| \leq Bn^{1/4-\eta}$ for some $0 < B <$

∞, gives the following bound for some constant $C < \infty$:

$$\sup_{\theta} V_n(\theta) < Cn^{1/2-2\eta} \left\{ \sum_{i=1}^{n} |w_{in}|^2 \|N' \ddot{g}(x_i, \beta_{mi}^*) N\|^2 \right\}^{1/2}. \tag{3.207}$$

By Assumption 3.A.9/1a, $\sum_{i=1}^{n} |w_{in}|^2 \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\|^2 = o_p(n^{4\eta-1})$, hence $\sup_{\theta} V_n(\theta) = o_p(1)$. ■

Assumption 3.A.10. Assume the following:

1. *The regressors and the function $g(\cdot)$:*
   (a) $\max_{1 \le i \le n} \|N' \dot{g}(x_i; \beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$
   (b) $\max_{1 \le i \le n} \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \le B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = O_p(n^{-1/2+2\eta-\kappa})$
   (c) $\mathsf{E} \sum_{i=1}^{n} |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \le B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = O(n^{1/4+2\eta})$
   (d) $\mathsf{E} \sum_{i=1}^{n} |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \le B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| \|N' \dot{g}(x_i; \beta_0)\| = O(n^{-1/4+2\eta})$
   (e) $\mathsf{E} \sum_{i=1}^{n} |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \le B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\|^2 = O(n^{-1/2+3\eta})$
2. *The mark function $m(u)$:*
   (a) (local Lipschitz) $\exists \omega > 0$ and a function $\mathring{m}(u) \ge 0$ such that if $\forall u^*, u$ such that $|u^* - u| \le (1 + |u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \le |u^* - u| \mathring{m}(u)$ and where:
      - $\sup_{u \in \mathbb{R}} (1 + |u|) \mathring{m}(u) f(u) < \infty$
      - $\int_{-\infty}^{+\infty} (1 + |u|) \mathring{m}(u) f(u) du < \infty \iff \mathsf{E}(1 + |u|) \mathring{m}(u) < \infty$.

Lemma 3.A.9. Under Assumption 3.A.10, the following holds:

$$\sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \left\{ \dot{m} \left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right\} \right| = o_p(1). \tag{3.208}$$

Proof. For the purposes of this proof, define

$$V_n(\theta) := n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \left\{ \dot{m} \left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right\} \right|. \tag{3.209}$$

For any x, y: $|xy| = |x||y|$, therefore the term is bounded by:

$$V_n(\theta) = n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right| \left| \dot{m} \left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right|. \tag{3.210}$$

Recall Definition 3.A.2 of set S_n, i.e. $S_n := S_{1,n} \cap S_{2,n}$, where $S_{1,n} := (\max_i \|N'\dot{g}(x_i, \beta_0)\| \leq B_1 n^{-1/4+\eta-\kappa})$ and $S_{2,n} := (\max_i \sup_{\beta_i} \|N'\ddot{g}(x_i, \beta_i)N\| \leq B_1 n^{-1/2+2\eta-\kappa})$. The supremum of a sum is no greater than the sum of suprema, hence:

$$\sup_\theta V_n(\theta) \leq \sup_\theta \{V_n(\theta)\} 1_{S_n} + \sup_\theta \{V_n(\theta)\} 1_{S_n^c}. \tag{3.211}$$

Define: $V_n^S(\theta) := \{V_n(\theta)\} 1_{S_n}$. Following Remark 3.A.2 and Assumptions 3.A.10/1a and 3.A.10/1b it suffices to consider the quantity of interest on the set S_n, i.e. it suffices to show that: $\sup_\theta V_n^S(\theta) = o_p(1)$.

To show this, we apply the local Lipschitz condition (Assumption 3.A.10/2a) to the term $\left|\dot{m}\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m}\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|$. First, we investigate the following term, where $u_i := \frac{\varepsilon_i}{\sigma}$ and $u_i^* := \frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m}$:

$$|u_i^* - u_i| = \left|\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m} - \frac{\varepsilon_i}{\sigma}\right|. \tag{3.212}$$

Following reasoning identical to that in the proof of Lemma 3.A.7 and equations (3.178) - (3.186), given Assumptions 3.A.10/1a and 3.A.10/1b, the following holds on the set S_n:

$$\left|\dot{m}\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m}\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| < \left|\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m} - \frac{\varepsilon_i}{\sigma}\right| \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|. \tag{3.213}$$

Therefore:

$$\sup_\theta V_n^S(\theta) < \sup_\theta \left\{ n^{-1/2} \sum_{i=1}^n |w_{in}| \left|\frac{\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right| \left|\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m} - \frac{\varepsilon_i}{\sigma}\right| \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \right\} 1_{S_n}. \tag{3.214}$$

Note that $1_{S_n} \leq 1_{S_n} + 1_{S_n^c} = 1$ and all terms are non-negative, hence:

$$\sup_\theta V_n^S(\theta) < \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \left|\frac{\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right| \left|\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m} - \frac{\varepsilon_i}{\sigma}\right| \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|. \tag{3.215}$$

Use Definition 3.A.1 of the scale, as well as the bound on the term $\left|\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma + n^{-1/2}a_m} - \frac{\varepsilon_i}{\sigma}\right|$ given in equation (3.180), to obtain the further bound:

$$\sup_\theta V_n^S(\theta) < \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \left|\frac{\frac{1}{s}\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma}\right| \left\{ \left|\frac{\varepsilon_i}{\sigma}\right| \left| n^{-1/2}\frac{1}{s}a_m/\sigma \right| + \left|\frac{\frac{1}{s}b_m'\dot{g}_{in}^0}{\sigma}\right| + \left|\frac{\frac{1}{s}\frac{1}{2}b_m'\bar{g}_m b_m}{\sigma}\right| \right\} \left|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \tag{3.216}$$

Using Lemma 3.A.1 and Remark 3.A.1, for sufficiently large n the above can be further bounded

by:

$$\sup_{\theta} V_n^S(\theta) < \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| 2 \left| \frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \right| \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| 2 \left| n^{-1/2} a_m / \sigma \right| + 2 \left| \frac{b'_m \dot{g}_{in}^0}{\sigma} \right| + 2 \left| \frac{\frac{1}{2} b'_m \bar{g}_m b_m}{\sigma} \right| \right\} \left| \mathring{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right|. \tag{3.217}$$

Note that $|\bar{g}_m| \leq |\ddot{g}_{in}^{m*}|$ and $\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \geq 0$, hence:

$$\sup_{\theta} V_n^S(\theta) < \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| 2 \left| \frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \right| \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| 2 \left| n^{-1/2} a_m / \sigma \right| + 2 \left| \frac{b'_m \dot{g}_{in}^0}{\sigma} \right| + 2 \left| \frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \right| \right\} \mathring{m} \left(\frac{\varepsilon_i}{\sigma} \right). \tag{3.218}$$

Use the definitions of $\dot{g}_{in}^0$ and $\ddot{g}_{in}^{m*}$, to express the bound as:

$$\begin{aligned} \sup_{\theta} V_n^S(\theta) < \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{|b'_m| \| N' \ddot{g}(x_i; \beta_{mi}^*) N \| |b_m|}{\sigma} \right| \cdot \\ \cdot \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| 2 \left| n^{-1/2} a_m / \sigma \right| + 2 \left| \frac{|b'_m| |N' \dot{g}(x_i; \beta_0)|}{\sigma} \right| + \left| \frac{|b'_m| \| N' \ddot{g}(x_i; \beta_{mi}^*) N \| |b_m|}{\sigma} \right| \right\} \mathring{m} \left(\frac{\varepsilon_i}{\sigma} \right). \end{aligned} \tag{3.219}$$

Due to the supremum with respect to θ being taken over $|\theta| \leq B n^{1/4-\eta}$ and incorporating functions of $2, B$ and $1/\sigma$ into the finite constant C, we can express the bound as:

$$\begin{aligned} \sup_{\theta} V_n^S(\theta) < C n^{-1/2} \sum_{i=1}^{n} |w_{in}| n^{1/2-2\eta} \| N' \ddot{g}(x_i; \beta_{mi}^*) N \| \cdot \\ \cdot \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| n^{-1/4-\eta} + n^{1/4-\eta} |N' \dot{g}(x_i; \beta_0)| + n^{1/2-2\eta} \| N' \ddot{g}(x_i; \beta_{mi}^*) N \| \right\} \mathring{m} \left(\frac{\varepsilon_i}{\sigma} \right). \end{aligned} \tag{3.220}$$

Collect the terms and define $\tilde{V}_n(\theta)$ as their sum:

$$\begin{aligned} \tilde{V}_n(\theta) := C \Bigg\{ & n^{-1/4-3\eta} \sum_{i=1}^{n} |w_{in}| \| N' \ddot{g}(x_i; \beta_{mi}^*) N \| \left| \frac{\varepsilon_i}{\sigma} \right| \mathring{m} \left(\frac{\varepsilon_i}{\sigma} \right) \\ & + n^{1/4-3\eta} \sum_{i=1}^{n} |w_{in}| \| N' \ddot{g}(x_i; \beta_{mi}^*) N \| |N' \dot{g}(x_i; \beta_0)| \mathring{m} \left(\frac{\varepsilon_i}{\sigma} \right) + \\ & + n^{1/2-4\eta} \sum_{i=1}^{n} |w_{in}| \| N' \ddot{g}(x_i; \beta_{mi}^*) N \|^2 \mathring{m} \left(\frac{\varepsilon_i}{\sigma} \right) \Bigg\}. \end{aligned} \tag{3.221}$$

Given the definition of convergence in probability and Markov's inequality, it is sufficient to show:

$$\mathsf{P}\left(\left| \tilde{V}_n(\theta) \right| > \epsilon \right) \leq \frac{1}{\epsilon} \mathsf{E} \left| \tilde{V}_n(\theta) \right| \to 0 \quad \text{as} \quad n \to \infty. \tag{3.222}$$

By definition of $\tilde{V}_n(\theta)$:

$$\begin{aligned}\frac{1}{\epsilon}\mathsf{E}\left|\tilde{V}_n(\theta)\right| = \frac{1}{\epsilon}C\mathsf{E}\Bigg\{ & n^{-1/4-3\eta}\sum_{i=1}^{n}|w_{in}|\|N'\ddot{g}(x_i;\beta^*_{mi})N\|\left|\frac{\varepsilon_i}{\sigma}\right|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \\ & + n^{1/4-3\eta}\sum_{i=1}^{n}|w_{in}|\|N'\ddot{g}(x_i;\beta^*_{mi})N\||N'\dot{g}(x_i;\beta_0)|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) + \\ & + n^{1/2-4\eta}\sum_{i=1}^{n}|w_{in}|\|N'\ddot{g}(x_i;\beta^*_{mi})N\|^2\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\Bigg\}.\end{aligned} \tag{3.223}$$

By the Law of Iterated Expectations:

$$\begin{aligned}\frac{1}{\epsilon}\mathsf{E}\left|\tilde{V}_n(\theta)\right| = \frac{1}{\epsilon}C\mathsf{E}\mathsf{E}_{i-1}\Bigg\{ & n^{-1/4-3\eta}\sum_{i=1}^{n}|w_{in}|\|N'\ddot{g}(x_i;\beta^*_{mi})N\|\left|\frac{\varepsilon_i}{\sigma}\right|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \\ & + n^{1/4-3\eta}\sum_{i=1}^{n}|w_{in}|\|N'\ddot{g}(x_i;\beta^*_{mi})N\||N'\dot{g}(x_i;\beta_0)|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) + \\ & + n^{1/2-4\eta}\sum_{i=1}^{n}|w_{in}|\|N'\ddot{g}(x_i;\beta^*_{mi})N\|^2\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\Bigg\}.\end{aligned} \tag{3.224}$$

Bound the term $|b'_m|\|N'\ddot{g}(x_i,\beta^*_{mi})N\||b_m|$ by its supremum over $\beta_i : \|N^{-1}(\beta_i-\beta_0)\| \le Bn^{1/4-\eta}$.

$$\begin{aligned}\frac{1}{\epsilon}\mathsf{E}\left|\tilde{V}_n(\theta)\right| \le \frac{1}{\epsilon}C\mathsf{E}\mathsf{E}_{i-1}\Bigg\{ & n^{-1/4-3\eta}\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)\left|\frac{\varepsilon_i}{\sigma}\right|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \\ & + n^{1/4-3\eta}\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)|N'\dot{g}(x_i;\beta_0)|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) + \\ & + n^{1/2-4\eta}\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)^2\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\Bigg\}.\end{aligned} \tag{3.225}$$

Note that w_{in} and x_i are included in the filtration $\mathcal{F}_{i-1,n}$ and $\frac{\varepsilon_i}{\sigma}$ is independent of $\mathcal{F}_{i-1,n}$. Therefore, terms which are functions of w_{in} and x_i only can pass through the inner expectation:

$$\begin{aligned}\frac{1}{\epsilon}\mathsf{E}\left|\tilde{V}_n(\theta)\right| \le & \frac{1}{\epsilon}Cn^{-1/4-3\eta}\mathsf{E}\left\{\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)\mathsf{E}_{i-1}\left|\frac{\varepsilon_i}{\sigma}\right|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right\} + \\ & + \frac{1}{\epsilon}Cn^{1/4-3\eta}\mathsf{E}\left\{\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)|N'\dot{g}(x_i;\beta_0)|\mathsf{E}_{i-1}\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)+\right\} \\ & + \frac{1}{\epsilon}Cn^{1/2-4\eta}\mathsf{E}\left\{\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)^2\mathsf{E}_{i-1}\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right\}.\end{aligned} \tag{3.226}$$

By Assumption 3.A.10/2a: $\mathsf{E}_{i-1}\left|\frac{\varepsilon_i}{\sigma}\right|\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) = C_1 < \infty$ and $\mathsf{E}_{i-1}\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) = C_2 < \infty$. Incorporating

constants C_1, C_2 into the constant C, gives:

$$\begin{aligned}\frac{1}{\epsilon}\mathsf{E}\left|\tilde{V}_n(\theta)\right| \leq \frac{1}{\epsilon}Cn^{-1/4-3\eta}\mathsf{E}\left\{\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)\right\}+ \\ +\frac{1}{\epsilon}Cn^{1/4-3\eta}\mathsf{E}\left\{\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)|N'\dot{g}(x_i;\beta_0)|\right\}+ \\ +\frac{1}{\epsilon}Cn^{1/2-4\eta}\mathsf{E}\left\{\sum_{i=1}^{n}|w_{in}|\left(\sup_{\beta_i}\|N'\ddot{g}(x_i,\beta_i)N\|\right)^2\right\}.\end{aligned} \tag{3.227}$$

Assumption 3.A.10/1c states that $\mathsf{E}\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\| = O(n^{1/4+2\eta})$. Hence, the first term in equation (3.227) is $O(n^{-\eta})$. By Assumption 3.A.10/1d: $\mathsf{E}\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|\|N'\dot{g}(x_i;\beta_0)\| = O(n^{-1/4+2\eta})$. Hence the second term in equation (3.227) is $O(n^{-\eta})$. By Assumption 3.A.10/1e: $\mathsf{E}\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|^2 = O(n^{-1/2+3\eta})$. Hence, the last term in equation (3.227) is $O(n^{-\eta})$. Thus:

$$\frac{1}{\epsilon}\mathsf{E}\left|\tilde{V}_n(\theta)\right| = O\left(n^{-\eta}\right). \tag{3.228}$$

Hence, $\lim_{n\to\infty}\frac{1}{\epsilon}\mathsf{E}\left|\tilde{V}_n(\theta)\right| = 0$ and:

$$\mathsf{P}(\left|\tilde{V}_n(\theta)\right| > \epsilon) \leq \frac{1}{\epsilon}\mathsf{E}\left|\tilde{V}_n(\theta)\right| \to 0, \tag{3.229}$$

which establishes that $\tilde{V}_n(\theta) = o_p(1)$, as required.

■

Assumption 3.A.11. Let $\tilde{v}$ represent each of $w_{in}n^{-1/4-\eta}$ and $w_{in}n^{1/4-\eta}|N'\dot{g}(x_i;\beta_0)|$. Let $\tilde{h}(u)$ represent each of $\dot{m}(u)$ and $u\dot{m}(u)$. Let r be such that $2^{r-1} > 1 + (1/4-\eta)(\dim b_1)$ and q such that $1 \leq q \leq r$. Further, define:

$$\tilde{H}_r(c) = \int_{-\infty}^{c}\left(1+\{\tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right)f\left(\frac{\varepsilon_i}{\sigma}\right)\mathrm{d}\left(\frac{\varepsilon_i}{\sigma}\right) = \mathsf{E}\left(1+\{\tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right)1_{(\varepsilon_i/\sigma\leq c)} \tag{3.230}$$

Assume the following:

1. *The weights* $\tilde{v}_{in}$:
   (a) The weights $\tilde{v}_{in}$ are $\mathcal{F}_{i-1}$-measurable.
   (b) $\mathsf{E}|\tilde{v}_{in}|^{2^r} < \infty$
   (c) For all $1 \leq q \leq r$: $\mathsf{E}\sum_{i=1}^{n}|\tilde{v}_{in}|^{2^q} = O(n)$

(d) $\sum_{i=1}^{n} |\tilde{v}_{in}| = O_p(n)$

2. *The regressors and the function $g(\cdot)$, for all $1 \leq q \leq r$:*

   (a) $\mathsf{E} \sum_{i=1}^{n} |\tilde{v}_{in}|^{2^q} \|N' \dot{g}(x_i; \beta_0)\| = O(n^{3/4-\eta})$

   (b) $\mathsf{E} \sum_{i=1}^{n} |\tilde{v}_{in}|^{2^q} \|N' \dot{g}(x_i; \beta_0)\|^2 = O(n^{1/4+\eta})$

   (c) $\mathsf{E} \sum_{i=1}^{n} |\tilde{v}_{in}|^{2^q} \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = O(n^{2\eta})$

   (d) $\sum_{i=1}^{n} |\tilde{v}_{in}| \|N' \dot{g}(x_i; \beta_0)\| = O_p(n^{1/2})$

   (e) $\sum_{i=1}^{n} |\tilde{v}_{in}| \|N' \dot{g}(x_i; \beta_0)\|^2 = o_p(n^{2\eta})$

   (f) $\sum_{i=1}^{n} |\tilde{v}_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{2\eta})$

3. *The mark function $\tilde{h}(\cdot)$:*

   (a) $\mathsf{E} \left| \tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right) \right|^{2^r} < \infty$

4. *The distance function $H_r(c)$:*

   (a) the support of $\dot{H}_r$, that is $\mathcal{S} = \{c : \dot{H}_r(c) > 0\}$, is an interval with endpoints $\underline{c}, \bar{c}$ such that $-\infty \leq \underline{c} < \bar{c} \leq \infty$;

   (b) $\sup_{x \in \mathcal{S}} 0 \leq \dot{H}_r(x) < \infty$;

   (c) $H_r = H_r(\infty) = \int_{\mathbb{R}} \dot{H}_r(c) dc < \infty$;

   (d) $\dot{H}_r(c)$ is Lipschitz, as defined in Definition 3.4.1;

   (e) $\dot{H}_r(c)$ has weakly unimodal bounds as defined in Definition 3.4.1.

Lemma 3.A.10. Under Assumption 3.A.11, the following holds:

$$\begin{aligned} \sup_{\theta} \sup_{c} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\varepsilon_i (1 - n^{-1/2} \frac{1}{s} a_m / \sigma) - \frac{1}{s} b_m' \dot{g}_{in}^0}{\sigma} - \frac{\varepsilon_i}{\sigma} \right| \cdot \\ \cdot \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma} + d_{in}^*\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma}\right)} \right| = o_p(1) \end{aligned} \tag{3.231}$$

Proof. Define the quantity of interest:

$$\begin{aligned} V_n(\theta) := & n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| \frac{\varepsilon_i (1 - n^{-1/2} \frac{1}{s} a_m / \sigma) - \frac{1}{s} b_m' \dot{g}_{in}^0}{\sigma} - \frac{\varepsilon_i}{\sigma} \right| \cdot \\ & \cdot \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma} + d_{in}^*\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma}\right)} \right| \end{aligned} \tag{3.232}$$

By the triangle inequality:

$$\begin{aligned}\sup_\theta \sup_c V_n(\theta) \leq & \sup_\theta \sup_c n^{-1/2} \sum_{i=1}^n |w_{in}| \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| n^{-1/2} \left| \frac{1}{s} a_m/\sigma \right| + \left| \frac{\frac{1}{s} b'_m \dot{g}^0_{in}}{\sigma} \right| \right\} \cdot \\ & \cdot \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma} + d^*_{in}\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma}\right)} \right|\end{aligned} \tag{3.233}$$

By Definition 3.A.1 of the scale and Remark 3.A.1:

$$\begin{aligned}\sup_\theta \sup_c V_n(\theta) < & \sup_\theta \sup_c n^{-1/2} \sum_{i=1}^n |w_{in}| \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| 2n^{-1/2} |a_m/\sigma| + 2 \left| \frac{b'_m \dot{g}^0_{in}}{\sigma} \right| \right\} \cdot \\ & \cdot \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma} + d^*_{in}\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma}\right)} \right|\end{aligned} \tag{3.234}$$

The supremum with respect to θ is taken over $\left\{|\theta| \leq Bn^{1/4-\eta}\right\}$. Let $B_1 < \infty$ incorporate $2, B$ and σ. Thus:

$$\begin{aligned}\sup_\theta \sup_c V_n(\theta) < & \sup_\theta \sup_c B_1 n^{-1/2} \sum_{i=1}^n |w_{in}| \left\{ \left| \frac{\varepsilon_i}{\sigma} \right| n^{-1/2} n^{1/4-\eta} + \left| \dot{g}^0_{in} \right| n^{1/4-\eta} \right\} \cdot \\ & \cdot \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma} + d^*_{in}\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma}\right)} \right|\end{aligned} \tag{3.235}$$

The supremum of a sum is no greater than the sum of the suprema, therefore:

$$\begin{aligned}\sup_\theta \sup_c V_n(\theta) < & \sup_\theta \sup_c n^{-1/2} \sum_{i=1}^n |w_{in}| B_1 n^{-1/4-\eta} \left| \frac{\varepsilon_i}{\sigma} \right| \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma} + d^*_{in}\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma}\right)} \right| + \\ & + \sup_\theta \sup_c n^{-1/2} \sum_{i=1}^n |w_{in}| \left| \dot{g}^0_{in} \right| B_1 n^{1/4-\eta} \left| \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right| \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma} + d^*_{in}\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b'_1 \dot{g}^0_{in}}{\sigma}\right)} \right|\end{aligned} \tag{3.236}$$

The first term in equation (3.236) is $o_p(1)$ by Lemma 3.A.2 with $v_{in} = w_{in} n^{-1/4-\eta}$ and $h(u) = u\dot{m}(u)$. The second term is $o_p(1)$ by Lemma 3.A.2 with $v_{in} = w_{in} n^{1/4-\eta} |\dot{g}^0_{in}|$ and $h(u) = \dot{m}(u)$. ■

Assumption 3.A.12.

Assume the following:

1. *The weights* w_{in}:
   (a) $\mathsf{E} \sum_{i=1}^n |w_{in}| = O(n)$
2. *The regressors and the function* $g(\cdot)$*:*

(a) $\max_{1\leq i\leq n}\|N'\dot{g}(x_i;\beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$

(b) $\mathsf{E}\sum_{i=1}^{n}|w_{in}|\,|N'\dot{g}(x_i;\beta_0)|^2 = O(n^{\eta})$

3. *The mark function* $m(u)$:

   (a) (local Lipschitz) $\exists\omega > 0$ and a function $\mathring{m}(u) \geq 0$ such that if $\forall u^*, u$ such that $|u^* - u| \leq (1+|u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u|\mathring{m}(u)$ and where:

   - $\int_{-\infty}^{+\infty}(1+|u|+u^2)\mathring{m}(u)f(u)du < \infty$.

Lemma 3.A.11. Let u^* be an intermediate point such that: $\left|u^* - \frac{\varepsilon_i}{\sigma}\right| \leq \left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma} - \frac{\varepsilon_i}{\sigma}\right|$. Then, under Assumption 3.A.12, the following holds as $n\to\infty$:

$$\begin{aligned}\sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right|\cdot \\ \cdot\left|\dot{m}\left(u^*\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|\left|1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}+d^*_{in}\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}\right)}\right| = o_p(1)\end{aligned} \tag{3.237}$$

Proof. For the purpose of this proof, define the term of interest as:

$$\begin{aligned}V_n(\theta) :=& n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right|\cdot \\ &\cdot\left|\dot{m}\left(u^*\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|\left|1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}+d^*_{in}\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}\right)}\right|\end{aligned} \tag{3.238}$$

Start by bounding the difference between the indicator functions by unity. This allows for the supremum over c to be dropped:

$$\sup_{\theta}\sup_{c}V_n(\theta) \leq \sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right|\left|\dot{m}\left(u^*\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \tag{3.239}$$

Simplifying the fraction:

$$\sup_{\theta}\sup_{c}V_n(\theta) \leq \sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\frac{\varepsilon_i(-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}\right|\left|\dot{m}\left(u^*\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \tag{3.240}$$

By the triangle inequality:

$$\sup_{\theta}\sup_{c}V_n(\theta) \leq \sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left\{\left|\frac{\varepsilon_i}{\sigma}\right|n^{-1/2}\left|\frac{1}{s}a_m/\sigma\right|+\left|\frac{\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}\right|\right\}\left|\dot{m}\left(u^*\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \tag{3.241}$$

Apply Lemma 3.A.1:

$$\sup_{\theta}\sup_{c} V_n(\theta) < \sup_{\theta} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left\{\left|\frac{\varepsilon_i}{\sigma}\right| 2n^{-1/2}\left|a_m/\sigma\right| + 2\left|\frac{b'_m\dot{g}^0_{in}}{\sigma}\right|\right\}\left|\dot{m}\left(u^*\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \tag{3.242}$$

The supremum over θ is taken over the set $|\theta| \leq Bn^{1/4-\eta}$. Let $C = 2B/\sigma < \infty$. Therefore:

$$\sup_{\theta}\sup_{c} V_n(\theta) < Cn^{-1/2}\sum_{i=1}^{n}|w_{in}|\left\{\left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/2}n^{1/4-\eta} + n^{1/4-\eta}\left|\dot{g}^0_{in}\right|\right\}\sup_{\theta}\left|\dot{m}\left(u^*\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \tag{3.243}$$

Recall Definition 3.A.2 of the set $S_{1,n} := (\max_i \|N'\dot{g}(x_i,\beta_0)\| \leq B_1 n^{-1/4+\eta-\kappa})$. The bound can be rewritten as follows, since the sum of suprema is no smaller than the supremum of a sum:

$$\sup_{\theta}\sup_{c} V_n(\theta) < \sup_{\theta}\sup_{c}\left\{V_n(\theta)\right\}1_{S_{1,n}} + \sup_{\theta}\sup_{c}\left\{V_n(\theta)\right\}1_{S^c_{1,n}} \tag{3.244}$$

By Assumption 3.A.12/2a and Remark 3.A.1: as $n \to \infty$, for any $\epsilon_1 > 0, P(S^c_{1,n}) < \epsilon_1$. Thus it suffices to show that the following term is $o_p(1)$:

$$V^S_n(\theta) := \left\{Cn^{-1/2}\sum_{i=1}^{n}|w_{in}|\left\{\left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/4-\eta} + n^{1/4-\eta}\left|\dot{g}^0_{in}\right|\right\}\sup_{\theta}\left|\dot{m}\left(u^*\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|\right\}1_{S_{1,n}} \tag{3.245}$$

The term $\left|\dot{m}\left(u^*\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|$ is bounded on the set $1_{S_{1,n}}$ under the local Lipschitz condition in Assumption 3.A.12/3a. To show this, we exploit the assumption that u^* is an intermediate point such that: $\left|u^* - \frac{\varepsilon_i}{\sigma}\right| \leq \left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma) - \frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma} - \frac{\varepsilon_i}{\sigma}\right|$, which implies that

$$\left|u^* - \frac{\varepsilon_i}{\sigma}\right| \leq \left|\frac{\varepsilon_i(-n^{-1/2}\frac{1}{s}a_m/\sigma) - \frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}\right|. \tag{3.246}$$

By the triangle inequality:

$$\left|u^* - \frac{\varepsilon_i}{\sigma}\right| \leq \left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/2}\left|\frac{1}{s}a_m/\sigma\right| + \left|\frac{\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}\right|. \tag{3.247}$$

By Remark 3.A.1, the above is strictly bounded by:

$$\left|u^* - \frac{\varepsilon_i}{\sigma}\right| < 2\left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/2}\left|a_m/\sigma\right| + 2\left|\frac{b'_m\dot{g}^0_{in}}{\sigma}\right|. \tag{3.248}$$

By definition of the set $S_{1,n}$, on the set $S_{1,n}$ it holds that $\max_i\left|\dot{g}^0_{in}\right| = O(n^{-1/4+\eta-\kappa})$. Further-

more, $|\theta| \leq Bn^{1/4-\eta}$. Hence, the above can be written as:

$$\left|u^* - \frac{\varepsilon_i}{\sigma}\right| < \left|\frac{\varepsilon_i}{\sigma}\right| C_{1,n} + C_{2,n}, \tag{3.249}$$

where $C_{1,n} = O(n^{-1/4-\eta})$, and $C_{2,n} = O(n^{-\kappa})$. Thus, on the set $S_{1,n}$ the local Lipschitz condition is satisfied and therefore:

$$\left\{\left|\dot{m}\left(u^*\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|\right\} 1_{S_{1,n}} \leq \left\{\left|u^* - \frac{\varepsilon_i}{\sigma}\right| \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right\} 1_{S_{1,n}} \tag{3.250}$$

By equation (3.248):

$$\left\{\left|\dot{m}\left(u^*\right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right|\right\} 1_{S_{1,n}} < \left\{\left[2\left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/2} \left|a_m/\sigma\right| + 2\left|\frac{b'_m \dot{g}^0_{in}}{\sigma}\right|\right] \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right\} 1_{S_{1,n}} \tag{3.251}$$

Thus, the bound on term $V_n^S(\theta)$ becomes:

$$\begin{aligned} V_n^S(\theta) < \Bigg\{ & Cn^{-1/2} \sum_{i=1}^{n} |w_{in}| \left\{\left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/4-\eta} + n^{1/4-\eta} \left|\dot{g}^0_{in}\right|\right\} \cdot \\ & \cdot \sup_{\theta} \left[\left\{2\left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/2} \left|a_m/\sigma\right| + 2\left|\frac{b'_m \dot{g}^0_{in}}{\sigma}\right|\right\} \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right] \Bigg\} 1_{S_{1,n}} \end{aligned} \tag{3.252}$$

The supremum with respect to θ is taken over $|\theta| \leq Bn^{1/4-\eta}$. Let $C < \infty$ incorporate the constants $B, 2, 1/\sigma$, then:

$$V_n^S(\theta) < \left\{ Cn^{-1/2} \sum_{i=1}^{n} |w_{in}| \left\{\left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/4-\eta} + n^{1/4-\eta} \left|\dot{g}^0_{in}\right|\right\}^2 \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right\} 1_{S_{1,n}}. \tag{3.253}$$

All terms above are non-negative and $1_{S_{1,n}} \leq 1_{S_{1,n}} + 1_{S_{1,n}^c} = 1$, therefore:

$$V_n^S(\theta) < Cn^{-1/2} \sum_{i=1}^{n} |w_{in}| \left\{\left|\frac{\varepsilon_i}{\sigma}\right| n^{-1/4-\eta} + n^{1/4-\eta} \left|\dot{g}^0_{in}\right|\right\}^2 \mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.254}$$

The aim is to show that $V_n^S(\theta) = o_p(1)$. By the definition of convergence in probability:

$$V_n^S(\theta) = o_p(1) \iff \forall \epsilon > 0, \lim_{n\to\infty} \mathsf{P}(\left|V_n^S(\theta)\right| > \epsilon) = 0. \tag{3.255}$$

By Markov's inequality:

$$\mathsf{P}(\left|V_n^S(\theta)\right| > \epsilon) \leq \frac{1}{\epsilon} \mathsf{E}\left|V_n^S(\theta)\right|. \tag{3.256}$$

By the Law of Iterated Expectations:

$$\mathsf{P}(\left|V_n^S(\theta)\right| > \epsilon) \leq \frac{1}{\epsilon}\mathsf{E}\mathsf{E}_{i-1}\left|V_n^S(\theta)\right|, \tag{3.257}$$

where the inner expectation is conditional on the filtration $\mathcal{F}_{i-1}$ and the outer expectation is with respect to that filtration. Thus:

$$\mathsf{P}(\left|V_n^S(\theta)\right| > \epsilon) < C\frac{1}{\epsilon}\mathsf{E}\mathsf{E}_{i-1}n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left\{\left|\frac{\varepsilon_i}{\sigma}\right|n^{-1/4-\eta}+n^{1/4-\eta}\left|\dot{g}_{in}^0\right|\right\}^2\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.258}$$

Use the linearity of the expectation operator, and note that $w_{in} \subset \mathcal{F}_{i-1}$, so w_{in} can be taken outside of the conditional expectation:

$$\mathsf{P}(\left|V_n^S(\theta)\right| > \epsilon) < C\frac{1}{\epsilon}\mathsf{E}n^{-1/2}\sum_{i=1}^{n}|w_{in}|\,\mathsf{E}_{i-1}\left\{\left|\frac{\varepsilon_i}{\sigma}\right|n^{-1/4-\eta}+n^{1/4-\eta}\left|\dot{g}_{in}^0\right|\right\}^2\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.259}$$

For any $y, z \in \mathbb{R}$, $(y+z)^2 \leq 2y^2 + 2z^2$. Furthermore, $x_i \subset \mathcal{F}_{i-1}$ and it can, therefore, be taken outside of the inner expectation. Thus, after incorporating the constant 2 into the constant $C < \infty$, we obtain the following bound:

$$\begin{aligned}\mathsf{P}(\left|V_n^S(\theta)\right| > \epsilon) &< C\frac{1}{\epsilon}\mathsf{E}n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left(n^{-1/2-2\eta}\right)\mathsf{E}_{i-1}\left|\frac{\varepsilon_i}{\sigma}\right|^2\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right)+\\ &+C\frac{1}{\epsilon}\mathsf{E}n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left(n^{1/2-2\eta}\left|\dot{g}_{in}^0\right|^2\right)\mathsf{E}_{i-1}\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right).\end{aligned} \tag{3.260}$$

By the local Lipschitz condition, $\mathsf{E}_{i-1}\left|\frac{\varepsilon_i}{\sigma}\right|^2\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) = C_1 < \infty$ and $\mathsf{E}_{i-1}\mathring{m}\left(\frac{\varepsilon_i}{\sigma}\right) = C_2 < \infty$; where C_1, C_2 are constants which do not depend on n. Let C incorporate $C_1, C2$. Thus:

$$\mathsf{P}(\left|V_n^S(\theta)\right| > \epsilon) < C\frac{1}{\epsilon}n^{-1-2\eta}\mathsf{E}\sum_{i=1}^{n}|w_{in}| + C\frac{1}{\epsilon}n^{-2\eta}\mathsf{E}\sum_{i=1}^{n}|w_{in}|\left|\dot{g}_{in}^0\right|^2. \tag{3.261}$$

By Assumption 3.A.12/1: $\mathsf{E}\sum_{i=1}^{n}|w_{in}| = O(n)$. Therefore, the first term is $O(n^{-2\eta})$. By Assumption 3.A.12/2b: $\mathsf{E}\sum_{i=1}^{n}|w_{in}|\left|N'\dot{g}(x_i;\beta_0)\right|^2 = O(n^{\eta})$. Therefore, the second term is $O(n^{-\eta})$. Hence, as $n \to \infty$: $\mathsf{P}(\left|V_n^S(\theta)\right| > \epsilon) \to 0$. Thus, the term $V_n^S(\theta) = o_p(1)$, which concludes the proof.

■

3.B Appendix 3.B: Proofs of Key Theorems

3.B.1 Structure of the Appendix

Appendix 3.B presents proofs of the two key results in this chapter: Theorems 3.4.1 and 3.4.2. Subsection 3.B.2 outlines the overall strategy for the proof. Subsections 3.B.3 and 3.B.4 provide outlines of the proofs of Theorems 3.4.1 and 3.4.2 respectively, while subsections 3.B.5 and 3.B.6 provide the details of the assumptions, required theorems, and their proofs.

3.B.2 Outline of the Approach

As detailed in Section 3.4, we wish to show that the following two results hold:

$$\sup_{\theta} \sup_{c} \left| n^{1/2} \{ \bar{F}_n^{w,m,g}(\theta, c) - \bar{F}_n^{w,m}(\theta, c) \} \right| = o_p(1) \tag{3.262}$$

and:

$$\sup_{\theta} \sup_{c} \left| n^{1/2} \{ F_n^{w,m,g}(\theta, c) - F_n^{w,m}(\theta, c) \} \right| = o_p(1), \tag{3.263}$$

where $F_n^{w,m}(\theta, c)$ is defined in Section 3.4 analogously to Berenguer-Rico and Nielsen (2017) with the linear regressor x_{in} replaced with the first derivative $\dot{g}(x_i, \beta_0)$.

Using the definitions of $\mathbb{F}_n^{w,m,g}(\theta, c)$ and $\mathbb{F}_n^{w,m}(\theta, c)$, results (3.262) and (3.263) imply the following result

$$\sup_{\theta} \sup_{c} |\mathbb{F}_n^{w,m,g}(\theta, c) - \mathbb{F}_n^{w,m}(\theta, c)| = o_p(1). \tag{3.264}$$

Additionally, using Theorem 2.1 of Berenguer-Rico and Nielsen (2017) and the definition of $\mathbb{F}_n^{w,m,g}(\theta, c)$, result (3.264) implies that

$$\sup_{\theta} \sup_{c} |\mathbb{F}_n^{w,m,g}(\theta, c) - \mathbb{F}_n^{w,m}(0, c)| = o_p(1). \tag{3.265}$$

While Section 3.A.4 lists the theorems and lemmas used in proving results (3.262) and (3.263), the mathematical details underpinning those are presented in the following Sections 3.B.3 and 3.B.4. Each section shows how the terms of interest can be broken down into smaller components, each of which can be shown to be at most $o_p(1)$ using lemmas discussed in Section 3.A.4.

3.B.3 Outline of the Proof of Result (3.19)

For the purpose of this proof, define $\bar{D}_n(\theta, c)$ as the following difference

$$\bar{D}_n(\theta, c) := n^{1/2} \{ \bar{F}_n^{w,m,g}(\theta, c) - \bar{F}_n^{w,m}(\theta, c) \}. \tag{3.266}$$

In order to prove Theorem 3.4.1, we must show that

$$\sup_{\theta}\sup_{c}\left|\bar{D}_n(\theta,c)\right| = \sup_{\theta}\sup_{c}\left|n^{1/2}\{\bar{F}_n^{w,m,g}(\theta,c) - \bar{F}_n^{w,m}(\theta,c)\}\right| = o_p(1). \tag{3.267}$$

By definition of $\bar{F}_n^{w,m,g}(\theta,c)$ and $\bar{F}_n^{w,m}(\theta,c)$, the term can be expanded as follows

$$\begin{aligned}
&\sup_{\theta}\sup_{c}\left|\bar{D}_n(\theta,c)\right| = \\
&= \sup_{\theta}\sup_{c}\left|n^{1/2}\left\{n^{-1}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right)1_{\left(\varepsilon_i \le \sigma c + n^{-1/2}a_1 c + b_1'\dot{g}_{in}^0 + \frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1\right)}\right\} - \right.\right. \\
&- \left.\left. n^{-1}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma + n^{-1/2}a_m}\right)1_{\left(\varepsilon_i \le \sigma c + n^{-1/2}a_1 c + b_1'\dot{g}_{in}^0\right)}\right\}\right\}\right|.
\end{aligned} \tag{3.268}$$

Recall the following definitions, which simplify the notation:

$$\begin{aligned}
c_{in}^0 &:= \frac{\sigma c + n^{-1/2}a_1 c + b_1'\dot{g}_{in}^0}{\sigma}, \\
d_{in}^* &:= \frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}.
\end{aligned} \tag{3.269}$$

Thus, the term of interest can be simplified to

$$\begin{aligned}
&\sup_{\theta}\sup_{c}\left|\bar{D}_n(\theta,c)\right| = \\
&= \sup_{\theta}\sup_{c}\left|n^{1/2}\left\{n^{-1}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right)1_{\left(\varepsilon_i/\sigma \le c_{in}^0 + d_{in}^*\right)}\right\} - \right.\right. \\
&- \left.\left. n^{-1}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma + n^{-1/2}a_m}\right)1_{\left(\varepsilon_i/\sigma \le c_{in}^0\right)}\right\}\right\}\right|.
\end{aligned} \tag{3.270}$$

Add and subtract $n^{-1/2}\sum_{i=1}^{n} w_{in}\mathsf{E}_{i-1}m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma + n^{-1/2}a_m}\right)1_{\left(\varepsilon_i/\sigma \le c_{in}^0 + d_{in}^*\right)}$. This is a compensator term in which the mark function is based on the first derivative only while the indicator function includes both the first and second derivatives. Further, the triangle inequality and the fact that the supremum of a sum is no greater than the sum of the suprema, imply that

$$\begin{aligned}
&\sup_{\theta}\sup_{c}\left|\bar{D}_n(\theta,c)\right| \le \sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\mathsf{E}_{i-1}m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma + n^{-1/2}a_m}\right)\left\{1_{\left(\varepsilon_i/\sigma \le c_{in}^0 + d_{in}^*\right)} - 1_{\left(\varepsilon_i/\sigma \le c_{in}^0\right)}\right\}\right| + \\
&+ \sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\left|\mathsf{E}_{i-1}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma + n^{-1/2}a_m}\right)\right\}1_{\left(\varepsilon_i/\sigma \le c_{in}^0 + d_{in}^*\right)}\right|.
\end{aligned} \tag{3.271}$$

Since the modulus is a convex function, it follows from Jensen's inequality that $\mathsf{E}\,|x| \geq |\mathsf{E}(x)|$. Hence

$$\sup_{\theta}\sup_{c}\left|\bar{D}_n(\theta,c)\right| \leq \sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\mathsf{E}_{i-1}\left|m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\left\{1_{\left(\varepsilon_i/\sigma\leq c_{in}^0+d_{in}^*\right)}-1_{\left(\varepsilon_i/\sigma\leq c_{in}^0\right)}\right\}\right| + \\ + \sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\mathsf{E}_{i-1}\left|\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)-m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right\}1_{\left(\varepsilon_i/\sigma\leq c_{in}^0+d_{in}^*\right)}\right|. \tag{3.272}$$

The proof that the first term is $o_p(1)$ is provided in Theorem 3.B.1. The proof that the second term is $o_p(1)$ is provided in Theorem 3.B.2.

3.B.4 Outline of the Proof of Result (3.21)

For the purpose of this proof, define $D_n(\theta,c)$ as the following difference

$$D_n(\theta,c) := n^{1/2}\{F_n^{w,m,g}(\theta,c) - F_n^{w,m}(\theta,c)\}. \tag{3.273}$$

By definition of $F_n^{w,m,g}(\theta,c)$ and $F_n^{w,m}(\theta,c)$ we can express the term as

$$\sup_{\theta}\sup_{c}|D_n(\theta,c)| = \sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in}m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)1_{\left(\varepsilon_i/\sigma\leq c_{in}^0+d_{in}^*\right)} - \right. \\ \left. - \sum_{i=1}^{n} w_{in}m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)1_{\left(\varepsilon_i/\sigma\leq c_{in}^0\right)}\right|. \tag{3.274}$$

Add and subtract $\sum_{i=1}^{n} w_{in}m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)1_{\left(\frac{\varepsilon_i}{\sigma}\leq c_{in}^0+d_{in}^*\right)}$. This is an additional compensator term in which the mark function is evaluated based on the first derivative while the indicator function includes both the first and second derivatives. Thus, the term can be re-expressed as

$$\sup_{\theta}\sup_{c}|D_n(\theta,c)| = \\ = \sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)-m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right\}1_{\left(\frac{\varepsilon_i}{\sigma}\leq c_{in}^0+d_{in}^*\right)} + \right. \\ \left. + \sum_{i=1}^{n} w_{in}m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\left\{1_{\left(\frac{\varepsilon_i}{\sigma}\leq c_{in}^0+d_{in}^*\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\leq c_{in}^0\right)}\right\}\right|. \tag{3.275}$$

Further, add and subtract $\sum_{i=1}^{n} w_{in}m\left(\frac{\varepsilon_i}{\sigma}\right)\left\{1_{\left(\frac{\varepsilon_i}{\sigma}\leq c_{in}^0+d_{in}^*\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\leq c_{in}^0\right)}\right\}$. In this additional term, the mark function contains only the true innovations ε_i/σ without any estimation errors, while

the term in braces is the difference between the indicator function based on both the first and second derivatives and the indicator function based on the first derivative only. Therefore, our term of interest can be expanded as

$$\begin{aligned}
&\sup_{\theta}\sup_{c}|D_n(\theta,c)| = \\
&= \sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right\}1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)}+\right. \\
&+ \sum_{i=1}^{n} w_{in}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i}{\sigma}\right)\right\}\left\{1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0\right)}\right\}+ \\
&\left.+ \sum_{i=1}^{n} w_{in} m\left(\frac{\varepsilon_i}{\sigma}\right)\left\{1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0\right)}\right\}\right|.
\end{aligned} \tag{3.276}$$

By the triangle inequality, and noting that the supremum of a sum is no greater than the sum of suprema, the above can be bounded by:

$$\begin{aligned}
&\sup_{\theta}\sup_{c}|D_n(\theta,c)| \le \\
&\le \sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right\}1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)}\right|+ \\
&+ \sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in}\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i}{\sigma}\right)\right\}\left\{1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0\right)}\right\}\right|+ \\
&+ \sup_{\theta}\sup_{c} n^{-1/2}\left|\sum_{i=1}^{n} w_{in} m\left(\frac{\varepsilon_i}{\sigma}\right)\left\{1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0+d_{in}^*\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma}\le c_{in}^0\right)}\right\}\right|.
\end{aligned} \tag{3.277}$$

The proof that the first term is $o_p(1)$ is provided in Theorem 3.B.3. The proof that the second term is $o_p(1)$ is given in Theorem 3.B.4. The proof that the third term is $o_p(1)$ is provided in Theorem 3.B.5.

3.B.5 Theorems Required for Result (3.19)

Assumption 3.B.1. Assume the following:

1. *The regressors and the function $g(\cdot)$:*
   (a) $\max_{1\le i\le n}\|N'\dot{g}(x_i;\beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$
   (b) $\sum_{i=1}^{n}|w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\le Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{2\eta})$

(c) $\sum_{i=1}^{n} |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq Bn^{1/4-\eta}} \|N'\dot{g}(x_i;\beta_0)\| \|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{-1/4+3\eta})$

(d) $\sum_{i=1}^{n} |w_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq Bn^{1/4-\eta}} \|N'\dot{g}(x_i;\beta_0)\|^2 \|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{-1/2+4\eta})$

2. *The mark function* $m(u)$:

(a) $\sup_{u \in \mathbb{R}} |m(u)| f(u) < \infty$

(b) $\sup_{u \in \mathbb{R}} (1 + |u|) |\dot{m}(u)| f(u) < \infty$

(c) (local Lipschitz) $\exists \omega > 0$ and a function $\mathring{m}(u) \geq 0$ such that if $\forall u^*, u$ such that $|u^* - u| \leq (1 + |u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u| \mathring{m}(u)$ and where:

- $\sup_{u \in \mathbb{R}} (1 + |u| + u^2) \mathring{m}(u) f(u) < \infty$

Theorem 3.B.1. Under Assumption 3.B.1, the following holds:

$$\sup_{\theta} \sup_{c} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \mathsf{E}_{i-1} \left| m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m}\right) \left\{ 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0 + \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1\right)} - 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0\right)} \right\} \right| = o_p(1). \tag{3.278}$$

Proof. For the purposes of this proof, define the quantity of interest as $V_n(\theta, c)$:

$$V_n(\theta, c) := n^{-1/2} \sum_{i=1}^{n} |w_{in}| \mathsf{E}_{i-1} \left| m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m}\right) \left\{ 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0 + \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1\right)} - 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0\right)} \right\} \right|. \tag{3.279}$$

Recall the following definitions of c_{in}^0 and d_{in}^*:

$$c_{in}^0 := \frac{\sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0}{\sigma} \quad \text{and} \quad d_{in}^* := \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma}. \tag{3.280}$$

Thus, the quantity of interest can be re-expressed as

$$\sup_{\theta} \sup_{c} V_n(\theta, c) = \sup_{\theta} \sup_{c} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \mathsf{E}_{i-1} \left| m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m}\right) \left\{ 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c_{in}^0 + d_{in}^*\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c_{in}^0\right)} \right\} \right|. \tag{3.281}$$

By integrating over a larger region, due to taking the absolute value of d_{in}^*, we obtain

$$\sup_{\theta} \sup_{c} V_n(\theta, c) \leq \sup_{\theta} \sup_{c} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{c_{in}^0 - |d_{in}^*|}^{c_{in}^0 + |d_{in}^*|} \left| m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m}\right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.282}$$

Use Definition 3.A.1 of the scale to simplify the denominators as follows:

$$\sup_{\theta}\sup_{c} V_n(\theta,c) \leq$$
$$\leq \sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\int_{c_{in}^0-|d_{in}^*|}^{c_{in}^0+|d_{in}^*|}\left|m\left\{\frac{\varepsilon_i}{\sigma}\left(1-n^{-1/2}\frac{1}{s}\frac{1}{\sigma}a_m\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\right| f\left(\frac{\varepsilon_i}{\sigma}\right)d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.283}$$

Define the integral as a function $K_i(|d_{in}^*|)$ of $|d_{in}^*|$:

$$K_i(|d_{in}^*|) := \int_{c_{in}^0-|d_{in}^*|}^{c_{in}^0+|d_{in}^*|}\left|m\left\{\frac{\varepsilon_i}{\sigma}\left(1-n^{-1/2}\frac{1}{s}\frac{1}{\sigma}a_m\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\right| f\left(\frac{\varepsilon_i}{\sigma}\right)d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.284}$$

We use a first-order Taylor-series expansion around $K_i(0)$. In order to do so, we use the Leibniz rule for differentiation under the integral sign. For a general function $h(x,t)$, the integral of $h(x,t)$ with limits of integration $a(x)$ and $b(x)$ can be decomposed as

$$\frac{d}{dx}\int_{a(x)}^{b(x)}h(x,t)dt = h(x,b(x))\frac{d(b(x))}{dx} - h(x,a(x))\frac{d(a(x))}{dx} + \int_{a(x)}^{b(x)}\frac{d}{dx}h(x,t)dt. \tag{3.285}$$

In our case, we set: $x := |d_{in}^*|, t = \frac{\varepsilon_i}{\sigma}$, $\frac{d(b(x))}{dx} = +1$, $\frac{d(a(x))}{dx} = -1$,
$h(|d_{in}^*|, \frac{\varepsilon_i}{\sigma}) = \left|m\left\{\frac{\varepsilon_i}{\sigma}\left(1-n^{-1/2}\frac{1}{s}\frac{1}{\sigma}a_m\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\right| f\left(\frac{\varepsilon_i}{\sigma}\right)$, which gives:

$$\frac{d}{dx}h\left(x,t\right) = 0, \tag{3.286}$$

because $h(|d_{in}^*|, \frac{\varepsilon_i}{\sigma}) = \left|m\left\{\frac{\varepsilon_i}{\sigma}\left(1-n^{-1/2}\frac{1}{s}\frac{1}{\sigma}a_m\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\right| f\left(\frac{\varepsilon_i}{\sigma}\right)$ does not contain $|d_{in}^*|$. Further, the function inside the integral evaluated at the limits of integration is

$$h(x,b(x)) = \left|m\left\{\left(c_{in}^0+|d_{in}^*|\right)\left(1-n^{-1/2}\frac{1}{s}\frac{1}{\sigma}a_m\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\right| f\left(c_{in}^0+|d_{in}^*|\right) \tag{3.287}$$

for the upper limit and

$$h(x,a(x)) = \left|m\left\{\left(c_{in}^0-|d_{in}^*|\right)\left(1-n^{-1/2}\frac{1}{s}\frac{1}{\sigma}a_m\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\right| f\left(c_{in}^0-|d_{in}^*|\right) \tag{3.288}$$

for the lower limit. Thus, using a first-order Taylor series expansion around $K_i(0)$, and applying the Leibniz formula, we can express the integral $K_i(|d_{in}^*|)$ as follows, for an intermediate point

$\left|\bar{d}\right|$ such that $0 \leq \left|\bar{d}\right| \leq |d_{in}^*| = \left|\frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma}\right|$:

$$
\begin{aligned}
K_i(|d_{in}^*|) &= K_i(0) + (|d_{in}^*| - 0)\, \dot{K}_i(\left|\bar{d}\right|) = \\
&= K_i(0) + |d_{in}^*| \left| m\left\{ \left(c_{in}^0 + \left|\bar{d}\right|\right) \left(1 - n^{-1/2} \frac{1}{s} \frac{1}{\sigma} a_m\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 \right\} \right| f\left(c_{in}^0 + \left|\bar{d}\right|\right) + \\
&+ |d_{in}^*| \left| m\left\{ \left(c_{in}^0 - \left|\bar{d}\right|\right) \left(1 - n^{-1/2} \frac{1}{s} \frac{1}{\sigma} a_m\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 \right\} \right| f\left(c_{in}^0 - \left|\bar{d}\right|\right).
\end{aligned}
\tag{3.289}
$$

Note that $K_i(0) = 0$ since it corresponds to integrating over an interval of zero length. Additionally, use $a_n^s = n^{-1/2} a_m / \sigma$ to simplify the notation, and apply the triangle inequality. The resulting bound is

$$
\begin{aligned}
\sup_\theta \sup_c V_n(\theta, c) \leq \sup_\theta \sup_c n^{-1/2} \sum_{i=1}^n |w_{in}| \left[|d_{in}^*| \left| m\left\{ \left(c_{in}^0 + \left|\bar{d}\right|\right) \left(1 - \frac{1}{s} a_n^s\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 \right\} \right| f\left(c_{in}^0 + \left|\bar{d}\right|\right) \right] + \\
+ \sup_\theta \sup_c n^{-1/2} \sum_{i=1}^n |w_{in}| \left[|d_{in}^*| \left| m\left\{ \left(c_{in}^0 - \left|\bar{d}\right|\right) \left(1 - \frac{1}{s} a_n^s\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 \right\} \right| f\left(c_{in}^0 - \left|\bar{d}\right|\right) \right].
\end{aligned}
\tag{3.290}
$$

We Taylor-expand $m\left\{\left(c_{in}^0 + \left|\bar{d}\right|\right)\left(1 - \frac{1}{s} a_n^s\right) - \frac{1}{s}\frac{1}{\sigma} b_m' \dot{g}_{in}^0\right\}$ around $m\left(c_{in}^0 + \left|\bar{d}\right|\right)$. This will allow us to use Assumption 3.B.1/2a by ensuring that arguments inside functions $m(\cdot)$ and $f(\cdot)$ are the same. Let u_+ be an intermediate point such that

$$
\left|u_+ - \left(c_{in}^0 + \left|\bar{d}\right|\right)\right| \leq \left| \left(c_{in}^0 + \left|\bar{d}\right|\right) \left(1 - \frac{1}{s} a_n^s\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 - \left(c_{in}^0 + \left|\bar{d}\right|\right) \right|,
$$

which is equivalent to

$$
\left|u_+ - \left(c_{in}^0 + \left|\bar{d}\right|\right)\right| \leq \left| \left(c_{in}^0 + \left|\bar{d}\right|\right) \left(-\frac{1}{s} a_n^s\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 \right|.
$$

Then, by a first-order Taylor series expansion:

$$
\begin{aligned}
& m\left\{ \left(c_{in}^0 + \left|\bar{d}\right|\right) \left(1 - \frac{1}{s} a_n^s\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 \right\} \\
&= m\left(c_{in}^0 + \left|\bar{d}\right|\right) + \left\{ \left(c_{in}^0 + \left|\bar{d}\right|\right) \left(1 - \frac{1}{s} a_n^s\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 - \left(c_{in}^0 + \left|\bar{d}\right|\right) \right\} \dot{m}(u_+) \\
&= m\left(c_{in}^0 + \left|\bar{d}\right|\right) + \left\{ \left(c_{in}^0 + \left|\bar{d}\right|\right) \left(-\frac{1}{s} a_n^s\right) - \frac{1}{s} \frac{1}{\sigma} b_m' \dot{g}_{in}^0 \right\} \dot{m}(u_+).
\end{aligned}
\tag{3.291}
$$

Following an analogous expansion of $m\left\{\left(c_{in}^0 - \left|\bar{d}\right|\right)\left(1 - \frac{1}{s} a_n^s\right) - \frac{1}{s}\frac{1}{\sigma} b_m' \dot{g}_{in}^0\right\}$ around $m\left(c_{in}^0 - \left|\bar{d}\right|\right)$

for an intermediate point u_- gives

$$m\left\{\left(c_{in}^0-|\bar{d}|\right)\left(1-\frac{1}{s}a_n^s\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}=m\left(c_{in}^0-|\bar{d}|\right)+\left\{\left(c_{in}^0-|\bar{d}|\right)\left(-\frac{1}{s}a_n^s\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\dot{m}(u_-). \tag{3.292}$$

Thus, the term of interest, $V_n(\theta,c)$, is bounded as follows:

$$\begin{aligned}&\sup_\theta\sup_c V_n(\theta,c)\leq\\&\leq\sup_\theta\sup_c n^{-1/2}\sum_{i=1}^n|w_{in}|\,|d_{in}^*|\left|m\left(c_{in}^0+|\bar{d}|\right)+\left\{\left(c_{in}^0+|\bar{d}|\right)\left(-\frac{1}{s}a_n^s\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\dot{m}(u_+)\right|f\left(c_{in}^0+|\bar{d}|\right)+\\&+\sup_\theta\sup_c n^{-1/2}\sum_{i=1}^n|w_{in}|\,|d_{in}^*|\left|m\left(c_{in}^0-|\bar{d}|\right)+\left\{\left(c_{in}^0-|\bar{d}|\right)\left(-\frac{1}{s}a_n^s\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\dot{m}(u_-)\right|f\left(c_{in}^0-|\bar{d}|\right).\end{aligned} \tag{3.293}$$

By the triangle inequality, we can express the bound as

$$\sup_\theta\sup_c V_n(\theta,c)\leq\sup_\theta\sup_c V_{1,n}(\theta,c)+\sup_\theta\sup_c V_{2,n}(\theta,c)+\sup_\theta\sup_c V_{3,n}(\theta,c)+\sup_\theta\sup_c V_{4,n}(\theta,c), \tag{3.294}$$

where terms $V_{1,n}(\theta,c), V_{2,n}(\theta,c), V_{3,n}(\theta,c), V_{4,n}(\theta,c)$ are defined for the purposes of this proof as

$$\begin{aligned}V_{1,n}(\theta,c)&:=n^{-1/2}\sum_{i=1}^n|w_{in}|\,|d_{in}^*|\left|m\left(c_{in}^0+|\bar{d}|\right)\right|f\left(c_{in}^0+|\bar{d}|\right),\\V_{2,n}(\theta,c)&:=n^{-1/2}\sum_{i=1}^n|w_{in}|\,|d_{in}^*|\left|\left\{\left(c_{in}^0+|\bar{d}|\right)\left(-\frac{1}{s}a_n^s\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\dot{m}(u_+)\right|f\left(c_{in}^0+|\bar{d}|\right),\\V_{3,n}(\theta,c)&:=n^{-1/2}\sum_{i=1}^n|w_{in}|\,|d_{in}^*|\left|m\left(c_{in}^0-|\bar{d}|\right)\right|f\left(c_{in}^0-|\bar{d}|\right),\end{aligned} \tag{3.295}$$

and

$$V_{4,n}(\theta,c):=n^{-1/2}\sum_{i=1}^n|w_{in}|\,|d_{in}^*|\left|\left\{\left(c_{in}^0-|\bar{d}|\right)\left(-\frac{1}{s}a_n^s\right)-\frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\dot{m}(u_-)\right|f\left(c_{in}^0-|\bar{d}|\right). \tag{3.296}$$

Terms $V_{1,n}(\theta,c)$ and $V_{3,n}(\theta,c)$.

The following proof focuses on term $V_{1,n}(\theta,c)$. The proof for term $V_{3,n}(\theta,c)$ follows analogously.

By definition of d_{in}^*,

$$\sup_\theta\sup_c V_{1,n}(\theta,c)=n^{-1/2}\sum_{i=1}^n|w_{in}|\left|\frac{\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1}{\sigma}\right|\left|m\left(c_{in}^0+|\bar{d}|\right)\right|f\left(c_{in}^0+|\bar{d}|\right). \tag{3.297}$$

Bound $\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1$ by its supremum over $\left\{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}\right\}$, and bound

$\left|m\left(c_{in}^0+\left|\bar{d}\right|\right)\right| f\left(c_{in}^0+\left|\bar{d}\right|\right)$ by its supremum over $u\in\mathbb{R}$. Then

$$\sup_\theta \sup_c V_{1,n}(\theta,c) \leq \sup_\theta n^{-1/2}\sum_{i=1}^n |w_{in}||b_1| \sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\||b_1| \sup_u |m\left(u\right)| f\left(u\right). \tag{3.298}$$

By Assumption3.B.1/2a: $\sup_{u\in\mathbb{R}} |m(u)| f(u) = C < \infty$. Thus

$$\sup_\theta \sup_c V_{1,n}(\theta,c) \leq \sup_\theta Cn^{-1/2}\sum_{i=1}^n |w_{in}||b_1| \sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\||b_1|. \tag{3.299}$$

Note that the supremum with respect to θ is taken over: $|\theta| \leq Bn^{1/4-\eta}$. Let the constant C incorporate the constant B. Thus, the bound can be expressed as

$$\sup_\theta \sup_c V_{1,n}(\theta,c) \leq Cn^{-1/2}\sum_{i=1}^n |w_{in}|n^{1/2-2\eta} \sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\| \tag{3.300}$$

By Assumption 3.B.1/1b: $\sum_{i=1}^n |w_{in}| \sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{2\eta})$. Hence $\sup_\theta \sup_c V_{1,n}(\theta,c) = o_p(1)$. By an analogous argument, $\sup_\theta \sup_c V_{3,n}(\theta,c) = o_p(1)$.

Terms $V_{2,n}(\theta,c)$ and $V_{4,n}(\theta,c)$.

Terms $V_{2,n}(\theta,c)$ and $V_{4,n}(\theta,c)$ are symmetric about c_{in}^0 since they only differ in the sign in front of $\left|\bar{d}\right|$. Thus, the following derivation is performed only for $V_{2,n}(\theta,c)$. The derivation for $V_{4,n}(\theta,c)$ is analogous.

Recall the definition of $V_{2,n}(\theta,c)$:

$$V_{2,n}(\theta,c) := n^{-1/2}\sum_{i=1}^n |w_{in}|\left|d_{in}^*\right| \left|\left\{\left(c_{in}^0+\left|\bar{d}\right|\right)\left(-\frac{1}{s}a_n^s\right) - \frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right\}\dot{m}(u_+)\right| f\left(c_{in}^0+\left|\bar{d}\right|\right). \tag{3.301}$$

Use the property that for any $x, y : |xy| = |x||y|$. Then, add and subtract $\dot{m}\left(c_{in}^0+\left|\bar{d}\right|\right)$ from $\dot{m}(u_+)$ to re-express the quantity as

$$\begin{aligned}\sup_\theta \sup_c V_{2,n}(\theta,c) = \sup_\theta \sup_c n^{-1/2}\sum_{i=1}^n |w_{in}|\left|d_{in}^*\right| \left|\left(c_{in}^0+\left|\bar{d}\right|\right)\left(-\frac{1}{s}a_n^s\right) - \frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^0\right| \cdot \\ \cdot \left|\dot{m}\left(u_+\right)+\dot{m}\left(c_{in}^0+\left|\bar{d}\right|\right)-\dot{m}\left(c_{in}^0+\left|\bar{d}\right|\right)\right| f\left(c_{in}^0+\left|\bar{d}\right|\right).\end{aligned} \tag{3.302}$$

Use the triangle inequality to bound the sum of the absolute value by two sums of absolute

values:

$$\begin{aligned}
&\sup_{\theta}\sup_{c} V_{2,n}(\theta,c) \leq \\
&\leq \sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\,|d^{*}_{in}|\left|\left(c^{0}_{in}+|\bar{d}|\right)\left(-\frac{1}{s}a^{s}_{n}\right)-\frac{1}{s}\frac{1}{\sigma}b'_{m}\dot{g}^{0}_{in}\right|\left|\dot{m}\left(u_{+}\right)-\dot{m}\left(c^{0}_{in}+|\bar{d}|\right)\right| f\left(c^{0}_{in}+|\bar{d}|\right)+ \\
&+\sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\,|d^{*}_{in}|\left|\left(c^{0}_{in}+|\bar{d}|\right)\left(-\frac{1}{s}a^{s}_{n}\right)-\frac{1}{s}\frac{1}{\sigma}b'_{m}\dot{g}^{0}_{in}\right|\left|\dot{m}\left(c^{0}_{in}+|\bar{d}|\right)\right| f\left(c^{0}_{in}+|\bar{d}|\right).
\end{aligned} \tag{3.303}$$

In the proof that follows we analyse each term in equation (3.303) separately. To simplify notation, we define $K_{1,n}$ and $K_{2,n}$ as the respective terms in equation (3.303):

$$\begin{aligned}
K_{1,n} &:= n^{-1/2}\sum_{i=1}^{n}|w_{in}|\,|d^{*}_{in}|\left|\left(c^{0}_{in}+|\bar{d}|\right)\left(-\frac{1}{s}a^{s}_{n}\right)-\frac{1}{s}\frac{1}{\sigma}b'_{m}\dot{g}^{0}_{in}\right|\left|\dot{m}\left(u_{+}\right)-\dot{m}\left(c^{0}_{in}+|\bar{d}|\right)\right| f\left(c^{0}_{in}+|\bar{d}|\right), \\
K_{2,n} &:= n^{-1/2}\sum_{i=1}^{n}|w_{in}|\,|d^{*}_{in}|\left|\left(c^{0}_{in}+|\bar{d}|\right)\left(-\frac{1}{s}a^{s}_{n}\right)-\frac{1}{s}\frac{1}{\sigma}b'_{m}\dot{g}^{0}_{in}\right|\left|\dot{m}\left(c^{0}_{in}+|\bar{d}|\right)\right| f\left(c^{0}_{in}+|\bar{d}|\right).
\end{aligned} \tag{3.304}$$

Note that $\sup_{\theta}\sup_{c} K_{2,n} = o_p(1)$ by Lemma 3.A.5 under Assumptions 3.B.1/1b, 3.B.1/1c and 3.B.1/2b, which are identical to Assumptions 3.A.6/1a, 3.A.6/1b, and 3.A.6/2a, respectively, and which form an exhaustive list of assumptions required for Lemma 3.A.5.

For term $K_{1,n}$, we make a Lipschitz argument as follows. Recall that u_{+} is an intermediate point such that:

$$\left|u_{+}-\left(c^{0}_{in}+|\bar{d}|\right)\right| \leq \left|\left(c^{0}_{in}+|\bar{d}|\right)\left(-n^{-1/2}\frac{1}{s}\frac{1}{\sigma}a_{m}\right)-\frac{1}{s}\frac{1}{\sigma}b'_{m}\dot{g}^{0}_{in}\right|. \tag{3.305}$$

Use the triangle inequality to bound the absolute value of the sum by the sum of absolute values:

$$\left|u_{+}-\left(c^{0}_{in}+|\bar{d}|\right)\right| \leq \left|c^{0}_{in}+|\bar{d}|\right| n^{-1/2}\left|\frac{1}{s}\frac{1}{\sigma}a_{m}\right|+\left|\frac{1}{s}\frac{1}{\sigma}b'_{m}\dot{g}^{0}_{in}\right|. \tag{3.306}$$

By Lemma 3.A.1 and Remark 3.A.2, for sufficiently large n,

$$\left|u_{+}-\left(c^{0}_{in}+|\bar{d}|\right)\right| < \left|c^{0}_{in}+|\bar{d}|\right| 2n^{-1/2}\left|\frac{1}{\sigma}a_{m}\right|+2\left|\frac{1}{\sigma}b'_{m}\dot{g}^{0}_{in}\right|. \tag{3.307}$$

Consider set $S_{1,n} := (\max_i \|N'\dot{g}(x_i,\beta_0)\| \leq B_1 n^{-1/4+\eta-\kappa})$, as given in Definition 3.A.2. Note

that:

$$\sup_{\theta}\sup_{c} K_{1,n} \leq \sup_{\theta}\sup_{c}\left\{K_{1,n}\right\} 1_{S_{1,n}} + \sup_{\theta}\sup_{c}\left\{K_{1,n}\right\} 1_{S_{1,n}^{c}}. \tag{3.308}$$

Define the term $K_{1,n}^{S}$ as

$$K_{1,n}^{S} = \left\{K_{1,n}\right\} 1_{S_{1,n}}. \tag{3.309}$$

Following Remark 3.A.2, and Assumption 3.B.1/1a, $P(S_{1,n}^{c}) < \epsilon_1$. Thus, it suffices to consider term $K_{1,n}$ the set $S_{1,n}$, i.e. it suffices to consider $K_{1,n}^{S}$. On the set $S_{1,n}$, the following holds:

$$\left\{\left|u_{+} - \left(c_{in}^{0} + \left|\bar{d}\right|\right)\right|\right\} 1_{S_{1,n}} < \left\{\left|c_{in}^{0} + \left|\bar{d}\right|\right| 2n^{-1/2}\left|\frac{1}{\sigma}a_m\right| + 2\left|\frac{1}{\sigma}b_m'\dot{g}_{in}^{0}\right|\right\} 1_{S_{1,n}}. \tag{3.310}$$

By definition of the set $S_{1,n}$ and by Assumption 3.B.1/1a, on the set $S_{1,n}$ it holds that $|\dot{g}_{in}^{0}| = O(n^{-1/4+\eta-\kappa})$. Further, note that $|\theta| \leq Bn^{1/4-\eta}$. Thus, the above bound can be written as

$$\left\{\left|u_{+} - \left(c_{in}^{0} + \left|\bar{d}\right|\right)\right|\right\} 1_{S_{1,n}} < \left\{\left|c_{in}^{0} + \left|\bar{d}\right|\right|\right\} 1_{S_{1,n}} C_{1,n} + C_{2,n}, \tag{3.311}$$

where $C_{1,n} = O(n^{-1/4-\eta})$ and $C_{2,n} = O(-\kappa)$. Thus, the local Lipschitz condition is satisfied on the set $S_{1,n}$ and

$$\left\{\left|\dot{m}\left(u_{+}\right) - \dot{m}\left(c_{in}^{0} + \left|\bar{d}\right|\right)\right|\right\} 1_{S_{1,n}} < \left\{\left|u_{+} - \left(c_{in}^{0} + \left|\bar{d}\right|\right)\right| \mathring{m}\left(c_{in}^{0} + \left|\bar{d}\right|\right)\right\} 1_{S_{1,n}}, \tag{3.312}$$

where $\mathring{m}\left(u\right) \geq 0$ is a function such that $\sup_u(1 + |u| + u^2)\mathring{m}(u)f(u) < \infty$, by Assumption 3.B.1/2c. Consequently, substituting for $\left|\dot{m}\left(u_{+}\right) - \dot{m}\left(c_{in}^{0} + \left|\bar{d}\right|\right)\right|$ gives, on the set $S_{1,n}$,

$$\begin{aligned}\sup_{\theta}\sup_{c} K_{1,n}^{S} < \sup_{\theta}\sup_{c}\Bigg\{ & n^{-1/2}\sum_{i=1}^{n}\left|w_{in}\right|\left|d_{in}^{*}\right|\left|\left(c_{in}^{0} + \left|\bar{d}\right|\right)\left(-\frac{1}{s}a_n^{s}\right) - \frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^{0}\right| \cdot \\ & \cdot \left|u_{+} - \left(c_{in}^{0} + \left|\bar{d}\right|\right)\right| \mathring{m}\left(c_{in}^{0} + \left|\bar{d}\right|\right) f\left(c_{in}^{0} + \left|\bar{d}\right|\right)\Bigg\} 1_{S_{1,n}}.\end{aligned} \tag{3.313}$$

Note that $1_{S_{1,n}} \leq 1_{S_{1,n}} + 1_{S_{1,n}^{c}} = 1$, hence

$$\sup_{\theta}\sup_{c} K_{1,n}^{S} \cdot \left|u_{+} - \left(c_{in}^{0} + \left|\bar{d}\right|\right)\right| \mathring{m}\left(c_{in}^{0} + \left|\bar{d}\right|\right) f\left(c_{in}^{0} + \left|\bar{d}\right|\right). \tag{3.314}$$

Using equation (3.305) we can bound $\left|u_{+} - \left(c_{in}^{0} + \left|\bar{d}\right|\right)\right|$ by $\left|\left(c_{in}^{0} + \left|\bar{d}\right|\right)\left(-\frac{1}{s}a_n^{s}\right) - \frac{1}{s}\frac{1}{\sigma}b_m'\dot{g}_{in}^{0}\right|$,

which results in the following bound:

$$\sup_{\theta}\sup_{c} K_{1,n}^{S} < \sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}\left|w_{in}\right|\left|d_{in}^{*}\right|\left|\left(c_{in}^{0}+\left|\bar{d}\right|\right)\left(-\frac{1}{s}a_{n}^{s}\right)-\frac{1}{s}\frac{1}{\sigma}b_{m}'\dot{g}_{in}^{0}\right|^{2}\mathring{m}\left(c_{in}^{0}+\left|\bar{d}\right|\right)f\left(c_{in}^{0}+\left|\bar{d}\right|\right). \tag{3.315}$$

Expand the squared term, noting that for any $y, z \in \mathbb{R}$, $(y+z)^2 \leq 2y^2 + 2z^2$. Further, use the triangle inequality. The resulting bound is

$$\begin{aligned}\sup_{\theta}\sup_{c} K_{1,n}^{S} < \sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}\left|w_{in}\right|\left|d_{in}^{*}\right|\left(c_{in}^{0}+\left|\bar{d}\right|\right)^{2}\left(\frac{1}{s}a_{n}^{s}\right)^{2}\mathring{m}\left(c_{in}^{0}+\left|\bar{d}\right|\right)f\left(c_{in}^{0}+\left|\bar{d}\right|\right)+\\ +\sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}\left|w_{in}\right|\left|d_{in}^{*}\right|\left|\frac{1}{s}\frac{1}{\sigma}b_{m}'\dot{g}_{in}^{0}\right|^{2}\mathring{m}\left(c_{in}^{0}+\left|\bar{d}\right|\right)f\left(c_{in}^{0}+\left|\bar{d}\right|\right).\end{aligned} \tag{3.316}$$

We analyse each term in equation (3.316) separately. To simplify notation, we denote the terms as $J_{1,n}$ and $J_{2,n}$ respectively, i.e.

$$J_{1,n} := n^{-1/2}\sum_{i=1}^{n}\left|w_{in}\right|\left|d_{in}^{*}\right|\left(c_{in}^{0}+\left|\bar{d}\right|\right)^{2}\left(\frac{1}{s}a_{n}^{s}\right)^{2}\mathring{m}\left(c_{in}^{0}+\left|\bar{d}\right|\right)f\left(c_{in}^{0}+\left|\bar{d}\right|\right) \tag{3.317}$$

and

$$J_{2,n} := n^{-1/2}\sum_{i=1}^{n}\left|w_{in}\right|\left|d_{in}^{*}\right|\left|\frac{1}{s}\frac{1}{\sigma}b_{m}'\dot{g}_{in}^{0}\right|^{2}\mathring{m}\left(c_{in}^{0}+\left|\bar{d}\right|\right)f\left(c_{in}^{0}+\left|\bar{d}\right|\right). \tag{3.318}$$

It suffices to show that $\sup_\theta \sup_c J_{1,n}(\theta, c) = o_p(1)$ and $\sup_\theta \sup_c J_{2,n}(\theta, c) = o_p(1)$. The terms are analysed in turn.

Term $J_{1,n}$.

Bound the function $\left(c_{in}^{0}+\left|\bar{d}\right|\right)^{2}\mathring{m}\left(c_{in}^{0}+\left|\bar{d}\right|\right)f\left(c_{in}^{0}+\left|\bar{d}\right|\right)$ by its supremum over $u \in \mathbb{R}$. Since c_{in}^0 is the only term which is a function of c, this allows for the supremum over c to be dropped:

$$\sup_{\theta}\sup_{c} J_{1,n} \leq \sup_{\theta} n^{-1/2}\sum_{i=1}^{n}\left|w_{in}\right|\left|d_{in}^{*}\right|\left(\frac{1}{s}a_{n}^{s}\right)^{2}\sup_{u} u^{2}\mathring{m}\left(u\right)f\left(u\right). \tag{3.319}$$

By Assumption 3.B.1/2c: $\sup_u u^2 \mathring{m}(u) f(u) = C < \infty$, where C is a finite constant which does not depend on n. This gives the following bound, where the definitions of $a_n^s := n^{-1/2} a_m/\sigma$ and $d_{in}^* := \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1/\sigma$ have been used and terms $1/2$ and $1/\sigma$ have been incorporated into the

constant C:

$$\sup_{\theta}\sup_{c} J_{1,n} \leq C \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left|b_1' \ddot{g}_{in}^{1*} b_1\right| \left(\frac{1}{s} n^{-1/2} a_m\right)^2. \tag{3.320}$$

Using Lemma 3.A.1 and Remark 3.A.1, for sufficiently large n, $\left|\frac{1}{s} a_m\right| < 2|a_m|$. This gives the further bound, where the new constant 2^2 has been incorporated into the constant C:

$$\sup_{\theta}\sup_{c} J_{1,n} < C \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left|b_1' \ddot{g}_{in}^{1*} b_1\right| \left(n^{-1/2} a_m\right)^2. \tag{3.321}$$

The supremum with respect to θ is taken over $|\theta| \leq B n^{1/4-\eta}$. Let the constant C incorporate the constant B, then

$$\sup_{\theta}\sup_{c} J_{1,n} < C \sup_{\theta} n^{-1/2-4\eta} \sum_{i=1}^{n} |w_{in}| \|N' \dot{g}(x_i; \beta_0)\|. \tag{3.322}$$

By Assumption 3.B.1/1b: $\sum_{i=1}^{n} |w_{in}| \sup_{\beta_i} \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{2\eta})$. Therefore, $\sup_\theta \sup_c J_{1,n} = o_p(n^{-1/2-2\eta})$. This result implies that $\sup_\theta \sup_c J_{1,n} = o_p(1)$, which we require. Therefore, it only remains to show that $\sup_\theta \sup_c J_{2,n} = o_p(1)$.

Term $J_{2,n}$.

Bound the function $\mathring{m}\left(c_{in}^0 + \left|\bar{d}\right|\right) f\left(c_{in}^0 + \left|\bar{d}\right|\right)$ by its supremum over $u \in \mathbb{R}$. Since c_{in}^0 is the only term which is a function of c, this allows for the supremum over c to be dropped:

$$\sup_{\theta}\sup_{c} J_{2,n} \leq \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left|d_{in}^*\right| \left|\frac{1}{s}\frac{1}{\sigma} b_m' \dot{g}_{in}^0\right|^2 \sup_u \mathring{m}(u) f(u). \tag{3.323}$$

By Assumption 3.B.1/2c: $\sup_u \mathring{m}(u) f(u) = C < \infty$, where C is a finite constant which does not depend on n. After substituting the definition of d_{in}^*, this gives the following bound:

$$\sup_{\theta}\sup_{c} J_{2,n} \leq C \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \frac{\frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1}{\sigma} \left|\frac{1}{s}\frac{1}{\sigma} b_m' \dot{g}_{in}^0\right|^2. \tag{3.324}$$

Lemma 3.A.1 and Remark 3.A.1 imply that, for sufficiently large n, $\left|\frac{1}{s} b_m' \dot{g}_{in}^0\right| < 2\left|b_m' \dot{g}_{in}^0\right|$. Let C incorporate the constants 2^2, $1/2$ and $1/\sigma$. Then

$$\sup_{\theta}\sup_{c} J_{2,n} < C \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left|b_1' \ddot{g}_{in}^{1*} b_1\right| \left|b_m' \dot{g}_{in}^0\right|^2. \tag{3.325}$$

The supremum with respect to θ is taken over $|\theta| \leq B n^{1/4-\eta}$. Let the constant C incorporate the

constant B. Further, bound $\|N'\ddot{g}(x_i;\beta_i)N\|$ by its supremum over β_i. Thus, the bound becomes

$$\sup_{\theta}\sup_{c} J_{2,n} < Cn^{1/2-4\eta}\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\ddot{g}(x_i;\beta_i)N\|\left|N'\dot{g}(x_i;\beta_0)\right|^2. \tag{3.326}$$

By Assumption 3.B.1/1d: $\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i}\|N'\dot{g}(x_i;\beta_0)\|^2\|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{-1/2+4\eta})$. Thus, $\sup_\theta \sup_c J_{2,n} = o_p(1)$, as required.

Since $\sup_\theta \sup_c J_{1,n} = o_p(n^{-1/2-2\eta})$ and $\sup_\theta \sup_c J_{2,n} = o_p(1)$, then by equation (3.316), it follows that $\sup_\theta \sup_c K^S_{1,n} = o_p(1)$. Consequently, all terms in equation (3.303) are at most $o_p(1)$. Thus, the premise of the lemma follows.

■

Assumption 3.B.2. Assume the following:

1. *The regressors and the function $g(\cdot)$:*
   (a) $\max_{1\le i\le n}\|N'\dot{g}(x_i;\beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$
   (b) $\max_{1\le i\le n}\sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\le Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\| = O_p(n^{-1/2+2\eta-\kappa})$
   (c) $\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\le Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{2\eta})$
   (d) $\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\le Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\|^2 = o_p(n^{-1/2+4\eta})$
   (e) $\sum_{i=1}^{n}|w_{in}|\sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\le Bn^{1/4-\eta}}\|N'\dot{g}(x_i;\beta_0)\|\|N'\ddot{g}(x_i;\beta_i)N\| = o_p(n^{-1/4+3\eta})$
2. *The mark function $m(u)$:*
   (a) $\int_{-\infty}^{\infty}|\dot{m}(u)|\,f(u)du < \infty$
   (b) (local Lipschitz) $\exists\omega > 0$ and a function $\mathring{m}(u) \ge 0$ such that if $\forall u^*, u$ such that $|u^*-u| \le (1+|u|)\omega$ then $|\dot{m}(u^*)-\dot{m}(u)| \le |u^*-u|\mathring{m}(u)$ and where:
      - $\sup_{u\in\mathbb{R}}(1+u)\mathring{m}(u)f(u) < \infty$
      - $\int_{-\infty}^{+\infty}(1+|u|)\mathring{m}(u)f(u)du < \infty \iff \mathsf{E}(1+|u|)\mathring{m}(u) < \infty$.

Theorem 3.B.2. Under Assumption 3.B.2, the following holds:

$$\sup_{\theta}\sup_{c} n^{-1/2}\sum_{i=1}^{n}|w_{in}|\mathsf{E}_{i-1}\left|\left\{m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right\}\right| \cdot 1_{\left(\varepsilon_i\le\sigma c+n^{-1/2}a_1c+b_1'\dot{g}_{in}^0+\frac{1}{2}b_1'\ddot{g}_{in}^{1*}b_1\right)} = o_p(1). \tag{3.327}$$

Proof. For the purposes of this proof, define the quantity of interest as $V_n(\theta, c)$. We wish to show that $\sup_\theta \sup_c V_n(\theta, c) = o_p(1)$:

$$V_n(\theta, c) := n^{-1/2} \sum_{i=1}^{n} |w_{in}| \mathsf{E}_{i-1} \left| \left\{ m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) - m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) \right\} \right| \cdot \\ \cdot 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0 + \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1\right)}. \tag{3.328}$$

We bound the indicator function by unity, which allows us to drop the supremum over c:

$$\sup_\theta \sup_c V_n(\theta, c) \leq \sup_\theta n^{-1/2} \sum_{i=1}^{n} |w_{in}| \mathsf{E}_{i-1} \left| m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) - m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) \right|. \tag{3.329}$$

Let $\bar{g}_m$ be an intermediate point such that $|\bar{g}_m| \leq |\ddot{g}_{in}^{m*}|$. Use a first-order Taylor series expansion of $m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m}\right)$ around $m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m}\right)$ to re-express the difference between the values of the mark function:

$$\begin{aligned} & m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \\ & = m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) + \left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} - \frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) \dot{m}\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) \\ & = m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) + \left(\frac{-\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \dot{m}\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right). \end{aligned} \tag{3.330}$$

The Taylor series expansion allows us to re-express the bound on $\sup_\theta \sup_c V_n(\theta, c)$ as a function of $\dot{m}(\cdot)$. Furthermore, express the expectation E_{i-1} as an integral with respect to the i.i.d. innovations ε_i/σ, which results in the bound:

$$\sup_\theta \sup_c V_n(\theta, c) \leq \sup_\theta n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \left| \left(\frac{-\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \dot{m}\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) \right| f\left(\frac{\varepsilon_i}{\sigma}\right) d\left(\frac{\varepsilon_i}{\sigma}\right). \tag{3.331}$$

Add and subtract $\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)$, in order to arrive at an expression which we can apply the Lipschitz condition to. Use the triangle inequality to bound the absolute value of a sum into a sum of

absolute values:

$$\begin{aligned}
&\sup_\theta \sup_c V_n(\theta, c) \leq \\
&\leq \sup_\theta \sup_c n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \left| \left(\frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \left\{ \dot{m} \left(\frac{\varepsilon_i - b'_m \dot{g}_{in}^0 - \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right\} \right| f\left(\frac{\varepsilon_i}{\sigma} \right) d\left(\frac{\varepsilon_i}{\sigma} \right) + \\
&+ \sup_\theta \sup_c n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \left| \left(\frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right| f\left(\frac{\varepsilon_i}{\sigma} \right) d\left(\frac{\varepsilon_i}{\sigma} \right).
\end{aligned} \tag{3.332}$$

We analyse each term in equation (3.332) in turn, with the aim of showing that both terms are $o_p(1)$. For ease of notation, we define the terms as $V_{1,n}(\theta)$ and $V_{2,n}(\theta)$ respectively:

$$\begin{aligned}
V_{1,n}(\theta) &:= n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \left| \left(\frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \left\{ \dot{m} \left(\frac{\varepsilon_i - b'_m \dot{g}_{in}^0 - \frac{1}{2} b'_m \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right\} \right| f\left(\frac{\varepsilon_i}{\sigma} \right) d\left(\frac{\varepsilon_i}{\sigma} \right), \\
V_{2,n}(\theta) &:= n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \left| \left(\frac{\frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \right) \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right| f\left(\frac{\varepsilon_i}{\sigma} \right) d\left(\frac{\varepsilon_i}{\sigma} \right).
\end{aligned} \tag{3.333}$$

Term $\sup_\theta \sup_c V_{1,n}(\theta)$ is $o_p(1)$ by Lemma 3.A.7 under Assumptions 3.B.2/1a, 1b, 1c, 1d, 1e and 2b, which together are equivalent to Assumption 3.A.8, which is required for Lemma 3.A.7.

For term $\sup_\theta V_{2,n}(\theta)$, use the definition of the scale to re-express the bound as

$$\sup_\theta V_{2,n}(\theta) = \sup_\theta n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \left| \frac{\frac{1}{s} \frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m}{\sigma} \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right| f\left(\frac{\varepsilon_i}{\sigma} \right) d\left(\frac{\varepsilon_i}{\sigma} \right). \tag{3.334}$$

From Lemma 3.A.1 and Remark 3.A.1, it follows that, for sufficiently large n, $\left| \frac{1}{s} \frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m \right| < 2 \left| \frac{1}{2} b'_m \ddot{g}_{in}^{m*} b_m \right|$. Thus, we obtain the strict bound

$$\sup_\theta V_{2,n}(\theta) < \sup_\theta n^{-1/2} \sum_{i=1}^{n} |w_{in}| \int_{-\infty}^{\infty} \frac{|b'_m \ddot{g}_{in}^{m*} b_m|}{\sigma} \left| \dot{m} \left(\frac{\varepsilon_i}{\sigma} \right) \right| f\left(\frac{\varepsilon_i}{\sigma} \right) d\left(\frac{\varepsilon_i}{\sigma} \right). \tag{3.335}$$

The final term is $o_p(1)$ by Lemma 3.A.6 under Assumptions 3.B.2/1c and 3.B.2/2a, which are equivalent to Assumption 3.A.7 required for Lemma 3.A.6.

Thus, both terms in equation (3.332) are $o_p(1)$, which concludes the proof.

■

3.B.6 Theorems Required for Result (3.21)

Assumption 3.B.3. Assume the following:

1. *The regressors and the function $g(\cdot)$:*

(a) $\max_{1\leq i\leq n}\|N'\dot{g}(x_i;\beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$

(b) $\max_{1\leq i\leq n}\sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\| = O_p(n^{-1/2+2\eta-\kappa})$

(c) $\sum_{i=1}^n |w_{in}|^2 \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\|^2 = o_p(n^{4\eta-1})$

(d) $\mathsf{E}\sum_{i=1}^n |w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\| = O(n^{1/4+2\eta})$

(e) $\mathsf{E}\sum_{i=1}^n |w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\|\|N'\dot{g}(x_i;\beta_0)\| = O(n^{-1/4+2\eta})$

(f) $\mathsf{E}\sum_{i=1}^n |w_{in}| \sup_{\beta_i:\|N^{-1}(\beta_i-\beta_0)\|\leq Bn^{1/4-\eta}}\|N'\ddot{g}(x_i;\beta_i)N\|^2 = O(n^{-1/2+3\eta})$

2. *The mark function* $m(u)$:

(a) $\mathsf{E}\,|\dot{m}(u)|^2 < \infty$

(b) (local Lipschitz) $\exists\omega > 0$ and a function $\mathring{m}(u) \geq 0$ such that if $\forall u^*, u$ such that $|u^* - u| \leq (1+|u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u|\mathring{m}(u)$ and where:

- $\sup_{u\in\mathbb{R}}(1+u)\mathring{m}(u)f(u) < \infty$
- $\int_{-\infty}^{+\infty}(1+|u|)\mathring{m}(u)f(u)du < \infty \iff \mathsf{E}(1+|u|)\mathring{m}(u) < \infty$.

Theorem 3.B.3. Under Assumption 3.B.3, as $n\to\infty$,

$$\sup_\theta n^{-1/2}\sum_{i=1}^n |w_{in}|\left|m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right| = o_p(1). \tag{3.336}$$

Proof.

For the purposes of this proof, denote the quantity of interest as $V_n(\theta)$. We wish to show that $\sup_\theta V_n(\theta) = o_p(1)$:

$$V_n(\theta) := n^{-1/2}\sum_{i=1}^n |w_{in}|\left|m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)\right|. \tag{3.337}$$

Let $\bar{g}_m$ be an intermediate point such that $|\bar{g}_m| \leq |\ddot{g}_{in}^{m*}|$. Analogously to equation (3.330) use a first-order Taylor series expansion of $m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)$ around $m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right)$, to obtain the following equality:

$$m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma + n^{-1/2}a_m}\right) - m\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0}{\sigma+n^{-1/2}a_m}\right) = \left(\frac{-\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\right)\dot{m}\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma+n^{-1/2}a_m}\right). \tag{3.338}$$

Thus, the quantity of interest can be re-expressed as a function of $\dot{m}(\cdot)$:

$$\sup_\theta V_n(\theta) = \sup_\theta n^{-1/2}\sum_{i=1}^n |w_{in}|\left|\frac{\frac{1}{2}b_m'\ddot{g}_{in}^{m*}b_m}{\sigma+n^{-1/2}a_m}\dot{m}\left(\frac{\varepsilon_i - b_m'\dot{g}_{in}^0 - \frac{1}{2}b_m'\bar{g}_m b_m}{\sigma+n^{-1/2}a_m}\right)\right|. \tag{3.339}$$

Add and subtract $\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)$, in order to arrive at an expression which we can apply the Lipschitz condition to. Use the triangle inequality to bound the absolute value of a sum by the sum of absolute values:

$$\sup_\theta V_n(\theta) \leq \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \left\{ \dot{m}\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right\} \right| + \\ + \sup_\theta n^{-1/2} \sum_{i=1}^n |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right|. \tag{3.340}$$

We analyse each term in equation (3.340) in turn to show that both are $o_p(1)$. For ease of notation, we denote the terms as $V_{1,n}(\theta)$ and $V_{2,n}(\theta)$ respectively, defined as:

$$\begin{aligned} V_{1,n}(\theta) &:= n^{-1/2} \sum_{i=1}^n |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \left\{ \dot{m}\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0 - \frac{1}{2} b_m' \bar{g}_m b_m}{\sigma + n^{-1/2} a_m} \right) - \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right\} \right|, \\ V_{2,n}(\theta) &:= n^{-1/2} \sum_{i=1}^n |w_{in}| \left| \frac{\frac{1}{2} b_m' \ddot{g}_{in}^{m*} b_m}{\sigma + n^{-1/2} a_m} \dot{m}\left(\frac{\varepsilon_i}{\sigma}\right) \right|. \end{aligned} \tag{3.341}$$

The term $\sup_\theta V_{1,n}(\theta)$ is $o_p(1)$ by Lemma 3.A.9 under Assumptions 3.B.3/1a, 1b, 1d, 1e, 1f, 2b, which together are equivalent to Assumption 3.A.10, which is required for Lemma 3.A.9. The term $\sup_\theta V_{2,n}(\theta)$ is $o_p(1)$ by Lemma 3.A.8 under Assumptions 3.B.3/1c and 2a, which are equivalent to Assumption 3.A.9, which is required for Lemma 3.A.8. Consequently, both terms in equation (3.340) are $o_p(1)$, which concludes the proof.

■

Assumption 3.B.4. Let $\tilde{v}$ represent each of $w_{in} n^{-1/4-\eta}$ and $w_{in} n^{1/4-\eta} |N' \dot{g}(x_i; \beta_0)|$. Let $\tilde{h}(u)$ represent each of $\dot{m}(u)$ and $u\dot{m}(u)$. Let r be such that $2^{r-1} > 1 + (1/4 - \eta)(\dim b_1)$ and q such that $1 \leq q \leq r$. Further, define:

$$\tilde{H}_r(c) = \int_{-\infty}^c \left(1 + \{\tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right) f\left(\frac{\varepsilon_i}{\sigma}\right) \mathrm{d}\left(\frac{\varepsilon_i}{\sigma}\right) = \mathsf{E}\left(1 + \{\tilde{h}\left(\frac{\varepsilon_i}{\sigma}\right)\}^{2^r}\right) 1_{(\varepsilon_i/\sigma \leq c)}. \tag{3.342}$$

Assume the following:

1. *The weights $\tilde{v}_{in}$*:
   (a) The weights $\tilde{v}_{in}$ are $\mathcal{F}_{i-1}$-measurable.
   (b) $\mathsf{E}|\tilde{v}_{in}|^{2^r} < \infty$
   (c) For all $1 \leq q \leq r$: $\mathsf{E} \sum_{i=1}^n |\tilde{v}_{in}|^{2^q} = O(n)$

(d) $\sum_{i=1}^{n} |\tilde{v}_{in}| = O_p(n)$

2. *The weights* w_{in}:

   (a) $\mathsf{E} \sum_{i=1}^{n} |w_{in}| = O(n)$

3. *The regressors and the function* $g(\cdot)$*, for all* $1 \leq q \leq r$*:*

   (a) $\max_{1 \leq i \leq n} \|N' \dot{g}(x_i; \beta_0)\| = O_p(n^{-1/4+\eta-\kappa})$

   (b) $\mathsf{E} \sum_{i=1}^{n} |w_{in}| \, |N' \dot{g}(x_i; \beta_0)|^2 = O(n^{\eta})$

   (c) $\mathsf{E} \sum_{i=1}^{n} |\tilde{v}_{in}|^{2^q} \|N' \dot{g}(x_i; \beta_0)\| = O(n^{3/4-\eta})$

   (d) $\mathsf{E} \sum_{i=1}^{n} |\tilde{v}_{in}|^{2^q} \|N' \dot{g}(x_i; \beta_0)\|^2 = O(n^{1/4+\eta})$

   (e) $\mathsf{E} \sum_{i=1}^{n} |\tilde{v}_{in}|^{2^q} \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = O(n^{2\eta})$

   (f) $\sum_{i=1}^{n} |\tilde{v}_{in}| \|N' \dot{g}(x_i; \beta_0)\| = O_p(n^{1/2})$

   (g) $\sum_{i=1}^{n} |\tilde{v}_{in}| \|N' \dot{g}(x_i; \beta_0)\|^2 = o_p(n^{2\eta})$

   (h) $\sum_{i=1}^{n} |\tilde{v}_{in}| \sup_{\beta_i : \|N^{-1}(\beta_i - \beta_0)\| \leq B n^{1/4-\eta}} \|N' \ddot{g}(x_i; \beta_i) N\| = o_p(n^{2\eta})$

4. *The mark function* $m(u)$:

   (a) (local Lipschitz) $\exists \omega > 0$ and a function $\mathring{m}(u) \geq 0$ such that if $\forall u^*, u$ such that $|u^* - u| \leq (1 + |u|)\omega$ then $|\dot{m}(u^*) - \dot{m}(u)| \leq |u^* - u| \mathring{m}(u)$ and where:

   - $\int_{-\infty}^{+\infty} (1 + |u| + u^2) \mathring{m}(u) f(u) du < \infty$.

5. *The mark function* $\tilde{h}(\cdot)$*:*

   (a) $\mathsf{E} \left| \tilde{h} \left(\frac{\varepsilon_i}{\sigma} \right) \right|^{2^r} < \infty$

6. *The distance function* $\tilde{H}_r(c)$:

   (a) the support of $\dot{\tilde{H}}_r$, that is $\mathcal{S} = \{c : \dot{\tilde{H}}_r(c) > 0\}$, is an interval with endpoints $\underline{c}, \bar{c}$ such that $-\infty \leq \underline{c} < \bar{c} \leq \infty$;

   (b) $\sup_{x \in \mathcal{S}} 0 \leq \dot{\tilde{H}}_r(x) < \infty$;

   (c) $\tilde{H}_r = \tilde{H}_r(\infty) = \int_{\mathbb{R}} \dot{\tilde{H}}_r(c) dc < \infty$;

   (d) $\dot{\tilde{H}}_r(c)$ is Lipschitz, as defined in Definition 3.4.1;

   (e) $\dot{\tilde{H}}_r(c)$ has weakly unimodal bounds as defined in Definition 3.4.1.

Theorem 3.B.4. Under Assumption 3.B.4, the following holds as $n \to \infty$,

$$\sup_{\theta} \sup_{c \in \mathbb{R}} n^{-1/2} \left| \sum_{i=1}^{n} w_{in} \left\{ m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) - m\left(\frac{\varepsilon_i}{\sigma} \right) \right\} \left\{ 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c_{in}^0 + d_{in}^* \right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c_{in}^0 \right)} \right\} \right| = o_p(1). \tag{3.343}$$

Proof. Recall the definition of $c_{in}^0 := \left(\sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0 \right) / \sigma$. Thus, the objective is to show the following.

$$\begin{aligned} &\sup_{\theta} \sup_{c} n^{-1/2} \left| \sum_{i=1}^{n} w_{in} \left\{ m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) - m\left(\frac{\varepsilon_i}{\sigma} \right) \right\} \cdot \right. \\ &\left. \cdot \left\{ 1_{\left(\frac{\varepsilon_i}{\sigma} \leq \frac{\sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0}{\sigma} + d_{in}^* \right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq \frac{\sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0}{\sigma} \right)} \right\} \right|. \end{aligned} \tag{3.344}$$

The proof relies on first showing that $\sup_{\theta} \sup_{c} V_n(\theta, c) = o_p(1)$, where $V_n(\theta, c)$ is defined as

$$V_n(\theta, c) := n^{-1/2} \left| \sum_{i=1}^{n} w_{in} \left\{ m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) - m\left(\frac{\varepsilon_i}{\sigma} \right) \right\} \left\{ 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma} + d_{in}^* \right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma} \right)} \right\} \right|. \tag{3.345}$$

In particular, note that the only difference between $V_n(\theta, c)$ and the term inside the supremum in equation (3.344) is that the term $V_n(\theta, c)$ does not include $(n^{-1/2} a_1 c)/\sigma$ in the indicator functions. In the following proof the result is initially proven for $V_n(\theta, c)$, and it is then extended to include the term $(n^{-1/2} a_1 c)/\sigma$ in the indicator function. For an initial bound on $\sup_{\theta} \sup_{c} V_n(\theta, c)$, use the triangle inequality, to take the modulus operator inside the summation. Further, note that for any $x, y \in \mathbb{R}$: $|xy| = |x||y|$. Thus

$$\begin{aligned} &\sup_{\theta} \sup_{c} V_n(\theta, c) \leq \\ &\leq \sup_{\theta} \sup_{c} n^{-1/2} \sum_{i=1}^{n} |w_{in}| \left| m\left(\frac{\varepsilon_i - b_m' \dot{g}_{in}^0}{\sigma + n^{-1/2} a_m} \right) - m\left(\frac{\varepsilon_i}{\sigma} \right) \right| \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma} + d_{in}^* \right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma} \right)} \right|. \end{aligned} \tag{3.346}$$

Apply linearisations as given in Definition 3.A.1. Use $a_n^s := n^{-1/2} a_m / \sigma$ to simplify the notation:

$$\begin{aligned} \sup_{\theta} \sup_{c} V_n(\theta, c) \leq \sup_{\theta} \sup_{c} n^{-1/2} \sum_{i=1}^{n} |w_{in}| &\left| m\left(\frac{\varepsilon_i (1 - \frac{1}{s} a_n^s) - \frac{1}{s} b_m' \dot{g}_{in}^0}{\sigma} \right) - m\left(\frac{\varepsilon_i}{\sigma} \right) \right| \cdot \\ &\cdot \left| 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma} + d_{in}^* \right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c + \frac{b_1' \dot{g}_{in}^0}{\sigma} \right)} \right|. \end{aligned} \tag{3.347}$$

Use a first order Taylor series expansion of $m\left(\frac{\varepsilon_i(1-\frac{1}{s}a_n^s)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}\right)$ around $m\left(\frac{\varepsilon_i}{\sigma}\right)$, for an intermediate point u^* such that: $\left|u^*-\frac{\varepsilon_i}{\sigma}\right| \leq \left|\frac{\varepsilon_i(1-\frac{1}{s}a_n^s)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right|$, to express the bound as

$$\begin{aligned}\sup_\theta \sup_c V_n(\theta,c) \leq \sup_\theta \sup_c n^{-1/2}\sum_{i=1}^n |w_{in}| \left|\frac{\varepsilon_i(1-\frac{1}{s}a_n^s)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right| \cdot \\ \cdot \left|\dot{m}\left(u^*\right)\right| \left|1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}+d^*_{in}\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}\right)}\right|.\end{aligned} \tag{3.348}$$

Add and subtract $\dot{m}\left(\varepsilon_i/\sigma\right)$, which will allow us to apply Lemma 3.A.11, which is based on the Lipschitz condition. Use the triangle inequality as well as the fact that the supremum of a sum is no greater than the sum of individual suprema, to obtain the bound

$$\sup_\theta \sup_c V_n(\theta,c) \leq \sup_\theta \sup_c V_{1,n}(\theta,c) + \sup_\theta \sup_c V_{2,n}(\theta,c), \tag{3.349}$$

where $V_{1,n}(\theta,c)$ and $V_{2,n}(\theta,c)$ are defined as

$$\begin{aligned}V_{1,n}(\theta,c) :=& n^{-1/2}\sum_{i=1}^n |w_{in}| \left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right| \left|\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \cdot \\ & \cdot \left|1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}+d^*_{in}\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}\right)}\right|,\end{aligned} \tag{3.350}$$

and

$$\begin{aligned}V_{2,n}(\theta,c) :=& n^{-1/2}\sum_{i=1}^n |w_{in}| \left|\frac{\varepsilon_i(1-n^{-1/2}\frac{1}{s}a_m/\sigma)-\frac{1}{s}b'_m\dot{g}^0_{in}}{\sigma}-\frac{\varepsilon_i}{\sigma}\right| \left|\dot{m}\left(u^*\right)-\dot{m}\left(\frac{\varepsilon_i}{\sigma}\right)\right| \cdot \\ & \cdot \left|1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}+d^*_{in}\right)}-1_{\left(\frac{\varepsilon_i}{\sigma}\leq c+\frac{b'_1\dot{g}^0_{in}}{\sigma}\right)}\right|.\end{aligned} \tag{3.351}$$

The first term, $\sup_\theta \sup_c V_{1,n}$, is $o_p(1)$ by Lemma 3.A.10 under Assumptions 3.B.4/1, 3c to 3h, 5 and 6, which together are equivalent to Assumption 3.A.11. The second term, $\sup_\theta \sup_c V_{2,n}$, is $o_p(1)$ by Lemma 3.A.11 under Assumptions 3.B.4/2, 3a, 3b and 4, which collectively are equivalent to Assumption 3.A.12, which is required for Lemma 3.A.11.

Thus, by equation (3.349), the term $\sup_\theta \sup_c V_n(\theta,c)$ is $o_p(1)$. In order to prove Theorem 3.B.4, the result must be extended to include the term $n^{-1/2}a_1c/\sigma$ in the indicator function.

Define $\tilde{a} := a_1/\sigma$, and $c_a := c + n^{-1/2}\tilde{a}c$. Note that the supremum with respect to a_1 is taken over $\left\{|a_1| \leq Bn^{1/4-\eta}\right\}$. Thus, $n^{-1/2}\tilde{a} = O(n^{-1/4-\eta})$. Consequently, for $\sigma < \infty$ there exists $n_0 < \infty$ such that $\forall n > n_0 : n^{-1/2}\tilde{a} < 1$.

Since $c_a \in \mathbb{R}$ then in order to show that term (3.343) is $o_p(1)$ for all $|a_1| \leq Bn^{1/4-\eta}$, it suffices to show that term (3.343) is $o_p(1)$ uniformly in $c_a \in \mathbb{R}$. Since the result has been proven to hold uniformly in $c \in \mathbb{R}$, it therefore holds for any $c_a \in \mathbb{R}$.

■

Assumption 3.B.5. Let Assumption 3.A.4 hold with $v_{in} = w_{in}$ and $h(u) = m(u)$.

Theorem 3.B.5. Under Assumption 3.B.5, the following holds as $n \to \infty$,

$$\sup_{\theta} \sup_{c} n^{-1/2} \left| \sum_{i=1}^{n} w_{in} m\left(\frac{\varepsilon_i}{\sigma}\right) \left\{ 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c_{in}^0 + d_{in}^*\right)} - 1_{\left(\frac{\varepsilon_i}{\sigma} \leq c_{in}^0\right)} \right\} \right| = o_p(1). \tag{3.352}$$

Proof. Use the triangle inequality to take the absolute value inside the summation sign. Additionally, apply the the definitions of $c_{in}^0 := \left(\sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0\right)/\sigma$ and $d_{in}^* := \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1/\sigma$. Thus, the result to be proven is

$$\sup_{c} \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} \left| w_{in} m\left(\frac{\varepsilon_i}{\sigma}\right) \{ 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0 + \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1\right)} - 1_{\left(\varepsilon_i \leq \sigma c + n^{-1/2} a_1 c + b_1' \dot{g}_{in}^0\right)} \} \right| = o_p(1). \tag{3.353}$$

Firstly, note that applying Lemma 3.A.3 with $v_{in} = w_{in}$ and $h(u) = m(u)$ gives the following result, which only differs from the result to be proven by the term $n^{-1/2} a_1 c$ in the indicator function:

$$\sup_{c} \sup_{\theta} n^{-1/2} \sum_{i=1}^{n} \left| w_{in} m\left(\frac{\varepsilon_i}{\sigma}\right) \{ 1_{\left(\varepsilon_i \leq \sigma c + b_1' \dot{g}_{in}^0 + \frac{1}{2} b_1' \ddot{g}_{in}^{1*} b_1\right)} - 1_{\left(\varepsilon_i \leq \sigma c + b_1' \dot{g}_{in}^0\right)} \} \right| = o_p(1). \tag{3.354}$$

In order to prove Theorem 3.B.5, the result must be extended to include the term $n^{-1/2} a_1 c$ in the indicator function. To this end, define $\tilde{a} := a_1/\sigma$, and $c_a := c + n^{-1/2} \tilde{a} c$. Note that the supremum with respect to a_1 is taken over $\left\{|a_1| \leq Bn^{1/4-\eta}\right\}$. Thus, $n^{-1/2}\tilde{a} = O(n^{-1/4-\eta})$. Consequently, for $\sigma < \infty$ there exists $n_0 < \infty$ such that $\forall n > n_0 : n^{-1/2}\tilde{a} < 1$. Since $c_a \in \mathbb{R}$ then in order to show that term (3.353) is $o_p(1)$ for all $|a_1| \leq Bn^{1/4-\eta}$, it suffices to show that term (3.353) is $o_p(1)$ uniformly in $c_a \in \mathbb{R}$. Since the result has been proven to hold uniformly in $c \in \mathbb{R}$, it therefore holds for any $c_a \in \mathbb{R}$.

■

Thesis Conclusion

This thesis presented three chapters which make contributions to the fields of applied and theoretical Econometrics.

The first two chapters focused on settings in which the restrictions imposed by the model are not sufficient to uniquely identify the parameters of interest. In Chapter 1, I presented a method, based on the maximum-entropy principle, which can be used to achieve point-identification of agent-specific parameter vectors in revealed-preference settings. I presented methods of model selection and heterogeneity analysis in this context, and I applied them to data from the Human Sociality Project to analyse the determinants of other-regarding preferences.

In Chapter 2, I analysed the computational aspects of implementing partial identification in practice, in the context of binary-outcome random-coefficients models with endogeneity analysed by Chesher and Rosen (2014). To aid the implementation of this novel set-identifying technique, I explained the key results of the model and I contrasted the set-identifying model with point-identifying alternatives. I illustrated, through pseudo-code and Matlab code, how the method can be implemented and I analysed its computational complexity. I applied the method to data from the British Household Panel Survey, to analyse female labour participation decisions.

In Chapter 3, I derived asymptotic theory for marked and weighted empirical processes of residuals from non-linear regressions with outlier detection. I showed that, under certain assumptions on the regressors, the non-linear function, the marks, the weights, and the errors; the empirical process based on non-linear regressions is asymptotically equivalent to that based on linear regressions. This, in turn, allowed me to show its asymptotic equivalence to empirical processes based on the true innovations, without estimation errors. I illustrated through an example how each of the assumptions made in the chapter can be verified and I suggested several applications of the theory.

Bibliography

Adams, A. (2013). 'The Nonparametric Approach to Demand Analysis: Essays in Revealed Preference Theory'. DPhil thesis, University of Oxford.

Alih, E. and Ong, H. C. (2015), 'An outlier-resistant test for heteroscedasticity in linear models', *Journal of Applied Statistics 42(8)*, 1617-1634.

Amemiya, T. (1983), 'Non-linear regression models', *Handbook of Econometrics*, 333-389.

Andreoni, J., Castillo, M., & Petrie, R. (2003). 'What do bargainers' preferences look like? Experiments with a convex ultimatum game'. *American Economic Review*, 672-685.

Andreoni, J., & Miller, J. (2002). 'Giving according to GARP: An experimental test of the consistency of preferences for altruism'. *Econometrica, 70(2)*, 737-753.

Andrews, D. W. K., & P. Guggenberger (2009). 'Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities'. *Econometric Theory*, 25, 669-709.

Andrews, D. W. K., & G. Soares (2010). 'Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection'. *Econometrica*, 78(1), 119-157.

Anscombe, F. J. and Tukey, J. W. (1963), 'The Examination and Analysis of Residuals', *Technometrics 5(2)*, 141-160.

Artstein, Z. (1983). 'Distributions of Random Sets and Random Selections'. *Israel Journal of Mathematics*, 46(4), 313-324.

Atkinson, A. C. (1981), 'Two graphical displays for outlying and influential observations in regression', *Biometrika 68*, 13-20.

Atkinson, A. C. (1994), 'Fast Very Robust Methods for the Detection of Multiple Outliers', *Journal of the American Statistical Association 89(428)*, 1329-1339.

Bajari, P., Fox, J., Kim, K.-i. & S. P. Ryan (2012). 'The Random Coefficients Logit Model is Identified', *Journal of Econometrics*, 166(2).

Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E. L. M, Henrich, J., Lesorogol, C., Marlowe, F., McElreath, R., Tracer, D. and Ziker, J. 'Human Sociality Project: Phase II'. [data collection]

Beatty, T. K., & Crawford, I. A. (2011). 'How demanding is the revealed preference approach to demand?'. *American Economic Review, 101(6)*, 2782-2795.

Becker, G. S. (1962). 'Irrational behavior and economic theory'. *Journal of Political Economy*, 1-13.

Bellemare, C., Kröger, S., & Van Soest, A. (2008). 'Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities'. *Econometrica, 76(4)*, 815-839.

Bellemare, C., Kröger, S., & van Soest, A. (2011). 'Preferences, intentions, and expectation violations: A large-scale experiment with a representative subject pool'. *Journal of Economic Behavior & Organization, 78(3)*, 349-365.

Belsley, D. A., Kuh, E. and Welsch, R. E. (2005), *Regression diagnostics: Identifying influential data and sources of collinearity*, Vol. 571, John Wiley & Sons.

Bercu, B. and Touati, A. (2008), 'Exponential inequalities for self-normalized martingales with applications', *The Annals of Applied Probability 18(5)*, 1848-1869.

Berenguer-Rico, V., Johansen, S. and Nielsen, B. (2019a), 'The Analysis of Marked and Weighted Empirical Processes of Estimated Residuals', Department of Economics Discussion Paper Series No. 870, University of Oxford.

Berenguer-Rico, V., Johansen, S. and Nielsen, B. (2019b), 'Uniform Consistency of Marked and Weighted Empirical Distributions of Residuals', Department of Economics Discussion Paper Series No. 871, University of Oxford.

Berenguer-Rico, V. and Nielsen, B. (2017), 'Marked and Weighted Empirical Processes of Residuals with Applications to Robust Regressions', University of Oxford Working Paper.

Berenguer-Rico, V. and Nielsen, B. (2020), 'Cumulated Sum of Squares Statistics for Nonlinear and Nonstationary Regressions', *Econometric Theory 36(1)*, 1-47.

Berenguer-Rico, V. and Wilms, I. (2021), 'Heteroscedasticity testing after outlier removal',

Econometric Reviews 40(1), 51-85.

Beresteanu, A., I. Molchanov, and F. Molinari (2012). 'Partial Identification Using Random Set Theory'. *Journal of Econometrics*, 166(1), 17-32.

Beresteanu, A., & Molinari, F. (2008). 'Asymptotic properties for a class of partially identified models'. *Econometrica, 76(4)*, 763-814.

Berry, S. T., & Haile, P. A. (2009). 'Nonparametric identification of multinomial choice demand models with heterogeneous consumers'. NBER Working Paper No. w15276. National Bureau of Economic Research.

Berry, S. T., & Haile, P. A. (2010). 'Identification in differentiated products markets using market level data'. NBER Working Paper No. w15641. National Bureau of Economic Research.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). 'A within-subject analysis of other-regarding preferences'. *Games and Economic Behavior, 72(2)*, 321-338.

Blundell, R., Dearden, L., and B. Sianesi (2001). 'Estimating the returns to education: Models, methods and results'. Centre for Economics of Education Working Paper.

Blundell, R. and J. L. Powell (2004). 'Endogeneity in Semiparametric Binary Response Models'. *Review of Economic Studies*, 71 (3): 655-679.

Bolton, G. E., & Ockenfels, A. (2000). 'ERC: A theory of equity, reciprocity, and competition'. *American Economic Review*, 166-193.

Briesch, R. A., Chintagunta, P. K., & Matzkin, R. L. (2010). 'Nonparametric discrete choice models with unobserved heterogeneity'. *Journal of Business and Economic Statistics, 28(2).*

Browning, M., Crawford, I., & Knoef, M. (2014). 'An entropy-based approach to the APC problem, with applications to female mortality and labour-force participation'. Nuffield College Working Paper. Nuffield College, University of Oxford.

Burgess, S. M. (1992), 'Nonlinear Dynamics in a Structural Model of Employment', *Journal of Applied Econometrics 7*, S101-S118.

Burgess, S.M. & Propper, C.(1998). 'Early Health-Related Behaviours and Their Impact on Later Life Chances: Evidence from the US', *Health Economics*, Vol. 7, pp. 381–399.

Charness, G., & Haruvy, E. (2002). 'Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach'. *Games and Economic Behavior, 40(2)*, 203-231.

Charness, G., & Rabin, M. (2002). 'Understanding social preferences with simple tests'. *Quarterly journal of Economics*, 817-869.

Charness, G., & Rabin, M. (2005). 'Expressed preferences and behavior in experimental games'. *Games and Economic Behavior, 53(2)*, 151-169.

Chernozhukov, V., D. Chetverikov, and K. Kato (2015). 'Testing Many Moment Inequalities'. CeMMAP working paper CWP42/16.

Chernozhukov, V., Hong, H., & Tamer, E. (2007). 'Estimation and confidence regions for parameter sets in econometric models'. *Econometrica, 75(5)*, 1243-1284.

Chernozhukov, V., W. Kim, S. Lee, and A. M. Rosen (2015). 'Implementing Intersection Bounds in Stata'. *Stata Journal*, 15(1), 21-44.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). 'Intersection Bounds: Estimation and Inference'. *Econometrica*, 81(2), 667-737.

Chesher, A., and A.M. Rosen (2014). 'An Instrumental Variable Random Coefficients Model for Binary Outcomes'. *The Econometrics Journal*, 17, S1-S19.

Chesher, A., and A.M. Rosen (2017). 'Generalized Instrumental Variable Models'. *Econometrica, 85 (3)*, 959-989.

Chesher, A., and A.M. Rosen (2020). 'Generalized instrumental variable models, methods, and applications' in *Handbook of Econometrics*, Vol. 7, Part A, 1-110.

Chesher, A., A. M. Rosen, and K. Smolinski (2013). 'An Instrumental Variable Model of Multiple Discrete Choice'. *Quantitative Economics*, 4(2), 157-196.

Chevalier, A. (2011). 'Subject Choice and Earnings of UK Graduates'. *Economics of Education Review*, 30(6), 1187-1201.

Cook, R. D. and Weisberg, S. (1982), *Residuals and influence in regression*, New York: Chapman and Hall.

Conrath, D. W., & Deci, E. L. (1969). 'The determination and scaling of a bivariate utility function'. *Behavioral Science, 14(4)*, 316-327.

Conte, A., & Moffatt, P. G. (2014). 'The econometric modelling of social preferences'. *Theory and Decision, 76(1)*, 119-145.

Cox, J. & Sadiraj, V (2007). 'Direct Tests of Models of Social Preferences and a New Model'. Georgia State University Department of Economics Working Paper No. 07-13.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*, 2nd edition. Hoboken, NJ: Wiley Interscience.

Csiszar, I. (1984). 'Sanov property, generalized I-projection and a conditional limit theorem'. *Annals of Probability*, 768-793.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. Oxford, UK: Oxford University Press.

Deck, C. A. (2001). 'A test of game-theoretic and behavioral models of play in exchange and insurance environments'. *American Economic Review*, 1546-1555.

Dickson, M. (2013). 'The Causal Effect of Education on Wages Revisited'. *Oxford Bulletin of Economics and Statistics*, 75, 4(2013), 477-498.

Dufwenberg, M., & Kirchsteiger, G. (2004). 'A theory of sequential reciprocity'. *Games and economic behavior, 47(2)*, 268-298.

Dunker, F., Hoderlein, S., & Kaido, H. (2014). 'Nonparametric identification of endogenous and heterogeneous aggregate demand models: complements, bundles and the market level'. Cemmap Working Paper No. CWP23/1. Centre for Microdata Methods and Practice.

Edgeworth, F. Y. (1887), 'On Observations Relating to Several Quantities', Hermathena 6(13), 279-285.

Engelmann, D., & Strobel, M. (2004). 'Inequality aversion, efficiency, and maximin preferences in simple distribution experiments'. *American Economic Review*, 857-869.

Escanciano, J. C. (2007), 'Model checks using residual marked empirical processes', *Statistica Sinica 17(1)*, 115-138.

Evans, W. N. and E. Montgomery (1994). 'Education and Health: Where there's Smoke there's an Instrument', Working Paper no. 4949, National Bureau of Economic Research, Cambridge, MA.

Falk, A., & Fischbacher, U. (2006). 'A theory of reciprocity'. *Games and Economic Behavior, 54(2)*, 293-315.

Fehr, E., & Schmidt, K. M. (2006). 'The economics of fairness, reciprocity and altruism —

experimental evidence and new theories'. *Handbook of the economics of giving, altruism and reciprocity*, 1, 615-691.

Fehr, E., & Schmidt, K. M. (1999). 'A theory of fairness, competition, and cooperation'. *Quarterly Journal of Economics*, 817-868.

Fisman, R., Kariv, S., & Markovits, D. (2007). 'Individual preferences for giving'. *American Economic Review, 97(5)*, 1858-1876.

Fox, J. T., and A. Gandhi (2012). 'Nonparametric Identification and Estimation of Random Coefficients in Multinomial Choice Models', working paper, University of Michigan.

Fox, J. T., & Lazzati, N. (2013). 'Identification of discrete choice models for bundles and binary games'. Cemmap Working Paper No. CWP04/13. Centre for Microdata Methods and Practice.

Freedman, D. A. (1975), 'On Tail Probabilities for Martingales', *The Annals of Probability 3(1)*, 100-118.

Frisch, R. (1934). *Statistical Confluence*. Publ. No. 55. Oslo: University Institute of Economics.

Fuchs, V. R. (1982). 'Time Preference and Health: an Exploratory Study', in Fuchs V. R. (ed.) *Economic Aspects of Health*, University of Chicago Press, Chicago, pp. 93-120.

Gautier, E. and Y. Kitamura (2013). 'Nonparametric Estimation in Random Coefficients Binary Choice Models.' *Econometrica*, 81:2, 581-607.

Golan, A., Judge, G. G., & Miller, D. (1996). *Maximum entropy econometrics: Robust estimation with limited data*. New York: Wiley.

Güth, W., Kliemt, H., & Ockenfels, A. (2003). 'Fairness versus efficiency: an experimental study of (mutual) gift giving'. *Journal of Economic Behavior and Organization, 50(4)*, 465-475.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). 'An Experimental Analysis of Ultimatum Bargaining'. *Journal of Economic Behavior and Organization, 3(4)*, 367-88.

Grossman, M. (2006). 'Education and Nonmarket Outcomes', in Hanushek, E. and Welch, F. (eds) *Handbook of the Economics of Education*, Vol. 1, Elsevier, Amsterdam, pp. 577–628.

Hadi, A. S. (1992), 'A new measure of overall potential influence in linear regression', *Computational Statistics & Data Analysis 14(1)*, 1-27.

Hadi, A. S. and Simonoff, J. S. (1993), 'Procedures for the Identification of Multiple Outliers in

Linear Models', *Journal of the American Statistical Association 88(424)*, 1264-1272.

Harmon, C., Oosterbeek, H., and I. Walker (2003). 'The Returns to Education: Microeconomics'. *Journal of Economic Surveys*,17(2), 115-156.

Heckman, J. J. and E. Vytlacil (1998). 'Instrumental Variable Methods for the Correlated Random Coefficient Model'. *Journal of Human Resources*, Vol. 33, No. 4, pp. 974-987.

Henderson, D. J., Polachek, S. W., and L. Wang (2011). 'Heterogeneity in Schooling Rates of Return'. *Economics of Education Review*, 30(6), 1202-1214.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. and McElreath, R. [2001], 'In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies', *The American Economic Review 91(2)*, 73-78.

Henrich, J., & Ensminger, J. (2014). 'Chapter 2: Theoretical Foundations: The Coevolution of Social Norms, Intrinsic Motivation, Markets, and the Institutions of Complex Societies'. In: Ensminger, J. & Henrich, J. (Eds.), *Experimenting with social norms: Fairness and punishment in cross-cultural perspective*, pp. 89-148. New York: Russell Sage Foundation.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. and Ziker, J. [2010], 'Markets, religion, community size, and the evolution of fairness and punishment', *Science 327(5972)*, 1480-1484.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. and Ziker, J. (2006), 'Costly punishment across human societies', *Science 312(5781)*, 1767-1770.

Henrich, J. P., Boyd, R., Bowles, S., Fehr, E., Camerer, C., Gintis, H. (2004), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford University Press.

Ho, K. and Rosen, A. (2017). 'Partial Identification in Applied Research: Benefits and Challenges'. In B. Honore, A. Pakes, M. Piazzesi, & L. Samuelson (Eds.), *Advances in Economics and Econometrics: Eleventh World Congress*, Econometric Society Monographs, pp. 307-359.

Hoaglin, D. C. and Welsch, R. E. (1978), 'The Hat Matrix in Regression and ANOVA', *The American Statistician 32(1)*, 17-22.

Hoderlein, S. (2009). 'Endogenous Semiparametric Binary Choice Models with Heteroscedastic-

ity'. CeMMAP working paper CWP34/09.

Hoderlein, S. and R.Sherman (2015). 'Identification and Estimation in a Correlated Random Coefficients Binary Response Model'. *Journal of Econometrics*, 188, 135-149.

Horowitz, J. (1992). 'A Smoothed Maximum Score Estimator for the Binary Response Model'. *Econometrica*, 60(3), 505-531.

Huber, P. J. (1964), 'Robust Estimation of a Location Parameter', *The Annals of Mathematical Statistics 35(1)*, 73-101.

Huber, P. J. (1973), 'Robust Regression: Asymptotics, Conjectures and Monte Carlo', *The Annals of Statistics 1(5)*, 799-821.

Iriberri, N., Rey-Biel, P. (2009). 'Elicited beliefs and social information in modified dictator games: what do dictators believe other dictators do?' Universitat Pompeu Fabra Economics Working Papers No. 1137.

Jarque, C. M. and Bera, A. K. (1987), 'A test for normality of observations and regression residuals', *International Statistical Review/Revue Internationale de Statistique*, pp. 163-172.

Jaynes, E. T. (1957). 'Information theory and statistical mechanics'. *Physical Review, 106(4)*, 620.

Jiao, X. and Nielsen, B. (2017), 'Asymptotic Analysis of Iterated 1-Step Huber-Skip M-Estimators with Varying Cut-Offs, *in* J. Antoch, J. Jureckova, M. Maciak and M. Pesta, eds, 'Analytical Methods in Statistics', Springer Proceedings in Mathematics & Statistics, Springer International Publishing, Cham, pp. 23-52.

Johansen, S. and Nielsen, B. (2013), 'Outlier Detection in Regression Using an Iterated One-Step Approximation to the Huber-Skip Estimator', *Econometrics 1(1)*, 53-70.

Johansen, S. and Nielsen, B. (2016a), 'Analysis of the Forward Search using some new results for martingales and empirical processes', *Bernoulli 22(2)*, 1131-1183.

Johansen, S. and Nielsen, B. (2016b), 'Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models', *Scandinavian Journal of Statistics 43(2)*, 321-348.

Jolliffe, I. (2002). *Principal component analysis*, 2nd edition. New York : Springer.

Judge, G. G., & Mittelhammer, R. C. (2012). *An information theoretic approach to econometrics*. Cambridge University Press.

Kalina, J. (2012), 'Testing Heteroscedasticity In Robust Regression', *Research Journal of Economics, Business and ICT*, 4.

Khmaladze, E. V. and Koul, H. L. (2004), 'Martingale transforms goodness-of-fit tests in regression models', *The Annals of Statistics 24(1)*, 380-404.

Koul, H. L. (1996), 'Asymptotics of some estimators and sequential residual empiricals in nonlinear time series', *The Annals of Statistics 24(1)*, 380-404.

Koul, H. L. (2002), *Weighted empirical processes in dynamic nonlinear models*, number 166 in 'Lecture Notes in Statistics', 2nd edn, Springer Science & Business Media, New York.

Koul, H. L. and Levental, S. (1989), 'Weak convergence of the residual empirical process in explosive autoregression', *The Annals of Statistics 17(4)*, 1784-1794.

Koul, H. L. and Ossiander, M. (1994), 'Weak Convergence of Randomly Weighted Dependent Residual Empiricals with Applications to Autoregression', *The Annals of Statistics 22(1)*, 540-562.

Koul, H. L. and Stute, W. (1999), 'Nonparametric model checks for time series', *The Annals of Statistics 27(1)*, 204-236.

Kritikos, A., & Bolle, F. (2001). 'Distributional concerns: equity-or efficiency-oriented?'. *Economics Letters, 73(3)*, 333-338.

Kulkarni, R. (2019), 'Big Data Goes Big'. *Forbes Magazine Online*.
URL: https://www.forbes.com/sites/rkulkarni/2019/02/07/big-data-goes-big/.
(Accessed: 01/11/2021).

LeBaron, B. (1992), 'Forecast Improvements Using a Volatility Index', *Journal of Applied Econometrics 7*, S137-S149.

Levine, D. K. (1998). 'Modeling altruism and spitefulness in experiments'. *Review of Economic Dynamics, 1(3)*, 593-622.

Lewbel, A. (2000). 'Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables'. *Journal of Econometrics*, 97, 145–177.

Li, Q., & Racine, J. (2003). 'Nonparametric estimation of distributions with categorical and continuous data'. *Journal of Multivariate Analysis*, 86(2), 266-292.

Loewenstein, G. F., Thompson, L., & Bazerman, M. H. (1989). 'Social utility and decision

making in interpersonal contexts'. *Journal of Personality and Social Psychology, 57(3)*, 426.

Mallows, C. L. (1975), 'On some topics in robustness', *Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ.*

Manski, C. (2008). 'Partial Identification in Econometrics'. In: Durlauf, N. and Blume, L. *Palgrave Dictionary of Economics*, 2nd Edition. Palgrave Macmillan.

Manski, C. F. (1975). 'Maximum score estimation of the stochastic utility model of choice'. *Journal of Econometrics, 3(3)*, 205-228.

Marciukaitis, M., Zutautaite, I., Martisauskas, L., Joksas, B., Gecevicius, G. and Sfetsos, A. (2017), 'Non-linear regression model for wind turbine power curve', *Renewable Energy*, 113, 732-741.

Mare, R. (1991). 'Five Decades of Educational Assortative Mating'. *American Sociological Review* 56, 15-32.

McFadden, D. (1973). 'Conditional Logit Analysis of Qualitative Choice Behavior'. In: *Frontiers in Econometrics*, Zarembka, P.(Ed.). New York, NY: Academic Press, pp. 105–142.

McFadden, D. (1974). 'The measurement of urban travel demand'. *Journal of Public Economics, 3(4)*, 303-328.

Molchanov, I. S. (2005). *Theory of Random Sets*. Springer Verlag, London.

Molchanov, I. S. (2017), *Theory of Random Sets*, number 87 in 'Probability Theory and Stochastic Modelling', 2nd edn, Springer, London.

Motulsky, H. J. and Brown, R. E. (2006), 'Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate', *BMC Bioinformatics 7(1).*

O'Leary, N. C., and P.J. Sloane (2005). 'The Return to a University Education in Great Britain'. *National Institute Economic Review*, 193(1), 75-89.

Nevo, A. (2000). 'A Practitioner's Guide to Estimation of Random Coefficients Logit Models of Demand', Journal of Economics and Management Strategy, 9(4), 513-548.

Önder, A. and Zaman, A. (2003), A test for normality based on robust regression residuals, *in* R. Dutter, P. Filzmoser, U. Gather and P. J. Rousseeuw, eds, 'Developments in robust statistics', Physica-Verlag HD, Heidelberg, pp. 296-306.

Oreopoulos, P. (2006). 'Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter', American Economic Review, Vol. 96, pp. 152–175.

Pakes, A. (2010). 'Alternative models for moment inequalities'. *Econometrica, 78(6)*, 1783-1822.

Pelligra, V., & Stanca, L. (2013). 'To give or not to give? Equity, efficiency and altruistic behavior in an artefactual field experiment'. *Journal of Socio-Economics, 46*, 1-9.

Pencavel, J. (1998). 'Assortative Mating by Schooling and the Work Behavior of Wives and Husbands'. *American Economic Review* 88:2, 326-29.

Qian, Z. (1998). 'Changes in Assortative Mating: the Impact of Age and Education, 1970-1990'. *Demography* 35:3, 279-92.

Rabin, M. (1993). 'Incorporating fairness into game theory and economics'. *American Economic Review*, 1281-1302.

Reiersol, O. (1941). 'Analysis by Means of Lag Moments and Other Methods of Confluence Analysis'. *Econometrica*, 9(1), 1-24.

Riazoshams, H., Midi, H. and Ghilagaber, G. 2018, *Robust Nonlinear Regression: with Applications using R*, John Wiley & Sons.

Rousseeuw, P. J. (1984), 'Least Median of Squares Regression', *Journal of the American Statistical Association 79(388)*, 871-880.

Selten, R. (1991). 'Properties of a measure of predictive success'. *Mathematical Social Sciences, 21(2)*, 153-67.

Selten, R., & Krischker, W. (1983). 'Comparison of two theories for characteristic function experiments'. In R. Tietz (Ed.), *Aspiration Levels in Bargaining and Economic Decision Making*, (pp. 259-64). Berlin: Springer.

Shannon, C. E. (1948). 'A Mathematical Theory of Communication'. *Bell System Technical Journal 27*, 379—423.

Smith, C. R., & Grandy Jr, W. (Eds.). (1985).*Maximum-entropy and Bayesian methods in inverse problems*. Springer Science & Business Media.

Sobel, J. (2005). 'Interdependent preferences and reciprocity'. *Journal of Economic Literature*, 392-436.

Srikantan, K. S. (1961), 'Testing for the Single Outlier in a Regression Model', *Sankhya: The Indian Journal of Statistics*, Series A (1961-2002) 23(3), 251-260.

Su, J. Q. and Wei, L. J. (1991), 'A Lack-of-Fit Test for the Mean Function in a Generalized Linear Model', *Journal of the American Statistical Association 86(414)*, 420-426.

Teräsvirta, T. and Anderson, H. M. (1992), 'Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models', *Journal of Applied Econometrics* 7, S119-S136.

Town, R. J. (1992), 'Merger Waves and the Structure of Merger and Acquisition Time-Series', *Journal of Applied Econometrics* 7, S83-S100.

University of Essex. Institute for Social and Economic Research. (2010). *British Household Panel Survey: Waves 1-18, 1991-2009.* [data collection]. 7th Edition. UK Data Service. SN: 5151, http://doi.org/10.5255/UKDA-SN-5151-1

Vankeerberghen, P., Smeyers-Verbeke, J., Leardi, R., Karr, C. L. and Massart, D. L. (1995), 'Robust regression and outlier detection for non-linear models using genetic algorithms', *Chemometrics and Intelligent Laboratory Systems 28(1)*, 73-87.

Vopson, M. (2021)'The world's data explained: how much we're producing and where it's all stored'. *The Conversation.* URL: http://theconversation.com/the-worlds-data-explained-how-much-were-producing- and-where-its-all-stored-159964. (Accessed: 01/11/2021).

Vytlacil, E. & N. Yildiz (2007). 'Dummy Endogenous Variables in Weakly Separable Models'. *Econometrica*, 75: 757–779.

Walker, I. & Y. Zhu (2011). 'Differences by Degree: Evidence of the Net Financial Rates of Return to Undergraduate Study for England and Wales'. *Economics of Education Review*, 30, 1177-1186.

Wang, Q. & Phillips, P. C. (2012), 'A specification test for nonlinear nonstationary models', *The Annals of Statistics 40(2)*, 727-758.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica 48(4)* 817-838.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT Press, Cambridge.

Wu, C.-F. (1981), 'Asymptotic theory of nonlinear least squares estimation', *The Annals of Statistics 9(3)*, 501-513.