MONASH University

Identification and Estimation of Causal Treatment Effects: Instrumental Variable Approaches

A thesis submitted for the degree of

Doctor of Philosophy

Lina Zhang

Master of Economics (University of California, Los Angeles)

Department of Econometrics and Business Statistics

Monash University

February, 2021

Copyright notice

Abstract

This thesis contributes to the literature regarding two main challenges in identifying and estimating causal treatment effects, namely, the treatment endogeneity and the treatment spillover. It particularly focuses on the instrumental variable approach which is a common solution to issues caused by these two challenges. This thesis includes five chapters. Chapter 1 provides an overview of the thesis. Chapters 2, 3 and 4 present three independent and self-contained research papers. Chapter 5 concludes the thesis.

Weak instrumental variable is a serious problem hindering the identification and estimation of causal effects when the treatment is endogenous. There is a series of well-developed literature on weak instrument tests for linear regression models. However, there is little theoretical development regarding a test for weak instruments in discrete choice models. **Chapter 2** aims to fill this gap. This chapter proposes a consistent test for weak instruments in the discrete choice models and demonstrates that if the null hypothesis of weak instruments is rejected, the standard Wald inference can be implemented in the usual manner. As a by-product of the proposed testing approach, we construct a generalised "concentration parameter" that allows us to extend the standard "rule-of-thumb" for linear models to discrete choice models. This generalised concentration parameter provides insights regarding instrument strength in a host of discrete choice models. A Monte Carlo analysis is used to compare our testing approach against several commonly applied weak instrument tests. The simulation results simultaneously demonstrate the good performance of our testing approach and the fundamental failure of the conventional linear weak instrument tests in this context. We compare our testing approach to those commonly applied tests in two empirical examples: married women's labour force participation, and the US food aid and civil conflicts.

Instrument strength is often studied in parametric models where the degree of instrument weakness can be captured by a drifting data generating process and where the causal effect is point identified. However, once the parametric assumptions are relaxed and less restrictive models are taken into account, the point identification may be lost unless a relatively strong restriction, such as "identification at infinity", is imposed on

the instrumental variable. **Chapter 3** focuses on the nonparametric models where the treatment effect can only be partially identified, and examines the role played by the instrumental variables in the identification. This chapter rigorously examines the contribution of the instruments and their interplays with other factors, such as the endogeneity degree and the covariates, to the identification gains for the treatment effect. The identification gains are decomposed into a sequence of measurable components, and a standardised quantitative measure is constructed for the instrument identification power (IIP). The decomposition and the IIP evaluation are illustrated using finite-sample simulation studies and an empirical example of childbearing and women's labour supply. The simulation results demonstrate that the IIP offers a potential criterion for detecting irrelevant instruments and sheds new light on instrument selection in high-dimension settings in conjunction with partial identification frameworks.

In the literature on treatment effects, it is widely assumed that the treatment of one unit does not affect others' socioeconomic behaviour (see Chapters 2 and 3). However, treatment spillovers under network interactions have been observed in many empirical studies. It is important to account for treatment spillovers, because ignoring them may lead to misinterpretations of the mechanism through which the treatment operates. Social networks are indispensable for studying treatment spillovers, because they act as the medium for the spillovers. However, there is increasing evidence that the network data is often mismeasured for various reasons, such as misreporting, survey fatigue, or drawbacks of sampling schemes. **Chapter 4** explores the identification and estimation of treatment spillovers with mismeasured networks. Unlike Chapters 2 and 3, Chapter 4 focuses on the spillover effects of a randomised treatment intervention. It proposes a nonparametric point-identification method for the effects of interest, and it exploits an instrumental variable approach to address the issues caused by the network measurement errors. A semiparametric estimation approach is provided and the estimator is shown to be consistent and asymptotically normal. The analysis of this chapter is also applied to study the spillover effects of an information program for a weather insurance on the insurance adoption decisions of rice farmers in rural China.

Declaration Statement

This thesis is an original work derived from my research. It contains no material that has been accepted for the award of any other degree or diploma at any university or equivalent institution. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Print Name: Lina Zhang

Date: November 2020

Acknowledgements

This thesis could have never been a success without the guidance and encouragement of my supervisors, families and friends. First, I would like to sincerely thank Professor Donald Poskitt, Professor Xueyan Zhao and Associate Professor David Frazier for being incredible supervisors, mentors and friends. Their endless enthusiasm for research, wealth of knowledge and productivity have motivated me throughout my PhD at Monash. I appreciate all their contributions of time and ideas that made my PhD candidature being productive and stimulating. It is because of working with them for the past four years that I have decided to devote myself to pursuing an academic career. I believe that they are the best supervisors that I could ever find in my lifelong research journey.

I also deeply acknowledge my coauthors Professor Eric Renault and Dr. Denni Tommasi. I will always consider it a tremendous honour to have worked with Professor Eric Renault. I appreciate his invaluable inspiration and guidance for my first paper which gives me remarkable confidence for the rest of the PhD period. I am grateful for Denni who although is not my supervisor, unreservedly shares his experience and expertise with me, which will definitely help me survive in the future.

I acknowledge my evaluation panel Professor Gael Martin, Professor Farshid Vahid, Associate Professor Vasilis Sarafidis, Assistant Professor Jun Sung Kim, and Associate Professor Anastasios Panagiotelis for their constructive comments. A special thanks goes to Professor Gael Martin and Professor Farshid Vahid for being wonderful PhD coordinators and providing unwavering support to make my candidature a pleasant and fulfilling journey. I take this opportunity express my sincere thanks to all other academic and administrative staff members of the department of econometrics and business statistics at Monash University, who provided me a supportive and friendly environment.

Thank you to all my fellow students, especially to Dr. Kanchana Nadarajah, Dr. Puwasala Gamakumara, Dr. Earo Wang, Dr. Yan Meng, Yuejun Zhao and my old friends Dr. Nan Liu and Dr. Hualei Shang, for your enduring friendship throughout my PhD journey. The past few years have not been an easy ride but your company makes

it enjoyable. I would also like to thank the Monash HPC Team and the MonARCH service provided by Monash University. I also acknowledge Capstone Editing who provided copyediting and proofreading services, following the guidelines laid out in the university-endorsed national "Guidelines for Editing Research Theses." Although it would be impossible to list every name, there is a huge community of others who helped me flounder my way through all the chapters.

This research was supported by the Dean's Excellence Award of Monash Business School, the Donald Cochrane Graduate Research Scholarship established in the memory by the family of the late Emeritus Professor Donald Cochrane through the trustees of the Cochrane-Schofield Charitable Fund ("the Trustees"), the Silver Jubilee Postgraduate Scholarship and the Postgraduate Publications Award provided by Monash University, and an Australian Government Research Training Program (RTP) Scholarship.

Finally, to my parents and my husband. Thank you all for always being there, and for your understanding and supporting of my unforgettable PhD life. I am fortunate to be your daughter and to be so loved.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis studies some methodological issues relating to the instrumental variable (IV) approach to causal effect models. It consists of three self-contained studies. All three studies relate to important econometric problems frequently encountered in empirical research in the identification and estimation of causal treatment effects. The first two studies focus on issues that have arisen from treatment endogeneity and the strength of instrument, and the third study examines the treatment spillover effect in the context of network interactions when the network is measured with errors.

The endogeneity of regressors is a common problem for economists hoping to establish causal effects, and the IV approach is widely used to solve this issue. However, it is well-known that the very feature that renders the instrument useful for estimating causal effects, namely the instrument's exogeneity with respect to the model error terms, can occur at the same time with IVs having little explanatory power to the endogenous regressors. Consequently, the so-called "weak instrument" has been of concern in many empirical studies. The resulting behaviour of the IV estimator in the presence of weak IVs has been extensively studied in the linear regression models (e.g., Staiger and Stock, 1997; Stock and Yogo, 2005). However, there is little theoretical evidence regarding the properties of the IV estimation in discrete choice models. Using Monte Carlo simulations, Dufour and Wilde (2018) demonstrate the poor behaviour of the Wald test and the likelihood-ratio test for the causal inference of a probit model with weak instruments. Magnusson (2007) considers the Wald test and the distance metric test for the probit

model and finds that, with weak instruments, both tests over-reject the null hypothesis even when the concentration parameter is larger than ten. Thus, the undesirable performance of the conventional inference procedure in the presence of weak IVs implies the necessity of a widespread two-stage decision rule: a pretest for weak IVs, and then, if the null of weak IVs is rejected, the standard inference procedures when estimating the causal effects in discrete choice models. However, there is no available weak IV test for discrete choice models.

Chapter 2 presents a test for weak IVs in discrete choice models. The test is related to Antoine and Renault (2009, 2012, 2020). These authors study identification failure within a nonlinear and non-separable generalised method of moments (GMM) setting. Similarly, the proposed weak IV test in this chapter conceptualises the identification failure using a drifting data generating process (DGP) that captures the rank deficiency of the limit Jacobian for the moment conditions. Our testing approach differs from that of Antoine and Renault (2020), however, because it allows for the detection of the actual instrument weakness; that is, whether the strength of the IVs in the first-stage regression is too weak to ensure the estimation consistency. This analysis sheds new light on the inappropriate application of the popular rule-of-thumb developed for linear models to the discrete choice models. Specifically, the test can be understood as a generalisation of the standard first-stage F-test, as such, enables the measurement of the genuine strength of the instruments. This chapter also demonstrates that the standard rule-of-thumb for linear models does not adequately capture the strength of instruments due to the nonlinearity of the discrete choice models.

Chapter 3 explores the IV identification power under less restrictive modelling assumptions. The literature on partially identified models offers a useful framework for the IV identification power analysis. The notion of partial identification relates to the idea that in certain situations such as limited observability, more than one DGP or model can produce the observed data. These models are referred to as "observationally equivalent" and the identified set of the causal parameter is then defined as the collection of all its possible values from different observationally equivalent models.

When studying the average treatment effect (ATE) with binary treatment and binary

outcome, there is a missing data problem because only one of each individual's potential outcomes is observed (depending on the treatment status). There is a notion of "identification by functional form" (Li, Poskitt, and Zhao, 2019), where such non-linear models can be point identified without any IVs, relying on restrictive parametric assumptions, such as a bivariate probit and large support of covariates. However, modelling assumptions such as the bivariate probit are overly restrictive and hard to verify in practice; thus, the resulting point identification has been described as "fragile" (Marra and Radice, 2011). When less restrictive assumptions are allowed, the IVs have been shown to play a crucial role for meaningful identification in partially identified models (see e.g., Chesher, 2005, 2010; Shaikh and Vytlacil, 2011; Li et al., 2019).

Chapter 3 rigorously examines the role of IVs and their interplays with other factors in the identification gains for the ATE in binary outcome models with an endogenous binary treatment. The concepts of *IV strength* and *IV identification power* in this context are distinguished. We find that the conventional IV strength, as measured by the explanatory power of IVs to the treatment variable, is crucial in the identification gains when conducting partial identification analysis. Importantly, we find that the identification gains are also significantly affected by the sign and degree of endogeneity. Therefore, the IV strength itself no longer provides a sufficient measure of the IV identification power in the nonlinear models considered in this chapter. As a result, the various pseudo R^2 goodness-of-fit measures (Veall and Zimmermann, 1992, 1996), which are designed for binary dependent variable models, are not appropriate for measuring the IV identification power. It is because that they fail to capture the critical fact that the IV's identification information varies with the endogeneity degree. Based on these findings, we propose a novel decomposition of the identification gains in the ATE bounding analysis, by disentangling the different sources of the overall identification gains. Given the decomposition, a standardised measure is constructed for the IV identification power, which is a useful index for indicating the IV relevance and selecting irrelevant IVs.

While Chapters 2 and 3 highlight that IVs are crucial when studying the treatment effect with endogenous regressors, **Chapter 4** utilises the IV approach from a different perspective. In the literature on treatment effects, the stable unit treatment value as-

sumption (SUTVA) (Rubin, 1990) is widely adopted for the causal inference. It states that the treatment of one unit does not affect others' outcomes. However, the spillover effects of the treatment via network interactions have been documented in many applications. Because of the increasing availability of network data, the economic research of treatment spillovers has increased dramatically in the past decades (Angelucci and Di Maro, 2016). Existing methods studying spillover effects typically assume that the network data is correctly observed (e.g., Leung, 2020b; Vazquez-Bare, 2019; Viviano, 2019). However, such a requirement of the network accuracy is difficult to satisfy in many empirical studies (Advani and Malde, 2018; Kossinets, 2006).

Chapter 4 provides a method for overcoming the potential failure of identifying and estimating the spillover effects in the presence of network measurement errors. It focuses on the spillover effects of a randomised intervention via a superpopulation model studied by Leung (2020b). However, unlike Leung (2020b), this chapter assumes that the observed network data is mismeasured. We first analytically characterise the bias in treatment and spillover effects caused by inaccurate network information. It has been found that when ignoring the network mismeasurement, not only the spillover effects, but also the treatment effects that are triggered by the correctly-observed and randomised treatment interventions, can be incorrectly identified.

Most importantly, this chapter proposes a novel strategy to nonparametrically point-identify the treatment and spillover effects, when two network proxies are available. This will be the case, for instance, in longitudinal data where network is elicited on multiple occasions over time; when networks under various interaction contexts are collected; or when both self-reported and administrative networks are available. In these situations, the second network proxy is used as an IV for the true latent network. There are several attractive features of the method proposed in this chapter. First, the nonparametric model allows flexible forms of heterogeneity in the treatment and spillover effects. It is important, especially for program evaluation and social planning, to inform how treatment response varies across populations (Manski, 2001). In addition, unlike studies of spillover effects relying on the two-stage experiments, the adjacency matrix used in this chapter need not be block-diagonal (i.e., we do not require the "partial interference"

in Sobel, 2006). Further, the proposed method can deal with network mismeasurement arising from missing and/or misreported network links and does not require modelling of the network formation or its misclassification probabilities.

Chapter 5 summarises the main findings, in light of which we discuss some potential extensions for future research.

Chapter 2

Weak Identification in Discrete Choice Models

2.1. Introduction

A prevalent econometric issue is the assessment of the causal impact of some economic variable on a qualitative feature of the economy. For example, there is a growing body of research that studies the causal impact of economic conditions on civil conflict in developing countries. In this context, economic conditions may be summarised by a given state variable like "economic growth" (see e.g. Miguel et al., 2004) or one may set the focus on a given policy tool, such as US Food Aid (see Nunn and Qian, 2014). In such settings, the most common modelling strategy is to characterise the distribution of a qualitative variable, say y_1, via some piecewise constant function of a latent quantitative variable, say y_1^*. This allows the researcher to view y_1^* as evolving according to a regression equation. If y_2 stands for the economic variable whose causal impact is at stake, we will consider a regression equation:

$$y_{1i}^* = \alpha y_{2i} + x_i'\beta + u_i,\ i = 1, \ldots, n, \tag{2.1}$$

where x_i denotes a vector of k_x exogenous variables and, for sake of expositional simplicity, we observe $i = 1, 2, ..., n$ independent and identically distributed (i.i.d.) cross-

sectional realisations of the random variables. In the introduction, we use the terminology "exogenous" to refer to the explanatory variables x_i and to the instrumental variables z_i. In Section 2.2.1 we define, following Newey et al. (1999), a precise concept of control variables that is more relevant than (and not equivalent to) the common concept of exogeneity.

The causal analysis of interest is conducted through statistical inference on the true unknown value of the parameter α, which must be carefully defined in order to account for the (possible) presence of simultaneity. The critical feature to recall about such a setting is that the structural model (2.1) can not be seen as a model for the conditional expectation of y_{1i}^* given y_{2i} and x_i, because more often than not the economic variable y_{2i} is not exogenous and thus such a conditional expectation does not have a causal interpretation.

To illustrate this point, considered the concrete example given by Nunn and Qian (2014) on the impact of US food aid in civil conflicts: let y_{2i} denote the amount of US food aid to country i, and assume we are interested in analysing if y_{2i} causes new civil conflicts and/or helps offset existing conflicts. In this setting, one must be concerned about the existence of reverse causality ("Do countries receive US aid precisely because they are doing well?") or common cause ("May US strategic objectives be a common cause for conflict and food aid receipts?") between these two measurements. For this reason, identification of the structural parameters in (2.1) will require a set of instrumental variables z_i that are assumed to be exogenous.

However, even with instrumental variables in hand, estimation of the key quantity of interest, α, is hindered by the fact that the scale of the latent variable y_{1i}^* is often unobservable, and at best we can hope to observe the sign of y_{1i}^*. As such, consider that we only observe a binary y_{1i} defined as

$$y_{1i} = 1[y_{1i}^* > 0].$$

At the cost of more involved notations, the methodology developed in this chapter can easily be extended to a wide variety of multinomial models, such as ordered probit

models. To some extent, the binary case considered here is the most extreme case of information loss with respect to the observation of the latent variable. Critically, while the case of binary y_{1i} is commonly treated in many economic applications with endogeneity, the potential impact on the relevance of the underlying instruments, i.e., their "strength", and their ability to identify the quantity of interest, have not been sufficiently well-studied in this setting.

The goal of this chapter is to understand, characterise, and quantify the concept of instrument strength as it pertains to discrete choice models. We make three primary contributions. First, we give a novel characterisation of instrument strength in discrete choice models which demonstrates that instrument strength can be significantly impacted by factors other than the linear correlation between the instruments and the endogenous variables. Our second contribution is to use this characterisation of instrument strength to propose a consistent test for the null hypothesis that "instruments are so weak that point estimators are inconsistent", while under the alternative consistent estimation is warranted. Our final contribution is to demonstrate that, once we have rejected the null of inconsistent estimation, Wald-based inference can be carried out in the standard manner.

We now discuss these contributions in more detail, and place them into the broader literature on weak instruments.

Testing for Instrument Strength: Existing Literature

Since the analysis of Staiger and Stock (1997), practitioners have used the well-regarded "rule-of-thumb" to measure instrument strength in the case of continuous y_{1i}. The magnitude of the F-statistic from the reduced form regression equation is arguably the most common measure for determining instrument strength in the linear regression model. Subsequent to the development of the rule-of-thumb, several influential refinements of this measure, and indeed the very concept of weak instruments in the linear model, have been put forward. Stock and Yogo (2005) provide a quantitative definition of weak instruments in the linear model, and use this definition to propose a formal test for instrument weakness. While the approach of Stock and Yogo (2005) relies on conditionally

homoscedastic and serially uncorrelated regression errors, an extension of the Stock and Yogo (2005) testing strategy to heteroscedastic and serially correlated errors is devised in Montiel Olea and Pflueger (2013).

However, when one moves to general nonlinear economic models, the impact of instrument weakness on the resulting estimates is more difficult to ascertain. As presented in Antoine and Renault (2009, 2012), and following the work of Hahn and Kuersteiner (2002) and Caner (2009), there can exist a range of identification strengths in nonlinear models, between the extreme cases of weak identification (when estimators are not consistent) and strong identification (when estimators are consistent and root-n asymptotically normal). Indeed, these authors have shown that the generalised method of moments (GMM) estimators can be consistent at a rate slower than the canonical rate of $n^{1/2}$, but only in the case of a convergence rate strictly larger than $n^{1/4}$ is standard inference based on the normal distribution approximation warranted. The key issue is that, when convergence is too slow and the model is nonlinear, second-order terms in Taylor expansions, which govern the behaviour of the estimator, may not be negligible in front of first-order terms, so that standard asymptotic inference may no longer be valid. Such slow rates of convergence have also been documented in the case of many weak instruments (see Newey and Windmeijer, 2009 and references therein) while a general study of nearly strong instruments is available in Andrews and Cheng (2012).

Using this characterisation of varying identification strength, Antoine and Renault (2020) have devised a testing strategy that is capable of detecting (certain levels of) instrument strength in nonlinear models estimated by GMM. The proposed test, dubbed the distorted J-test (DJ test), is based on computing the GMM J-test statistic at a slightly perturbed value of the continuously updated GMM (CUGMM) estimator. The logic behind the test is that, if the instruments are truly weak, a small perturbation of the J-statistic will not significantly alter its value, while if the instruments are not weak this perturbation will result in a significant increase in the value of the J-statistic. Similar to other inference strategies robust to weak identification, the approach explicitly relies on the nature of the CUGMM objective function, which, as originally pointed out by Stock and Wright (2000), automatically controls the behaviour of the GMM objective

function under weak identification.

Interestingly, Antoine and Renault (2020) have demonstrated that their DJ test is akin to the standard rule-of-thumb when the model is linear and homoscedastic. In contrast, they stress (see also Windmeijer, 2019 for related work in the context of clustering) that this DJ test differs from standard "robustified" versions of the rule-of-thumb in case of a heteroscedastic linear model. We note, in particular, that when using linear probability models, one is faced (besides the well-known criticisms of this approach) with a severely heteroscedastic linear model.

Herein, we adapt the general testing strategy of Antoine and Renault (2020) to the case of discrete choice models and construct a consistent test for the null hypothesis that the instruments are too weak to allow consistent point estimation. Following the nomenclature of Antoine and Renault (2020), we refer to this test as a distorted J-test (DJ test). Similar to Antoine and Renault (2020), we demonstrate that our DJ test can be interpreted as a natural "generalised rule-of-thumb" in the context of discrete choice models, in the sense that this test appropriately modifies the standard approach to account for both heteroscedasticity and non-linearity.

We compare the performance of this test with the aforementioned existing approaches both through Monte Carlo experiments and an empirical analysis. Monte Carlo results show that our DJ test, albeit conservative, has respectable power. However, the crucial feature of this approach is its ability to discern that the underlying estimator may not be reliable, while the standard rule-of-thumb, because it overlooks information lost due to the nonlinearity of the model, will severely over-reject the null of weak identification.

When applied to real data, our DJ test is able to unambiguously determine when the null of weak identification should be rejected (as in the textbook example of the causal effect of education of married women on their labour force participation, with strong instruments like parents education), while it rightly questions the use of standard inference approaches when identification appears weak. In particular, by contrast with the naive rule-of-thumb, the DJ test casts some doubt on the consistency of the estimator of the parameter α in (2.1) applied to US food aid and offset of civil conflicts, which is key for the reliability

of the important conclusion of Nunn and Qian (2014) that "the primary effect of food aid is to prolong the duration of smaller-scale conflicts".

In addition to the development of our DJ test, this chapter also reinforces the asymptotic theory developed in Antoine and Renault (2009, 2012) regarding inference with nearly-strong instruments. By characterising the strength of instruments in terms of a drifting data generating process, a la Staiger and Stock (1997) and Stock and Wright (2000), we demonstrate that once the null hypothesis of estimator inconsistency has been rejected, Wald-based inference can be performed as usual. This result is in stark contrast to the existing results for general nonlinear models under weak identification, where it has been shown that standard inference is only warranted once the rate of convergence is strictly larger than $n^{1/4}$. In this setting, our ability to perform standard inference stems from the fact that discrete choice models are built from latent linear models, albeit non-linear, which we demonstrate are close enough to linear models to validate standard inference once the underlying estimator is consistent. While the convergence rate of the resulting estimator may be very slow, the studentisation performed in computing Wald test statistics make their behaviour consistent with the standard critical values. In short, if our DJ test rejects the null of estimator inconsistency, which will be accomplished asymptotically with probability one under the alternative, the practitioner can safely apply standard inference procedures.

In this respect, our recommendation remains true to the widespread practice of a two-stage decision rule: a pretest for weak IV followed by standard inference when the null of weak identification is rejected. Of course, an alternative would be to use more computationally demanding inference strategies that are robust to weak identification. The robust approach proposed by Kleibergen (2005) has been extended by Magnusson (2010) to the context of limited dependent variable models. More generally, while the existence of weak IV is a common phenomena, there is little theoretical evidence regarding the properties of GMM estimators in endogenous discrete choice models. Using Monte Carlo simulations, Dufour and Wilde (2018) demonstrate the poor behaviour of Wald and Likelihood Ratio tests in the presence of weak instrument. Finlay and Magnusson (2009) considers the Wald test for the probit model and find that, with weak instruments, the

test significantly over-rejects the null hypothesis (the truth).

We note that the development of a consistent test for weak instrument in discrete choice models is particularly important since the similarity between the linear model and common discrete choice models, such as the probit model, have led researcher to apply tests that are appropriate for linear models to this nonlinear context. In particular, it is relatively common to apply the rule-of-thumb developed by Staiger and Stock (1997) in the linear context to detect the present of weak instruments in discrete choice models: see, e.g., Miguel et al. (2004), Arendt (2005), McKenzie and Rapoport (2011), Cawley and Meyerhoefer (2012), Block et al. (2013) and Goto and Iizuka (2016). However, the above studies do not question the validity of this rule-or-thumb in discrete choice models. Some other researchers may prefer to abandon the discrete choice framework in favor of the linear probability models; see, e.g., Lochner and Moretti (2004), Powell et al. (2005), Kinda (2010), Ruseski et al. (2014). Besides the fact that they are heavily heteroscedastic, linear probability models are by definition misspecified. Since our DJ test is based on a distortion of the standard J-test statistic for misspecification, it should not be used in the context of misspecified moment models.

The remainder of the chapter is organised as follows.

Section 2.2 introduces our model setup and assumptions. The key maintained assumption is the existence of a control function, in which the conditional probability distribution of the structural error term, given all the variables in the reduced form regression, coincides with the conditional distribution of the structural error term conditional on the reduced form error term. The control function approach for probit with endogeneity has been pioneered by Rivers and Vuong (1988) and led them to put forward a two-stage conditional maximum likelihood (2SCML) approach. In this section, we note that a GMM framework allows us to obtain asymptotically equivalent estimators for the structural parameters without necessarily resorting to a two-stage approach. Moreover, we show that our GMM approach is also versatile enough to encompass the Quasi-LIML approach of Wooldridge (2014).

In Section 2.3, we present our DJ test and prove its asymptotic properties: size control

(for the null of weak identification) and consistency (under the alternative). We further demonstrate that as long as the estimators are consistent, i.e., under the alternative to the null hypothesis of weak identification, standard Wald-style inference can be applied. This stands in contrast to the general case of identification strength for nonlinear models considered in Antoine and Renault (2009, 2012) and Andrews and Cheng (2014), where it is shown that in nonlinear models standard inference approaches are warranted only when the rate of convergence is faster than the $n^{1/4}$ rate. Lastly, we demonstrate that, in the context of a discrete choice model, the DJ test can be interpreted as a generalised rule-of-thumb that accounts for the nonlinear nature of the probit model.

Monte Carlo experiments in Section 2.4 compare the finite-sample properties of our proposed DJ test as well as the performance of other weak IV tests. Section 2.5 applies our weak IV test to two empirical examples: Wooldridge (2010) married women's labour force participation, and Nunn and Qian (2014) US food aid and civil conflicts. Section 2.6 concludes.

2.2. General Framework

Blundell and Powell (2004) propose a control function (hereafter, CF) approach to conduct inference on the structural parameters of endogenous binary choice models. In this and the next section, we examine the impact of weak instruments on such a CF approach to inference. However, we first demonstrate the general point that a CF approach allows us to see both the 2SCML of Rivers and Vuong (1988) and the Quasi-LIML approach of Wooldridge (2014) as particular cases of a class of GMM estimators, which we discuss in Section 2.2.2. While these GMM estimators can always be characterised by a one-step minimisation problem, using similar arguments to those in Section 6 of Newey and McFadden (1994), we can also interpret the estimator of the structural parameters as a two-step estimator, whereby a preliminary plug-in estimator (obtained from a reduced form regression equation) is used within the moments. After establishing the general framework, in Section 2.2.3 we then sketch the weak IV issue in the context of probit models.

2.2.1. Model and Control Function Approach

Newey et al. (1999) suggest that the key for a CF approach is to start from a triangular simultaneous equations model. In the context of endogenous binary choice models, this entails specifying structural and reduced form regression equations, and the mechanism generating the binary responses.

The structural equation characterises the response of an unobservable endogenous variable y_{1i}^*, conditional on a scalar-valued endogenous variable y_{2i} and a k_x-dimensional vector of explanatory variables x_i, as the sum of an unknown structural function $g(y_{2i}, x_i)$ and a structural error term u_i:

$$y_{1i}^* = g(y_{2i}, x_i) + u_i, \quad \mathbb{E}[u_i] = 0. \tag{2.2}$$

While Imbens and Newey (2009) propose an even more general structural model where the error term u_i may not be additively separable at the cost of more restrictive independence assumptions, such an extension is beyond the scope of this chapter. For sake of expositional simplicity, we will maintain the following linear specification for the structural function

$$g(y_{2i}, x_i) = \alpha y_{2i} + x_i'\beta,$$

but we note that the analysis remains applicable to any situation where $g(y_{2i}, x_i)$ is a parametric function of (y_{2i}, x_i); the case of nonparametric $g(\cdot)$ is beyond the scope of this current chapter, and is left for future research. Our primary focus of interest is the case where only the sign of the quantitative structural variable y_{1i}^* is observable, which yields the structural equation defining the observed binary outcome y_{1i}:

$$y_{1i} = 1[y_{1i}^* > 0].$$

The binary choice model allows us to address the issue of weak identification in the case of maximum information loss going from the quantitative latent variable y_{1i}^* to the observed variable y_{1i}. However, we note that the general methodology developed in this chapter would be similarly relevant for any observation scheme that would define y_{1i} as a known function of y_{1i}^* and x_i (see e.g., Tobit model, Gompit model, disequilibrium

model, etc.). A reduced form, or first stage, regression equation relates the endogenous explanatory variable y_{2i} to a k_z-dimensional vector of valid instrumental variables, z_i, and the explanatory variables x_i:

$$y_{2i} = \pi(x_i, z_i) + v_i, \quad \mathbb{E}[v_i \mid x_i, z_i] = 0. \tag{2.3}$$

Remark 2.2.1 While we have chosen to view the reduced form regression equation (2.3) as the specification of a conditional expectation, we could alternatively follow the quasi-LIML estimation approach of Wooldridge (2014). In his approach, the reduced form regression equation is only required to be a linear projection of y_{2i} onto x_i and z_i. We will always assume that x_i includes a constant, so that the reduced form error term v_i has a zero mean. That is, instead of (2.3), we could have assumed

$$y_{2i} = x_i'\pi + z_i'\xi + v_i, \ \mathbb{E}[v_i] = 0, \ \text{with Cov}\left(\begin{bmatrix} x_i \\ z_i \end{bmatrix}, v_i\right) = 0. \tag{2.4}$$

Remark 2.2.2 As noted by Blundell and Powell (2004), the reduced form error term v_i often appears to be conditionally heteroscedastic. Taking this possibility into account will allow us to devise more efficient estimators when the reduced form error term is deduced from a conditional expectation rather then from only a linear projection. We will actually combine the advantages of both approaches (2.3) and (2.4) by assuming that:

$$y_{2i} = x_i'\pi + z_i'\xi + v_i, \ \mathbb{E}[v_i \,|x_i, z_i] = 0 \tag{2.5}$$

However, it must be acknowledged that the linearity assumption for the conditional expectation is restrictive, and prevents us from considering cases where the endogenous explanatory variable y_{2i} is itself qualitative. We also note that, while Blundell and Powell (2004) propose a nonparametric estimator of the possibly nonlinear regression function $\pi(x_i, z_i)$, a given nonlinear parametric form of this regression function would not result either in a significant change in our proposed methodology.

As stressed by Newey et al. (1999), the CF approach does not assume that x_i and z_i are

valid instruments, in that the approach does not require

$$\mathbb{E}[u_i \,|x_i, z_i] = 0, \tag{2.6}$$

but instead only that

$$\mathbb{E}[u_i \,|v_i, x_i, z_i] = \mathbb{E}[u_i \,|v_i] \,. \tag{2.7}$$

Moreover, it is worth realising that neither equation (2.6) or equation (2.7) implies the other. While we will eventually maintain a stronger version of equation (2.7), i.e., u_i conditionally independent of x_i, z_i given v_i, there is no reason to believe that v_i is itself independent of (x_i, z_i), which jointly with the former conditional independence would be tantamount to joint independence of (u_i, v_i) and (x_i, z_i), and would in turn imply (2.6). In particular, such independence would rule out the possibility of conditional heteroscedasticity for the error term v_i in the reduced form regression equation (2.5).

As clearly defined by Wooldridge (2015), "a control function is a variable that, when added to a regression, renders a policy variable appropriately exogenous." Typically, the restriction in (2.7) allows us to rewrite equation (2.2) as

$$y_{1i}^* = g\left(y_{2i}, x_i\right) + \mathbb{E}[u_i \,|v_i] + \varepsilon_i, \tag{2.8}$$

where

$$\varepsilon_i = y_{1i}^* - \mathbb{E}\left[y_{1i}^* \mid v_i, x_i, z_i\right] = u_i - \mathbb{E}[u_i \,|v_i] \,,$$

which ensures, by definition, that the policy variable is appropriately exogenous; i.e.,

$$\mathbb{E}\left[\varepsilon_i | y_{2i}, x_i, v_i\right] = 0.$$

In their seminal work, Rivers and Vuong (1988) note that the only assumption needed to obtain valid inference in the probit model is that the conditional distribution of u_i given v_i is normal with a mean that is linear in v_i and with a fixed variance. While this condition is satisfied if (u_i, v_i) is jointly normal, joint normality is not required in general. Similarly, for general discrete choice models a CF approach can be constructed by assuming that $\mathbb{E}\left[u_i \mid v_i\right]$ is linear in v_i and that $\varepsilon_i = u_i - \mathbb{E}\left[u_i \mid v_i\right]$ is independent

of v_i, along with an assumption that ε_i has a known continuous cumulative distribution function denoted by Φ. We assume that this probability distribution is symmetric, i.e., $\Phi(\varepsilon) = 1 - \Phi(-\varepsilon)$, which, together with (2.7), allows us to write

$$\begin{aligned}\Pr\left[y_{1i} = 1 \mid v_i, x_i, z_i\right] &= \Pr\left\{\varepsilon_i > -g\left(y_{2i}, x_i\right) - \mathbb{E}\left[u_i \mid v_i\right] \mid v_i, x_i, z_i\right\} \\ &= \Phi\left\{g\left(y_{2i}, x_i\right) + \mathbb{E}[u_i \,|v_i]\right\}.\end{aligned}$$

We now collect the maintained assumptions on the general model in (2.2)-(2.3).

Assumption 2.2.1 *The following conditions are satisfied.*

(a) (Observation scheme) The observed data $\{s_i\}_{i=1}^n = \{(y_{1i}, y_{2i}, x_i', z_i')'\}_{i=1}^n$ *is an i.i.d. sample and for some* $\kappa > 0, \mathbb{E}\left[\|s_i\|^{2+\kappa}\right] < \infty$.

(b) (Reduced form regression): $y_{2i} = \pi\left(x_i, z_i\right) + v_i$, *and* $\mathbb{E}[v_i \mid x_i, z_i] = 0$.

(c) (Structural equation): (i) $\mathbb{E}[u_i \,|v_i, x_i, z_i] = \mathbb{E}[u_i \,|v_i]$; *(ii)* Φ *is a known cumulative distribution function, twice continuously differentiable and strictly increasing, such that* $\Phi(\varepsilon) = 1 - \Phi(-\varepsilon)$; *(iii) for some unknown parameter* $\tilde{\rho} \in \mathbb{R}$,

$$\Pr[y_{1i} = 1 \mid v_i, x_i, z_i] = \Phi[g\left(y_{2i}, x_i\right) + \tilde{\rho} v_i].$$

(d) (Linearity): The unknown functions $g(\cdot, \cdot)$ *and* $\pi(\cdot, \cdot)$ *are linear:*

(i) For unknown parameters $\alpha \in \mathbb{R}$ *and* $\beta \in \mathbb{R}^{k_x}$, $g\left(y_{2i}, x_i\right) = \alpha y_{2i} + x_i'\beta$;

(ii) For unknown parameters $\pi \in \mathbb{R}^{k_x}$ *and* $\xi \in \mathbb{R}^{k_z}$, $\pi\left(x_i, z_i\right) = x_i'\pi + z_i'\xi$.

(e) (Parameters) The unknown parameters $\theta = (\theta_1', \theta_2')'$, *where* $\theta_1 := (\tilde{\rho}, \alpha, \beta')'$ *and* $\theta_2 := (\pi', \xi')'$, *are of dimension* $p = 2 + 2k_x + k_z$. *We have* $\theta_1 \in \Theta_1 \subset \mathbb{R}^{k_x+2}$, $\theta_2 \in \Theta_2 \subset \mathbb{R}^{k_x+k_z}$, $\Theta := \Theta_1 \times \Theta_2$ *and* Θ *is compact. For* θ^0 *denoting the unknown true value of* θ, *we have* $\theta^0 \in Int(\Theta)$.

As already mentioned, the linearity in Assumption 2.2.1 (d) is innocuous and what follows

can be extended to settings where $g(y_{2i}, x_i)$ has any parametric single-index structure and to cases where $\pi(x_i, z_i)$ has any parametric form. In the more general nonparametric setting, Newey et al. (1999) demonstrate that identification by CF of the structural model is tantamount to assuming that there is no functional relationship between the random variables y_{2i}, x_i and v_i (see Newey et al. 1999 for a precise definition of this concept). With a linear structural function $g(y_{2i}, x_i)$, identification of the structural parameter α is equivalent to assuming that y_{2i} is not a linear combination of x_i and v_i, meaning that the reduced form regression depends on z_i, i.e., $\xi \neq 0$.

To give a more concise treatment, throughout the remainder we restrict our analysis to the case where Φ is the CDF of the standard normal distribution and refer to the model:

$$\Pr[y_{1i} = 1 \,|v_i, x_i, z_i] = \Phi\left[\alpha y_{2i} + x_i'\beta + \tilde{\rho} v_i\right]$$

as a probit model. Since only the sign of the latent variable y_{1i}^* is observed, the probit model generally requires the normalisation condition $\mathrm{Var}(u_i) = 1$. However, it is without loss of generality to instead consider the normalisation condition

$$\mathrm{Var}[u_i \,|v_i] = \mathrm{Var}(\varepsilon_i) = 1.$$

If ρ denotes the linear correlation coefficient between u_i and v_i, the above normalisation ensures that

$$\mathrm{Var}(u_i) = \tilde{\rho}^2 \mathrm{Var}(v_i) + 1 = \rho^2 \mathrm{Var}(u_i) + 1,$$

where $\sigma_v = \sqrt{\mathrm{Var}(v_i)}$,

$$\mathrm{Var}(u_i) = \frac{1}{1-\rho^2}, \;\; \tilde{\rho} = \frac{\rho}{\sigma_v \sqrt{1-\rho^2}},$$

and where we have that $\tilde{\rho}$ is monotonic in ρ. Of course, conditional on $\alpha \neq 0$, the simultaneity/endogeneity problem is in evidence if and only if $\rho \neq 0$ or equivalently $\tilde{\rho} \neq 0$. It is worth to emphasise that the endogeneity is governed by ρ, when the regressor y_2 has a nonzero impact on the outcome variable.

2.2.2. Estimating Equations

Throughout the remainder, we partition the parameter vector as $\theta = (\theta_1', \theta_2')'$, where

$$\theta_1 = (\tilde{\rho}, \alpha, \beta')', \quad \theta_2 = (\pi', \xi')'.$$

The vector θ_1 (resp., θ_2) represents the vector of structural (resp., reduced-from) parameters. Following Assumption 2.2.1, the true value of the reduced form parameters θ_2 is defined by the conditional moment restrictions

$$\mathbb{E}[r_{2i}(\theta_2)|x_i, z_i] = 0, \text{ where } r_{2i}(\theta_2) = y_{2i} - x_i'\pi - z_i'\xi. \tag{2.9}$$

For fixed θ_2, the true value of the structural parameters θ_1 is defined by the conditional moment restrictions

$$\mathbb{E}[r_{1i}(\theta_1, \theta_2)|y_{2i}, x_i, z_i] = 0, \text{ where } r_{1i}(\theta_1, \theta_2) = y_{1i} - \Phi[\alpha y_{2i} + x_i'\beta + \tilde{\rho} v_i(\theta_2)], \tag{2.10}$$

and where

$$v_i(\theta_2) = r_{2i}(\theta_2) = y_{2i} - x_i'\pi - z_i'\xi.$$

As usual, we will handle conditional moment restrictions by choosing vectors of instrumental functions, denoted respectively as $\tilde{b}(x_i, z_i)$ for (2.9) and $\tilde{a}(y_{2i}, x_i, z_i)$ for (2.10), and where it is assumed that the moments $\mathbb{E}[\|\tilde{a}(y_{2i}, x_i, z_i)\|^{2+\kappa}]$ and $\mathbb{E}[\|\tilde{b}(x_i, z_i)\|^{2+\kappa}]$ are finite for some $\kappa > 0$. For a given choice of instrumental functions $\tilde{a}(\cdot,\cdot,\cdot)$ and $\tilde{b}(\cdot,\cdot)$, we maintain the following identification assumption.

Assumption 2.2.2 *(Identification): The true unknown value $\theta^0 = (\theta_1^{0\prime}, \theta_2^{0\prime})' \in Int(\Theta)$ is the unique solution $\theta \in \Theta$ to the following moment restrictions:*

$$\begin{aligned} \textbf{\textit{Reduced form}}: \quad & \mathbb{E}[\tilde{b}(x_i, z_i) r_{2i}(\theta_2)] = 0 \quad \Longleftrightarrow \quad \theta_2 = \theta_2^0, \\ \textbf{\textit{Structural}}: \quad & \mathbb{E}[\tilde{a}(y_{2i}, x_i, z_i) r_{1i}(\theta_1, \theta_2^0)] = 0 \quad \Longleftrightarrow \quad \theta_1 = \theta_1^0. \end{aligned}$$

We can summarise the unconditional moment conditions in Assumption 2.2.2 as follows: for $H \geq p$, and H-dimensional vectors a_i and b_i of the same dimension, define

$$g_i(\theta) = a_i r_{1i}(\theta_1, \theta_2) + b_i r_{2i}(\theta_2), \text{ where } a_i = \begin{bmatrix} \tilde{a}(y_{2i}, x_i, z_i) \\ \mathbf{0} \end{bmatrix}, \quad b_i = \begin{bmatrix} \mathbf{0} \\ \tilde{b}(x_i, z_i) \end{bmatrix}$$

then Assumption 2.2.2 implies that the moment function $g_i(\theta)$ satisfies

$$\mathbb{E}[g_i(\theta)] = 0 \iff \theta = \theta^0.$$

A GMM estimator of θ^0 can then be constructed using the moment function

$$g_i(\theta) = (g_{1i}(\theta)', g_{2i}(\theta)')', \text{ where } g_{1i}(\theta) = \tilde{a}(y_{2i}, x_i, z_i) r_{1i}(\theta), \quad g_{2i}(\theta) = \tilde{b}(x_i, z_i) r_{2i}(\theta_2). \tag{2.11}$$

In particular, for W_n a sequence of positive-definite $H \times H$ weighting matrix, we can estimate θ^0 using the GMM estimator

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \bar{g}_n(\theta)' W_n \bar{g}_n(\theta), \text{ where } \bar{g}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\theta) \equiv \left(\bar{g}_{1n}(\theta)' \quad \bar{g}_{2n}(\theta)' \right)'.$$

Remark 2.2.3 In general, imposing that some components of the vectors a_i and b_i are zero prevents us from choosing optimal instruments, and ultimately results in $\hat{\theta}_n$ being an inefficient estimator of θ^0. The characterisation of optimal instrumental functions for the joint set (2.9) and (2.10) of conditional moment restrictions is non-standard because they correspond to different conditioning variables. The optimal instrumental functions in this case have been characterised by Kawaguchi et al. (2017) (see also Ai and Chen (2003) for a general study). Their result implies that in case of overidentification and simultaneity ($\tilde{\rho} \neq 0$ and $\alpha \neq 0$), the first set $r_{1i}(\theta)$ of moment conditions is also informative about θ_2, so that a more efficient estimator of θ_2 (and in turn θ_1) is obtained by an appropriate choice of a_i in which all of its components are non-zero.

While the specific choice of instrumental functions a_i and b_i may be sub-optimal, this choice allows us to demonstrate the equivalence between a GMM-based approach and the 2SCML approach of Rivers and Vuong (1988). In particular, for $g_{1i}(\theta)$ and $g_{2i}(\theta)$

defined as in equation (2.11), we have that

$$\begin{aligned}\text{Cov}\left[g_{1i}(\theta^0), g_{2i}(\theta^0)\right] &= \mathbb{E}\left[\tilde{a}\left(y_{2i}, x_i, z_i\right)\tilde{b}'\left(x_i, z_i\right) r_{1i}\left(\theta^0\right) r_{2i}\left(\theta_2^0\right)\right] \\ &= \mathbb{E}\left\{\tilde{a}\left(y_{2i}, x_i, z_i\right)\tilde{b}'\left(x_i, z_i\right) r_{2i}\left(\theta_2^0\right)\mathbb{E}[r_{1i}\left(\theta^0\right)|y_{2i}, x_i, z_i]\right\} = 0.\end{aligned}$$

Thus, an efficient GMM estimator based on the moment functions in (2.11) can be defined as

$$\begin{aligned}\hat{\theta}_n &= \arg\min_{\theta\in\Theta} \bar{g}_n(\theta)' \begin{bmatrix} W_{1n} & 0 \\ 0 & W_{2n} \end{bmatrix} \bar{g}_n(\theta) \\ &= \arg\min_{\theta\in\Theta} \left\{\bar{g}_{1n}(\theta)' W_{1n} \bar{g}_{1n}(\theta) + \bar{g}_{2n}(\theta)' W_{2n} \bar{g}_{2n}(\theta)\right\},\end{aligned}$$

for an appropriate choice of the weighting matrices W_{1n} and W_{2n}. Consequently, the components of the first-order conditions for the structural parameters θ_1 are given by

$$\frac{\partial \bar{g}_{1n}(\hat{\theta}_n)'}{\partial \theta_1} W_{1n} \bar{g}_{1n}(\hat{\theta}_n) = 0. \tag{2.12}$$

Equation (2.12) allows us to see the estimator $\hat{\theta}_{1n}$ as a two-step estimator based on the moment conditions

$$\mathbb{E}[r_{1i}\left(\theta_1, \theta_2^0\right)|y_{2i}, x_i, z_i] = 0, \tag{2.13}$$

where the nuisance parameter θ_2^0 is replaced by a consistent first-step estimator $\hat{\theta}_{2n}$. From (2.12), we can see that the estimator $\hat{\theta}_{1n}$ is the solution in $\theta_1 = (\tilde{\rho}, \alpha, \beta')'$ to the $(2 + k_x)$ orthogonality conditions

$$\sum_{i=1}^{n} \gamma_{i,n} \left\{y_{1i} - \Phi\left[\alpha y_{2i} + x_i'\beta + \tilde{\rho} v_i\left(\hat{\theta}_{2n}\right)\right]\right\} = 0, \text{ for } \gamma_{i,n} = \frac{\partial \bar{g}_{1n}(\hat{\theta}_n)'}{\partial \theta_1} W_{1n} \tilde{a}\left(y_{2i}, x_i, z_i\right). \tag{2.14}$$

The optimal instruments associated with estimation of θ_1^0 in equation (2.13) (i.e., where θ_2^0 is known) are given by any consistent estimator of:

$$\gamma_i^* = \left[\text{Var}\left(r_{1i}\left(\theta_1^0, \theta_2^0\right)\right)|y_{2i}, x_i, z_i)\right]^{-1} \mathbb{E}\left[\left.\frac{\partial r_{1i}\left(\theta_1^0, \theta_2^0\right)}{\partial \theta_1}\right| y_{2i}, x_i, z_i\right]$$

$$\equiv \frac{\phi_i\left(\theta^0\right)}{\Phi_i\left(\theta^0\right)\left[1-\Phi_i\left(\theta^0\right)\right]}\begin{bmatrix} v_i\left(\theta_2^0\right) \\ y_{2i} \\ x_i \end{bmatrix}$$

where

$$\begin{aligned}\Phi_i\left(\theta^0\right) &= \Phi\left[\alpha^0 y_{2i} + x_i'\beta^0 + \tilde{\rho}^0 v_i\left(\theta_2^0\right)\right] \\ \phi_i\left(\theta^0\right) &= \phi\left[\alpha^0 y_{2i} + x_i'\beta^0 + \tilde{\rho}^0 v_i\left(\theta_2^0\right)\right]\end{aligned}$$

and $\phi(x) = d\Phi(x)/dx$ is the probability density function associated to Φ. Therefore, if one were to choose a consistent estimator of γ_i^* as instruments, the estimator $\hat{\theta}_{1n}$ can be seen as the solution in $\theta_1 = (\tilde{\rho}, \alpha, \beta')'$ to the equations:

$$\sum_{i=1}^{n} \frac{\phi_i\left(\theta_1, \hat{\theta}_{2n}\right)}{\Phi_i\left(\theta_1, \hat{\theta}_{2n}\right)\left[1-\Phi_i\left(\theta_1, \hat{\theta}_{2n}\right)\right]}\begin{bmatrix} v_i\left(\hat{\theta}_{2n}\right) \\ y_{2i} \\ x_i \end{bmatrix}\left\{y_{1i} - \Phi\left[\alpha y_{2i} + x_i'\beta + \tilde{\rho} v_i\left(\hat{\theta}_{2n}\right)\right]\right\} = 0. \tag{2.15}$$

Equation (2.15) shows that, for any choice of a consistent first-step estimator $\hat{\theta}_{2n}$, the estimator $\hat{\theta}_{1n}$ is a 2SCML estimator a la Rivers and Vuong (1988).

2.2.3. The Weak IV Issue in the Probit Model

The representation in equation (2.15) demonstrates that the general class of GMM estimators for θ_1 defined in equation (2.14) contains both 2SCML and Quasi-LIML estimators as particular cases. Therefore, we can ascertain the impact of instrument weakness, on these and related methods, by studying instrument weakness in this general class of GMM estimators.

However, before moving to a general study, we give some intuition on the potential impacts of instrument weakness in the probit model. These implications are most easily elucidated in the infeasible case where we replace the optimal instruments in equation (2.15) with their infeasible counterpart γ_i^*, and where we replace the estimator $\hat{\theta}_{2n}$ by the true value θ_2^0.

Under these simplification, and under the one-to-one transformation of θ_1 defined by

$$\eta_1 = \tilde{\rho}, \quad \eta_2 = \alpha + \tilde{\rho}, \quad \eta_3 = \beta - \tilde{\rho}\pi^0,$$

the infeasible estimator $\tilde{\eta}_n$ of η^0 (and thus θ_1^0) can be defined as the solution to

$$\begin{aligned} &\sum_{i=1}^{n} \gamma_i^* \left\{ y_{1i} - \Phi\left[\eta_1\left(-z_i'\xi^0\right) + \eta_2 y_{2i} + x_i'\eta_3\right]\right\} \\ =&\sum_{i=1}^{n} w_i D_i \left\{ y_{1i} - \Phi\left[\eta_1\left(-z_i'\xi^0\right) + \eta_2 y_{2i} + x_i'\eta_3\right]\right\} = 0, \end{aligned}$$

where $\gamma_i^* = w_i D_i$, $w_i = 1/\Phi_i(\theta^0)[1 - \Phi_i(\theta^0)]$ and $D_i = \phi_i(\theta^0)(-z_i'\xi^0, y_{2i}, x_i')'$. The simplification made in the term D_i, i.e., replacing $v_i(\theta_2^0)$ by $-z_i'\xi^0$, follows from the row operation on γ_i^* which does not affect the solution of the linear equations in (2.15) asymptotically. A Taylor expansion allows us to heuristically write

$$\begin{aligned} &y_{1i} - \Phi\left[\eta_1\left(-z_i'\xi^0\right) + \eta_2 y_{2i} + x_i'\eta_3\right] \\ &\quad \approx y_{1i} - \Phi_i\left(\theta^0\right) - \phi_i\left(\theta^0\right)\left[\left(-z_i'\xi^0\right)\left(\eta_1 - \eta_1^0\right) + y_{2i}\left(\eta_2 - \eta_2^0\right) + x_i'\left(\eta_3 - \eta_3^0\right)\right]. \end{aligned}$$

Using this expansion within the infeasible estimating equations, $\tilde{\eta}_n$ can be seen to solve

$$\sum_{i=1}^{n} w_i D_i \left(\tilde{y}_{1i} - D_i'\eta\right) = 0, \text{ where } \tilde{y}_{1i} = y_{1i} - \Phi(\theta^0) + D_i'\eta^0.$$

Consequently, $\tilde{\eta}_n$ is obtained from a weighted least squares regression of $\tilde{y}_{1i}$ on the explanatory variables $D_i = \phi_i(\theta^0)(-z_i'\xi^0, y_{2i}, x_i')'$. While the above estimating equations are not identical to those in equation (2.15), it is clear from comparing the two that they are of a similar form, and therefore whatever implications are drawn about the later will be sustained by the former.

This regression-based viewpoint yields two important, and interrelated, implications for inference in endogenous binary choice models. First, the linear regression that is considered is not the one suggested by a linear probability model, which would be based on explanatory variables $z_i'\xi^0, y_{2i}, x_i$, and not the weighted versions in D_i. Second, since the explanatory variables in the regression are weighted by $\phi_i(\theta^0)$, it is inappropriate to

focus solely on the contribution of $z_i'\xi^0$ in the reduced form regression as a measure of instrument strength.

Remark 2.2.4 Before moving on, we note that the above type of estimation approach has been dubbed "two-stage residual inclusion" (2SRI) estimation by Terza et al. (2008). In particular, using the first stage consistent estimators $\hat{\theta}_{2n} = (\hat{\pi}_n', \hat{\xi}_n')'$, the estimated first stage residual

$$\hat{v}_i = y_{2i} - x_i'\hat{\pi}_n - z_i'\hat{\xi}_n$$

is included in the computation of the generalised residual

$$r_{1i}(\theta_1, \theta_2) = y_{1i} - \Phi\left[\alpha y_{2i} + x_i'\beta + \tilde{\rho}v_i(\theta_2)\right].$$

We know from Hausman (1978) that, in a fully linear model and as far as estimation of structural parameters α and β is concerned, 2SRI is equivalent to 2SLS. The inclusion of the residual $\hat{v}_i$ in the regression equation ensures that naive OLS would coincide with 2SLS. In addition, Terza et al. (2008) dub "Two-stage predictor substitution" (2SPS) the direct generalisation of 2SLS to our nonlinear context, meaning that in the structural equation, the endogenous variable is simply replaced by its first stage adjusted value, leading to the generalised residual:

$$\begin{aligned} \hat{u}_i &= y_{1i} - \Phi\left[\alpha\hat{y}_{2i} + x_i'\beta\right] \\ \hat{y}_{2i} &= x_i'\hat{\pi}_n + z_i'\hat{\xi}_n \end{aligned}$$

Not surprisingly, Terza et al. (2008) show that in a nonlinear model, 2SPS is not equivalent anymore to 2SRI and only the latter provides a consistent estimator of structural parameters. The intuition is quite clear. Due to the non-linearity of the function $\Phi(.)$, plugging in $\hat{y}_{2i}$ to instrument y_{2i} does not fix satisfactorily the endogeneity bias problem.

As alluded to above, it can be misleading to set the focus on the contribution of $z_i'\xi^0$ in the reduced form regression to gauge the instrument strength, as is done when using the standard rule-of-thumb. Doing so is akin to overlook the impact of nonlinearity in the same way that it is wrong to confuse the correct 2SRI and the flawed 2SPS. Indeed, as

the above arguments clarify, the relevant variable for capturing instrument strength is not z_i, as in the standard linear case, but $\phi_i(\theta^0)z_i$. Thus, the assessment of identification strength should rather be based on the variability of $\phi_i(\theta^0)z_i'\xi^0$.

We can easily illustrate the impact of moving from $z_i'\xi^0$ to $\phi(\theta^0)z_i'\xi^0$ in terms of instrument strength in the probit model, so that $\phi(\cdot)$ is the probability density function of the Gaussian distribution. The conclusions given below will remain valid for any other probability distribution with thin tails, such that the variability of the $\phi_i(\theta^0)z_i$ is drastically different from the one of z_i. First we recall that that for a real valued variable ν and any given number c, the absolute value of the function $h(\nu) = \nu\phi(c+\nu)$ is decreasing in $|\nu|$ when the latter value is larger than the absolute value of the roots of the polynomial $[1 - c\nu - \nu^2]$. Moreover, the rate of this decrease is sharp (converging swiftly to zero) due to the thin tails of the Gaussian distribution.

Using this argument, one may realise that the multiplication of $z_i'\xi^0$ by

$$\phi_i(\theta^0) = \phi\left[\alpha^0 y_{2i} + x_i'\beta^0 + \tilde{\rho}^0\left(y_{2i} - x_i'\pi^0 - z_i'\xi^0\right)\right]$$

erases the variability of $z_i'\xi^0$, by pruning all its large values. For $Z \sim \mathbb{N}(0, \sigma_z^2)$, it is useful to illustrate the above point by comparing the variance of $Z\phi(1+Z)$ as a percentage of the variance of Z. For various values of σ_z^2 , we collect these ratios in Table 2.2.1 below.

Table 2.2.1: Comparison of Variance of Z to Variance of $W = Z\phi(1+Z)$

σ_z^2	1	2	5	10	50	100
Rel. %	100%	79.03%	30.18%	28.13%	7.42%	3.83%

Note: For $\sigma_w^2 = \text{Var}(W)$, we first calculate $l_z = \sigma_w^2/\sigma_z^2$, i.e., the variance of W as a percentage of the variance of Z, for various values of σ_z^2. The value of Rel % in the table is the value of l_z expressed as a percentage of $\sigma_w^2/1$, i.e., we report the results relative to the case where $\sigma_z^2 = 1$.

The results in Table 2.2.1 constitute compelling evidence on the likely flaws of the standard rule-of-thumb in the probit context. It is also worth stressing that, while Table 2.2.1 only displays results with the normalised function $\phi(1+Z)$, the pruning impact of large values of $z_i'\xi^0$ within the function $\phi(\cdot)$ may actually be magnified in finite sample by a

large value of the parameter $\tilde{\rho}^0$. We may then expect that the pruning effect documented in Table 2.2.1 will be even more detrimental for small values of σ_v and/or a large degree of endogeneity ρ, with both cases corresponding to a large value of $\tilde{\rho}$. These possible perverse effects for the naive rule-of-thumb will be confirmed by the Monte Carlo experiments in Section 2.4. These experiments will show that the standard rule-of-thumb will be more prone to over-reject the null of weak instruments in the case of strong simultaneity (ρ close to one) and/or a large signal to noise ratio σ_z/σ_v in the reduced form regression.

2.3. A Test for Instruments Weakness

2.3.1. Intuition

Several authors, such as Kleibergen (2005), Caner (2009), Chaudhuri and Renault (2020), Stock and Wright (2000), and Antoine and Renault (2020), have discussed the advantages of a continuously updated GMM (CUGMM) approach to efficient GMM estimation in case of possible weak identification. Following the latter two authors, in our context the advantage of the CUGMM approach is that, irrespective of identification weakness, the asymptotic behaviour of the CUGMM criterion is always controlled. This feature of the CUGMM criterion will ultimately allow us to obtain a test for instrument weakness that is size controlled and consistent.

To see that this key feature remains true in our setting, recall the specific moment conditions underlying this analysis given by equation (2.11); namely, for $\theta_1 = (\tilde{\rho}, \alpha, \beta')'$ and $\theta_2 = (\pi', \xi')'$, and $g_{1i}(\theta) = \tilde{a}(y_{2i}, x_i, z_i) r_{1i}(\theta_1, \theta_2)$, $g_{2i}(\theta) = \tilde{b}(x_i, z_i) r_{2i}(\theta_2)$,

$$g_i(\theta) = r_{1i}(\theta) a\left(y_{2i}, x_i, z_i\right) + r_{2i}(\theta_2) b\left(x_i, z_i\right) = \left(g_{1i}(\theta)', \quad g_{2i}(\theta)'\right)'.$$

Defining the weighting matrix

$$S_n(\theta) = \begin{bmatrix} S_{11,n}(\theta) & 0 \\ 0 & S_{22,n}(\theta) \end{bmatrix}, \quad S_{jj,n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[g_{j,i}(\theta) - \bar{g}_{j,n}(\theta)\right] \left[g_{j,i}(\theta) - \bar{g}_{j,n}(\theta)\right]',$$

for $j = 1, 2$, we consider a version of the CUGMM estimator (hereafter, CUE) that takes

into account the block diagonal structure of the population variance matrix. Then our CUE of θ^0 based on $\bar{g}_n(\theta) = (\bar{g}_{1n}(\theta)', \bar{g}_{2n}(\theta)')'$ is defined as

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} J_n(\theta,\theta), \text{ for } J_n(\theta,\tilde{\theta}) = n\bar{g}_n(\theta)' S_n^{-1}(\tilde{\theta})\bar{g}_n(\theta),$$

where the notation $J_n(\theta,\tilde{\theta})$ differentiates the occurrences of θ in the moments, $\bar{g}_n(\theta)$, from those in the weighting matrix, $S_n^{-1}(\tilde{\theta})$.

The critical feature of the criterion $J_n(\theta,\tilde{\theta})$ is that, by definition,

$$J_n(\theta^0,\theta^0) \geq J_n(\hat{\theta}_n,\hat{\theta}_n), \tag{2.16}$$

while, since $\text{Cov}\left[g_{1i}(\theta^0), g_{2i}(\theta^0)\right] = 0$, it follows that $J_n\left(\theta^0,\theta^0\right)$ converges in distribution to a chi-square random variable with H degrees of freedom, denoted throughout as $\chi^2(H)$.

The general validity of this upper bound, regardless of the instrument strength, and, hence consistency of $\hat{\theta}_n$, is the reason why we resort to CUGMM. This upper bound will allow us to control the size of our test for weak identification.[1]

The key intuition for our test of weak identification is the following observation. Under weak identification, there are certain directions of the parameter space where the CUGMM objective function $J_n(\cdot,\hat{\theta}_n)$ is flat in the neighbourhood of $\hat{\theta}_n$. In these directions, if we distort $\hat{\theta}_n$ by some "small" value, say $\Delta_n \in \mathbb{R}^p$, and evaluate $J_n(\cdot,\hat{\theta}_n)$ at $\hat{\theta}_n^\delta = \hat{\theta}_n + \Delta_n$, then the value of $J_n(\hat{\theta}_n^\delta,\hat{\theta}_n)$ should not differ "significantly" from that of $J_n(\hat{\theta}_n,\hat{\theta}_n)$. Herein, the concept of "significance" means that $J_n(\hat{\theta}_n^\delta,\hat{\theta}_n)$ exceeds some pre-specified quantile of the $\chi^2(H)$ distribution.

Critically, however, since the objective function scales the squared norm of the sample

[1]The upper bound (2.16) is generally invalid if a first-step estimator of θ^0 is used to estimate the optimal instrumental functions. The only way to incorporate optimal instrumental functions for $a(y_{2i},x_i,z_i)$ and $b(x_i,z_i)$ would be to use them with a free value of θ like in the weighting matrix of CUGMM. The discussion of this alternative approach is left for future research. Also, we note that in the just identified case, the minimum $J_n(\hat{\theta}_n,\hat{\theta}_n)$ of $J_n(\theta)$ is asymptotically, with probability one, equal to zero and $S_n^{-1}(\theta)$ is immaterial. In particular, when using the first-order conditions of some M-estimator, including two-stage conditional maximum likelihood or quasi-LIML, the weighting matrix is irrelevant.

mean $\bar{g}_n(\theta)$, by the factor n, when identification is not weak the distortion introduces a wedge between $\bar{g}_n(\hat{\theta}_n^\delta)$ and $\bar{g}_n(\hat{\theta}_n)$. Therefore, if identification is not weak, so long as the distortion goes to zero sufficiently slowly with n, the criterion $J_n(\hat{\theta}_n^\delta, \hat{\theta}_n)$ diverges asymptotically and thus exceeds (with probability going to one) the chosen quantile of the $\chi^2(H)$ distribution. Throughout the remainder, we refer to this testing procedure as a distorted J-test. It is worth noting that this test is dubbed the "distorted J-test" because it uses the J statistic proposed by Hansen (1982) in the overidentified case to test for the validity of a set of moments. The terminology is a bit misleading since our test may work even in the just identified case ($H = p$). There are actually two possible points of view: either one chooses to perform the distorted J-test test in a just identified setting ($H = p$), or in the overidentified setting ($H > p$).

2.3.2. The null hypothesis of weak identification

As already discussed in Section 2.2.3, weak instruments impact estimation of the structural parameters through the structural moment function

$$g_{1i}(\theta) = \tilde{a}(y_{2i}, x_i, z_i)\, r_{1i}(\theta_1, \theta_2), \text{ where } r_{1i}(\theta_1, \theta_2) = y_{1i} - \Phi\left[(\tilde{\rho} + \alpha) y_{2i} + x_i'(\beta - \tilde{\rho}\pi) - \tilde{\rho} z_i'\xi\right].$$

The impact of weak instruments can be most easily disentangled under the parameterisation

$$\eta = (\eta_1, \eta_2, \eta_3')' = (\tilde{\rho}, \tilde{\rho} + \alpha, \beta' - \tilde{\rho}\pi')', \tag{2.17}$$

which allows us to restate the moment function as

$$g_{1i}(\eta, \theta_2) = \tilde{a}(y_{2i}, x_i, z_i)\tilde{r}_{1i}(\eta, \theta_2), \text{ where } \tilde{r}_{1i}(\eta, \theta_2) = y_{1i} - \Phi\left[-\eta_1 z_i'\xi + \eta_2 y_{2i} + x_i'\eta_3\right].$$

Following Staiger and Stock (1997) and Stock and Wright (2000), we use a drifting data generating process (DGP) to capture instrument weakness, so that population expectations are viewed as being n-dependent. However, to paraphrase Lewbel (2019), we do not actually believe that the DGP is changing as n changes, but use the drifting DGP concept in order to obtain more reliable asymptotic approximations in the context of weak identification. To this end, we consider that the population expectation of

$\bar{g}_{1n}(\eta, \theta_2)$ is defined as

$$m_{1n}(\eta, \theta_2) = \frac{1}{n}\mathbb{E}_n\left[\sum_{i=1}^{n} \tilde{a}(y_{2i}, x_i, z_i)\, \tilde{r}_{1i}(\eta, \theta_2)\right].$$

Under this drifting DGP, we are obliged to see θ_2^0, and hence η^0, as n-dependent, so that the maintained identification assumption should technically be recast as

$$m_{1n}(\eta, \theta_2) = 0 \iff (\eta, \theta_2) = (\eta_n^0, \theta_{2n}^0).$$

However, to keep the notational burned to a minimum, we only make the true-values dependence on n explicit when absolutely necessary.

Following the approach of Stock and Wright (2000) (see their Section 2.3), the following decomposition of $m_{1n}(\eta, \theta_2)$ will ultimately allow us to isolate the impact of instrument weakness

$$\begin{aligned} m_{1n}\left(\eta, \theta_2^0\right) &= m_{1n}\left(\eta^0, \theta_2^0\right) + \left[m_{1n}\left(\eta, \theta_2^0\right) - m_{1n}\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right)\right] \\ &+ \left[m_{1n}\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right) - m_{1n}\left(\eta^0, \theta_2^0\right)\right]. \end{aligned}$$

In particular, since $m_{1n}(\eta^0, \theta_2^0) = 0$, we have

$$m_{1n}\left(\eta, \theta_2^0\right) = \left[m_{1n}\left(\eta, \theta_2^0\right) - m_{1n}\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right)\right] + m_{1n}\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right). \tag{2.18}$$

As explained in the Section 2.2.3, instrument weakness is encapsulated by the explanatory variable $\phi_i(\theta^0) z_i'\xi^0$. The impact of this explanatory variable on instrument strength can be directly obtained by linearising $m_{1n}(\eta, \theta_2^0)$ around η_1^0 to obtain

$$\begin{aligned} m_{1n}\left(\eta, \theta_2^0\right) - m_{1n}\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right) &= \left(\eta_1 - \eta_1^0\right) \frac{\partial m_{1n}}{\partial \eta_1}\left(\eta_{1n}^*, \eta_2, \eta_3, \theta_2^0\right) \\ &= \left(\eta_1 - \eta_1^0\right) \frac{1}{n}\mathbb{E}_n\left[\sum_{i=1}^{n} \tilde{a}(y_{2i}, x_i, z_i)\, \phi_i\left(\eta_{1n}^*, \eta_2, \eta_3, \theta_2^0\right) z_i'\xi^0\right], \end{aligned} \tag{2.19}$$

where η_{1n}^* denotes a component-by-component intermediate value, which can vary according to the components of the function $\tilde{a}(.)$.

Equation (2.19) allows us to write the decomposition in equation (2.18) in the following semi-separable form, which clearly partitions the directions of weakness in the parameter space: for some real, positive, and deterministic sequence $\varsigma_n \to \infty$ as $n \to \infty$, with $\varsigma_n = O(\sqrt{n})$, possibly $o(\sqrt{n})$,

$$m_{1n}\left(\eta, \theta_2^0\right) = q_{11,n}\left(\eta\right)/\varsigma_n + q_{12,n}\left(\eta_2, \eta_3\right), \tag{2.20}$$

where

$$\begin{aligned} q_{11,n}\left(\eta\right) &= \varsigma_n \left[m_{1n}\left(\eta, \theta_2^0\right) - m_{1n}\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right)\right], \\ q_{12,n}\left(\eta_2, \eta_3\right) &= m_{1n}\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_n\left[\tilde{a}\left(y_{2i}, x_i, z_i\right)\tilde{r}_{1i}\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right)\right]. \end{aligned}$$

Given this decomposition of $m_{1n}\left(\eta, \theta_2^0\right)$, the identification strength of η_1 is entirely determined by equation (2.19) and therefore $q_{11,n}(\eta)/\varsigma_n$. In particular, the rate ς_n can be thought of as encapsulating the speed with which the curvature of the moments approaches zero in the η_1 direction, and thus ς_n determines the degree of identification weakness. If ς_n diverges like $\sqrt{n}$, the speed at which this curvature vanishes is matched by the rate at which information accumulates in the sample, i.e., $\sqrt{n}$, and there is no hope that η_1^0 can be identified from sample information; i.e., η_1^0 is weakly identified. In contrast, the identification of η_2, η_3 is determined by $q_{12,n}(\eta_2, \eta_3)$ and is not afflicted by identification weakness. That is, in this rotated parameter space of η, identification weakness only occurs in the η_1 direction and does not permeate the remaining directions in the parameter space. The representation in equation (2.20) is conformable, but not equivalent, to the decomposition employed by Stock and Wright (2000) to study the behaviour of GMM under weak identification (see Remark 2.3.2 for details). We maintain the following conditions on $m_{1n}(\eta, \theta_2^0)$, which has the same form as Assumption C in Stock and Wright (2000).

Assumption 2.3.1 *For $\varsigma_n = O(\sqrt{n})$, possibly $o(\sqrt{n})$, $m_{1n}\left(\eta, \theta_2^0\right) = q_{11,n}\left(\eta\right)/\varsigma_n + q_{12,n}\left(\eta_2, \eta_3\right)$:*

(a) $q_{11,n}\left(\eta\right) \to q_{11}\left(\eta\right)$ as $n \to \infty$ uniformly in η, where $q_{11}\left(\eta^0\right) = 0$, and $q_{11}(\cdot)$ is uniformly continuous (and hence bounded) in η.

(b) $q_{12,n}(\eta_2, \eta_3) \to q_{12}(\eta_2, \eta_3)$ *as* $n \to \infty$ *uniformly in* η_2, η_3. *For all* $n \geq 1$, $q_{12,n}(\eta_2, \eta_3)$ *satisfies* $q_{12,n}(\eta_2, \eta_3) = 0 \iff (\eta_2, \eta_3) = (\eta_2^0, \eta_3^0)$, *and is continuously differentiable, with* $\partial q_{12,n}(\eta_2, \eta_3)/\partial(\eta_2, \eta_3')'$ *full column rank at* $(\eta_2^0, \eta_3^{0'})'$.

Remark 2.3.1 Assumption 2.3.1 (a) is justified by the decomposition in equation (2.19) and Assumptions 2.2.1 and 2.2.2. Secondly, we note that Assumption 2.3.1 is natural in our context. Assumption 2.3.1 (b) enforces that, for

$$q_{12,n}(\eta_2, \eta_3) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_n \left\{ \tilde{a}(y_{2i}, x_i, z_i) \left[y_{1i} - \Phi\left(-\eta_1^0 z_i' \xi^0 + \eta_2 y_{2i} + x_i' \eta_3 \right) \right] \right\},$$

we have that

$$-\frac{\partial q_{12,n}(\eta_2, \eta_3)}{\partial(\eta_2, \eta_3')'} = \frac{1}{n}\mathbb{E}_n \left\{ \sum_{i=1}^{n} \tilde{a}(y_{2i}, x_i, z_i)\, \phi_i\left(\eta_1^0, \eta_2, \eta_3, \theta_2^0\right) (y_{2i} \vdots x_i') \right\}$$

has full column rank at $(\eta_2^0, \eta_3^{0'})'$. This is tightly related to the requirement that the components of $(y_{2i} \vdots x_i')$ be linearly independent, since they coincide with the explanatory variables of the latent structural equation.

For the set,

$$\Upsilon(\theta_2^0) := \left\{ \eta \in \mathbb{R}^{k_x+2} \ : \ \eta = (\tilde{\rho}, \alpha + \tilde{\rho}, \beta' - \tilde{\rho}\pi^{0'})', \text{ for some } \theta_1 = (\tilde{\rho}, \alpha, \beta)' \in \Theta_1 \right\},$$

we state the null hypothesis of weak identification as follows.

Null Hypothesis of Weak Identification:

$$\mathrm{H}_0\left(\varsigma_n = \sqrt{n}\right) : \sup_{\eta \in \Upsilon(\theta_2^0)} \frac{1}{n} \left\| \mathbb{E}_n \left[\sum_{i=1}^{n} \tilde{a}(y_{2i}, x_i, z_i)\, \phi_i\left(\eta, \theta_2^0\right) z_i' \xi^0 \right] \right\| = O\left(\frac{1}{\sqrt{n}}\right). \tag{2.21}$$

The set $\Upsilon(\theta_2^0)$ denotes the set of structural parameters under the parametrisation in (2.17), so that the supremum over η in (2.21) is akin to a supremum over the structural parameters θ_1, given the true value θ_2^0 of the reduced form parameters. Both sets of structural parameters, the initial one Θ_1 and the reparameterised one $\Upsilon(\theta_2^0)$ are compact

subsets of $\mathbb{R}^{k_x+2}$. Based on the decomposition of (2.20), the identification strength of η_1 is determined by the rate ς_n, and $\varsigma_n = O(\sqrt{n})$ implies that even asymptotically, the population objective function is nearly flat in η_1. Such asymptotic behaviour of the objective function will lead to inconsistent estimation of η_1^0 in the rotated parameter space and for the structural parameter θ_1 in the original parameter space Θ_1.

Remark 2.3.2 It is worth noting that this definition of weak identification is a generalisation of Stock and Wright (2000) since it is considered at the true value θ_2^0 of the parameters of the reduced form regression equation. This must be seen as the relevant extension of the concept of weak instruments for the context of control variables. As explained in Section 2.2.3, the relevant explanatory variables for the structural equation are $\phi_i(\eta, \theta_2^0)(z_i'\xi^0, y_{2i}, x_i')'$. In particular, it is the impact $z_i'\xi^0$, at the true value ξ^0, that matters for identification and the pruning effect of $\phi_i(\eta, \theta_2^0)$, also at the true value $\theta_2^0 = (\pi^{0\prime}, \xi^{0\prime})'$. This extension is made possible by the reinforced identification condition in Assumption 2.2.2 (identification of θ_2^0 by the second set of moment conditions in isolation) and the choice of block-diagonal weighting matrix.

2.3.3. A Distorted J-test (DJ test) for the Null of Weak Identification

The decomposition in equation (2.20), along with Assumption 2.3.1, clarifies and confines the weak identification issue to the η_1 direction. Therefore, to construct a distorted testing approach for weak identification along the lines proposed in Section 2.3.1, it is precisely this direction, and only this direction, that should be distorted. To this end, and given a CUE $\hat{\eta}_n$, consider distorting the first component of $\hat{\eta}_n$ as

$$\hat{\eta}_n^\delta = \hat{\eta}_n + \begin{bmatrix} \delta_n & 0 & \dots & 0 \end{bmatrix}'.$$

Under the change of basis in equation (2.17), this is equivalent to distorting the CUE $\hat{\theta}_n$ as

$$\hat{\theta}_n^\delta = \begin{bmatrix} \hat{\theta}_{1n} \\ \hat{\theta}_{2n} \end{bmatrix} + \begin{bmatrix} \Delta_{1n} \\ \mathbf{0} \end{bmatrix}, \text{ where } \Delta_{1n} = \begin{bmatrix} \delta_n \\ -\delta_n \\ \delta_n \pi_0 \end{bmatrix},$$

which leads to a distortion of the entire structural parameter vector θ_1.

As explained in Section 2.3.1, under weak identification, if we distort the CUE $\hat{\theta}_n$ by some small value in the directions of weak identification, i.e., η_1, the value of the GMM criterion at $\hat{\theta}_n^\delta$ should not differ significantly from the criterion evaluated at $\hat{\theta}_n$. More precisely, recalling the definitions of $\bar{g}_n(\theta)$ and $S_n(\theta)$ given in Section 2.3.1,

$$J_n(\theta, \tilde{\theta}) = n\bar{g}_n(\theta)' S_n^{-1}(\tilde{\theta})\bar{g}_n(\theta), \quad J_n(\hat{\theta}_n, \hat{\theta}_n) = \min_{\theta \in \Theta} J_n(\theta, \theta),$$

we introduce the distorted J-test statistic:

$$J_n^\delta = n\bar{g}_n(\hat{\theta}_n^\delta)' S_n^{-1}(\hat{\theta}_n)\bar{g}_n(\hat{\theta}_n^\delta).$$

To deduce the behaviour of J_n^δ under the null of weak identification, we must maintain a regularity condition on the Jacobian of the moments. However, given that our null of weak identification is local about η_1, at the fixed value of θ_2^0, we are only required to maintain the following assumption.

Assumption 2.3.2 *Uniformly over* $\Upsilon(\theta_2^0)$, $\sqrt{n}\left\{\partial\bar{g}_n(\eta, \theta_2^0)/\partial\eta_1 - \mathbb{E}_n\left[\partial\bar{g}_n(\eta, \theta_2^0)/\partial\eta_1\right]\right\} \Rightarrow \Psi(\eta, \theta_2^0)$, *for* $\Psi(\eta, \theta_2^0)$ *a mean-zero Gaussian process, and where* $\Rightarrow$ *denotes weak convergence in the sup-norm.*

We note that Assumption 2.3.2 is guaranteed under Assumption 2.2.1 and a functional central limit theorem. See the proof of Lemma 2.7.2 in the Appendix for details. We state this result as an assumption to ease the comparison with standard results.

Proposition 2.3.1 (Lack of Consistency) *If Assumptions 2.2.1-2.3.2 are satisfied, and if* $\mathbb{E}_n[\|\tilde{a}(y_{2i}, x_i, z_i)z_i'\|^2] < \infty$, *then under the null of weak identification, for any* $\delta_n = o(1)$,

$$\operatorname*{plim}_{n\to\infty} \sqrt{n}\left[\bar{g}_n(\hat{\theta}_n^\delta) - \bar{g}_n(\hat{\theta}_n)\right] = 0.$$

In addition, if $\sup_{\theta\in\Theta}\|S_n^{-1}(\theta)\| = O_p(1)$, *then*

$$\operatorname*{plim}_{n\to\infty}\left[J_n^\delta - J_n(\hat{\theta}_n, \hat{\theta}_n)\right] = 0.$$

Proposition 2.3.1 demonstrates that under the null of weak identification, the curvature of the objective function is insensitive to a small departure from the CUE, indicating the lack of consistency of $\hat{\theta}_n$. By adapting the general testing approach of Antoine and Renault (2020), Proposition 2.3.1 paves the way for a testing strategy for weak instruments in discrete choice models. Recall that the number of model parameters is $p = 2 + 2k_x + k_z$, and H denotes the number of moments.

Theorem 2.3.2 (Distorted J-test: Under the Null) *Under Assumptions 2.2.1-2.3.2 and the null of weak identification, for any deterministic sequence $\delta_n = o(1)$, define the distorted J-test by the rejection region:*

$$W_n^\delta = \left\{ J_n^\delta > \chi^2_{1-\alpha}\left(H + 1 - p\right) \right\},$$

where $\chi^2_{1-\alpha}\left(H + 1 - p\right)$ is the $(1-\alpha)$ quantile of the Chi-square distribution with $(H + 1 - p)$ degrees of freedom. Under the null hypothesis of weak identification, W_n^δ has asymptotic size of at most α.

As discussed in Section 2.3.1, the CUGMM framework allows us to control the size of our test by ensuring that we can obtain a convenient upper bound for J_n^δ under the null of weak identification. Since there is only a single direction of weakness in the rotated parameter space, this bound can be based on the $\chi^2(H + 1 - p)$ distribution; please see the proof of Theorem 2.3.2 for details. While the test statistic J_n^δ coincides with the one given in Section 2.3.1, we have improved the asymptotic power of the test W_n^δ by using a critical value calculated from $\chi^2\left(H + 1 - p\right)$ instead of $\chi^2\left(H\right)$. This power gain is obviously important since we may be afraid that our test would be overly conservative.

2.3.4. Estimation and Testing Under the Alternative

In this section, we prove that W_n^δ, the distorted J-test based on J_n^δ, is consistent under the alternative. Before presenting this result, we first discuss the asymptotic behaviour of the CUE under the alternative.

Estimation Under the Alternative

We first deduce the properties of the CUE under the parameterisation defined by (2.17), and then translate these results to the original parameters. To this end, and for η as defined in equation (2.17), define $\zeta = (\eta', \theta_2')'$ with true value ζ^0. The vector ζ represents the following change of basis in the parameter space:

$$\theta = R\zeta, \text{ where } R = \begin{pmatrix} R_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{k_x+k_z} \end{pmatrix}, \; R_1 = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ \pi^0 & \mathbf{0} & \mathbf{I}_{k_x} \end{pmatrix}.$$

In this rotated space, the CUE of ζ^0 is given by

$$\hat{\zeta}_n = \underset{\zeta \in R^{-1}\Theta}{\operatorname{argmin}} \, \bar{g}_n(R\zeta)' S_n^{-1}(R\zeta) \bar{g}_n(R\zeta).$$

To deduce the properties of $\hat{\zeta}_n$ under the alternative, we first recall that the null of weak identification, defined by (2.21), implies that

$$\begin{aligned} \sup_{\eta \in \Upsilon(\theta_2^0)} \left\| \frac{1}{n} \mathbb{E}_n \left\{ \sum_{i=1}^n \frac{\partial g_i(\eta, \theta_2^0)}{\partial \eta_1} \right\} \right\| &= \sup_{\eta \in \Upsilon(\theta_2^0)} \left\| \frac{1}{n} \mathbb{E}_n \left\{ \sum_{i=1}^n \tilde{a}\left(y_{2i}, x_i, z_i\right) \phi_i\left(\eta, \theta_2^0\right) z_i' \xi^0 \right\} \right\| \\ &= O(1/\sqrt{n}). \end{aligned}$$

The alternative hypothesis to this null implies the existence of a deterministic sequence $\varsigma_n = o(\sqrt{n})$ such that

$$\limsup_{n \to \infty} \sup_{\eta \in \Upsilon(\theta_2^0)} \left\| \frac{1}{n} \mathbb{E}_n \left\{ \sum_{i=1}^n \tilde{a}\left(y_{2i}, x_i, z_i\right) \phi_i\left(\eta, \theta_2^0\right) z_i' \xi^0 \right\} \varsigma_n \right\| > 0.$$

To deduce the behaviour of the CUE $\hat{\zeta}_n$ under the alternative, we slightly reinforce this condition as follows.

Assumption 2.3.3 *Under the alternative hypothesis, there exists a deterministic sequence $\varsigma_n = o(\sqrt{n})$ and a continuous, and deterministic vector function $V^0(\eta)$ such that,*

$\inf_{\eta\in\Upsilon(\theta_2^0)} \|V^0(\eta)\| > 0$, *and*

$$\lim_{n\to\infty} \sup_{\eta\in\Upsilon(\theta_2^0)} \left\| \frac{1}{n}\mathbb{E}_n \left\{ \sum_{i=1}^{n} \tilde{a}\left(y_{2i}, x_i, z_i\right) \phi_i\left(\eta, \theta_2^0\right) z_i'\xi^0 \right\} \varsigma_n - V^0(\eta) \right\| = 0.$$

Remark 2.3.3 Even though Assumption 2.3.3 arguably limits the scope of the alternative hypothesis, it is more general than if we were to follow the approach of Staiger and Stock (1997) and characterise identification strength only through the reduced form regression equation. In the latter case, one would consider that the reduced form regression evolves according to the drifting DGP

$$\mathbb{E}_n[y_{2i}\,|x_i, z_i] = x_i'\pi^0 + z_i'\xi_n^0.$$

Under the null of weak identification, we have that $\xi_n^0 = O(1/\sqrt{n})$. In contrast, Assumption 2.3.3 would require that, for some $\gamma^0 \in \mathbb{R}^{k_z}$ with $\|\gamma^0\| > 0$ and some $\varsigma_n = o(\sqrt{n})$,

$$\lim_{n\to\infty} \varsigma_n\xi_n^0 = \gamma^0, \quad \text{and } V^0(\eta) = \mathbb{E}_n\left[\tilde{a}(y_{2i}, x_i, z_i)\phi_i(\eta, \theta_2^0)z_i'\right]\gamma^0 \neq 0.$$

However, as explained in Section 2.2.3, this approach to characterise identification strength is not sufficient in our opinion, since it only accounts for the instrument strength in the reduced form regression, ξ_n^0, and does not account for the interactions between the instrumental function $\tilde{a}(y_{2i}, x_i, z_i)$ and $\phi_i\left(\eta, \theta_{2n}^0\right) z_i'\xi_n^0$, which may result in the pruning of large realisations of the instruments via the behaviour of $\phi_i\left(\eta, \theta_{2n}^0\right)$.

By defining the alternative hypothesis using Assumption 2.3.3, we clearly partition the two possible cases for estimation of ζ^0: (i) if identification is weak, $\hat{\zeta}_n$ is not consistent (as implied by Proposition 2.3.1), nor are other commonly applied estimators such as 2SCML or Quasi-LIML estimators; (ii) when identification is not weak, $\hat{\zeta}_n$ is consistent.

Proposition 2.3.3 (Consistency)
If Assumptions 2.2.1-5 are satisfied, and if $\sup_{\zeta\in R^{-1}\Theta} \|S_n^{-1}(\zeta)\| = O_p(1)$, *then* $\|\hat{\zeta}_n - \zeta^0\| = o_p(1)$.

The asymptotic distribution of $\hat{\zeta}_n$ depends on the behaviour of the Jacobian for the

moments. Under Assumptions 2.3.1 and 2.3.3, the scaled Jacobian of the moment functions, as defined below in Lemma 2.3.4, is full rank under the following mild assumption, which, if we take $\tilde{b}(x_i, z_i) = (x_i' \ : \ z_i')'$ is nothing but the standard rank condition on the reduced form regression.

Assumption 2.3.4 *For all $n \geq 1$, $\mathbb{E}_n[\tilde{b}(x_i, z_i)(x_i' \ : \ z_i')]$ has column rank $(k_x + k_z) = dim(\theta_2)$.*

Lemma 2.3.4 *Under **Assumptions 2.2.1-2.3.4**, for a given sequence $\varsigma_n = o(\sqrt{n})$, the matrix*

$$M = \operatorname*{plim}_{n\to\infty} \left\{ \frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'} \right\} \Lambda_n, \ \text{where } \Lambda_n = \begin{bmatrix} \varsigma_n & \boldsymbol{O}_{p-1} \\ \boldsymbol{O}_{p-1} & \boldsymbol{I}_{p-1} \end{bmatrix},$$

exists and is full column rank.

Given the full-rank nature of the scaled Jacobian, we would expect the CUE to be asymptotically normal. In particular, under the alternative (as defined by Assumptions 2.3.1 and 2.3.3), we can then deduce the following result.

Theorem 2.3.5 (Asymptotic Normality) *If **Assumptions 2.2.1-2.3.4** are satisfied then*

$$\sqrt{n}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0) \xrightarrow{d} \mathbb{N}\left(0, [M'S^{-1}M]^{-1}\right), \ \text{where } S := \operatorname*{plim}_{n\to\infty} S_n(\zeta^0).$$

As expected, all entries of ζ, save for η_1, are $\sqrt{n}$-consistent and asymptotically normal GMM estimators. In contrast, the direction η_1 converges at the $\{\sqrt{n}/\varsigma_n\}$-rate, which is possibly slower than $\sqrt{n}$. Of course, our goal is not to conduct inference on ζ^0, but on θ^0. By the change of basis in (2.17), $\theta = R\zeta$, and Theorem 2.3.5 implies that the feasible CUGMM estimator $\hat{\theta}_n$ satisfies

$$\sqrt{n}\Lambda_n^{-1}R^{-1}(\hat{\theta}_n - \theta^0) \xrightarrow{d} \mathbb{N}\left(0, [M'S^{-1}M]^{-1}\right). \tag{2.22}$$

Importantly, since the matrix R is not diagonal, the slower rate of $\{\sqrt{n}/\varsigma_n\}$ pollutes the entire vector of structural parameters $\theta_1 = (\tilde{\rho}, \alpha, \beta')'$, which follows from the change

of basis in (2.17). Therefore, all structural parameter estimates in the probit model converge at the slower $\{\sqrt{n}/\varsigma_n\}$-rate.

Equation (2.22) itself does not directly provide a feasible inference strategy since the matrix R depends on the unknown parameter π^0. Of course the matrix R may be consistently estimated. However, as explained by Antoine and Renault (2012) (see the discussion of their Theorem 4.5), a sufficient condition to ensure that the estimation of R does not pollute the asymptotic distribution in (2.22) is that the matrix R is estimable at a rate faster than $n^{1/4}$. In the case of the probit model, the matrix R only depends on the unknown true reduced form parameter π^0, which is strongly identified and consistently estimable at the $\sqrt{n}$-rate. Therefore, if $\hat{R}_n$ denotes the matrix R when π^0 is replaced by $\hat{\pi}_n$, we can conclude that, following Theorem 4.5 in Antoine and Renault (2012),

$$\sqrt{n}\Lambda_n^{-1}\hat{R}_n^{-1}(\hat{\theta}_n - \theta^0) \xrightarrow{d} \mathbb{N}\left(0, [M'S^{-1}M]^{-1}\right). \tag{2.23}$$

Remark 2.3.4 The result in equation (2.23) implies that $\sqrt{n}\Lambda_n^{-1}\hat{R}_n^{-1}(\hat{\theta}_n - \theta^0)$ behaves like a mean-zero Gaussian random variable, whose variance can be consistently estimated by

$$[\Lambda_n\hat{R}_n'\{\partial\bar{g}_n(\hat{\theta}_n)/\partial\theta'\}'S_n^{-1}(\hat{\theta}_n)\{\partial\bar{g}_n(\hat{\theta}_n)/\partial\theta'\}\hat{R}_n\Lambda_n]^{-1}.$$

However, Theorem 2.3.5 *does not say* that the common estimator of the variance-matrix of $\sqrt{n}(\hat{\theta}_n - \theta^0)$, obtained using the standard formula

$$[\{\partial\bar{g}_n(\hat{\theta}_n)/\partial\theta'\}'S_n^{-1}(\hat{\theta}_n)\{\partial\bar{g}_n(\hat{\theta}_n)/\partial\theta'\}]^{-1},$$

is well-behaved, which follows by noting that the matrix $\frac{\partial\bar{g}_n(\theta^0)}{\partial\theta}S_n^{-1}(\theta^0)\frac{\partial\bar{g}_n(\theta^0)}{\partial\theta'}$ is asymptotically singular unless $\varsigma_n = O(1)$. Fortunately, Theorem 5.1 in Antoine and Renault (2012) allows us to conclude that standard formulas for Wald inference based on the GMM estimator $\hat{\theta}_n$ are asymptotically valid. The main intuition is that the Studentisation implied by Wald inference cancels out the required rescaling terms. This is all the more important given that the rescaling factor ς_n is unknown in practice.

We stress that this result is in contrast to the general non-linear case where the asymptotic normality requires faster than $n^{1/4}$ convergence rate, and it is only due to the

specificities of the probit model that we are able to conduct valid Wald inference as soon as identification is not genuinely weak. That is, any near weakness, even as severe as ς_n being arbitrarily close to $\sqrt{n}$, will still allow us to compute a consistent GMM estimator and apply standard formulas for Wald inference based on this estimator.

The Power of the Distorted J-Test

The key to ensuring that the size of W_n^δ is asymptotically controlled is the equivalence between J_n, the usual J-statistic, and J_n^δ, the distorted J-statistic, that obtains under the null of weak identification. However, as demonstrated by Proposition 2.3.3 and Theorem 2.3.5, under the alternative hypothesis the CUE is consistent and asymptotically normal. Therefore, there is no reason to suspect that J_n and J_n^δ will be asymptotically equivalent under the alternative, at least under reasonable choices for the tuning parameter δ_n.

The following result demonstrates that under the alternative, the distorted J-test, W_n^δ, is a consistent test for the null of weak instruments across a wide range of choices for the perturbation sequence δ_n.

Theorem 2.3.6 (Distorted J-test: Under the Alternative) *If **Assumptions 2.2.1-2.3.4** are satisfied, then W_n^δ is consistent under the alternative so long as $\{\sqrt{n}/\varsigma_n\}\delta_n \to \infty$ as $n \to \infty$.*

Remark 2.3.5 Theorem 2.3.6 implies that our choice of δ_n has important consequences for the power of the distorted J-test. All else equal, the test is more powerful the slower δ_n goes to zero. However, it is also helpful to understand how fast δ_n can converge to zero before the result of Theorem 2.3.6 is invalidated. To this end, consider the rate requirement on δ_n that results from parametrising ς_n as $\varsigma_n = n^\lambda$ for some $0 < \lambda < 1/2$. Using this parametrisation, we see that the distorted J-test is consistent so long as $\delta_n n^{1/2-\lambda} \to \infty$, and clarifies that if δ_n goes to zero too fast, i.e., if $\delta_n \ll n^{\lambda-1/2}$, the test can not be consistent.

Remark 2.3.6 It is worth keeping in mind that Assumption 2.2.2 maintains that both the structural and reduced form moments are correctly specified. Thus, when the observed data lead to a rejection of W_n^δ, we immediately conclude that it is *not due to*

misspecification of the moment conditions but *due to their identification power*. However, if the model is misspecified, but we reject the null of weak identification, then we can actually consistently test for model misspecification. Indeed, under the alternative, the standard overidentification test

$$\{J_n(\hat{\theta}_n, \hat{\theta}_n) > \chi^2_{1-\alpha}(H-p)\},$$

remains a consistent test for model misspecification. As such, if we reject the null of weak identification, we can compare the value of $J_n(\hat{\theta}_n, \hat{\theta}_n)$ against $\chi^2_{1-\alpha}(H-p)$ to deduce a consistent test for model misspecification.

2.3.5. Testing Procedure

We now explain one approach to implement our distorted J-test in practice. The key step in the testing procedure is to choose the perturbation (tuning parameter) δ_n. To this end, we take $\delta_n = \delta/r_n$, and fix $r_n = \log\{\log(n)\}$. It is then possible to choose δ using a data-driven approach.

To present our approach to choosing δ, first recall that the perturbation $\delta_n = \delta/\log\{\log(n)\}$ can be thought of as only being applied to the single direction of weakness in the rotated parameter space; namely, the parameter η_1, which by equation (2.17) is nothing but $\tilde{\rho}$. Therefore, it is with respect to the magnitude of $\hat{\tilde{\rho}}_n$ that the perturbation δ_n should be chosen.

To ensure the value of δ_n is sufficiently close to the magnitude of $\hat{\tilde{\rho}}_n$, we design a grid of m candidate points for δ by dissecting the *standard confidence interval* of $\hat{\tilde{\rho}}_n$ into m equal regions, with δ then taken to be the midpoint of the m-th region. This results in m different values for δ, denoted by δ_i, $i = 1, \ldots, m$, and produces a grid of perturbations.

Whilst it is possible to use any given $\delta_{n,i}$ to conduct the test, we suggest carrying out the test across the entire grid of $\delta_{n,i}$ values and then appropriately modify the critical value via a Bonferroni correction. In particular, let $J^{\delta}_{n,i}$ denote the test statistic J^{δ}_n calculated under the perturbation $\delta_{n,i}$. This approach would lead us to reject the null of weak

identification if

$$\max_{i\in\{1,\ldots,m\}} J^{\delta}_{n,i} > \chi^2_{1-\alpha/m}(H+1-p).$$

Using the above decision rule, our approach can be implemented using the following four steps.

(1) Compute $\hat{\theta}_n = \text{argmin}_{\theta\in\Theta} J_n(\theta,\theta)$;

(2) For a given choice of m, choose the sequence of tuning parameter $\delta_n = \delta/r_n$, as described above;

(3) For each $i = 1,\ldots,m$, compute the test statistic $J^{\delta}_{n,i}$, as defined in Section 2.3.3;

(4) Rejection rule: reject if $\max_{i\in\{1,\ldots,m\}} J^{\delta}_{n,i} > \chi^2_{1-\alpha/m}(H+1-p)$.

Under the null hypothesis, the testing procedure is size controlled for any choice of $\delta_{n,i} = o(1)$, while under the alternative the choice of $\delta_{n,i}$ only has implications for the power of the test. Moreover, since the values of δ_i are chosen from some compact set, dividing by $\log\{\log(n)\}$ ensures that $\delta_{n,i} = o(1)$ under both the null and alternative.

2.3.6. Generalising the Rule-of-Thumb to Probit Models

We begin our discussion on the so-called "rule-of-thumb", initially inspired by the work of Staiger and Stock (1997), in the infeasible situation where the latent endogenous variable y^*_{1i} is observable, meaning that we would consider a bivariate linear model. For sake of expositional simplicity, let us consider a simplification of this model whereby the vector x_i only contains a constant, so that the model becomes

$$\begin{aligned} y^*_{1i} &= \alpha y_{2i} + \beta + u_i \\ y_{2i} &= \pi + z_i'\xi + v_i. \end{aligned} \tag{2.24}$$

The rule-of-thumb starts from the reduced form regression and its OLS estimator for ξ,

$$\hat{\xi}_n = (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{Y}_2,$$

where for $\mathbf{1}_n$ a $(n \times 1)$-vector of ones

$$Y_2 = (y_{21}, \ldots, y_{2n})', \ \tilde{Y}_2 = Y_2 - \bar{y}_{2n}\mathbf{1}_n,$$
$$Z = (z_1', \ldots, z_n')', \ \tilde{Z} = Z - \bar{Z}_n,$$

and where $\bar{y}_{2n} = \frac{1}{n}\sum_{i=1}^{n} y_{2i}$ and $\bar{Z}_n$ denotes the $(n \times k_z)$ matrix whose j^{th}-column has all its entries equal to

$$\bar{z}_{j,n} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}.$$

Let F_n denote the F-test statistic for testing the null hypothesis that the vector ξ of coefficients for the variables z_i in the reduced form regression are zero. Under the assumption of conditional homoscedasticity for the error term v_i, the F-test statistic can be written as

$$F_n = \frac{n - k_z}{nk_z}\frac{1}{\hat{\sigma}^2_{v,n}}\left[\hat{\xi}_n'\left(\tilde{Z}'\tilde{Z}\right)\hat{\xi}_n\right],$$

with $\hat{\sigma}^2_{v,n}$ a consistent estimator of variance of v_i, σ^2_v. The rule-of-thumb amounts to conclude that instruments are strong (i.e., consistent estimation is feasible) if F_n exceeds a pre-specified threshold value, which differs from the standard critical value used to test the null hypothesis $\mathrm{H}_0 : \xi = 0$, and which has been extensively documented by Stock and Yogo (2005). The rationale for this rule can be understood from the drifting DGP considered in Remark 2.3.3. Under the alternative hypothesis to the null of weak identification, for n large,

$$\xi_n^0 \sim \frac{\gamma^0}{\varsigma_n} \Longrightarrow k_z F_n \sim \frac{n}{\varsigma_n^2}\frac{1}{\sigma_v^2}\gamma^{0\prime}\mathrm{Var}\left(z_i\right)\gamma^0. \tag{2.25}$$

Therefore, under the null of weak identification ($\varsigma_n = \sqrt{n}$), F_n in equation (2.25) has a finite limit, whilst under the alternative ($\varsigma_n = o\left(\sqrt{n}\right)$) the statistic F_n diverges to infinity with a slope defined by the squared norm of γ^0 and a weighting matrix that is proportional to $\mathrm{Var}(z_i)/\mathrm{Var}(v_i)$. This sounds like a natural criterion to measure instrument strength in the infeasible model (2.24), since the reduced form regression will lead to the control variable $v_i = y_{2i} - \pi - z_i'\xi$ and endogeneity in the structural equation will be

controlled thanks to the two-stage residual inclusion (2SRI):

$$y_{1i}^{*} = \alpha y_{2i} + \beta + \tilde{\rho}\left[y_{2i} - \pi - z_i'\xi\right] + \varepsilon_i. \tag{2.26}$$

Since identification of $\eta_1 = \tilde{\rho}$ in equation (2.26) depends on the variation of

$$z_i'\xi_n^0 \sim \frac{z_i'\gamma^0}{\varsigma_n},$$

it may sound natural to assess the magnitude of γ^0 after normalisation by the variance of z_i. As noted by Stock and Andrews (2005), "IVs can be weak and the F-statistic small, either because γ is close to zero or because the variability of z_i is low relative to the variability of v_i." However, the F-test statistic follows a Fisher distribution (and asymptotically a distribution $\chi^2\left(k_z\right)/k_z$) under the null $\mathrm{H}_0 : \xi = 0$ only when the reduced form error term v_i is conditionally homoscedastic. When one is concerned with the presence of conditional heteroscedasticity in this equation (i.e., non-constant $\mathrm{Var}[v_i \mid z_i]$), one may consider the heteroscedasticity corrected Fisher test statistic

$$F_n^{*} = \frac{n - k_z}{k_z}\left[\hat{\xi}_n'\hat{\Sigma}_n^{-1}\hat{\xi}_n\right],$$

where $\hat{\Sigma}_n$ is a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\xi}_n - \xi_n^0)$. While Stock and Yogo (2005) propose to extend the use of the rule-of-thumb by using instead F_n^{*} in case of conditional heteroscedasticity, several authors, including Andrews (2018) and Montiel Olea and Pflueger (2013), have documented the disappointing performance of the heteroscedasticity corrected rule-of-thumb. One may help to clarify this issue by noting that, denoting $\tilde{z}_i$ to be the i-th column vector of the matrix $\tilde{Z}'$, for n large and for $\sigma_v^2(z_i) = \mathrm{Var}[v_i \mid z_i]$,

$$\xi_n^0 \sim \frac{\gamma^0}{\varsigma_n} \Longrightarrow k_z F_n^{*} \sim \frac{n}{\varsigma_n^2}\gamma^{0\prime}\mathrm{Var}\left(z_i\right)\left[\mathbb{E}\left(\tilde{z}_i\tilde{z}_i'\sigma_v^2(z_i)\right)\right]^{-1}\mathrm{Var}\left(z_i\right)\gamma^0. \tag{2.27}$$

Equation (2.27) is a straightforward extension of a result provided by Antoine and Renault (2020), and makes explicit how robustifying the test statistic for heteroscedasticity modifies the rule-of-thumb. This modification is arguably puzzling since what really matters for identification power, namely the residual inclusion of v_i in the structural equation

(2.26), is not fully captured by $\sigma_v^2(z_i)$. More precisely, the conditional heteroscedasticity that intuitively matters in the structural equation is instead

$$\sigma_u^2(z_i) = \text{Var}[u_i \,|z_i] = \tilde{\rho}^2 \text{Var}[v_i \,|z_i] + \text{Var}[\varepsilon_i \,|z_i] \,.$$

This intuition is confirmed by Antoine and Renault (2020) who show that, when nesting the IV estimation procedure in a GMM framework, the distorted J-test leads to a decision rule based on the following weighted norm of γ^0:

$$\frac{n}{\varsigma_n^2} \gamma^{0\prime} \text{Var}\left(z_i\right) \left[\mathbb{E}\left(\tilde{z}_i \tilde{z}_i' \sigma_u^2(z_i)\right)\right]^{-1} \text{Var}\left(z_i\right) \gamma^0.$$

In the context of the probit model, where only the sign y_{1i} of y_{1i}^* is observed, the 2SRI equation becomes

$$y_{1i} = \Phi\left[\alpha y_{2i} + \beta + \tilde{\rho}\left(y_{2i} - \pi - z_i'\xi\right)\right] + \varepsilon_i,$$

for some error term ε_i, and the conditional heteroscedasticity in the structural equation takes the form

$$\text{Var}[\varepsilon_i \,|y_{2i}, z_i] = \Phi_i\left(\theta^0\right)\left[1 - \Phi_i\left(\theta^0\right)\right], \quad \text{where } \Phi_i\left(\theta\right) = \Phi\left[\alpha y_{2i} + \beta + \tilde{\rho}\left(y_{2i} - \pi - z_i'\xi\right)\right].$$

One may then expect that any generalised rule-of-thumb for probit models must account not only for this conditional heteroscedasticity but also the impact of the non-linearity in the structural equation. In the simple context of Remark 2.3.3, we may then expect that the key element to obtain a decision rule about weak instruments in the probit model is the magnitude of the vector

$$V^0\left(\eta\right) = \mathbb{E}_n\left[\tilde{a}\left(y_{2i}, z_i\right) \phi_i\left(\eta, \theta_2^0\right) z_i'\right] \gamma^0, \text{ where } \|\gamma^0\| > 0.$$

More generally, since the alternative to weak identification, defined by Assumption 2.3.3, is tantamount to the non-nullity of the vector $V^0\left(\eta\right)$, the generalised rule-of-thumb should be based on a norm of $V^0\left(\eta\right)$. We argue that we do have a well-suited generalization for the standard rule-of-thumb when applying a decision rule that rejects the null of weak identification if the norm $\|U\|$, of a certain vector U, exceeds a specified threshold with the following definition for U.

(i) $U = \sqrt{n}\text{Var}\left(z_i\right)^{1/2}/\sigma_v\xi^0$ for a linear model with conditional homoscedasticity (i.e. the standard rule-of-thumb);

(ii) $U = \sqrt{n}\left[\mathbb{E}\left(\tilde{z}_i\tilde{z}_i'\sigma_u^2(z_i)\right)\right]^{-1/2}\text{Var}\left(z_i\right)\xi^0$ for a linear model with conditional heteroscedasticity (i.e. the generalisation of the standard rule-of-thumb proposed by Antoine and Renault, 2020);

(iii) $U = \sqrt{n}S_{11,n}^{-1/2}\left(\theta^0\right)\mathbb{E}_n\left[\tilde{a}\left(y_{2i},z_i\right)\phi_i\left(\eta,\theta_2^0\right)z_i'\right]\xi^0\delta_n$ for the probit model (2.24) (in the context of Remark 2.3.3) and more generally $U = \sqrt{n}S_{11,n}^{-1/2}\left(\theta^0\right)V^0\left(\eta\right)\delta_n/\varsigma_n$, where the perturbation term δ_n is introduced by the design of the distorted J-test.

It is worth realising that this generalised rule-of-thumb is, for n large, precisely what is performed by our test for the null of weak identification based on the distorted J-test statistic. To see this, we extend the argument of Antoine and Renault (2020) by noting that under the alternative, our distorted J-test statistic sets the focus on the norm of

$$U = S_n^{-1/2}\left(\theta^0\right)\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta),$$

where

$$\bar{g}_n(\hat{\theta}_n^\delta) = \bar{g}_n(\hat{\theta}_n) + \begin{bmatrix} \bar{g}_{1n}(\hat{\theta}_n^\delta) - \bar{g}_{1n}(\hat{\theta}_n) \\ \mathbf{0} \end{bmatrix}.$$

Noting that,

$$\sqrt{n}\left[\bar{g}_{1n}(\hat{\theta}_n^\delta) - \bar{g}_{1n}(\hat{\theta}_n)\right] = \sqrt{n}\frac{\partial\bar{g}_{1n}}{\partial\eta_1}(\eta_{1n}^*,\hat{\eta}_{2n},\hat{\eta}_{3n},\hat{\theta}_{2n})\delta_n,$$

where η_{1n}^* denotes a component-by-component intermediate value between the first coefficient of $\hat{\theta}_n$ and $\hat{\theta}_n^\delta$, under the alternative hypothesis to the null of weak identification

$$\begin{aligned}\frac{\partial\bar{g}_{1n}}{\partial\eta_1}\left(\eta_{1n}^*,\hat{\eta}_{2n},\hat{\eta}_{3n},\hat{\theta}_{2n}\right) &= \mathbb{E}_n\left[\frac{\partial\bar{g}_{1n}}{\partial\eta_1}\left(\eta^0,\theta_2^0\right)\right] + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{1}{n}\mathbb{E}_n\left\{\sum_{i=1}^n\tilde{a}\left(y_{2i},z_i\right)\phi_i\left(\eta^0,\theta_2^0\right)z_i'\xi^0\right\} + O_p\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

and where[2]

$$\frac{1}{n}\mathbb{E}_n\left\{\sum_{i=1}^{n}\tilde{a}\left(y_{2i},z_i\right)\phi_i\left(\eta,\theta_2^0\right)z_i'\xi^0\right\}\sim\frac{V^0\left(\eta\right)}{\varsigma_n}$$

is the dominant term since $\varsigma_n = o\left(\sqrt{n}\right)$. To summarise, under the alternative hypothesis to the null of weak identification, and for a δ_n such that $\{\sqrt{n}/\varsigma_n\}\delta_n \to \infty$,

$$\|U\| = \left\|S_n^{-1/2}\left(\theta^0\right)\sqrt{n}\bar{g}_n\left(\hat{\theta}_n^{\delta}\right)\right\| \sim \left\|S_{11,n}^{-1/2}\left(\theta^0\right)V^0\left(\eta^0\right)\right\|\frac{\sqrt{n}}{\varsigma_n}\delta_n,$$

which diverges as $n \to \infty$ and yields a natural generalisation of the rule-of-thumb to probit models.

2.4. Monte Carlo: Conventional Weak IV Tests v.s. Distorted J-test

In this section, we verify the properties of the distorted J-test (hereafter, DJ test) and compare this test against three commonly used weak IV tests, which, even though they are not designed for discrete choice models, have been widely applied in the literature on discrete choice modelling: (i) the Staiger and Stock (1997) standard rule-of-thumb (SS); (ii) Stock and Yogo (2005) (SY); (iii) the robust weak IV test of Montiel Olea and Pflueger (2013) (Robust).

We generate observed data according to

$$y_{1i} = 1[\beta + \alpha y_{2i} + u_i > 0], \quad y_{2i} = \pi + \xi z_i + v_i, \quad i = 1, 2, ..., n \tag{2.28}$$

where $z_i \sim \mathbb{N}(0, \sigma_z^2)$ is i.i.d. univariate, $(u_i, v_i)'$ is i.i.d. homoscedastic and normally distributed, and $(u_i, v_i)'$ is independent of z_i. We set $\beta = 0.5$, $\alpha = 1$ and $\pi = 0.3$. In addition, we take $\rho = corr(u_i, v_i) \in \{0.5,\ 0.95\}$, and $\sigma_u = 1/\sqrt{1-\rho^2}$ (to ensure the normalisation of $\text{Var}[u_i|y_{2i}, z_i] = 1$). To characterise the potential instrument weakness, we adjust the value of ξ to restrict the correlation between the endogenous regressor y_{2i} and the instrument z_i to be $corr(y_{2i}, z_i) = \gamma/n^{\lambda}$, with $\gamma = 1.5$ and we consider a grid of

[2]The $O_p(1/\sqrt{n})$ term in the expansion of $\partial\bar{g}_{1n}(\eta_{1n}^*, \hat{\eta}_{2n}, \hat{\eta}_{3n}, \hat{\theta}_{2n})/\partial\eta_1$ can be deduced via a Taylor series expansion, re-arranging terms, and noting that the derivative of the Jacobian, in the η_1 direction, is also degenerate at the ς_n-rate.

values for $\lambda \in \{0.5,\ 0.4,\ 0.3,\ 0.2,\ 0.1\}$.

Since the performance of the DJ test and the standard weak IV tests may depend on σ_z and σ_v, we simulate data using the following grids: $\sigma_z \in \{0.2,\ 0.5,\ 1,\ 5,\ 10\}$ and $\sigma_v \in \{0.2,\ 0.5,\ 1,\ 5,\ 10\}$. For each Monte Carlo trial, we take the sample size to be one of $n = 500,\ 5000,\ 10000$ and consider $N = 1000$ Monte Carlo replications.

Across each Monte Carlo design, $\theta = (\tilde{\rho}, \alpha, \beta, \pi, \xi)'$ is estimated by CUGMM with a single degree of over-identification. We choose the instrument functions $a_i = a(y_{2i}, z_i) = (1, y_{2i}, z_i, z_i^2, 0, 0)'$ and $b_i = b(z_i) = (0, 0, 0, 0, 1, z_i)'$. The DJ test is implemented following the procedure presented in Section 2.3.5. For computational simplicity, in the Monte Carlo simulations, we adopt the perturbation $\delta_n = \hat{\tilde{\rho}}/\log(\log(n))$, where $\hat{\tilde{\rho}}$ is the CUGMM estimate of $\tilde{\rho}$ in each Monte Carlo replication. This procedures is a simplified version of the data-driven approach developed in Section 2.3.5. Using a 5% significant level, we reject the null hypothesis of weak instruments in accordance to Theorem 2.3.6; i.e., we reject the null if $J_n^\delta > \chi^2_{0.95}(H + 1 - p)$, where in this case $H = 6$, $p = 5$ and $\chi^2_{0.95}(H + 1 - p) = 5.99$. Theoretically, the hypotheses of the DJ test corresponds to $H_0 : \lambda = 0.5$, the alternative to $\lambda < 0.5$.[3] However, we note that in finite samples, it is hardly the case that λ alone determines the behaviour of the CUEs.

Given this, to compare the behaviour of the DJ test with the conventional linear tests, we introduce two sets of criteria to assess the potential impact of instrument weakness in finite samples: the behaviour of the CUE and the size distortions of the associated Wald statistic. Specifically, we compute the bias, standard deviation (s.d.) and relative root mean square error (rrmse) as below (take α as an example) to measure the estimation performance under different designs:

$$\text{bias} = \bar{\hat{\alpha}} - \alpha^0, \quad \text{s.d.} = \sqrt{\frac{1}{N}\sum_{l=1}^{N}(\hat{\alpha}_l - \bar{\hat{\alpha}})^2}, \quad \text{rrmse} = \sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(\frac{\hat{\alpha}_l - \alpha^0}{\alpha^0}\right)^2} \tag{2.29}$$

[3] We note that the null hypothesis of each test are slightly different: DJ- $H_0 : \lambda = 0.5$; SS- $F_n < 10$ as an informal null hypothesis; SY- the triple $\{\xi, \sigma_v^2, \sigma_z^2\}$ is such that 2SLS relative bias or Wald test size distortion is larger than a given tolerance using the Cragg-Donald statistic; the Robust test regards that the Nagar bias exceeds a fraction of the benchmark as null. Although the definitions of the weak instrument are different for each test, their null hypothesis are consistent in the sense to capture situations under which the instrument is weak.

where $\bar{\hat{\alpha}} = 1/N \sum_{l=1}^{N} \hat{\alpha}_l$, $\hat{\alpha}_l$ stands for the l-th Monte Carlo CUGMM estimate and α^0 is the true value. As proven in Sections 2.3.3 and 2.3.4, under the null the CUE is consistent, while under the alternative, the estimator will be consistent and asymptotically normal, albeit with non-standard rates. Unlike Stock and Yogo (2005), who choose the relative bias of 2SLS to OLS as one criterion to detect weak instruments, here we consider the bias, the s.d. and the rrmse defined in (2.29) instead, for the following reasons. For the IV probit model (2.28), the CUE (and other commonly adopted estimation methods) does not have a closed-from expression. Therefore, the usual notion of 'bias towards OLS' under potential IV weakness in linear models is not valid in this nonlinear context, with the potential impact of the IV weakness now being complicated by the nonlinear features of the model. In this case, there is no guarantee that the positive and negative biases will not offset each other, and lead to a spuriously small overall bias. Therefore, to capture the instrument strength and the resulting performance of the CUGMM estimation procedure, we rely on the bias, standard deviation and rrmse of the estimator.

In addition, to better understand weakness in this discrete choice model, we conduct a Wald test of $\mathrm{H}_0 : \alpha = \alpha^0$ and compute its size distortion, relative to the 5% significant level, across all the Monte Carlo designs. We carry out this Wald test for two different estimation methods: the CUE considered in this chapter and the 2SCML estimator proposed by Rivers and Vuong (1988). The size distortion of the Wald test is widely used to capture instrument weakness, see e.g. Staiger and Stock (1997) and Stock and Yogo (2005). This measure not only reflects the performance of the hypothesis test, but also the coverage rate of confidence intervals associated with the two estimation methods.

Under the null hypothesis of $\lambda = 0.5$, the performance of the CUE and the rejection probabilities for the different testing procedures are collected in Table 2.4.1 ($\rho = 0.5$) and Table 2.4.2 ($\rho = 0.95$). For brevity, we only report the estimation results for the structural parameter of interest, α, and Wald test size distortions under five designs: $(\sigma_z, \sigma_v) \in \{(1, 0.2), (1, 10), (1, 1), (0.2, 1), (10, 1)\}$. Additional results for all designs can be obtained from the authors. Figures 2.4.1 and 2.4.2 display the empirical distribution of the 1000 CUEs of $\hat{\alpha}$.

Table 2.4.1: Estimation and Rejection Rates under $\lambda = 0.5$ (Significant Level 5%, $\rho = 0.50$)

		$\sigma_z = 1$ $\sigma_v = 0.2$	$\sigma_z = 1$ $\sigma_v = 10$	$\sigma_z = 1$ $\sigma_v = 1$	$\sigma_z = 0.2$ $\sigma_v = 1$	$\sigma_z = 10$ $\sigma_v = 1$
	bias	0.690	-0.045	-0.050	-0.058	-0.048
	s.d.	4.982	0.627	1.307	1.455	1.501
	rrmse	5.027	0.628	1.308	1.456	1.501
	Wald size distortion (2SCML)	-0.003	-0.004	-0.003	-0.004	0.000
	Wald size distortion (CUGMM)	-0.026	-0.036	-0.037	-0.031	-0.031
n=500	SS	0.061	0.056	0.061	0.060	0.063
	SY (5%)	0.007	0.005	0.009	0.008	0.004
	SY (10%)	0.091	0.085	0.090	0.076	0.080
	Robust (5%)	0.000	0.000	0.000	0.000	0.000
	Robust (10%)	0.000	0.000	0.000	0.001	0.000
	DJ	0.018	0.022	0.016	0.010	0.017
	bias	0.550	-0.128	0.002	-0.076	0.124
	s.d.	4.526	0.301	1.078	1.091	1.260
	rrmse	4.557	0.327	1.078	1.093	1.266
	Wald size distortion (2SCML)	-0.005	-0.023	-0.009	-0.016	0.017
	Wald size distortion (CUGMM)	-0.030	-0.047	-0.033	-0.040	-0.023
n=5000	SS	0.099	0.069	0.057	0.070	0.085
	SY (5%)	0.015	0.008	0.009	0.010	0.003
	SY (10%)	0.132	0.088	0.095	0.091	0.119
	Robust (5%)	0.000	0.000	0.000	0.000	0.000
	Robust (10%)	0.001	0.002	0.001	0.001	0.000
	DJ	0.013	0.025	0.012	0.013	0.022
	bias	0.581	-0.103	0.046	-0.002	0.130
	s.d.	4.354	0.266	1.050	0.993	1.191
	rrmse	4.391	0.285	1.051	0.992	1.197
	Wald size distortion (2SCML)	0.007	-0.016	0.001	-0.004	0.022
	Wald size distortion (CUGMM)	-0.026	-0.047	-0.030	-0.032	-0.026
n=10000	SS	0.130	0.072	0.091	0.088	0.103
	SY (5%)	0.023	0.012	0.016	0.006	0.008
	SY (10%)	0.174	0.098	0.129	0.116	0.151
	Robust (5%)	0.000	0.000	0.000	0.000	0.000
	Robust (10%)	0.001	0.000	0.001	0.000	0.001
	DJ	0.013	0.019	0.019	0.010	0.018

Note: (a) SS rejects the null if $F_n > 10$. SY (5%) and SY (10%) reject the null if the Cragg-Donald statistic is larger than the critical value of a maximal 5% and 10% size distortion of a 5% Wald test, respectively.
(b) For the Robust (5%) and Robust (10%) tests, reject rates are computed based on critical values in Table 1 of Montiel Olea and Pflueger (2013), corresponding to the effective degree of freedom one and tolerance thresholds 5% and 10%, respectively, where the tolerance is the fraction that the Nagar bias relative to the benchmark.
(c) The reject rates of DJ test are computed based on perturbation $\hat{\rho}/\log\{\log(n)\}$ and critical value $\chi^2_{0.95}(2) = 5.99$.

Table 2.4.2: Estimation and Rejection Rates under $\lambda = 0.5$ (Significant Level 5%, $\rho = 0.95$)

		$\sigma_z = 1$ $\sigma_v = 0.2$	$\sigma_z = 1$ $\sigma_v = 10$	$\sigma_z = 1$ $\sigma_v = 1$	$\sigma_z = 0.2$ $\sigma_v = 1$	$\sigma_z = 10$ $\sigma_v = 1$
	bias	2.422	-0.023	-0.117	-0.045	0.008
	s.d.	10.316	0.758	2.866	3.145	2.883
	rrmse	10.591	0.758	2.867	3.144	2.881
	Wald size distortion (2SCML)	0.168	0.003	0.110	0.128	0.126
	Wald size distortion (CUGMM)	0.110	-0.022	0.073	0.074	0.090
n=500	SS	0.072	0.049	0.053	0.062	0.061
	SY (5%)	0.006	0.004	0.004	0.010	0.007
	SY (10%)	0.105	0.066	0.076	0.088	0.088
	Robust (5%)	0.000	0.000	0.000	0.000	0.000
	Robust (10%)	0.000	0.000	0.000	0.000	0.000
	DJ	0.040	0.044	0.039	0.042	0.047
	bias	3.480	-0.072	0.304	0.170	0.405
	s.d.	8.506	0.444	2.232	2.119	2.259
	rrmse	9.187	0.449	2.251	2.124	2.294
	Wald size distortion (2SCML)	0.236	0.014	0.151	0.121	0.156
	Wald size distortion (CUGMM)	0.158	-0.012	0.091	0.076	0.102
n=5000	SS	0.113	0.050	0.076	0.063	0.087
	SY (5%)	0.013	0.005	0.012	0.009	0.007
	SY (10%)	0.158	0.068	0.099	0.085	0.120
	Robust (5%)	0.000	0.000	0.000	0.000	0.000
	Robust (10%)	0.001	0.000	0.000	0.000	0.000
	DJ	0.034	0.017	0.026	0.034	0.031
	bias	3.329	-0.073	0.533	0.472	0.677
	s.d.	8.826	0.459	2.167	1.915	1.988
	rrmse	9.429	0.465	2.230	1.971	2.099
	Wald size distortion (2SCML)	0.271	0.019	0.164	0.140	0.177
	Wald size distortion (CUGMM)	0.171	-0.008	0.112	0.094	0.122
n=10000	SS	0.138	0.047	0.079	0.077	0.090
	SY (5%)	0.016	0.004	0.006	0.008	0.008
	SY (10%)	0.185	0.074	0.106	0.102	0.116
	Robust (5%)	0.000	0.000	0.000	0.000	0.000
	Robust (10%)	0.000	0.000	0.000	0.000	0.001
	DJ	0.027	0.013	0.031	0.021	0.031

Note: (a) SS rejects the null if $F_n > 10$. SY (5%) and SY (10%) reject the null if the Cragg-Donald statistic is larger than the critical value of a maximal 5% and 10% size distortion of a 5% Wald test, respectively.
(b) For the Robust (5%) and Robust (10%) tests, reject rates are computed based on critical values in Table 1 of Montiel Olea and Pflueger (2013), corresponding to the effective degree of freedom one and tolerance thresholds 5% and 10%, respectively, where the tolerance is the fraction that the Nagar bias relative to the benchmark.
(c) The reject rates of DJ test are computed based on perturbation $\hat{\rho}/\log\{\log(n)\}$ and critical value $\chi^2_{0.95}(2) = 5.99$.

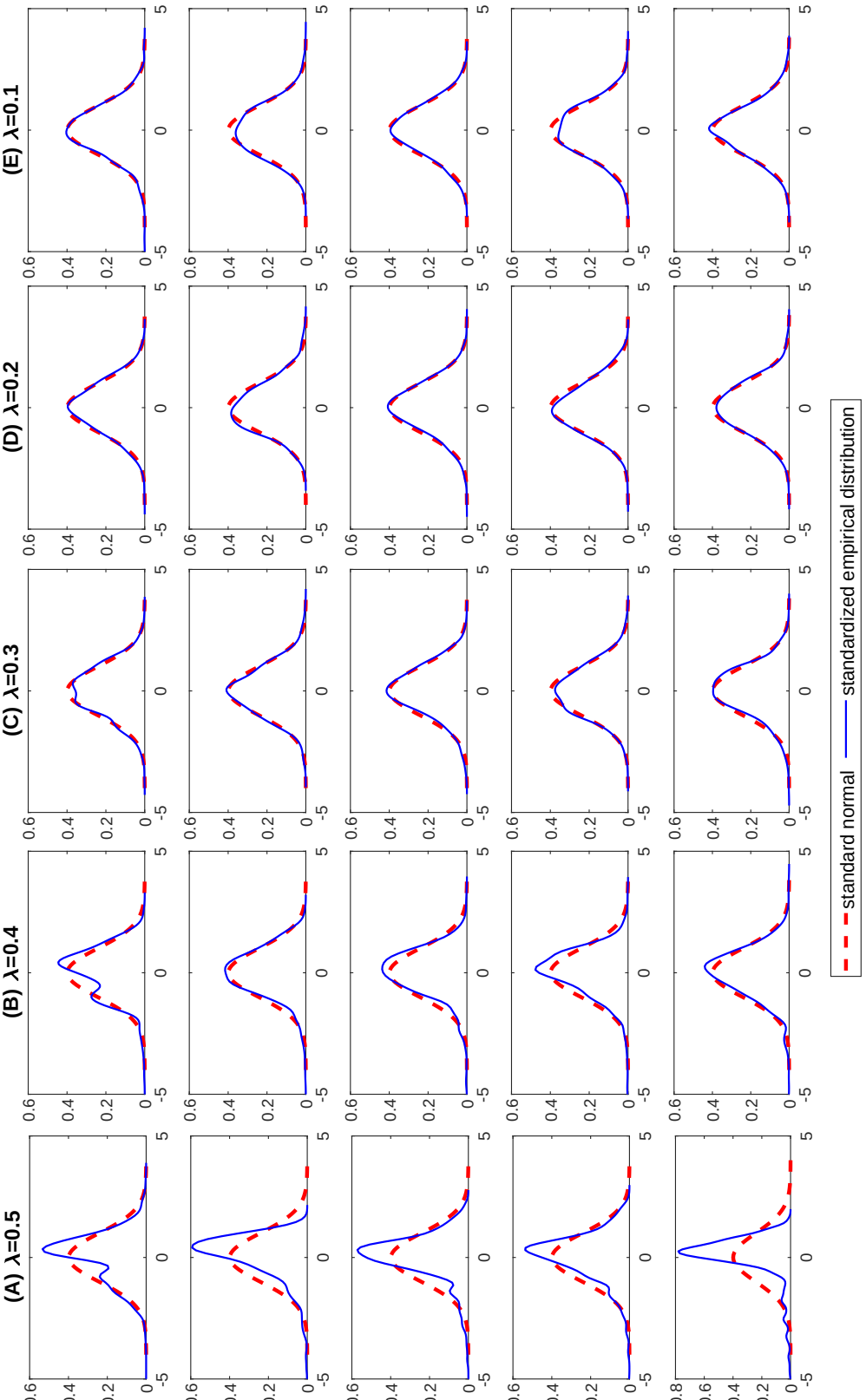
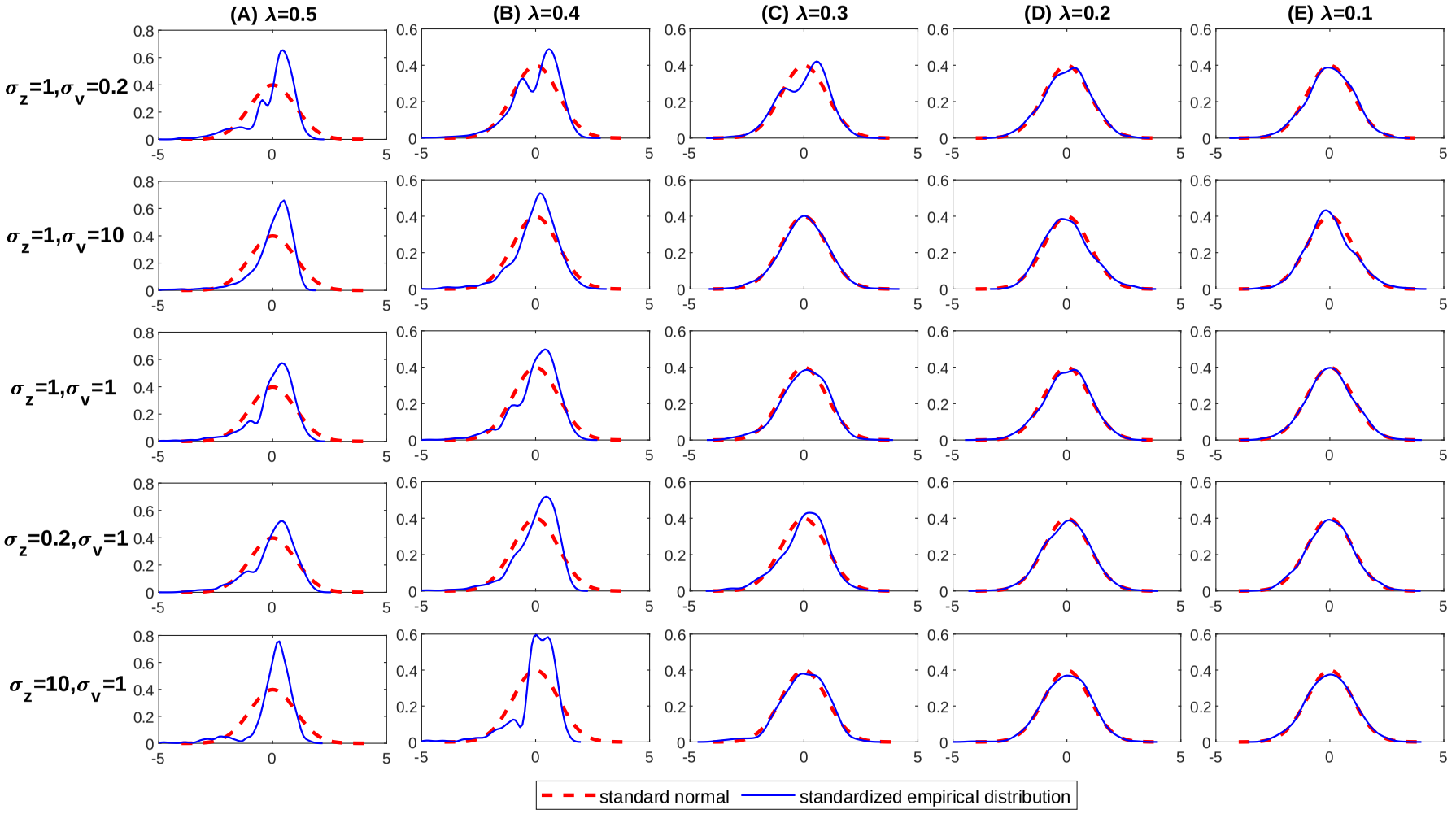


Figure 2.4.1: Kernel Density of Standardised CUE for α ($n = 10000, \rho = 0.50$)

Note: The standardised CUE for α is $(\hat{\alpha} - \bar{\hat{\alpha}})/s.d(\hat{\alpha})$, where $\bar{\hat{\alpha}} = 1/N \sum_{l=1}^{N} \hat{\alpha}_l$, $\hat{\alpha}_l$ stands for the l-th Monte Carlo CUGMM estimates, and $s.d(\hat{\alpha})$ is the standard deviation defined in (2.29).

Figure 2.4.2: Kernel Density of Standardised CUE for α ($n = 10000, \rho = 0.95$)



Note: The standardised CUE for α is $(\hat{\alpha} - \bar{\hat{\alpha}})/s.d(\hat{\alpha})$, where $\bar{\hat{\alpha}} = 1/N \sum_{l=1}^{N} \hat{\alpha}_l$, $\hat{\alpha}_l$ stands for the l-th Monte Carlo CUGMM estimates, and $s.d(\hat{\alpha})$ is the standard deviation defined in (2.29).

Simulation results in Tables 2.4.1 and 2.4.2 confirm our asymptotic results. When $\lambda = 0.5$, CUGMM estimation of α^0 is inconsistent and behaves poorly in general. More specifically, the biases are unstable, and the s.d. and rrmse do not decrease (in any noticeable way) as the sample size increases, especially when the endogeneity degree is high ($\rho = 0.95$). However, under the alternative, $\lambda < 0.5$, the s.d. and rrmse drop dramatically as n increases. In addition, the asymptotic normality of the CUE under $\lambda < 0.5$ is verified by viewing the standardised sampling distributions of the estimators across the Monte Carlo replications, which is given in Figures 2.4.1 and 2.4.2. The sampling distributions exhibit easily detectable bi-modality when λ is 0.5, or close to 0.5, especially when σ_v is small and ρ is large, indicating that a standard inference approach, relying on the normal approximation, is likely to perform poorly in those cases.

The results in Tables 2.4.1 and 2.4.2 also show that the behaviour of the Wald test varies across the different designs even when $\lambda = 0.5$. For a moderate level of endogeneity ($\rho = 0.5$), we see relative small size distortions, less than 5%, in most cases for the Wald tests based on both 2SCML and CUEs. However, for a high degree of endogeneity ($\rho = 0.95$), the Wald tests are significantly over-sized, with the size distortions for the Wald test based on 2SCML being much larger than those based on CUGMM. One exception, however, is the case of $(\sigma_z, \sigma_v) = (1, 10)$ and $\rho = 0.95$, where the Wald size distortions based on both estimation methods is less than 5%. For $(\sigma_z, \sigma_v) = (1, 10)$ and $\rho = 0.95$ case, the size distortion based on the CUGMM is 0.008 when $n = 10000$, indicating that the 95% confidence interval coverage rate is quite accurate even though $\lambda = 0.5$ ($\text{corr}(y_{2i}, z_i) = 0.015$). As such, this design constitutes additional evidence that the value of λ is not the only key in determining inference performance in weakly identified discrete choice models.

The false rejection rates of SS, SY, Robust and DJ under $\lambda = 0.5$ are displayed in Tables 2.4.1 and 2.4.2. Firstly, as expected, the DJ test is asymptotically conservative, i.e., the size is less than the significance level of 5%. The size of the DJ test varies between 1.0% and 1.9% under $\rho = 0.5$, and between 1.3% and 3.1% when $\rho = 0.95$. However, we note that the DJ test is much less conservative than the Robust approach of Montiel Olea and

Pflueger (2013), which is extremely conservative, and gives virtually zero rejections across all designs where identification is weak. Therefore, while the DJ test is conservative, we can conclude that it is much less conservative than the Robust approach, and can be relatively close to the nominal level (5%) when the degree of endogeneity is large.

In addition, we see that blindly applying conventional weak instrument tests can lead to poor outcomes. For example, for the design with $(\sigma_z, \sigma_v) = (1, 0.2)$ and a high level of endogeneity ($\rho = 0.95$ in Table 2.4.2), the rejection rates of SS and SY (10%) are all larger than 10% across different sample sizes, and are 13.8% and 18.5% respectively when $n = 10000$. The rejection rate of SY (10%) is computed based on the critical value of a maximal 10% size distortion of a 5% Wald test, provided by Stock and Yogo (2005). However, the rrmse in this case does not decrease as n increases, and the rrmese for the estimated α is between 910% and 1060% of the true value. Moreover, both the Wald size distortions exceed their nominal size by at least 10%. In particular, the 2SCML size distortion is between 17% and 27%, while the CUGMM size distortion is between 11% to 17%. Therefore, the identification is weak, while the SS and SY approaches can suggest the opposite, and hence fail to control size. In addition, false rejection rates for other designs, not reported here for brevity, demonstrate a similar pattern of over-rejection for SS and SY tests. Hence, in line with the analysis in Section 2.3.6, when assessing identification strength in discrete choice models, the conventional weak IV tests of SS and SY may fail to provide reliable conclusions regarding identification strength, especially if the degree of endogeneity is high.

Figure 2.4.3 ($\rho = 0.5$) and Figure 2.4.4 ($\rho = 0.95$) display the power of the four tests. Due to the conservativeness of DJ test, size adjusted power of DJ and of the conventional tests are also computed and compared in Figures 2.4.5 and 2.4.6. Size adjusted power is computed as follows: obtain the 95% quantile of the test statistic from the 1000 Monte Carlo replications when $\lambda = 0.5$ and use it as the critical value for cases when $\lambda < 0.5$. The resulting power curves show that the DJ test is consistent as the sample size diverges, and as identification strength increases. Moreover, in cases with high endogeneity (Figure 2.4.4), the unadjusted power of the DJ is higher than that of the Robust test across most designs. Furthermore, Figures 2.4.5 and 2.4.6 demonstrate that

the DJ-test displays non-negligible power even when identification is close to being weak, i.e, when $\lambda = 0.4$ or $\lambda = 0.3$, which gives convincing numerical evidence of the results in Theorem 2.3.6.

2.5. Empirical Application

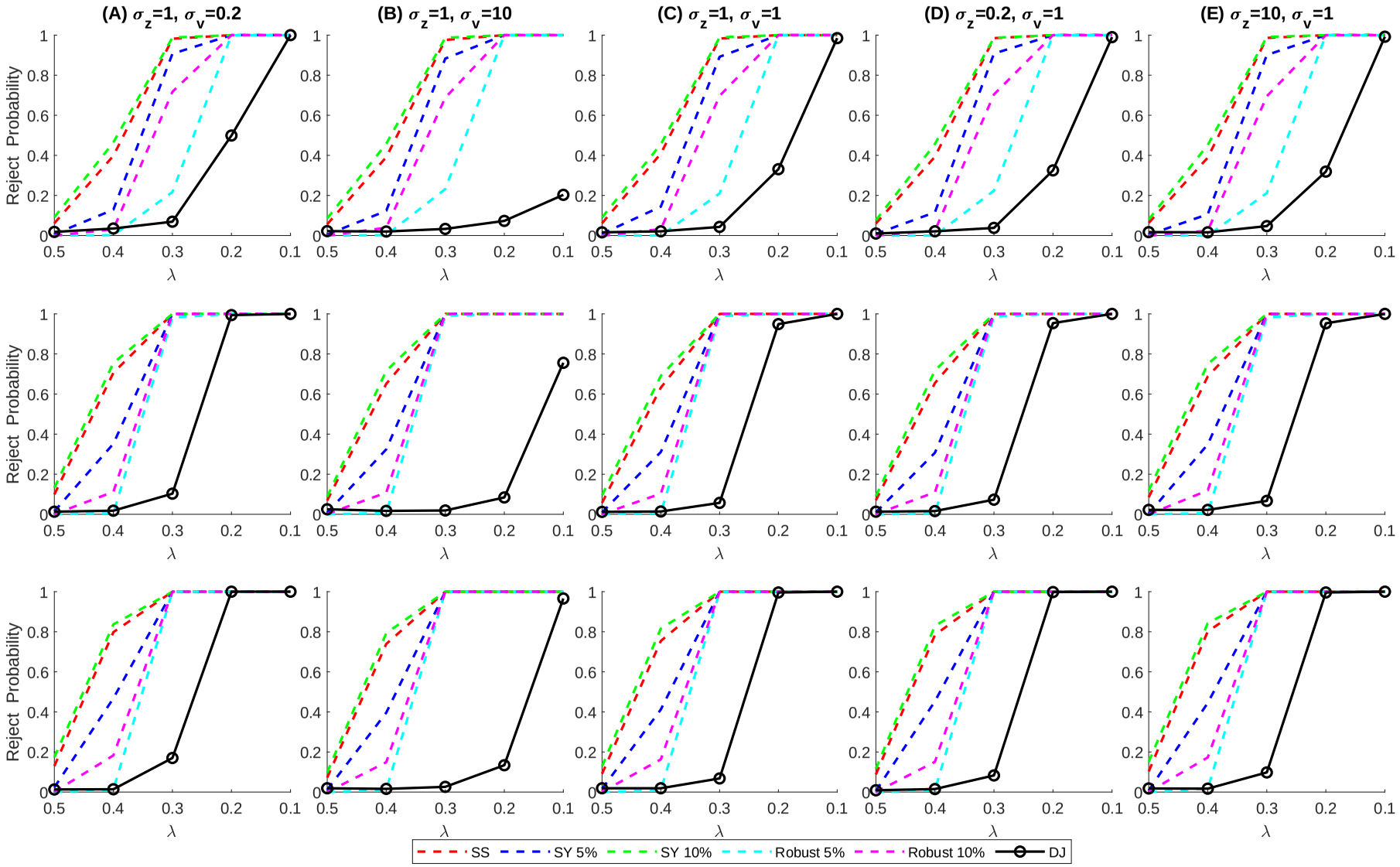
In this section, we apply our distorted J-test in two well-known empirical examples to test for the presence of weak instruments. We then contrast the results of our tests with those obtained from conventional weak IV tests for linear models, namely the SS, the SY, and the Robust tests.

2.5.1. Labour Force Participation of Married Women

We first study married women's labour force participation (hereafter LFP) when education, measured as the women's years of schooling, is treated as an endogenous regressor. We use data from the University of Michigan Panel Study of Income Dynamics (PSID) for the year 1975, which have been used in several studies. The data is public and available at Wooldridge (2010) Supplemental Content. Mroz (1987) provides an extensive analysis of the women's hours of labour supply, and considers a range of specifications including potential endogeneity of several regressors, the use of different instrumental variables and controls for self-selection into labour force participation. As a text book example, Wooldridge (2010) used the same dataset to study women's LFP decisions, and the potential endogeneity of education is tested after estimating an IV probit model using Rivers and Vuong (1988) two-step conditional maximum likelihood estimator (2SCML). In what follows we use similar specification as in Wooldridge (2010).

The PSID consists of data on 753 married, Caucasian women who are between 30 and 60 years of age at the time the sample was conducted. The dependent variable LFP is a binary response that equals unity if the respondent worked at some time during the year, and zero otherwise. Exogenous regressors include spousal income, the individual's work experience and its square, age, the number of children less than six years old, and the number of children older than six years old. The individual's education, measured as years of schooling, is considered to be endogenous. Following the strategy in Wooldridge

Figure 2.4.3: Rejection Rates under $\lambda < 0.5$ ($\rho = 0.50$)



Note: x-axis is IV strength λ. First row $n = 500$, second row $n = 5000$, third row $n = 10000$. The reject rates are computed using critical value $\chi^2_{0.95}(2) = 5.99$.

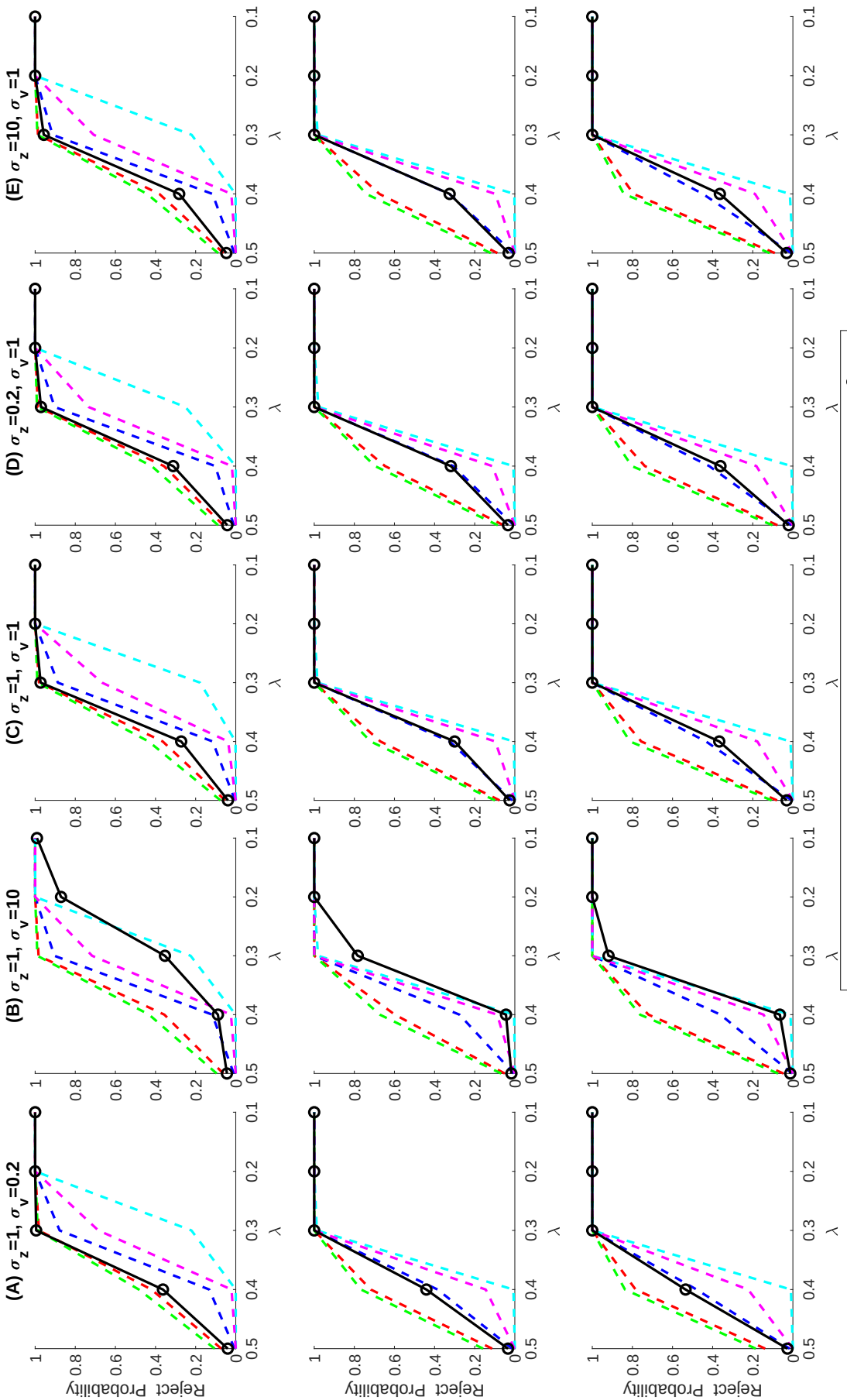
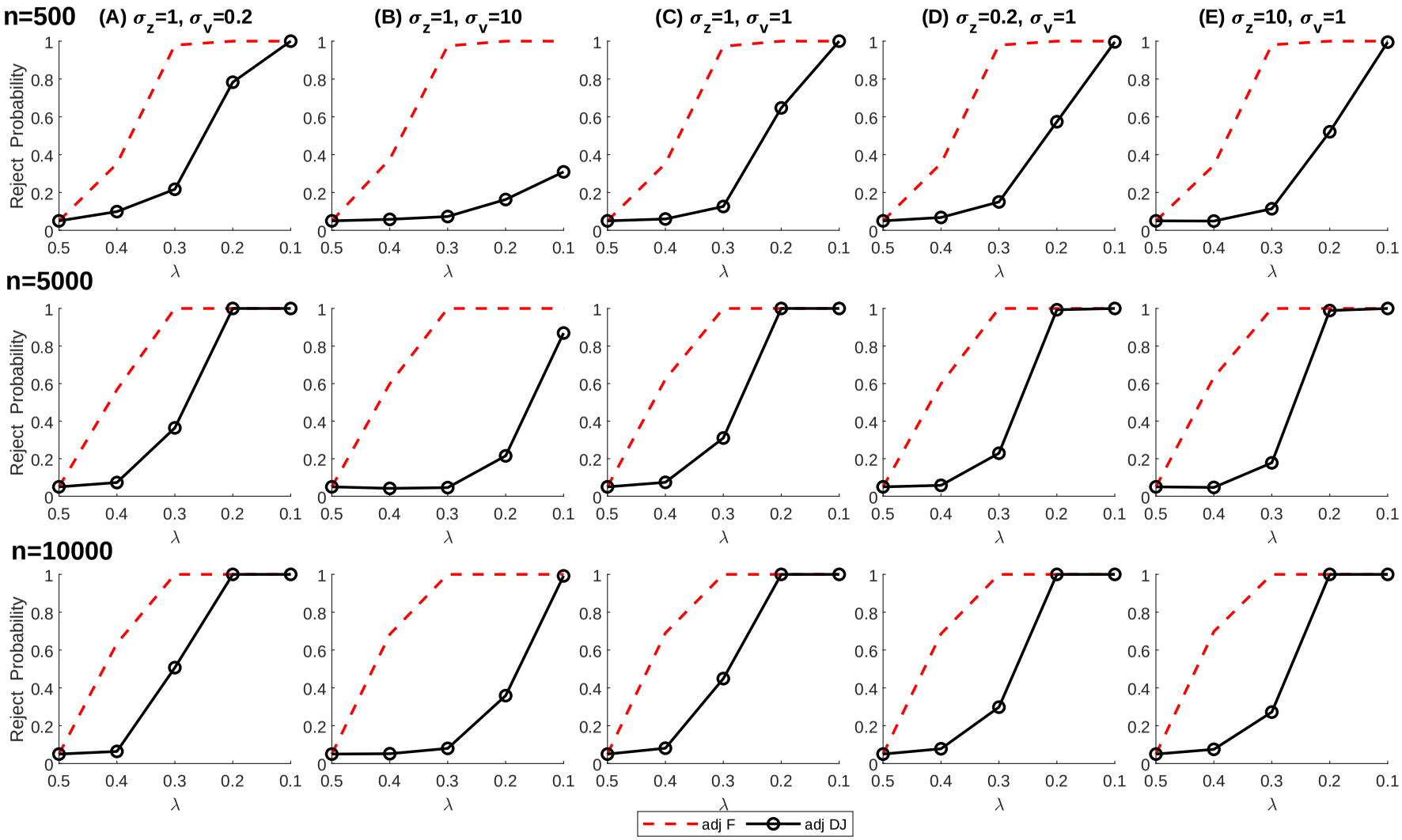


Figure 2.4.4: Rejection Rates under $\lambda < 0.5$ ($\rho = 0.95$)

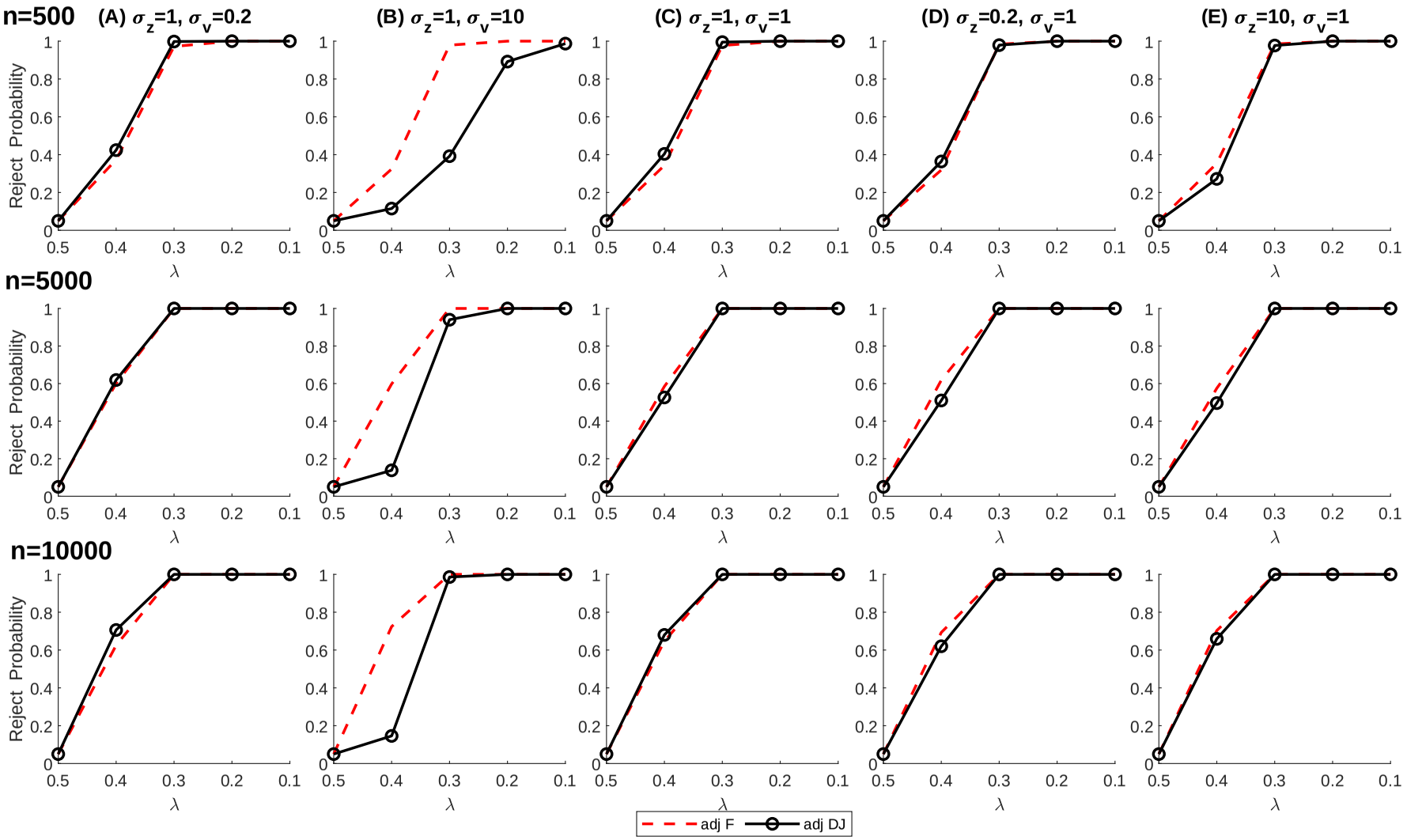
Note: x-axis is IV strength λ. First row $n = 500$, second row $n = 5000$, third row $n = 10000$. The reject rates are computed using critical value $\chi^2_{0.95}(2) = 5.99$.

Figure 2.4.5: Size Adjusted Rejection Rates under $\lambda < 0.5$ ($\rho = 0.50$)



Note: x-axis is IV strength λ. First row $n = 500$, second row $n = 5000$, third row $n = 10000$. The test statistic of SS, SY and Robust under one endogenous regressor, one instrument and homoscedastic errors, are the same, i.e. the reduced form regression F-statistic. The size adjusted power curve is therefore the same for SS, SY and Robust.

Figure 2.4.6: Size Adjusted Rejection Rates under $\lambda < 0.5$ ($\rho = 0.95$)



Note: x-axis is IV strength λ. First row $n = 500$, second row $n = 5000$, third row $n = 10000$. The test statistic of SS, SY and Robust under one endogenous regressor, one instrument and homoscedastic errors, are the same, i.e. the reduced form regression F-statistic. The size adjusted power curve is therefore the same for SS, SY and Robust.

(2010), the individual's family education, which are recorded as the years of schooling for both the individual's father and mother, are used as instruments for education.

Table 2.5.1: Data Summary of Married Women LFP (Obs. 753)

	Mean	Std. Dev.	Min	Max
LFP	0.57	0.50	0	1
Education	12.29	2.28	5	17
Father educ.	8.81	3.57	0	17
Mother educ.	9.25	3.37	0	17
Experience	10.63	8.07	0	45
Exper. square	178.04	249.63	0	2025
Nonwife income ($1000)	20.13	11.64	-0.029	96
Age	42.54	8.07	30	60
# Kids $<$ 6 years old	0.24	0.52	0	3
# Kids $>$ 6 years old	1.35	1.32	0	8

Note: Education, father/mother education and experience are measured in years.

Estimated coefficients and the average partial effects on the probability of LFP for all regressors are presented in Table 2.5.2 using two estimation methods: 2SCML as used in Wooldridge (2010) and CUGMM. More specifically, for the 2SCML, the first step is to regress the endogenous regressor on the instruments and all other exogenous regressors to obtain the reduced form residual. The second step is to run a probit maximum likelihood estimation of the binary response on the endogenous and the exogenous regressors, and the reduced form residual. The CUGMM estimation with over-identification degree one is conducted using $a_i = (1, y_{2i}, x_i', z_i', \mathbf{0}_{k+2}')'$ and $b_i = (\mathbf{0}_{k+3}', 1, x_i', z_i')'$, where y_{2i}, x_i and z_i denote the standardised variables corresponding to the women's education, exogenous regressors and two instruments, and k is the number of exogenous regressors and the intercept. The first step estimation of the 2SCML and the reduced form of the CUGMM are listed in the first and fourth columns of Table 2.5.2 respectively. Both the two IVs are highly significant based on both estimation methods. The CUGMM estimation results are reported in columns four through six. Broadly speaking, the CUGMM and 2SCML results are similar, with both methods providing evidence that education has a significant positive effect: one extra year of education increasing the probability of LFP by 5.87 percentage points for both the 2SCML and the CUGMM. Hansen's J-statistic is 0.122 which is less than $\chi^2_{0.95}(1) = 3.84$, therefore we fail to reject the null that all the

moments are valid.

Table 2.5.2: Regression Results of Labour Force Participation (LFP)

	2SCML Probit			CUGMM		
	1st step	2nd step	margin	reduced form	structural eq.	margin
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Var.	Education	LFP		Education	LFP	
Education		0.1503***	0.0568***		0.1500***	0.0569***
		(0.0539)	(0.0210)		(0.0538)	(0.0206)
Experience	0.0930***	0.1213***	0.0310***	0.0929***	0.1208***	0.0309***
	(0.0251)	(0.0194)	(0.0033)	(0.0249)	(0.0195)	(0.0033)
Exper. square	-0.0016*	-0.0018***		-0.0016*	-0.0018***	
	(0.0009)	(0.0006)		(0.0009)	(0.0006)	
Nonwife income ($1000)	0.0452***	-0.0132**	-0.0050**	0.0453***	-0.0139**	-0.0053**
	(0.0071)	(0.0061)	(0.0022)	(0.0070)	(0.0061)	(0.0023)
Age	-0.0217**	-0.0518***	-0.0196***	-0.0218**	-0.0514***	-0.0195***
	(0.0109)	(0.0087)	(0.0034)	(0.0109)	(0.0088)	(0.0034)
# Kids <6 years old	0.2268	-0.8733***	-0.3303***	0.2268	-0.8727***	-0.3307***
	(0.1570)	(0.1176)	(0.0456)	(0.1561)	(0.1210)	(0.0469)
# Kids >6 years old	-0.0934*	0.0395	0.0149	-0.0933*	0.0396	0.0150
	(0.0554)	(0.0459)	(0.0168)	(0.0551)	(0.0475)	(0.0180)
Father educ.	0.1552***			0.1551***		
	(0.0237)			(0.0236)		
Mother educ.	0.1721***			0.1724***		
	(0.0252)			(0.0250)		
Correlation ρ		-0.0453			-0.0453	
		(0.1105)			(0.1102)	
J-statistic	–	–	–	–	0.122	
Obs.	753	753	753	753	753	753

Note: [a] Standard errors (s.e.) in parentheses. Significance *** p<0.01, ** p<0.05, * p<0.1. The s.e. in columns (1)-(3) are heteroscedastic-robust. The s.e. in columns (4)-(6) are computed based on Theorem 2.3.5. According to Antoine and Renault (2020), when DJ rejects the null, standard inference procedures still work for all practical purpose.
[b] For CUGMM estimation, overidentification degree is one. Hansen's J-statistic 0.122 is less than $\chi^2_{0.95}(1) = 3.84$. Overidentification test fails to reject the null hypothesis that moments are all valid.
[c] Correlation ρ is the correlation of errors (u_i, v_i) in structural equation and reduced form.
[d] Margins in columns (3) and (6) are computed using the sample average of explanatory variables and IVs.

The weak IV test results are collected in Table 2.5.3 for all four tests, SS, SY, Robust and DJ. The Kleibergen-Paap F-statistic (Kleibergen and Paap, 2006) is 81.89, based on which the SS rule-of-thumb and the SY test both reject the null that IVs are weak.[4]

[4]The Kleibergen-Paap F-statistic is utilised when allowing for heteroscedastic standard error. The reduced form regression F-statistic and the Cragg-Donald statistic are 95.70, when assuming homoscedastic standard error. SY rejects its null according to the critical values of the maximal desired size distortion 5% and 10% of a 5% Wald test.

For the Robust test, the effective F-statistic is 91.44, the critical values for the tolerance thresholds $\{5\%, 10\%\}$ are 11.59 and 8.58, respectively.[5] Comparing the effective F-statistic 91.44 to the critical values, the Robust test also rejects the null of weak IV.

Table 2.5.3: Tests of Weak Instruments (Significance level 5%)

	SS	SY (5%)	SY (10%)	Robust (5%)	Robust(10%)	DJ (min & max)
Statistic	81.89	81.89	81.89	91.44	91.44	0.14 & 17.44
Critical value	10	19.93	11.59	8.58	6.17	11.98
Reject H_0	Reject	Reject	Reject	Reject	Reject	Reject

Note: (a) SS and SY test statistics 81.89 are Kleibergen-Paap F-statistic, which is heteroscedastic-robust. When assuming homoscedastic standard error, the reduced form F-statistic and the Cragg-Donald F-statistic is 95.70. SS critical value 10 is the rule-of-thumb. SY (5%) and SY (10%) critical values 19.93 and 11.59 are for i,i.d. errors, the maximal desired size distortions 5% and 10% of a 5% Wald test, respectively.

(b) Robust test statistics and critical values are computed using Stata command "weakivtest" (Pflueger and Wang, 2015) based on heteroscedastic-robust s.e. Robust (5%) and Robust (10%) critical values 8.58 and 6.17 are for 2SLS with 5% and 10% tolerance of the Nagar bias over benchmark, respectively. The estimated effective degrees of freedom with the tolerance {5%,10%} are 1.82 and 1.84.

(c) The perturbation of DJ test is chosen using the approach in Section 2.3.3. The critical value is $\chi^2_{1-0.05/20}(H-p+1) = 11.98$.

Finally, for the DJ test, the perturbation δ_n is computed as in Section 2.3.5, using $m = 20$ candidate grid points. This choice of m leads us to use the critical value $\chi^2_{1-0.05/20}(H+1-p) = 11.98$, where we note that we have used $H = 19$ moments and estimated $p = 18$ parameters. Of the candidate grid points, three lead to a value of the DJ statistic larger than 11.98, leading us to soundly reject the null of weak identification. The rejection conclusion of the DJ test is quite straightforward: when perturbing the CUE $\hat{\theta}_n$ by δ_n, the value of the J-statistic increases dramatically from 0.122 to a maximum of 17.44, implying that the CUGMM criterion is sensitive to even small departures. Overall, results reported in Table 2.5.3 suggest that the DJ test and the three conventional tests for the linear model agree in this example.

[5] The estimated effective degrees of freedom of the Robust test for the tolerance thresholds $\{5\%, 10\%\}$ are both about 1.8. See Montiel Olea and Pflueger (2013) for the definitions of the effective F-statistic, the tolerance threshold and the effective degrees of freedom. The Robust test statistic and the critical values are obtained using the Stata command "weakivtest" (Pflueger and Wang, 2015) under heteroscedastic-robust estimation.

2.5.2. US Food Aid and Civil Conflicts

In the second example we examine the impact of US food aid on the incidence of civil conflicts in recipient countries. The research in Nunn and Qian (2014) was motivated by concerns that humanitarian food aid may be ineffective and may even promote civil conflicts. The main challenge of this study is the potential endogeneity of US food aid due to reverse causality and joint determination. Their identification strategy relies on using the product of the lagged US wheat production and the average probability of receiving any US food aid for each country as the instrumental variable for wheat aid. Nunn and Qian (2014) estimate many variations of the basic binary model and consider different kinds of wars, different controls and alternative specifications.

Herein, we focus on the simple cross-sectional specification considered in Nunn and Qian (2014). More specifically, we estimate the impact of US wheat aid on the probability of civil war *onset* after a period of peace (column (3), Table 7, Nunn and Qian (2014)), using precisely the same model specification as in Nunn and Qian (2014).[6] We examine the IV strength by applying our DJ test to the model, as well as the three conventional weak IV tests for linear models.

The dataset in this analysis involves observations on 78 non-OECD countries from 1971 to 2006, and the observations used for the onset analysis are those country-year observations that have no intra-state civil conflict in the previous period. The event indicator for civil war onset is set to be one if it is the first period of a intra-state conflict episode, and zero otherwise. Nunn and Qian (2014) estimate a logistic discrete time hazard model for the probability of onset of war, controlling for the previous duration of peace up to the third degree of polynomial. The US wheat aid in year t is instrumented by the product of US wheat production in year $t-1$ and the probability of receiving any US food aid between 1971 and 2006 for each country. To be consistent with the setup of the chapter, we estimate a probit link model rather than a logit. Summary statistics for the data used in the onset analysis can be found in part (a) of Table 2.5.4.

[6]Data sets used to construct the incidence of conflict, US food aid, US wheat production and other variables include the UCDP/PRIO Armed Conflict Dataset Version 4-2010, the Food and Agriculture Organization's FAOSTAT database and the data from the United States Department of Agriculture. See Nunn and Qian (2014) for more detailed information.

Table 2.5.4: Data Summary of US Food Aid and Civil Conflict

(a) Civil Conflict Onset (obs. 1454)

	Mean	Std. Dev.	Min	Max
Onset of intra-state conflict	0.063	0.244	0	1
US wheat aid (1000 metric tons)	21.08	59.42	0	791.60
Lagged US wheat production (1000 metric tons)	59187	8754	36787	75813
Average US food aid probability 1971-2006	0.387	0.328	0	1
Peace duration (years)	11.59	9.48	1	46
Instrument	22936	19924	0	75813

(b) Civil Conflict Offset (obs. 709)

	Mean	Std. Dev.	Min	Max
Offset of intra-state conflict	0.185	0.388	0	1
US wheat aid (1000 metric tons)	56.07	123.58	0	854.7
Lagged US wheat production (1000 metric tons)	60374	8626	36787	75813
Average US food aid probability 1971-2006	0.503	0.313	0	1
Conflict duration (years)	8.70	8.45	1	42
Instrument	30413	19676	0	75813

Note: An observation is a country and year. Instrument is lag of US wheat production times average probability of receiving any US food aid during 1971 to 2006.

Part (a) of Table 2.5.5 presents results for the estimated coefficients and average partial effects from both 2SCML probit and CUGMM with the degree of overidentification equal to unity. The 2SCML in this example allows intragroup correlation for standard errors, clustered by countries. For comparison purposes, column (1) of Table 2.5.5 gives the estimated average partial effect of US wheat aid reported by Nunn and Qian (2014) using a 2SCML logit approach, which is a key result for their analysis. For CUGMM, we use $a_i = (x_i', z_i, z_i^2, z_i^3, z_i x_{1i}, \mathbf{0}_{k+1}')'$ and $b_i = (\mathbf{0}_{k+3}', 1, z_i, x_i')'$ to construct moments. The variables x_i and z_i denote the standardised variables of exogenous regressors and the instrument, x_{1i} is the non-standardised onset duration, and k is the number of exogenous regressors and the intercept. Columns (2) and (5) of Table 2.5.5 demonstrate that the IV is significantly related to the endogenous regressor of wheat aid at the 1% significant level by both estimation methods. However, it is worth noting that the estimates of interest, the effects of the US wheat aid on onset, based on the three different estimation approaches, are quite unstable and even differ in signs. The statistical insignificance of the US food aid on onset of civil conflict is pointed out by Nunn and Qian (2014). However, without reliable evidence of the instrument strength, we should be cautious

of drawing any inference conclusions according to the standard inference procedures. Estimates for other coefficients are quite stable and similar across the three sets of results. Finally, Hansen's J-statistic is 0.553, less than the critical value $\chi^2_{0.95}(1) = 3.84$, thus we cannot reject the null that moments are all valid. This evidence leads to the suspicion that the potential weakness of the IV could be one of the possible reasons for the unstable estimates of the US wheat aid coefficient.

Table 2.5.5: Regression Results of US Food Aid and Civil Conflict

(a) Civil Conflict Onset

	Nunn & Qian (2014)	2SCML Probit			CU-GMM		
	margin	1st step	2nd step	margin	reduced form	structural eq.	margin
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Var.	Onset	Wheat aid	Onset	Onset	Wheat aid	Onset	Onset
Wheat aid	0.000036		0.0011	0.000067		-0.0013	-0.000071
	(0.00015)		(0.0025)	(0.00016)		(0.0028)	(0.00027)
Peace dur.	-0.00164***	-1.66	-0.18***	-0.0018***	-1.66	-0.1815***	-0.0017*
	(0.00066)	(1.18)	(0.041)	(0.00072)	(1.21)	(0.045)	(0.0010)
Peace dur.^2		0.053	0.0087***		0.053	0.0085***	
		(0.066)	(0.0026)		(0.072)	(0.0031)	
Peace dur.^3		-0.00042	-0.00012***		-0.00042	-0.00011	
		(0.0011)	(0.00005)		(0.0012)	(0.00014)	
Instrument		0.0012***			0.0012***		
		(0.0002)			(0.0002)		
Correlation ρ			-0.0837			0.3109**	
			(0.1318)			(0.1408)	
J-statistic	–	–	–	–	–	0.553	–
Obs.	1454	1454	1454	1454	1454	1454	1454

(b) Civil Conflict Offset

	Nunn & Qian (2014)	2SCML Probit			CU-GMM		
	margin	1st step	2nd step	margin	reduced form	structural eq.	margin
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Var.	Offset	Wheat aid	Offset	Offset	Wheat aid	Offset	Offset
Wheat aid	-0.000256*		-0.0019*	-0.000284*		-0.0013	-0.000302
	(0.00016)		(0.0011)	(0.00017)		(0.0021)	(0.00029)
Conflict dur.	-0.00801***	4.97	-0.2794***	-0.00904***	4.97	-0.2998***	-0.0133***
	(0.0021)	(4.65)	(0.0525)	(0.00221)	(4.34)	(0.0690)	(0.0053)
Conflict dur.^2		-0.406	0.0164***		-0.406	0.0184**	
		(0.288)	(0.0046)		(0.371)	(0.0084)	
Conflict dur.^3		0.007	-0.0003***		0.007	-0.0003	
		(0.005)	(0.0001)		(0.009)	(0.0003)	
Instrument		0.003***			0.003***		
		(0.0007)			(0.0006)		
Correlation ρ			0.1277			0.1768	
			(0.1238)			(0.1585)	
J-statistic	–	–	–	–	–	1.500	–
Obs.	709	709	709	709	709	709	709

Note: [a] Standard errors (s.e.) in parentheses. Significance *** p<0.01, ** p<0.05, * p<0.1. For both panels (a) (b), the s.e. in column (1) is from Nunn and Qian (2014). The s.e. in columns (2)-(4) are clustered s.e. by countries, based on the 2SCML probit estimation. The s.e. in columns (5)-(7) are calculated by bootstrap with 1000 replications. Since DJ test fails to reject its null, implying standard inference procedures may no longer hold, we should be cautious of drawing any inference conclusions based on those s.e reported in the above tables.

[b] For CU-GMM estimation, overidentification degree is one. Hansen's J-statistics are less than $\chi^2_{0.95}(1) = 3.84$. Overidentification test fails to reject the null hypothesis that moments are all valid in both onset and offset cases.

[c] Correlation ρ is the correlation of errors (u_i, v_i) in structural equation and reduced form.

[d] Margins in columns (4) and (7) are computed based on sample average of explanatory variables and IVs.

This suspicion is verified by the DJ test. The perturbation for the onset analysis is chosen as described in Section 2.3.5, again using $m = 20$ candidate grid points. Panel (a) of Table 2.5.7 demonstrates that the DJ test cannot reject the null of weak instrument. In contrast to the earlier example in Section 2.5.1, in this example perturbing the estimators by δ_n does not lead to a significant change in the corresponding J-statistic, which indicates a lack of curvature and thus identification weakness. Across the entire grid of candidate δ_n values, the maximum of the DJ statistics is 7.5, which is less than the corresponding 5% critical value given by $\chi^2_{1-0.05/20}(H + 1 - p) = 11.98$, and which is based on using $H = 12$ moments to estimate $p = 11$ parameters. However, when we apply the conventional SS, SY and Robust tests to the onset of the civil conflict case, the SS and SY tests return a rejection of the weak IV hypothesis and the Robust test also rejects the null if the tolerance threshold is greater than 10%. As shown in Table 2.5.5, the reduced form regression Kleibergen-Paap F-statistic for SS and SY is 26.07, much larger than 10 and the critical values 16.38 and 8.96 of SY.[7] The Robust test effective F-statistic 26.39 is also larger than its 10% tolerance critical value 23.11.[8] In summary, for this example, the conventional weak IV tests and the DJ test suggest opposite results. This serves as a reminder that applying conventional weak IV tests for linear models to binary outcome models can lead to incorrect decisions in certain circumstances.

[7]To be consistent with Nunn and Qian (2014), standard errors (s.e.) are computed using clustered s.e. by countries. Kleibergen-Paap F-statistic (Kleibergen and Paap, 2006) is utilised when allowing for intragroup correlation s.e. The critical values 16.38 and 8.96 of SY test are based on the desired maximal size distortion 5% and 10% of a 5% Wald test, respectively.

[8]The effective F-statistic and critical values are computed using the Stata command "weakivtest" (Pflueger and Wang, 2015). The critical value 23.11 is for the case of effective degrees of freedom one and the tolerance threshold 10%. Robust test fails to reject the weak instrument based on the critical value of 5% tolerance.

Table 2.5.7: Tests of Weak Instrument (Significance level 5%)

(a) Civil Conflict Onset

	SS	SY (5%)	SY (10%)	Robust (5%)	Robust (10%)	DJ (min & max)
Statistic	26.07	26.07	26.07	26.39	26.39	0.57 & 7.50
Critical value	10	16.38	8.96	37.42	23.11	11.98
Reject H_0	Reject	Reject	Reject	Not Reject	Reject	Not Reject

(b) Civil Conflict Offset

	SS	SY (5%)	SY (10%)	Robust (5%)	Robust (10%)	DJ (min & max)
Statistic	17.29	17.29	17.29	17.49	17.49	1.50 & 9.46
Critical value	10	16.38	8.96	37.42	23.11	11.98
Reject H_0	Reject	Reject	Reject	Not Reject	Not Reject	Not Reject

For both onset and offset data, SS and SY test statistics are Kleibergen-Paap F-statistic (Kleibergen and Paap, 2006) based on clustered s.e. by countries, to be consistent with Nunn and Qian (2014). SS critical value 10 is the rule-of-thumb. SY (5%) and SY (10%) critical values 16.38 and 8.96 are for i.i.d. errors, one endogenous regressor and one IV, desired maximal size distortion 5% and 10% of a 5% Wald test.
Robust test statistics and critical values are computed using Stata command "weakivtest" (Pflueger and Wang, 2015) based on clustered s.e. by countries. For both onset and offset data, Robust (5%) and Robust (10%) critical values 37.42 and 23.11 are for 2SLS with 5% and 10% tolerance of the Nagar bias over benchmark, respectively. The estimated effective degrees of freedom with the tolerance $\{5\%, 10\%\}$ are both 1.
For the offset data, the Robust test rejects weak IV when tolerance is larger than 20%.
The perturbation of DJ test is chosen based on the process in Section 2.3.3. The critical value is $\chi^2_{1-0.05/20}(H - p + 1) = 11.98$.

Subsequently, we have repeated this analysis within the other 5 models considered in Nunn and Qian (2014) (columns (4)-(8) of Table 7, Nunn and Qian 2014), which include different specifications and exogenous regressors. In most of the cases, our DJ test shows that we cannot rule out the possibility of weak instruments, whilst the SS and SY tests all result in a rejection of weak instrument hypothesis. Results are not reported due to space limitation. SS and SY tests are based on Kleibergen-Paap F-statistic (Kleibergen and Paap, 2006). DJ test is implemented using the same a_i and b_i with those used to get the CUGMM in Table 2.5.5. The perturbation for each model is again chosen as in Section 2.3.5 with $m = 20$. The DJ test rejects the null of weak instrument in column (7), but fails to reject in the remaining columns. The Robust test also rejects the null in some cases. Based on the critical value 23.11 ($\tau = 10\%$), the Robust test rejects weak IV of the analysis in columns (4) and (8), but fails to reject in columns (5), (6) and (7). Results are obtained by using the Stata command "weakivtest" (Pflueger and Wang, 2015) and clustered s.e.. In part (b) of Table 2.5.7 and Table 2.5.5, we report the estimation results for the probability of *offset* of civil war after a period of war (column

(6) of Table 7, Nunn and Qian 2014), as well as the test results for weak instrument.

One important result to note is that Nunn and Qian (2014) estimate a significant and negative effect for offset of war, indicating that aid prolongs civil wars with 1,000 MT extra of US wheat aid reducing the probability of civil war offset by 0.04 percentage point. However, in this context, if one applies the DJ test using the same methodology as above, the DJ statistic varies between 1.50 and 9.46, which is again less than the corresponding critical value of 11.98, and indicates that identification may be weak in this example. If identification is indeed weak, as the DJ test suggests, conducting standard inference on the estimated treatment effect is no longer valid, therefore, the conclusion that US food aid prolongs civil conflict must be interpreted with caution.

2.6. Conclusion

Estimating the causal effects of policy relevant treatment variables is the key goal of many empirical analyses in economics and other diverse fields. Instrumental variables play a crucial role in the identification and estimation of treatment effects when the treatment is endogenous, but weak instruments have been identified as a potentially serious problem, with consequences including inconsistent estimation and invalid inference. Consequences and detection of weak identification due to instrument weakness have been extensively studied for linear models, but similar issues have not been thoroughly studied for non-linear models, such as discrete dependent variable models. In search for a suitable weak identification test, empirical researchers have often resorted to the inappropriate use of linear model weak IV tests for discrete outcome models, or the use of a linear probability model with a 2SLS estimator treating the discrete variables as continuous. The suitability of these linear tests in this nonlinear setting is not usually questioned in many empirical studies (see Dufour and Wilde, 2018 and Li et al., 2019 for additional analysis on the performance of the Stock and Yogo, 2005 testing approach in binary models).

This chapter proposes a much needed weak identification test in endogenous discrete choice models. The proposed test has desirable asymptotic properties including size control under the null of weak identification, and consistency under the alternative.

Moreover, we demonstrate that once the null of weak identification is rejected, standard Wald-based inference can be applied as usual. Our Monte Carlo results demonstrate that, whilst the conventional Stock and Yogo (2005) and Staiger and Stock (1997) tests are often over-sized and fail to detect weakness, our test always controls size and has reasonable power. We apply this testing approach to two empirical examples in the literature, and demonstrate that there are importance instances where our approach produces contradictory conclusions to the commonly applied linear testing approaches. Analysing the causal impact of US food aid on civil conflict, our approach fails to reject the null of weak identification, however, several commonly applied linear testing approaches all conclude that identification is not weak.

Another key contribution of the chapter is the construction of comprehensive concept of weak identification in discrete choice models, based not only on the convergence rate of drifting moments, but also on the respective weight of the key parameters, including variances of error terms and the level of simultaneity. This allows us to provide a unified GMM estimation framework for examining both linear and non-linear models, and for comparing the asymptotic properties of GMM estimators against other conventional two-step estimators for endogenous discrete models. While building on the general testing strategy of Antoine and Renault (2020), the test proposed in this chapter is based on a null hypothesis of genuine identification weakness, and not the nearly-strong identification null hypothesis analysed in, e.g., Andrews and Cheng (2012) and Antoine and Renault (2020).

The conclusions that this chapter gives to empirical researchers wishing to evaluate identification weakness in discrete models are clear. The canonical tests developed for linear models are not suitable for non-linear models, are likely to be overly optimistic, and can fail to detect genuinely weak identification. Our recommendation is a two-step approach. Conduct our testing approach in a first step, then, if the null is rejected, one can be very confident that identification is not weak, and conventional inference can proceed as usual. If the null of weak identification cannot be rejected, identification robust inference methods (as proposed in Stock and Wright, 2000 or Magnusson, 2010) would be more suitable to assert the significance of any estimated causal effects.

Furthermore, our asymptotic theory is conformable with the point of view on weak identification defended by Stock and Andrews (2005): "weak instruments should not be thought of as merely a small-sample problem, and the difficulties associated with weak instruments can arise even if the sample size is very large." We do see weak identification as a population problem (i.e. independent of the sample size): either the GMM estimator is not consistent (under the null of weak identification) or it is consistent (under the alternative). In this respect, the trick of using a drifting DGP, as contemplated in the weak identification literature, can be seen as a way to disentangle point identification (a maintained hypothesis in the framework of weak identification) and existence of a consistent estimator. This point of view may look at odds with the one put forward by Lewbel (2019) where it is stated that: "a parameter that is weakly identified (meaning that standard asymptotics provide a poor finite sample approximation to the actual distribution of the estimator) when $n = 100$ may be strongly identified when $n = 1000$." However, for all practical purpose, the methodological recommendation may not be so different: in our case, it is only for a large enough sample size that our consistent test may allow us to reject the null of weak identification. In these circumstances, the researcher can trust the consistency of the estimator and confidently use Wald inference.

2.7. Appendix

The appendix contains proofs for the main results in the chapter.

2.7.1. Lemmas

We first give several lemmas that are used to prove the main results.

Lemma 2.7.1 *Under Assumption 2.2.1, for* $\nu_n(\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (g_i(\theta) - \mathbb{E}[g_i(\theta)])$,

$$\nu_n(\theta) \Rightarrow \nu(\theta),$$

for $\nu(\theta)$ *a mean-zero Gaussian process with (uniformly) bounded covariance kernel* $S(\theta, \tilde{\theta})$.

Proof of Lemma 2.7.1. First, recall that for $g_i(\theta) = [a_i, b_i]r_i(\theta)$, with $r_i(\theta) := [r_{1i}(\theta), r_{2i}(\theta_2)]'$ so that

$$\|g_i(\theta)\| = \|[a_i, b_i]r_i(\theta)\| \leq \|[a_i, b_i]\|\|r_i(\theta)\|.$$

Under Assumption 2.2.1 (a), $[a_i, b_i]$ is i.i.d. and $\mathbb{E}[\|[a_i, b_i]\|^2] < \infty$. The result then follows if we can demonstrate that $r_i(\theta)$ is Donsker.

Consider the re-parameterisation $\vartheta = (\vartheta_1', \vartheta_2')'$, where $\vartheta_1 := (\alpha + \tilde{\rho}, \beta' - \tilde{\rho}\pi', \tilde{\rho}\xi')'$, and $\vartheta_2 := (\pi', \xi')'$. By compactness of Θ, the new parameter space $V := \{\vartheta = (\vartheta_1', \vartheta_2')' : \theta \in \Theta\}$ is also compact. Denote $w_{1i} = (y_{2i}, x_i', -z_i')'$ and $w_{2i} = (x_i', z_i')'$. Rewrite $\Phi[y_{2i}(\alpha + \tilde{\rho}) + x_i'(\beta' - \tilde{\rho}\pi') - z_i'\xi\tilde{\rho}] = \Phi(w_{1i}'\vartheta_1)$. By abuse of notation, define $r_{1i}(\vartheta_1) = y_{1i} - \Phi(w_{1i}'\vartheta_1)$, $r_{2i}(\vartheta_2) := y_{2i} - w_{2i}'\vartheta_2$, and define the class of functions

$$\mathcal{F} := \big\{r_i(\vartheta) = (r_{1i}'(\vartheta_1), r_{2i}'(\vartheta_2))' : \vartheta \in V\big\},$$

from the compactness of V, $(\mathcal{F}, \|\cdot\|)$ is totally bounded with $\|\cdot\|$ the Euclidean norm.

First, focus on $r_{1i}(\vartheta_1)$. For every w_{1i} and for $\vartheta_1, \bar{\vartheta}_1 \in V_1$, with V_1 a subspace of V associated with ϑ_1, without loss of generality, suppose $w_{1i}'\vartheta \geq w_{1i}'\bar{\vartheta}$. Then,

$$\begin{aligned}\|r_{1i}(\vartheta_1) - r_{1i}(\bar{\vartheta}_1)\| &= |\Phi(w_{1i}'\vartheta_1) - \Phi(w_{1i}'\bar{\vartheta}_1)| \\ &= \left|\int_{w_{1i}'\bar{\vartheta}_1}^{w_{1i}'\vartheta_1} \phi(t)dt\right| = \phi(c)|w_{1i}'(\vartheta_1 - \bar{\vartheta}_1)| \leq C\|w_{1i}\|\|\vartheta_1 - \bar{\vartheta}_1\|,\end{aligned}$$

for $c \in (w_{1i}'\bar{\vartheta}_1, w_{1i}'\vartheta_1)$ and some constant $C > 0$. For P the law of (w_{1i}', w_{2i}'), by Assumption 2.2.1 (a), we know that

$$\mathbb{E}_P[\|w_{1i}\|^2] < \infty.$$

Now, consider $r_{2i}(\vartheta_2)$ and note that, for $\vartheta_2, \bar{\vartheta}_2 \in V_2$, with V_2 a subspace of V associated with ϑ_2,

$$\|r_{2i}(\vartheta_2) - r_{2i}(\bar{\vartheta}_2)\| \leq \|w_{2i}\|\|\vartheta_2 - \bar{\vartheta}_2\|.$$

It then follows from Assumption 2.2.1 (a) that

$$\mathbb{E}_P[\|w_{2i}\|^2] < \infty.$$

Defining $L = \max\{\|w_{1i}\|, \|w_{2i}\|\}$, $\vartheta = (\vartheta_1', \vartheta_2')'$ and $\bar{\vartheta} = (\bar{\vartheta}_1', \bar{\vartheta}_2')'$, we have that $\mathbb{E}[L] < \infty$ and

$$\|r_i(\vartheta) - r_i(\bar{\vartheta})\| \leq L\|\vartheta - \bar{\vartheta}\|.$$

This Lipschitz property, together with the compactness of V implies that, by Theorem 2.7.11 of van der Vaart and Wellner (1996), $\mathcal{F}$ is P-Donsker. For $g_i(\theta) = [a_i, b_i]r_i(\theta)$, we then have that

$$\nu_n(\theta) := \sqrt{n}\,(\bar{g}_n(\theta) - \mathbb{E}[g_i(\theta)]) \Rightarrow \nu(\theta),$$

for $\theta \in \Theta$ where $\nu(\theta)$ denotes a Gaussian process with zero mean and variance kernel

$$S(\theta, \tilde{\theta}) := \mathbb{E}\left\{(g_i(\theta) - \mathbb{E}[g_i(\theta)])(g_i(\tilde{\theta}) - \mathbb{E}[g_i(\tilde{\theta})])'\right\}.$$

By the continuity of $S(\theta, \theta)$ in θ, Assumption 2.2.1 (a), and the compactness of Θ, we have

$$0 < \sup_{\theta, \tilde{\theta} \in \Theta} \|S(\theta, \tilde{\theta})\| < \infty.$$

■

The following results demonstrates that Assumption 2.3.2 in the main text is satisfied under Assumption 2.2.1.

Lemma 2.7.2 *Under Assumption 2.2.1, if* $\tilde{a}_i := \tilde{a}(y_{2i}, z_i, x_i)$ *satisfies* $\mathbb{E}_n\left[\|\tilde{a}_i z_i' \xi^0(y_{2i}, z_i', x_i')'\|^2\right] < \infty$, *for* $\Psi_n(\eta, \theta_2^0) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\tilde{a}_i \phi_i(\eta, \theta_2^0) z_i' \xi^0 - \mathbb{E}_n[\tilde{a}_i \phi_i(\eta, \theta_2^0) z_i' \xi^0]\}$,

$$\Psi_n(\eta, \theta_2^0) \Rightarrow \Psi(\eta, \theta_2^0),$$

for $\Psi(\eta, \theta_2^0)$ *a mean-zero Gaussian process over* $\Upsilon(\theta_2^0)$.

Proof of Lemma 2.7.2. Similar to the proof of Lemma 2.7.1, it suffices to show that

the class of functions

$$\mathcal{F} := \left\{ r_{3i}(\eta) = \tilde{a}_i \phi(\eta, \theta_2^0) z_i' \xi^0 : \ \eta \in \Upsilon(\theta_2^0) \right\},$$

is Donsker, where $\eta := (\tilde{\rho}, \tilde{\rho} + \alpha, \beta' - \tilde{\rho}\pi^{0'})'$. Hence, we only sketch the details.

Let $w_i := (-z_i'\xi^0, y_{2i}, x_i')'$. For every w_i and for $\eta, \bar{\eta} \in \Upsilon(\theta_2^0)$, without loss of generality, suppose $w_i'\eta \geq w_i'\bar{\eta}$. Let $\phi'(x)$ denote the derivative of the density function $\phi(x)$. Then, for $c \in (w_i'\bar{\eta}, w_i'\eta)$,

$$\|\tilde{a}_i \phi(\eta, \theta_2^0) z_i'\xi^0 - \tilde{a}_i \phi(\bar{\eta}, \theta_2^0) z_i'\xi^0\| = \phi'(c) \|\tilde{a}_i z_i'\xi^0 w_i'(\eta - \bar{\eta})\| \leq C \|\tilde{a}_i z_i'\xi^0 w_i\| \|\eta - \bar{\eta}\|,$$

for some constant $C > 0$, and where the equality follows by the intermediate value theorem, and the inequality from Cauchy-Schwartz. For P the joint law of w_i, by Assumption 2.2.1 (a), the compactness of Θ_2 (Assumption 2.2.1 (d)), and the moment hypothesis for $\tilde{a}_i$.,

$$\mathbb{E}_P[\|\tilde{a}_i z_i'\xi^0 w_i'\|^2] < \infty.$$

The remainder of the proof follows that of Lemma 2.7.1 and is omitted for brevity. ■

For $A_n = R\Lambda_n$, the following result demonstrates that, regardless of the interpretation for instrument weakness, for any consistent estimator the sample estimator $\partial \bar{g}_n(\theta_n)/\partial\theta' A_n$ is a consistent estimator of M in Assumption 2.3.3.

Lemma 2.7.3 *If $\{\theta_n\}$ is such that $\|\theta_n - \theta^0\| = o_p(1)$, then under Assumptions 2.2.1-2.3.4:*

$$M = \underset{n\to\infty}{plim} \frac{\partial \bar{g}_n(\theta_n)}{\partial \theta'} A_n, \ \textit{where } A_n = R\Lambda_n.$$

Proof of Lemma 2.7.3. Let $\bar{g}_n(\theta) = (\bar{g}_{1n}(\theta), \bar{g}_{2n}(\theta), ..., \bar{g}_{H,n}(\theta))'$. The mean value expansion of $\frac{\partial \bar{g}_{l,n}(\theta_n)}{\partial \theta'}$ at θ^0 yields

$$\frac{\partial \bar{g}_{l,n}(\theta_n)}{\partial \theta'} = \frac{\partial \bar{g}_{l,n}(\theta^0)}{\partial \theta'} + (\theta_n - \theta^0)' \frac{\partial^2 \bar{g}_{l,n}(\widetilde{\theta}_n)}{\partial \theta' \partial \theta}, \qquad l = 1, 2, ..., H$$

where $\widetilde{\theta}_n$ is component-by-component between θ^0 and θ_n. By the structure of the moment

$\bar{g}_n(\theta)$, the smoothness conditions on $\Phi(\cdot)$ and its derivatives, a_i and b_i are all measurable, it is not hard to prove that $\|\theta_n - \theta^0\| = o_p(1)$ implies the Hessian multiplied by A_n, $\frac{\partial^2 \bar{g}_{l,n}(\widetilde{\theta}_n)}{\partial\theta'\partial\theta} A_n = O_p(1)$ for $l = 1, 2, ..., H$. Therefore, $\|\theta_n - \theta^0\| = o_p(1)$ and Lemma 2.3.4 implies the result is satisfied. ■

Lemma 2.7.4 *Under Assumptions 2.2.1-2.3.4, and for Λ_n as in Lemma 2.3.4,* $\sqrt{n}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0) = O_p(1)$.

Proof of Lemma 2.7.4. The result is a consequence of Proposition 2.3.3 and Lemma 2.7.3, and the following inequality:

$$J_n(\zeta^0, \zeta^0) \geq J_n(\hat{\zeta}_n, \hat{\zeta}_n) = J_n(\hat{\zeta}_n, \zeta^0)\{1 + o_p(1)\},$$

which follows from the definition of $\hat{\zeta}_n$ and the consistency of $\hat{\zeta}_n$ in Proposition 2.3.3. For some component-by-component intermediate value ζ_n^*,

$$\sqrt{n}\bar{g}_n(\hat{\zeta}_n) = \sqrt{n}\bar{g}_n(\zeta^0) - \sqrt{n}\frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta'}(\zeta^0 - \hat{\zeta}_n),$$

and we can apply the inequality $\|a - b\| \geq -\|a\| + \|b\|$ to obtain

$$J_n^{1/2}(\hat{\zeta}, \zeta^0) \geq -\|\sqrt{n}\bar{g}_n(\zeta^0)\|_{\Omega_n} + \|\sqrt{n}\partial\bar{g}_n(\zeta_n^*)/\partial\zeta'(\zeta^0 - \hat{\zeta}_n)\|_{\Omega_n},$$

where $\Omega_n = S_n^{-1}(\zeta^0)$, $\|x\|_{\Omega_n} := (x'\Omega_n x)^{1/2}$ and where we have used the fact that (with probability converging to unity) $\lambda_{\min}(\Omega_n) > 0$. By the consistency of $\hat{\zeta}_n$ proved in Proposition 2.3.3 and Lemma 2.7.3, and for M as defined in Lemma 2.3.4, we have

$$\begin{aligned}
\|\sqrt{n}\partial\bar{g}_n(\zeta_n^*)/\partial\zeta'(\zeta^0 - \hat{\zeta}_n)\|_{\Omega_n} &= \|\partial\bar{g}_n(\zeta_n^*)/\partial\zeta'\Lambda_n\sqrt{n}\Lambda_n^{-1}(\zeta^0 - \hat{\zeta}_n)\|_{\Omega_n} \\
&= \|M\sqrt{n}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0) + o_p\left(\sqrt{n}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0)\right)\|_{\Omega_n} \\
&\geq C\|\sqrt{n}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0)\{1 + o_p(1)\}\|
\end{aligned}$$

for some constant $C > 0$, where the last inequality follows from the fact that M is full column rank and the fact that $\lambda_{\min}(\Omega_n) > 0$ (with probability converging to unity). Applying the above inequality into the first inequality, and using the fact that

$J_n(\zeta^0, \zeta^0) = O_p(1)$, we obtain

$$O_p(1) \geq C\|\sqrt{n}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0)\{1 + o_p(1)\}\|.$$

■

2.7.2. Proofs

Proof of Proposition 2.3.1. First, note that

$$\sqrt{n}\left[\bar{g}_n\left(\hat{\theta}_n^{\delta}\right) - \bar{g}_n\left(\hat{\theta}_n\right)\right] = \sqrt{n}\frac{\partial \bar{g}_n}{\partial \eta_1}\left(\eta_{1n}^*, \hat{\eta}_{2n}, \hat{\eta}_{3n}, \hat{\theta}_{2n}\right)\delta_n,$$

where η_{1n}^* denotes a component-by-component intermediate value between the first coefficients of $\hat{\theta}_n$ and $\hat{\theta}_n^{\delta}$. Recall $\delta_n \to 0$ as $n \to \infty$. Thus, we only have to prove that

$$\sqrt{n}\frac{\partial \bar{g}_n}{\partial \eta_1}\left(\eta_{1n}^*, \hat{\eta}_{2n}, \hat{\eta}_{3n}, \hat{\theta}_{2n}\right) = O_p\left(1\right).$$

For this purpose, we write the Taylor expansion

$$\begin{aligned}\sqrt{n}\frac{\partial \bar{g}_n}{\partial \eta_1}\left(\eta_{1n}^*, \hat{\eta}_{2n}, \hat{\eta}_{3n}, \hat{\theta}_{2n}\right) =& \sqrt{n}\frac{\partial \bar{g}_n}{\partial \eta_1}\left(\eta_{1n}^*, \hat{\eta}_{2n}, \hat{\eta}_{3n}, \theta_2^0\right) \\ &+ \frac{\partial^2 \bar{g}_n}{\partial \eta_1 \partial \theta_2'}\left(\eta_{1n}^*, \hat{\eta}_{2n}, \hat{\eta}_{3n}, \theta_{2n}^*\right)\sqrt{n}\left(\hat{\theta}_{2n} - \theta_2^0\right), \qquad (2.30)\end{aligned}$$

for some intermediate value θ_{2n}^*. By construction, the separation of estimators of θ_1 (or η_1) and θ_2 (see Remark 2.2.3 in Section 2.2.2) implies that $\sqrt{n}(\hat{\theta}_{2n} - \theta_2^0) = O_p(1)$. It is also worth noting that application of Lemma A1 of Stock and Wright (2000) would allow us to prove this result in an even more general context.

To see that the second part of the RHS of (2.30) is $O_p(1)$, note the following: (i), $\partial^2 \bar{g}_n / \partial \eta_1 \partial \theta_2'$ is continuous in η and θ_2; (ii), $\Upsilon(\theta_2^0) \times \Theta_2$ is compact; (iii), verify that $\|\partial^2 \bar{g}_n / \partial \eta_1 \partial \theta_2'\| \leq 2\|\tilde{a}(y_{2i}, z_i, x_i) z_i'\|$, where $\mathbb{E}_n[\|\tilde{a}(y_{2i}, z_i, x_i) z_i'\|] < \infty$ by hypothesis. From the i.i.d. nature of the data, the uniform law of large number (ULLN) then implies that the second derivative in question converges uniformly, and together with the fact that $\sqrt{n}(\hat{\theta}_{2n} - \theta_2^0) = O_p(1)$ implies that the second term on the RHS of (2.30) is $O_p(1)$.

Finally, it is straightforward to deduce that

$$\sup_{\eta\in\Upsilon\left(\theta_2^0\right)}\left\|\sqrt{n}\frac{\partial\bar{g}_n}{\partial\eta_1}\left(\eta,\theta_2^0\right)\right\| \leq \sup_{\eta\in\Upsilon\left(\theta_2^0\right)}\left\|\mathbb{E}_n\left\{\sqrt{n}\frac{\partial\bar{g}_n}{\partial\eta_1}\left(\eta,\theta_2^0\right)\right\}\right\| + \sup_{\eta\in\Upsilon\left(\theta_2^0\right)}\left\|\sqrt{n}\frac{\partial\bar{g}_n}{\partial\eta_1}\left(\eta,\theta_2^0\right)-\mathbb{E}_n\left\{\sqrt{n}\frac{\partial\bar{g}_n}{\partial\eta_1}\left(\eta,\theta_2^0\right)\right\}\right\|.$$

The first term is $O(1)$ under the null, while the second term is $O_p(1)$ under Assumption 2.3.2 (or Assumption 2.2.1 and Lemma 2.7.2). ■

Proof of Theorem 2.3.2. The result follows direction from **Proposition 2.3.1**. To see this, note that, by definition,

$$J_n\left(\hat{\theta}_n,\hat{\theta}_n\right) \leq J_n\left[\left(\eta_1^0,\tilde{\eta}_{2n},\tilde{\eta}_{3n},\tilde{\theta}_{2n}\right),\theta^0\right], \tag{2.31}$$

where $(\tilde{\eta}_{2n},\tilde{\eta}_{3n},\tilde{\theta}_{2n})$ denotes the infeasible CUGMM estimator of (η_2,η_3,θ_2) that would result if we knew η_1^0; i.e.,

$$\left(\tilde{\eta}_{2n},\tilde{\eta}_{3n},\tilde{\theta}_{2n}\right) = \underset{(\eta_2,\eta_3,\theta_2)}{\operatorname{argmin}} J_n\left[\left(\eta_1^0,\eta_2,\eta_3,\theta_2\right),\left(\eta_1^0,\eta_2,\eta_3,\theta_2\right)\right].$$

However, under Assumptions 2.2.1-2.3.1, the standard theory of the J-test for over-identification test for estimation of (η_2,η_3,θ_2) yields

$$J_n\left[\left(\eta_1^0,\tilde{\eta}_{2n},\tilde{\eta}_{3n},\tilde{\theta}_{2n}\right),\theta^0\right] \xrightarrow{d} \chi^2\left(H+1-p\right),$$

where $\xrightarrow{d}$ denotes convergence in distribution. Hence, the result in Proposition 2.3.1 implies that J_n^δ is asymptotically bounded above by a $\chi^2(H+1-p)$ random variable, which yields the necessary size control for the test W_n^δ. ■

Proof of Proposition 2.3.3. We work in the rotated parameter space, collected as $\zeta := (\eta',\theta_2')'$, but note that the result can be moved to the original parameters through the parameterisation $\theta = R\zeta$, and the fact that R can be consistently estimated.

Firstly, we demonstrate that there exist a deterministic diagonal matrix $\widetilde{\Lambda}_n$, a vector function $\gamma(\zeta)$, continuous in ζ, and a vector function $q_2(\eta_2,\eta_3)$, continuous in (η_2,η_3),

such that under our drifting DGP,

$$\mathbb{E}_n\left[\bar{g}_n(\zeta)\right] = \frac{\widetilde{\Lambda}_n}{\sqrt{n}}\gamma(\zeta) + q_2(\eta_2, \eta_3),$$

and

$$\gamma(\zeta) = 0 \text{ and } q_2(\eta_2, \eta_3) = 0 \quad \Longleftrightarrow \quad \zeta = 0,$$

where $\widetilde{\Lambda}_n$ has minimal and maximal eigenvalues, denoted by $\lambda_{\min}[\widetilde{\Lambda}_n]$ and $\lambda_{\max}[\widetilde{\Lambda}_n]$, respectively, that satisfy:

$$\lim_{n\to\infty} \lambda_{\min}[\widetilde{\Lambda}_n] = \infty \text{ and } \lim_{n\to\infty} \lambda_{\max}[\widetilde{\Lambda}_n]/\sqrt{n} < \infty.$$

After this, we can apply a similar strategy to Theorem 2.1 of Antoine and Renault (2012) to establish estimation consistency for the parameters $\zeta^0 := (\eta^{0\prime}, \theta_2^{0\prime})'$.

To simplify the calculations, we establish this result in the case where $x_i = 1$, for all i, and scalar z_i, which yields the moment functions: $g_i(\theta) = (g_{1i}(\eta, \theta_2)', g_{2i}(\theta)')'$, where

$$g_{1i}(\eta, \theta_2) = a_i\left(y_{1i} - \Phi\left[-\eta_1 z_i \xi + \eta_2 y_{2i} + \eta_3\right]\right), \; g_{2i}(\theta) = \begin{pmatrix} y_{2i} - \pi - \xi z_i \\ z_i\left(y_{2i} - \pi - \xi z_i\right) \end{pmatrix}.$$

From the identification condition in Assumption 2.2.2, $\theta_2^0 = (\pi^0, \xi^0)'$ can be directly identified from $\mathbb{E}_n[g_{2i}(\theta)] = 0$, which would yield least square estimators

$$\hat{\theta}_2 := \begin{pmatrix} \hat{\pi}_n \\ \hat{\xi}_n \end{pmatrix} = \begin{pmatrix} \bar{y}_{2n} - \hat{\xi}_n \bar{z}_n \\ \sum_{i=1}^n (z_i - \bar{z}_n)(y_{2i} - \bar{y}_{2n}) / \sum_{i=1}^n (z_i - \bar{z}_n)^2 \end{pmatrix},$$

for $\bar{z}_n = \sum_{i=1}^n z_i/n$ and $\bar{y}_{2n} = \sum_{i=1}^n y_{2i}/n$, which are clearly $\sqrt{n}$-consistent and asymptotically normal under Assumptions 2.2.1 and 2.2.2.

Now, define the stochastic process $\nu_n(\eta, \theta_2) = (\nu_{1n}(\eta, \theta_2)', \nu_{2n}(\theta_2))'$ to be conformable to $g_i(\eta, \theta_2) = (g_{1i}(\eta, \theta_2)', g_{2i}(\theta_2)')'$, where by abuse of notation, we write $g_{2i}(\theta)$ as $g_{2i}(\theta_2)$. From the $\sqrt{n}$-consistency of $(\hat{\pi}_n, \hat{\xi}_n)'$ and stochastic equicontinuity of $\nu_{1n}(\eta, \theta_2)$, we can restrict our analysis on the uniform behaviour of $\nu_{1n}(\eta, \theta_2)$ to the set $\Upsilon_n := \{(\eta, \theta_2) : \eta \in \Upsilon(\theta_2), \theta_2 \in \Theta_{2,n}\}$, for $\Upsilon(\theta_2)$ as defined above equation (2.21), and where for some $\delta > 0$

and $\delta = o(1)$,

$$\Theta_{2,n} := \left\{\theta_{2n} : \|\theta_{2n} - \theta_2^0\| \leq \delta/\sqrt{n}\right\}.$$

In the remainder, we take θ_{2n} to be an arbitrary sequence in $\Theta_{2,n}$.

For θ_{2n} as above, recall that, using the decomposition in equation (2.18), for some $\bar{\eta}_1$ such that $\eta_1^0 \leq \bar{\eta}_1 \leq \eta_1$,

$$\begin{aligned} m_{1n}(\eta, \theta_{2n}) &= m_{1n}(\eta, \theta_0^2) + m_{1n}(\eta, \theta_{2n}) - m_{1n}(\eta, \theta_0^2) \\ &= q_{11,n}(\eta)/\varsigma_n + q_{12,n}(\eta_2, \eta_3) + o_p(n^{-1/2}) \\ &= (\eta_1 - \eta_1^0)\mathbb{E}_n\left[\frac{1}{n}\sum_{i=1}^n \tilde{a}_i\phi_i(\bar{\eta}_1, \eta_2, \eta_3; \theta_2^0)z_i\xi^0\right] + q_{12,n}(\eta_2, \eta_3) + o_p(n^{-1/2}). \end{aligned} \tag{2.32}$$

Moreover, by Assumption 2.3.3, uniformly over $\bar{\eta} = (\bar{\eta}_1, \eta_2', \eta_3')' \in \Upsilon(\theta_2^0)$,

$$\left\|\mathbb{E}_n\left[\frac{1}{n}\sum_{i=1}^n \tilde{a}_i\phi_i(\bar{\eta}, \theta_2^0)z_i\xi^0\right]\varsigma_n - V^0(\bar{\eta})\right\| = o(1)$$

so that

$$m_{1n}(\eta, \theta_{2n}) = \varsigma_n^{-1}(\eta_1 - \eta_1^0)V^0(\bar{\eta}) + q_{12,n}(\eta_2, \eta_3) + o_p(n^{-1/2}). \tag{2.33}$$

Now, decompose $\sqrt{n}\bar{g}_{1n}(\eta, \theta_{2n})$ as

$$\sqrt{n}\bar{g}_{1n}(\eta, \theta_{2n}) = \sqrt{n}\left\{\bar{g}_{1n}(\eta, \theta_{2n}) - m_{1n}(\eta, \theta_{2n})\right\} + \sqrt{n}m_{1n}(\eta, \theta_{2n}),$$

and apply equation (2.33) to obtain

$$\begin{aligned} \sqrt{n}\bar{g}_{1n}(\eta, \theta_{2n}) &= \nu_{1n}(\eta, \theta_2^0) + \sqrt{n}m_{1n}(\eta, \theta_2^0) + o_p(1) \\ &= \nu_{1n}(\eta, \theta_2^0) + \frac{\sqrt{n}}{\varsigma_n}V^0(\bar{\eta})(\eta_1 - \eta_1^0)\{1 + o_p(1)\} + \sqrt{n}q_{12,n}(\eta_2, \eta_3). \end{aligned}$$

Recall that by Lemma 2.7.1, $\nu_n(\eta, \theta_2^0) \Rightarrow \nu(\eta, \theta_2^0)$, and hence is $O_p(1)$ uniformly for $\eta \in \Upsilon(\theta_2^0)$.

Define $\bar{\lambda}_n := \sqrt{n}/\varsigma_n$, which satisfies $\bar{\lambda}_n \to \infty$, as $n \to \infty$, where $\bar{\lambda}_n = o(\sqrt{n})$ by the definition of ς_n in Assumption 2.3.3. Now, define the matrix

$$\widetilde{\Lambda}_n := \begin{bmatrix} \bar{\lambda}_n \mathbf{I}_{\dim(g_1)} & \mathbf{O} \\ \mathbf{O} & n^{1/2}\mathbf{I}_{\dim(g_2)} \end{bmatrix}$$

and the vectors

$$\gamma(\zeta) = \begin{pmatrix} V^0(\eta)(\eta_1 - \eta_1^0) \\ \mathbb{E}_n\left[\bar{g}_{2n}(\theta_2)\right] \end{pmatrix}, \quad q_2(\eta_2, \eta_3) = \begin{pmatrix} q_{12,n}(\eta_2, \eta_3) \\ \mathbf{0} \end{pmatrix}.$$

Then, up to $o_p(1)$ terms,

$$\begin{aligned} \sqrt{n}\bar{g}_n(\eta, \theta_2) &= \sqrt{n}\left\{\bar{g}_n(\eta, \theta_2) - \mathbb{E}_n[\bar{g}_n(\eta, \theta_2)]\right\} + \sqrt{n}\mathbb{E}_n[\bar{g}_n(\eta, \theta_2)] \\ &= \nu_n(\eta, \theta_2) + \widetilde{\Lambda}_n\gamma(\zeta) + \sqrt{n}q_2(\eta_2, \eta_3). \end{aligned}$$

The remainder of the result follows a similar strategy to Theorem 2.1 in Antoine and Renault (2012). Let W be a positive-definite $H \times H$ matrix, and define $\|x\|_W^2 := x'Wx$. For $\nu_n(\zeta)$, $\widetilde{\Lambda}_n$ and $\gamma(\zeta)$ as above, we can rewrite the CUGMM objective function in the rotated parameter space as

$$J_n[\zeta, \zeta]/n = \left\| \frac{\nu_n(\zeta)}{\sqrt{n}} + \frac{\widetilde{\Lambda}_n}{\sqrt{n}}\gamma(\zeta) + q_2(\eta_2, \eta_3) \right\|_{\Omega_n(\zeta)}^2, \text{ for } \Omega_n(\zeta) := S_n^{-1}(\zeta).$$

By definition of $\hat{\zeta}_n$, $J_n[\zeta^0, \zeta^0] \geq J_n[\hat{\zeta}_n, \hat{\zeta}_n]$ which implies

$$\left\|\nu_n(\zeta^0)/\sqrt{n}\right\|_{\Omega_n(\zeta^0)}^2 \geq \left\|\nu_n(\hat{\zeta}_n)/\sqrt{n} + \widetilde{\Lambda}_n\gamma(\hat{\zeta}_n)/\sqrt{n} + q_2(\hat{\eta}_{2n}, \hat{\eta}_{3n})\right\|_{\Omega_n(\hat{\zeta}_n)}^2. \tag{2.34}$$

Define $\Omega_n^0 := \Omega_n(\zeta^0)$, $\hat{\Omega}_n := \Omega_n(\hat{\zeta}_n)$, $x_n := \nu_n(\hat{\zeta}_n)$, $y_n := \widetilde{\Lambda}_n\gamma(\hat{\zeta}_n) + \sqrt{n}q_2(\hat{\eta}_{2n}, \hat{\eta}_{3n})$ and $d_n := \nu_n(\hat{\zeta}_n)'\hat{\Omega}_n\nu_n(\hat{\zeta}_n) - \nu_n(\zeta^0)'\Omega_n^0\nu_n(\zeta^0)$. Denote $\lambda_{\min}[A]$ and $\lambda_{\max}[A]$ as the smallest and the largest eigenvalue of a matrix A, respectively. Then, from (2.34), we obtain

$$\begin{aligned} 0 &\geq J_n[\hat{\zeta}_n, \hat{\zeta}_n] - J_n[\zeta^0, \zeta^0] = d_n + \|y_n\|_{\hat{\Omega}_n}^2 + 2(\hat{\Omega}_n x_n)'y_n \\ &\geq d_n + \|y_n\|^2\lambda_{\min}\left[\hat{\Omega}_n\right] - 2\|y_n\|\|\hat{\Omega}_n x_n\|. \end{aligned} \tag{2.35}$$

Defining $z_n := \|y_n\|$, and for $\lambda_{\min}\left[\hat{\Omega}_n\right] > 0$, we can re-arrange equation (2.35) as

$$z_n^2 - 2z_n \frac{\|\hat{\Omega}_n x_n\|}{\lambda_{\min}\left[\hat{\Omega}_n\right]} + \frac{d_n}{\lambda_{\min}\left[\hat{\Omega}_n\right]} \leq 0$$

Solving the above equation for z_n yields:

$$B_n - \left[B_n^2 - C_n\right]^{1/2} \leq z_n \leq B_n + \left[B_n^2 - C_n\right]^{1/2}, \;\; B_n := \frac{\|\hat{\Omega}_n x_n\|}{\lambda_{\min}\left[\hat{\Omega}_n\right]}, \;\; C_n := \frac{d_n}{\lambda_{\min}\left[\hat{\Omega}_n\right]}, \tag{2.36}$$

where by definition of C_n and B_n we know that $B_n^2 - C_n \geq 0$. From (2.36), the result follows if

$$B_n = O_p(1), \text{ and } C_n = O_p(1).$$

Consider first, B_n and note that

$$B_n \leq \|x_n\| \frac{\lambda_{\max}\left[\hat{\Omega}_n\right]}{\lambda_{\min}\left[\hat{\Omega}_n\right]} \leq \sup_{\zeta \in R^{-1}\Theta} \|\nu_n(\zeta)\| \frac{\sup_{\zeta \in R^{-1}\Theta} \lambda_{\max}\left[\Omega_n(\zeta)\right]}{\inf_{\zeta \in R^{-1}\Theta} \lambda_{\min}\left[\Omega_n(\zeta)\right]}.$$

By the result of Lemma 2.7.1, $\sup_{\zeta \in R^{-1}\Theta} \|\nu_n(\zeta)\| = O_p(1)$. It then follows that $B_n = O_p(1)$ so long as, for all n large enough, with probability approaching one,

$$0 < \inf_{\zeta \in R^{-1}\Theta} \lambda_{\min}\left[\Omega_n(\zeta)\right] \leq \sup_{\zeta \in R^{-1}\Theta} \lambda_{\max}\left[\Omega_n(\zeta)\right] < \infty,$$

which is guaranteed to be satisfied for n large enough under the assumptions of the result. For C_n, recalling that $d_n = \|\nu_n(\hat{\zeta}_n)\|^2_{\hat{\Omega}_n} - \|\nu_n(\zeta^0)\|^2_{\Omega_n^0}$, we obtain

$$|C_n| \leq 2 \sup_{\zeta \in R^{-1}\Theta} \|\nu_n(\zeta)\|^2 \frac{\sup_{\zeta \in R^{-1}\Theta} \lambda_{\max}\left[\Omega_n(\zeta)\right]}{\inf_{\zeta \in R^{-1}\Theta} \lambda_{\min}\left[\Omega_n(\zeta)\right]}.$$

Repeating the same argument for C_n as for B_n yields $C_n = O_p(1)$. Applying $B_n = O_p(1)$, $C_n = O_p(1)$ to equation (2.36), we have

$$z_n = \|y_n\| = \|\widetilde{\Lambda}_n \gamma(\hat{\zeta}_n) + \sqrt{n} q_2(\hat{\eta}_{2n}, \hat{\eta}_{3n})\| = O_p(1)$$

It then follows that,

$$\|\gamma(\hat{\zeta}_n) + q_2(\hat{\eta}_{2n}, \hat{\eta}_{3n})\| = O_p\left(1/\bar{\lambda}_n\right).$$

Consistency of $\hat{\zeta}_n$ now follows by modifying the standard argument (see, e.g., Newey and McFadden (1994), page 2132). By continuity of $\gamma(\zeta) + q_2(\eta_2, \eta_3)$, for any $\epsilon > 0$, there exists some δ_ϵ such that

$$\Pr\left[\|\hat{\zeta}_n - \zeta^0\| > \epsilon\right] \leq \Pr\left[\left\|\left\{\gamma(\hat{\zeta}_n) + q_2(\hat{\eta}_{2n}, \hat{\eta}_{3n})\right\} - \gamma(\zeta^0) - q_2(\eta_2^0, \eta_3^0)\right\| > \delta_\epsilon\right].$$

However, by Assumption 2.3.3, $V^0(\eta)$ is non-zero uniformly for $\eta \in \Upsilon(\theta_2^0)$, so that under the identification condition in Assumption 2.2.2 and the identification of $q_{12,n}(\eta_2, \eta_3)$ in Assumption 2.3.1, we can conclude:

$$\|\gamma(\zeta)+q_2(\eta_2,\eta_3)\| \leq \sup_{\eta\in\Upsilon(\theta_2^0)} \|V^0(\eta)\|\|\eta_1-\eta_1^0\|+\|\mathbb{E}_n[\bar{g}_{2n}(\theta_2)]\|+\|q_{12,n}(\eta_2,\eta_3)\| = 0 \quad \Longleftrightarrow \quad \zeta = \zeta^0.$$

Therefore,

$$\Pr\left[\|\hat{\zeta}_n - \zeta^0\| > \epsilon\right] \leq \Pr\left[\delta_\epsilon < \left\|\gamma(\hat{\zeta}_n) + q_2(\hat{\eta}_{2n}, \hat{\eta}_{3n})\right\|\right] = o(1),$$

where the last equality follows from the fact that $\|\gamma(\hat{\zeta}_n)+q_2(\hat{\eta}_{2n}, \hat{\eta}_{3n})\| = O_p\left(1/\bar{\lambda}_n\right)$, and $\bar{\lambda}_n \to \infty$ as $n \to \infty$. ■

Proof of Lemma 2.3.4. In the rotated parameter space, the moment function is given by

$$g_i(\zeta) = a_i r_{1i}(\zeta) + b_i r_{2i}(\theta_2) = \begin{pmatrix} \tilde{a}_i(y_{2i}, x_i, z_i) r_{1i}(\zeta) \\ \tilde{b}_i(x_i, z_i) r_{2i}(\theta_2) \end{pmatrix} = \begin{pmatrix} g_{1i}(\zeta) \\ g_{2i}(\theta_2) \end{pmatrix}.$$

The $(H \times p)$-dimensional Jacobian matrix $\partial g_i(\zeta)/\partial \zeta'$ is given by

$$\partial g_i(\zeta)/\partial \zeta' = \begin{pmatrix} \partial g_{1i}(\zeta)/\partial \eta' & \partial g_{1i}(\zeta)/\partial \theta_2' \\ \mathbf{O} & \partial g_{2i}(\theta_2)/\partial \theta_2' \end{pmatrix}.$$

For Λ_n as in the statement of the result,

$$\begin{aligned}\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'}\Lambda_n &= \left\{\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'} - \mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'}\right]\right\}\Lambda_n + \left\{\mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'}\right]\right\}\Lambda_n \\ &= O_p(\varsigma_n/\sqrt{n}) + o_p(1) + \left\{\mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'}\right]\right\}\Lambda_n \\ &= o_p(1) + \left\{\mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'}\right]\right\}\Lambda_n. \end{aligned} \tag{2.37}$$

The second equality follows from Assumption 2.3.3, and the uniform convergence of the remaining derivatives, which follows from Assumptions 2.2.1, 2.2.2 and a ULLN for iid data. The third equation follows from the fact that $\varsigma_n/\sqrt{n} = o(1)$. For Λ_{1n} denoting the diagonal matrix

$$\Lambda_{1n} := \begin{pmatrix} \varsigma_n & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{k_x+1} \end{pmatrix}$$

we decompose the $(p \times p)$-dimensional matrix Λ_n as

$$\Lambda_n = \begin{pmatrix} \Lambda_{1n} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{k_x+k_z} \end{pmatrix}.$$

From this definition, the last term in equation (2.37) can be stated as

$$\mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'}\right]\Lambda_n = \mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \eta'} \vdots \frac{\partial \bar{g}_n(\zeta^0)}{\partial \theta_2'}\right]\Lambda_n = \mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \eta'}\Lambda_{1n} \vdots \frac{\partial \bar{g}_n(\zeta^0)}{\partial \theta_2'}\right]. \tag{2.38}$$

Recalling the functions $q_{11,n}(\eta)$ and $q_{12}(\eta_2, \eta_3)$ underlying Assumption 2.3.1, the first component in equation (2.38) can be seen to be given by

$$\begin{aligned}\mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \eta'}\right]\begin{pmatrix} \varsigma_n & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{k_x+1} \end{pmatrix} &= \begin{pmatrix} \frac{\partial q_{11,n}(\eta^0)}{\partial \eta_1}\varsigma_n & \mathbf{O} \\ \mathbf{O} & \frac{\partial q_{12}(\eta_2^0, \eta_3^0)}{\partial (\eta_2, \eta_3')'} \end{pmatrix} = \begin{pmatrix} V^0(\eta^0) & \mathbf{O} \\ \mathbf{O} & \frac{\partial q_{12}(\eta_2^0, \eta_3^0)}{\partial (\eta_2, \eta_3')'} \end{pmatrix} \\ &= M_1(\eta^0). \end{aligned}$$

By Assumption 2.3.1 (b) the south-east block of $M_1(\eta^0)$ has column rank $1 + k_x$, while by Assumption 2.3.3 the north-east block of $M_1(\eta^0)$ is of column rank 1. Therefore, since

$M_1(\eta^0)$ is block diagonal, conclude that

$$\lim_{n\to\infty} \text{column rank}\left[M_1(\eta^0)\right] = 2 + k_x.$$

For the second term in (2.38), recalling the Jacobian of $\partial g_i(\zeta)/\partial\zeta'$, we have that

$$\mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \theta_2'}\right] = \mathbb{E}_n\left[\begin{pmatrix} \partial \bar{g}_{1n}(\eta^0,\theta_2^0)/\partial\theta_2' \\ \partial \bar{g}_{2n}(\theta_2^0)/\partial\theta_2' \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}_n\left[(\mathbf{O} : \tilde{a}(y_{2i}, x_i, z_i)\phi_i(\eta^0,\theta_2^0)\eta_1^0 z_i')\right] \\ \mathbb{E}_n\left[\tilde{b}(x_i,z_i)\,(x_i' \ : \ z_i')\right] \end{pmatrix}$$

By Assumption 2.3.3, the matrix $\mathbb{E}_n\left[\tilde{b}(x_i,z_i)\,(x_i' \ : \ z_i')\right]$ has column rank $(k_x + k_z)$.

Combing the two Jacobian terms, the $H \times p$ dimensional Jacobian matrix in equation (2.38) can be seen as

$$\mathbb{E}_n\left[\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'}\right]\Lambda_n = \begin{pmatrix} M_1(\eta^0) & \mathbb{E}_n\left[(\mathbf{O} : \tilde{a}(y_{2i}, x_i, z_i)\phi_i(\eta^0,\theta_2^0)\eta_1^0 z_i')\right] \\ \mathbf{O} & \mathbb{E}_n\left[\tilde{b}(x_i,z_i)\,(x_i' \ : \ z_i')\right] \end{pmatrix}.$$

The matrix

$$M = \operatorname*{plim}_{n\to\infty}\left\{\frac{\partial \bar{g}_n(\zeta^0)}{\partial \zeta'}\Lambda_n\right\},$$

then exists and satisfies

$$\begin{aligned}\text{column rank}[M] &= \lim_{n\to\infty} \text{column rank}\left[M_1(\eta^0)\right] + \lim_{n\to\infty} \text{column rank}\left\{\mathbb{E}_n\left[\tilde{b}(x_i,z_i)\,(x_i' \ : \ z_i')\right]\right\} \\ &= (2 + k_x) + (k_x + k_z) = p.\end{aligned}$$

■

Proof of Theorem 2.3.5. From the first order condition of the CUGMM objective function, $\hat{\zeta}_n$ satisfies

$$n\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1}\bar{g}_n(\hat{\zeta}_n) - W \cdot n\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1}\bar{g}_n(\hat{\zeta}_n) = 0 \tag{2.39}$$

for W defined as

$$W \cdot \sqrt{n}\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} = \text{Cov}\left(\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta}, \bar{g}_n(\hat{\zeta}_n)\right)\left(\mathbf{I}_H \otimes \left[S_n(\hat{\zeta}_n)^{-1}\sqrt{n}\bar{g}_n(\hat{\zeta}_n)\right]\right), \tag{2.40}$$

and where $\text{Cov}(\cdot)$

$$\text{Cov}\left(\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta}, \bar{g}_n(\hat{\zeta}_n)\right) := \left[\text{Cov}\left(\frac{\partial \bar{g}_{1n}(\hat{\zeta}_n)}{\partial \zeta}, \bar{g}_n(\hat{\zeta}_n)\right), \cdots, \text{Cov}\left(\frac{\partial \bar{g}_{H,n}(\hat{\zeta}_n)}{\partial \zeta}, \bar{g}_n(\hat{\zeta}_n)\right)\right]. \tag{2.41}$$

Substituting (3.22) into (2.39), and multiplying both sides of the equation (2.39) by $n^{-1/2}$, we obtain

$$\begin{aligned}\sqrt{n}\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1}\bar{g}_n(\hat{\zeta}_n) - \text{Cov}\left(\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta}, \bar{g}_n(\hat{\zeta}_n)\right) \\ \times \left(\mathbf{I}_H \otimes \left[S_n(\hat{\zeta}_n)^{-1}\sqrt{n}\bar{g}_n(\hat{\zeta}_n)\right]\right) S_n(\hat{\zeta}_n)^{-1}\bar{g}_n(\hat{\zeta}_n) = 0.\end{aligned} \tag{2.42}$$

Apply the mean value theorem to $\bar{g}_n(\hat{\zeta}_n)$,

$$\begin{aligned}\bar{g}_n(\hat{\zeta}_n) &= \bar{g}_n(\zeta^0) + \frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta'}(\hat{\zeta}_n - \zeta^0) \\ &= \bar{g}_n(\zeta^0) + n^{-1/2}\frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta'}\Lambda_n n^{1/2}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0).\end{aligned}$$

By Proposition 2.3.3, $\hat{\zeta}_n$ is consistent and by Lemma 2.7.4, $\sqrt{n}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0) = O_p(1)$. Then Lemma 2.7.3 and Assumption 2.3.3 yield

$$n^{-1/2}\frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta'}\Lambda_n n^{1/2}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0) = n^{-1/2}MO_p(1) + o_p(n^{-1/2}) = O_p(n^{-1/2}),$$

so that we can conclude

$$\bar{g}_n(\hat{\zeta}_n) = \bar{g}_n(\zeta^0) + O_p(n^{-1/2}). \tag{2.43}$$

From (2.43), the convergence rate of $\bar{g}_n(\hat{\zeta}_n)$ is determined by $\bar{g}_n(\zeta^0)$, and by Lemma

2.7.1, and the fact $\mathbb{E}_n[g_i(\zeta^0)] = 0$ (under Assumption 2.2.2),

$$\sqrt{n}\bar{g}_n(\zeta^0) \Rightarrow \nu(\zeta^0),$$

where $\nu(\zeta^0)$ is a Gaussian process with mean-zero and variance matrix $S(\zeta^0)$. Therefore, $\bar{g}_n(\zeta^0) = O_p(n^{-1/2})$ and together with (2.43), we have that $\bar{g}_n(\hat{\zeta}_n) = O_p(n^{-1/2})$. Given Lemmas 2.7.1 and 2.7.2, and the fact that $\sup_{\zeta \in R^{-1}\Theta} \|S_n^{-1}(\zeta)\| < \infty$, the above result then yields:

$$\text{Cov}\left(\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta}, \bar{g}_n(\hat{\zeta}_n)\right) = O_p(1), \quad \text{and} \quad \mathbf{I}_H \otimes \left[S_n(\hat{\zeta}_n)^{-1}\sqrt{n}\bar{g}_n(\hat{\zeta}_n)\right] = O_p(1). \tag{2.44}$$

From $\bar{g}_n(\zeta^0) = O_p(n^{-1/2})$ and the results in (2.44), the second term on the left hand side of (2.42) is $O_p(n^{-1/2})$. Then, (2.42) becomes

$$\sqrt{n}\frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1}\bar{g}_n(\hat{\zeta}_n) = O_p(n^{-1/2}). \tag{2.45}$$

Plugging (2.43) into (2.45) and multiplying both sides by Λ_n', we obtain

$$\begin{aligned} O_p(n^{-1/2})\Lambda_n' = {} & \sqrt{n}\Lambda_n' \frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1}\bar{g}_n(\zeta^0) \\ & + \sqrt{n}\Lambda_n' \frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1} \frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta'} \Lambda_n \Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0). \end{aligned} \tag{2.46}$$

In addition, from to the uniform convergence of $S_n(\zeta)$ to $S(\zeta)$ over $\zeta \in R^{-1}\Theta$, which follows from compactness of $R^{-1}\Theta$, continuity of $g_i(\zeta)$, Assumption 2.2.1, and the consistency of $\hat{\zeta}_n$,

$$\begin{aligned} \|S_n(\hat{\zeta}_n) - S(\zeta^0)\| &= \|S_n(\hat{\zeta}_n) - S(\hat{\zeta}_n) + S(\hat{\zeta}_n) - S(\zeta^0)\| \\ &\leq \|S_n(\hat{\zeta}_n) - S(\hat{\zeta}_n)\| + \|S(\hat{\zeta}_n) - S(\zeta^0)\| \\ &\leq \sup_{\zeta \in R^{-1}\Theta} \|S_n(\zeta) - S(\zeta)\| + \|S(\hat{\zeta}_n) - S(\zeta^0)\| \\ &= o_p(1). \end{aligned} \tag{2.47}$$

Moreover, by the consistency of $\hat{\zeta}_n$, Lemma 2.7.3 and equation (2.47) imply that

$$\frac{\partial \bar{g}_n(\hat{\zeta}_n)}{\partial \zeta'}\Lambda_n \xrightarrow{p} M, \quad \Lambda_n' \frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1} \frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta'}\Lambda_n \xrightarrow{p} M' S^{-1} M.$$

Because the $H \times p$ matrix M is full column rank under Assumption 2.3.3, then the non-singularity of S and the rank condition of M imply that $\Lambda_n' \frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1} \frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta'}\Lambda_n$ is invertible for large enough n. Hence, from (2.46) and $\Lambda_n' O_p(n^{-1/2}) = O_p(\|\Lambda_n/\sqrt{n}\|) = o_p(1)$, we obtain

$$\begin{aligned}
&\sqrt{n}\Lambda_n^{-1}(\hat{\zeta}_n - \zeta^0) \\
&= -\left[\Lambda_n' \frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1} \frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta'}\Lambda_n\right]^{-1} \Lambda_n' \frac{\partial \bar{g}_n(\hat{\zeta}_n)'}{\partial \zeta} S_n(\hat{\zeta}_n)^{-1} \sqrt{n}\bar{g}_n(\zeta^0) + o_p(1). \quad (2.48)
\end{aligned}$$

Therefore, based on (2.47), (2.48) and the asymptotic normality of $\sqrt{n}\bar{g}_n(\zeta^0)$ from Lemma 2.7.1, the desired results follow. ■

Proof of Theorem 2.3.6. First, we recall the parameter rotation $\zeta = (\eta', \theta_2')'$, where $\zeta := R^{-1}\theta$. For $\hat{\zeta}_n = R^{-1}\hat{\theta}_n$, the perturbed parameter values are defined by

$$\begin{aligned}
\hat{\zeta}_n^{\delta} &:= \begin{pmatrix} \hat{\eta}_{1n} \\ \hat{\eta}_{2n} \\ \vdots \\ \hat{\theta}_{2n} \end{pmatrix} + \begin{pmatrix} \delta_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} = R^{-1}\hat{\theta}_n + \begin{pmatrix} \delta_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \\
\hat{\theta}_n^{\delta} &:= R\hat{\zeta}_n^{\delta}. \quad (2.49)
\end{aligned}$$

A mean value expansion of $\bar{g}_n(\hat{\theta}_n^{\delta})$ yields, for $A_n := R\Lambda_n$,

$$\begin{aligned}
\sqrt{n}\bar{g}_n(\hat{\theta}_n^{\delta}) &= \sqrt{n}\bar{g}_n(\theta^0) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial g_i(\theta_n^*)}{\partial \theta'}(\hat{\theta}_n^{\delta} - \theta^0) \\
&= \sqrt{n}\bar{g}_n(\theta^0) + \frac{\partial \bar{g}_n(\theta_n^*)}{\partial \theta'} A_n \sqrt{n} A_n^{-1}(\hat{\theta}_n - \theta^0) + \sqrt{n}\frac{\partial \bar{g}_n(\theta_n^*)}{\partial \theta'} R \begin{pmatrix} \delta_n \\ \mathbf{0}_{p-1} \end{pmatrix} \quad (2.50)
\end{aligned}$$

where θ_n^* is component-by-component between $\hat{\theta}_n$ and θ^0. We now analyse each of the terms in (2.50).

For the first term in (2.50), by Lemma 2.7.1, $\sqrt{n}\bar{g}_n(\theta^0) = O_p(1)$. For the second term, first note that, under the alternative hypothesis, $\|\hat{\theta}_n - \theta^0\| = o_p(1)$ (by Proposition 2.3.3), which further implies $\|\zeta_n^* - \zeta^0\| = o_p(1)$ for $\zeta_n^* = R^{-1}\theta_n^*$. Then, it follows that

$$\begin{aligned}\frac{\partial \bar{g}_n(\theta_n^*)}{\partial \theta'} A_n \sqrt{n} A_n^{-1}(\hat{\theta}_n - \theta^0) &= \frac{\partial \bar{g}_n(\zeta_n^*)}{\partial \zeta} \Lambda_n \sqrt{n} \Lambda_n^{-1} \left(\hat{\zeta}_n - \zeta^0\right) \\ &= M \cdot O_p(1) \\ &= O_p(1),\end{aligned}$$

where the second equality follows from Lemma 2.7.3, and the third from the fact that $\|M\| < \infty$. Therefore, the second term in (2.50) is $O_p(1)$.

Focusing on the last term in (2.50), we have

$$\begin{aligned}\sqrt{n}\frac{\partial \bar{g}_n(\theta_n^\star)}{\partial \theta'} R \begin{pmatrix} \delta_n \\ \mathbf{0}_{p-1} \end{pmatrix} &= \frac{\partial \bar{g}_n(\zeta_n^\star)}{\partial \zeta'} \Lambda_n \sqrt{n} \Lambda_n^{-1} \begin{pmatrix} \delta_n \\ \mathbf{0}_{p-1} \end{pmatrix} \\ &= M \begin{pmatrix} \delta_n \{\sqrt{n}/\varsigma_n\} \\ \mathbf{0}_{p-1} \end{pmatrix} + o_p(1), \\ &= \begin{pmatrix} V^0(\eta^0)\delta_n \{\sqrt{n}/\varsigma_n\} \\ \mathbf{0}_{p-1} \end{pmatrix} + o_p(1), \end{aligned} \tag{2.51}$$

where the second line follows from Lemma 2.7.3 and the last from the fact that M is full rank (Lemma 2.3.4). Applying these order results for the three terms in (2.50), we obtain

$$\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta) = O_p(1) + \begin{pmatrix} V^0(\eta^0)\delta_n \{\sqrt{n}/\varsigma_n\} \\ \mathbf{0}_{p-1} \end{pmatrix} + o_p(1).$$

Since $\|V^0(\eta^0)\| > 0$ by Assumption 2.3.3, conclude that $\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta)$ diverges if $\{\sqrt{n}/\varsigma_n\}\delta_n \to \infty$.

Using the above result, we can now show that J_n^δ diverges under the alternative. From the proof of Lemma 2.7.1,

$$n^{1/2}\{\bar{g}_n(\theta) - \mathbb{E}_n[\bar{g}_n(\theta)]\} \Rightarrow \nu(\theta), \tag{2.52}$$

where $\nu(\theta)$ is a Gaussian stochastic process on Θ with mean-zero and bounded covariance kernel $S(\theta, \theta)$. Since $\hat{\theta}_n^\delta \xrightarrow{p} \theta^0$ under Assumption 2.3.3, the uniform convergence (2.52) indicates that the sample covariance matrix satisfies $S_n(\hat{\theta}_n^\delta) \xrightarrow{p} S(\theta^0)$. Thus, for n large enough, $S_n(\hat{\theta}_n^\delta)$ is positive-definite with bounded maximal eigenvalue. Therefore,

$$J_n^\delta \geq \lambda_{\min}\left[S_n^{-1}(\hat{\theta}_n^\delta)\right] \left\|\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta)\right\|^2, \tag{2.53}$$

where $\lambda_{\min}\left[S_n^{-1}(\hat{\theta}_n^\delta)\right] > 0$ for large enough n. Thus, $\{\sqrt{n}/\varsigma_n\}\delta_n \to \infty$ implies $\operatorname{plim}_{n\to\infty} J_n^\delta \to \infty$. ■

Chapter 3

Decomposing Identification Gains and Evaluating Instrument Identification Power for Partially Identified ATE

3.1. Introduction

This chapter investigates the identification power of instrumental variables for the average treatment effect (ATE) in partially identified binary outcome models. Binary outcome models with binary endogenous treatment are widely used in empirical studies. The role played by the instrumental variable in identifying the ATE in such models has long been a controversial topic and has been discussed in many papers (see e.g., Freedman and Sekhon, 2010; Han and Vytlacil, 2017; Heckman, 1978; Li, Poskitt, and Zhao, 2019; Maddala, 1986; Mourifié and Méango, 2014; Wilde, 2000). In particular, there is a notion of "*identification by functional form*"(Li et al., 2019), where such non-linear models can be point identified even without any IVs based on restrictive parametric assumptions such as a bivariate probit. However, such identification has been described as "fragile" (Marra and Radice, 2011), as models such as the bivariate probit are restrictive

and hard to verify in practice. Once less restrictive assumptions are allowed, the IVs have been shown to play a crucial role for meaningful identification in partially identified models (see e.g., Chesher, 2005, 2010; Li et al., 2019; Shaikh and Vytlacil, 2011).

The literature on partially identified models offers a useful framework for the analysis of IV identification power. The identified set for the ATE, defined as the collection all possible values of the ATE from different observationally equivalent structures that can give rise to the observed data, offers an obvious measure for identification power. For example, Kitagawa (2009) and Swanson et al. (2018) use the size of the identified set to measure the identification power of model assumptions. Naturally, the width of the ATE identified set can also provide a measure to examine the IV contribution to the identification gains. In this chapter, we use the reduction in the width of the identified set as a measure for identification gains. Since the pioneering work of Manski (1990), most of the ATE partial identification studies with an endogenous treatment have relied on the IVs to bound the ATE (see Chesher, 2010; Chiburis, 2010; Flores and Chen, 2018; Heckman and Vytlacil, 1999, 2001; Shaikh and Vytlacil, 2011; Vytlacil and Yildiz, 2007; Vuong and Xu, 2017). Both Chesher (2010) and Li et al. (2018) show that the existence and the strength of the IVs can significantly affect the identification of the ATE for discrete outcome models. However, the mechanism through which the IV strength translates to identification gains in such non-linear models has not been well understood by researchers.

In endogenous treatment effect models, the IVs exert their influence through their impact on the treatment propensity score. Heckman, Urzua, and Vytlacil (2006) provide a comprehensive study of the properties of IVs in models with continuous outcomes, and point out the central role of the propensity scores in such models. Other works that establish the important role of the propensity score include Rosenbaum and Rubin (1983), Heckman and Robb (1985, 1986), Heckman (1990), and Ahn and Powell (1993). In continuous outcome models, it is well known that the "identification at infinity", namely the existence of values of the IVs that can produce propensity scores of zero and one, leads to the point identification of the ATE (Heckman and Vytlacil, 1999, 2001; Imbens and Angrist, 1994). However, this condition is rarely guaranteed in practice, especially

when available IVs have limited variation. Thus, it is important to understand how the achievable variation of the conditional propensity scores determines the ability of the IVs to shrink the size of the ATE identified set.

The crucial role played by the IVs has also been noted for discrete outcome models. In particular, it is commonly accepted that Manski's ATE bounds (Manski, 1990), which employ no IVs and have the support of the "hypothetical propensity score" as an empty set, can be uninformative. Chesher (2010) has pointed out that the support and the strength of the IVs play an important role in determining the ATE bounds. Li et al. (2018) use a version of pseudo R^2 to measure IV strength and show that the ATE bound width decreases as the pseudo R^2 increases. As with linear models, it is natural to expect that the propensity score variation is also a key component that governs the ability of the IVs to identify the ATE. However, to the authors' knowledge, no rigorous examinations have yet been conducted to investigate the factors contributing to the identification gains of the ATE for discrete outcome models when "identification at infinity" fails. It is part of the purpose of this chapter to investigate this lacuna.

This chapter presents a rigorous examination of the role of IVs and their interplay with other factors in the identification gains for the ATE in binary outcome models with an endogenous binary treatment. Using the bivariate joint threshold crossing model proposed by Shaikh and Vytlacil (2011) (henceforth referred to as the SV model or SV bounds) as an example, we study the identification gains achieved by the SV bounds against those from an ATE bounds benchmark, the bounds of Manski (1990) (hereafter Manski bounds). The rationale for using Manski's bounds as a benchmark follows from the observation that if the IVs are irrelevant, then the SV bounds collapse to Manski bounds. Using this framework, we disentangle the various impacts of IVs on identification gains, which yields a novel decomposition of the ATE SV bounds identification gains. This decomposition provides useful insights into the different sources and nature of identification gains.

Our chapter makes several contributions. Firstly, we distinguish the concepts of *IV strength* and *IV identification power* for binary dependent variables models. It can be shown that, as in the case of linear models, the IV strength, measured by the range of the

conditional propensity score (CPS) that are attributable to the IVs, plays a crucial role in the identification gains when bounding the ATE. More importantly, we demonstrate that unlike linear models, the IV identification power is also determined by the interplay of the IVs with the sign and the degree of treatment endogeneity. This is because in such non-linear models, the ATE bounds are governed by the joint probabilities of the outcome and the treatment, which are non-linear functions of the endogeneity degree. Thus, the same information contained in the IVs may be correspondingly scaled up or down via the leverage induced by the endogeneity. Therefore, the conventional notion of IV strength no longer provides the full picture of IV identification power, and is not the sole arbiter of instrument usefulness. Our second contribution is to propose a novel decomposition of the identification gains into three components. These components are governed by the IV validity, the IV strength, and the impact of the exogenous covariates via matching. The proposed decomposition of the ATE bounds is implemented by comparing the SV bounds (Shaikh and Vytlacil, 2011) to the benchmark of the Manski bounds (Manski, 1990), and by disentangling the different sources of the overall identification gains. This allows us to analyse the ATE partial identification mechanism and to thereby characterise the structure of the overall identification gains.

Based on the decomposition, the third contribution of this chapter is to propose a designated measure for the instrument identification power (hereafter IIP). The IIP measures the IV contribution to identification gains by quantifying the reduction in the size of the ATE identified set that can be attributed to the instruments alone. Works that aim to provide measures of the explained variation in limited dependent variable models, such as Veall and Zimmermann (1992, 1996), are already available and Windmeijer (1995) provides a comprehensive review of various pseudo R^2 goodness-of-fit measures. In general, pseudo R^2 statistics are developed for single equation limited dependent variable models, rather than for triangular systems with a binary endogenous treatment. Although such pseudo R^2 statistics will yield a measure of the *IV strength* (as used in Li et al., 2018), they are not appropriate measures for *IV identification power*, as they fail to capture the critical fact that the IV identification information pertaining to the ATE varies with the endogeneity degree. Consequently, any suggestion that pseudo R^2 statistics will be an indicator of the IV identification power would be misplaced. In contrast,

the IIP proposed in this chapter is specifically designed to evaluate the identification gains that can be solely attributed to the IVs.

Finally, our study also provides potential insights into the literature on instrument relevancy, weak instruments and instrument selection. The importance of this IIP measure is that it enables a ranking of alternative IVs by their identification power, thereby offering a potential criterion for detection of irrelevant IVs and for selection of sets of IVs for constructing the ATE bounds. In this way, our measure is akin to existing approaches in the generalised methods of moment (GMM) literature that seek to determine instrument "relevancy". The ability of our approach to determine and rank sets of IVs by their identification gains leads us to document, we believe for the first time, a critically important feature of binary triangular equations systems: while in the population, adding irrelevant IVs can not increase the IV identification power, in finite-samples, using such IVs to partially identify the ATE could lead to a loss in IV identification power, which may result in wider ATE bounds especially when the variation of covariates is limited. We liken this phenomena to the well-known problem of irrelevant moment conditions in GMM (see Breusch et al., 1999; Hall and Peixe, 2003; Hall, 2005; Hall et al., 2007, among others) and leave a more rigorous study of this topic for future research.

3.2. Model Setup and the ATE SV Bounds

Following the potential outcome framework, let Y be a binary outcome such that

$$Y = DY_1 + (1 - D)Y_0,$$

where $D \in \{0, 1\}$ is a treatment indicator with $D = 1$ denoting being treated and $D = 0$ denoting being untreated. The pair $Y_0, Y_1 \in \{0, 1\}$ are two potential outcomes in the untreated and treated states. We observe (Y, D, X, Z), where X denotes a vector of exogenous covariates and Z represents a vector of instruments that can be either continuous or discrete. Suppose we are interested in the conditional ATE, defined as

$$\text{ATE}(x) = \mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x].$$

Because only one of the potential outcomes is observed, we are faced with a missing data problem. If the potential outcomes are independent of the treatment D then it can be shown that the ATE(x) is point identified. However, in many empirical studies D is endogenous and hence correlated with the potential outcomes. Nevertheless, with the help of IVs we may partially identify the ATE(x) and construct an identified set for the ATE under mild conditions that are satisfied by a wide range of data generating processes.

For notational simplicity, henceforth we will use $\Pr(A|w)$ to represent $\Pr(A|W = w)$ for any event A, random variable W and its possible value w unless otherwise stated. For any generic random variables A and B, the support of A is denoted as Ω_A and the support of A conditional on $B = b$ is given by $\Omega_{A|b}$. Let $F_{A,B}$ denote the joint cumulative distribution function (CDF) of (A, B), F_A the marginal CDF of A, and $F_{A|B}$ the conditional CDF of A given B. Corresponding density functions will be denoted using a lower case f with associated subscript in an obvious way.

We now introduce the model and the identified set of the ATE studied in Shaikh and Vytlacil (2011), based on which we explore the factors determining the ATE bounds and how they impact the ATE bound width. Consider a joint threshold crossing model

$$\begin{aligned} Y &= 1[\nu_1(D, X) > \varepsilon_1], \\ D &= 1[\nu_2(X, Z) > \varepsilon_2], \end{aligned} \tag{3.1}$$

where $\nu_1(\cdot, \cdot)$ and $\nu_2(\cdot, \cdot)$ are unknown functions, and $(\varepsilon_1, \varepsilon_2)'$ is an unobservable error term with joint CDF $F_{\varepsilon_1,\varepsilon_2}$. Threshold crossing models are often used in treatment evaluation studies (see Heckman and Vytlacil, 1999, 2001, for example), and have been shown to be informative in the sense that the sign of the ATE can be recovered from the observable data, and the ATE can even be point identified in certain circumstances; see Shaikh and Vytlacil (2005, 2011), Vytlacil and Yildiz (2007) and Vuong and Xu (2017) among others. Bhattacharya et al. (2012) demonstrate that the SV bounds still hold under a rank similarity condition, a weaker property that allows heterogeneity in the sign of the ATE(x). Furthermore, as mentioned in Vytlacil and Yildiz (2007), it is possible to achieve the ATE point identification via the SV bounds if X contains a

continuous element or the exclusion restriction holds in both equations. Moreover, tests for the applicability of threshold crossing also have been developed (Bhattacharya et al., 2012; Heckman and Vytlacil, 2005; Kitagawa, 2015; Machado et al., 2013) for example. The following assumption summarises the conditions imposed by Shaikh and Vytlacil (2011).

Assumption 3.2.1 *The model in* (3.1) *is assumed to satisfy the following conditions:*

(a) The distribution of error term $(\varepsilon_1, \varepsilon_2)'$ *has a strictly positive density with respect to the Lebesgue measure on* $\mathbb{R}^2$.

(b) (X, Z) *is independent of* $(\varepsilon_1, \varepsilon_2)$.

(c) The distribution of $\nu_2(X, Z)|X$ *is non-degenerate.*

(d) The support of the distribution of (X, Z), $\Omega_{X,Z}$, *is compact.*

(e) $\nu_1 : \Omega_{D,X} \to \mathbb{R}$, $\nu_2 : \Omega_{X,Z} \to \mathbb{R}$ *are continuous in both arguments.*

Assumption 3.2.1 ensures that the instruments in Z satisfy the exclusion restriction, is independent of the error term $(\varepsilon_1, \varepsilon_2)'$ and relevant to the treatment D. In addition, Assumption 3.2.1 (a) and (b) are such that Z enters the outcome Y only through the propensity score, which is called index sufficiency. Conditions (d) and (e) are required to establish the sharpness of the identified set, and are imposed for analytical simplicity.

Denote random variable $P = \Pr[Y = 1|X, Z]$ with support Ω_P. Under Assumption 3.2.1 (a)-(c), Shaikh and Vytlacil (2011) show that the sign of the ATE(x) is identified: for any p and p' in Ω_P such that $p > p'$,

$$\text{sgn}[\text{ATE}(x)] = \text{sgn}[\nu_1(1, x) - \nu_1(0, x)] = \text{sgn}\left[\Pr[Y = 1|x, p] - \Pr[Y = 1|x, p']\right], \quad (3.2)$$

where sgn$[\cdot]$ is the conventional signum function. Given (3.2), it is apparent that the sign of the ATE(x) is recovered from the observables if Z is valid in the sense that Z is independent to $(\varepsilon_1, \varepsilon_2)$ and it has nonzero prediction power for the treatment, meaning that there exist two different values of $p, p' \in \Omega_{P|x}$ such that $p = \Pr[D = 1|x, z]$ and

$p' = \Pr[D = 1|x, z']$.

More importantly, Assumption 3.2.1 is sufficient to construct bounds for the ATE, referred to as SV bounds. Let P and P' are two independent random variables with the same distribution, and let x, x' be any two values in Ω_X. Now, define $H(x, x') = \mathbb{E}[h(x, x', P, P')|P > P']$ where

$$\begin{aligned} h(x, x', p, p') =& \Pr[Y = 1, D = 1|x', p] - \Pr[Y = 1, D = 1|x', p'] \\ & - \Pr[Y = 1, D = 0|x, p'] + \Pr[Y = 1, D = 0|x, p]. \end{aligned}$$

Let $\mathbf{X}_{0+}(x) = \{x' : H(x, x') \geq 0\}$, $\mathbf{X}_{0-}(x) = \{x' : H(x, x') \leq 0\}$, $\mathbf{X}_{1+}(x) = \{x' : H(x', x) \geq 0\}$, and $\mathbf{X}_{1-}(x) = \{x' : H(x', x) \leq 0\}$. Then the SV lower bound is

$$\begin{aligned} L^{SV}(x) = & \sup_{p \in \Omega_{P|x}} \left\{ \Pr[Y = 1, D = 1|x, p] + \sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[Y = 1, D = 0|x', p] \right\} \\ & - \inf_{p \in \Omega_{P|x}} \left\{ \Pr[Y = 1, D = 0|x, p] + p \inf_{x' \in \mathbf{X}_{0+}(x)} \Pr[Y = 1|x', p, D = 1] \right\}, \end{aligned} \tag{3.3}$$

and the SV upper bound is

$$\begin{aligned} U^{SV}(x) = & \inf_{p \in \Omega_{P|x}} \left\{ \Pr[Y = 1, D = 1|x, p] + (1 - p) \inf_{x' \in \mathbf{X}_{1-}(x)} \Pr[Y = 1|x', p, D = 0] \right\} \\ & - \sup_{p \in \Omega_{P|x}} \left\{ \Pr[Y = 1, D = 0|x, p] + \sup_{x' \in \mathbf{X}_{0-}(x)} \Pr[Y = 1, D = 1|x', p] \right\}. \end{aligned} \tag{3.4}$$

The SV bounds in (3.3) and (3.4) consist of two layers of intersection evaluations. The first layer is to intersect all possible values of the conditional propensity score, or equivalently, of the IVs. The second layer is to utilise the identifying information contained in covariates. In particular, for given x, the second layer of intersections are taken over values of the covariates other than x, say x', which lies in a certain subset of Ω_X, and there exists a $z' \in \Omega_{Z|x}$ such that $p = \Pr[D = 1|x, z] = \Pr[D = 1|x', z']$. Thus, both the IVs and the covariates contribute to the identification gains of SV bounds. It is understood that in (3.3) and (3.4) the supremum and infimum operators are only taken over regions where all conditional probabilities are well defined. The probabilities $\Pr[Y = y, D = d|x', p]$ and $\Pr[Y = y|x', p, D = d]$ are well defined for $y \in \{0, 1\}$ and $d \in \{0, 1\}$, if there exists

a value $z' \in \Omega_{Z|x}$ such that $\Pr[D = 1|x', z'] = p$. The supremum over an empty set is defined as 0, and the infimum over an empty set is defined as 1. Given (3.3) and (3.4), the width of SV bounds can be defined as

$$\omega^{SV}(x) = U^{SV}(x) - L^{SV}(x)\,.$$

In the next section, we study the factors that impact the SV bounds and $\omega^{SV}(x)$.

3.3. The Determinants of ATE Bounds

As discussed in the introduction, for binary dependent variables the propensity of being treated is a key factor that carries the identification information in the IVs. Therefore, we start from the conditional propensity score (CPS) of the treatment, defined as $\Pr[D = 1|X = x, Z]$, which is a random variable (function) of IV Z, and study the features of the CPS that are crucial in determining the SV bound width.

3.3.1. The Conditional Propensity Score

In the following proposition, for the sake of completeness, we first restate the sharpness result in Shaikh and Vytlacil (2011) under a stronger support condition $\Omega_{X,P} = \Omega_X \times \Omega_P$, and then introduce our new results about the connections between $P = \Pr[D = 1|X, Z]$ and the SV bound width. Denote the two extreme values of the support of variable P by $\underline{p} := \inf\{p \in \Omega_P\}$ and $\overline{p} := \sup\{p \in \Omega_P\}$ respectively.

Proposition 3.3.1 *Let Assumption 3.2.1 hold. If $\Omega_{X,P} = \Omega_X \times \Omega_P$, then the SV bounds in (3.3) and (3.4) are sharp. In addition, for any given $\forall x \in \Omega_X$,*

(a) $L^{SV}(x)$ is weakly increasing as $\underline{p}$ decreases or as $\overline{p}$ increases;

(b) $U^{SV}(x)$ is weakly decreasing as $\underline{p}$ decreases or as $\overline{p}$ increases;

and hence

(c) $\omega^{SV}(x)$ is weakly decreasing as $\underline{p}$ decreases or as $\overline{p}$ increases.

Notice that under the restriction $\Omega_{X,P} = \Omega_X \times \Omega_P$, the support of P is the same to the support of the CPS $\Pr[D = 1|X = x, Z]$ for $\forall x \in \Omega_X$. Proposition 3.3.1 shows that the locations of the lower and upper SV bounds are determined by the extreme values of the CPS, i.e. $\underline{p}$ and $\overline{p}$. Moreover, the width of the SV bounds $\omega^{SV}(x)$ weakly decreases as the support of the CPS "expands". It means that when the IVs are good predictors of the treatment status, the identified set of the ATE(x) (SV bounds) is likely to be informative.

The feature revealed by Proposition 3.3.1 is significant. It indicates that in partially identified models with binary dependent variables, the property of IVs that determines their contribution to identification gains is different from that which has hitherto been held to be important. Key ingredients of conventional measures of IV strength are the correlation between the IVs and the endogenous regressors (as evaluated via the first-stage F-statistic for continuous endogenous regressors, or the pseudo-R^2 for binary response variables), as well as the variation of the IVs to that of the random noise. However, Proposition 3.3.1 indicates that two IV sets that have the same CPS end points will make identical contributions to identification gains when partially identifying the ATE, irrespective of their correlation with the endogenous regressors or their variability.

The restriction $\Omega_{X,P} = \Omega_X \times \Omega_P$ in Proposition 3.3.1 is utilised in Shaikh and Vytlacil (2011) to simplify the expression of the SV bound and to prove the sharp result. It is also one of the sufficient conditions that ensures global identification in a parametric triangular system model with binary endogenous treatment, see Han and Vytlacil (2017) Theorem 5.1. Without $\Omega_{X,P} = \Omega_X \times \Omega_P$, the SV bound need not be sharp. Chiburis (2010) shows that under joint threshold crossing the sharp ATE bounds can only be implicitly determined by a copula, so that neither a closed form expression nor a computationally feasible linear programming algorithm that solves this problem exists. We therefore maintain the support restriction. The condition $\Omega_{X,P} = \Omega_X \times \Omega_P$ is saying that for any $x, x' \in \Omega_X$, we have $\Omega_{P|x} = \Omega_{P|x'}$; i.e. there exist possible realisations z, z' of Z such that $\Pr[D = 1|x, z] = \Pr[D = 1|x', z']$, which might fail to hold in practice especially when the variation in Z is limited. One sufficient condition for $\Omega_{X,P} = \Omega_X \times \Omega_P$ to hold is that X is mean independence of D given Z. The necessity of the condition

$\Omega_{X,P} = \Omega_X \times \Omega_P$ here is that without this support restriction, the SV bound may not exhibit a monotonic relationship with the extreme values of the CPS.

Fortunately, although Proposition 3.3.1 is derived using the support constraint, from the simulations in Section 3.7 we can see that the SV bound width decreases, on average, whenever the extreme values of the CPS changes to their endpoints (zero and one). In fact, as we will now show, without the imposition of the support condition $\Omega_{X,P} = \Omega_X \times \Omega_P$, a "widest bound" under Assumption 3.2.1 that restricts the size of $\omega^{SV}(x)$ can be derived for any given $x \in \Omega_X$. Define the two extremes of the CPS as $\underline{p}(x) := \inf_{z \in \Omega_{Z|x}}\{p \in \Omega_{P|x,z}\}$ and $\overline{p}(x) := \sup_{z \in \Omega_{Z|x}}\{p \in \Omega_{P|x,z}\}$.

Proposition 3.3.2 *Let Assumption 3.2.1 hold. There exists a function $\overline{\omega} : \Omega_X \mapsto [0,1]$ such that $0 \leq \omega^{SV}(x) \leq \overline{\omega}(x)$ for any given $x \in \Omega_X$. In addition,*

$$\text{if } ATE(x) > 0\,, \text{ then } \overline{\omega}(x) = Pr\left[Y = 1, D = 1 | x, \underline{p}(x)\right] + Pr\left[Y = 0, D = 0 | x, \overline{p}(x)\right] ;$$
$$\text{if } ATE(x) < 0\,, \text{ then } \overline{\omega}(x) = Pr\left[Y = 1, D = 0 | x, \overline{p}(x)\right] + Pr\left[Y = 0, D = 1 | x, \underline{p}(x)\right] .$$

Moreover, $\overline{\omega}(x)$ is weakly decreasing as $\underline{p}(x)$ decreases or as $\overline{p}(x)$ increases.

The explicit expressions of the widest bounds, with width $\overline{\omega}(x)$, can be found in (3.14) and (3.16); see the proof of Proposition 3.3.2. From Proposition 3.3.2 we can see that $\overline{\omega}(x)$ is monotone in the extreme values of CPS, i.e. $(\underline{p}(x), \overline{p}(x))$, and we are able to conclude that the extreme values of the CPS govern the size of the SV bound width even without the support restriction. Moreover, under the extreme case of perfect prediction, Proposition 3.3.2 implies that the ATE(x) is point identified by the SV bounds. Suppose $p^*, p^{**} \in \Omega_{P|x}$ are such that $\Pr[D = 0|x, p^*] = 1$ and $\Pr[D = 1|x, p^{**}] = 1$. By the definition of $\underline{p}(x), \overline{p}(x)$, we have that $p^* = \underline{p}(x)$ and $p^{**} = \overline{p}(x)$. Proposition 3.3.2 then yields that $\overline{\omega}(x) = 0$ whatever the sign of the ATE(x), indicating that the ATE(x) is point identified. From the above discussion it is apparent that perfect prediction in the binary dependent variables model is equivalent to "identification at infinity". Similar discussion can also be found when partially identifying the ATE in models with discrete outcomes in Chesher (2010).

3.3.2. The Degree of Endogeneity

The importance of IVs in determining the ATE bounds via the CPS has been recognised in several studies, but it seems that another crucial determinant, the degree of endogeneity, has so far received little attention. The ATE bounds are constructed using the joint probabilities of the outcome and the treatment, and the IVs affect those joint probabilities not only directly through the CPS but also indirectly through the co-movements of the outcome and the treatment due to the endogeneity. Thus, it is reasonable to expect that the information contained in the IVs may be correspondingly scaled via the leverage induced by the degree of endogeneity.

To facilitate obtaining interpretable relationships between the degree of endogeneity and the SV bound width, we introduce a family of bivariate single parameter copulae that specifies the joint distribution of the stochastic error terms in (3.1), while we do not require the copula nor the marginal distributions to be known. Denote a copula as $C(\cdot,\cdot;\rho) : (0,1)^2 \mapsto (0,1)$, where $\rho \in \Omega_\rho$ is a scalar dependence parameter that fully describes the joint dependence between ε_1 and ε_2, and their dependence increases as ρ increases. In the special case of a normal bivariate probit model ρ represents the correlation between the error terms and $\Omega_\rho = (-1,1)$, but the parameter space of ρ is not necessary $(-1,1)$. It differs along with the copula. It is worth noting that in our setting, for any given copula, the dependence parameter ρ can be understood as indicating the level of endogeneity. We also impose additional dependence structure, the concordance ordering, on the copula $C(\cdot,\cdot;\rho)$. Let $F_{\varepsilon_1,\varepsilon_2}$ and $\widetilde{F}_{\varepsilon_1,\varepsilon_2}$ be two distinct CDFs. Following Joe (1997), we define $\widetilde{F}_{\varepsilon_1,\varepsilon_2}$ as being *more concordant* than $F_{\varepsilon_1,\varepsilon_2}$, denoted by $F_{\varepsilon_1,\varepsilon_2} \prec_c \widetilde{F}_{\varepsilon_1,\varepsilon_2}$, as

$$F_{\varepsilon_1,\varepsilon_2}(e_1,e_2) \leq \widetilde{F}_{\varepsilon_1,\varepsilon_2}(e_1,e_2), \quad \forall\ (e_1,e_2) \in \mathbb{R}^2.$$

For $\rho_1 \neq \rho_2$ and $u_1, u_2 \in (0,1)^2$, we say that the copula $C(\cdot,\cdot;\rho)$ satisfies the concordant ordering with respect to ρ, denoted as $C(u_1,u_2;\rho_1) \prec_c C(u_1,u_2;\rho_2)$, if

$$C(u_1,u_2;\rho_1) \leq C(u_1,u_2;\rho_2), \text{ for any } \rho_1 < \rho_2. \tag{3.5}$$

The concordant ordering with respect to ρ is a stochastic dominance restriction. The concordant ordering is embodied in many well-known copulae, including the normal copula; see Joe (1997) Section 5.1 for the copulae families where (3.5) holds. Similar stochastic dominance conditions are employed in, for example, Han and Vytlacil (2017) and Han and Lee (2019), to derive identification and estimation results for the parametric bivariate probit model and its generalisations.

Assumption 3.3.1 *The joint distribution of* $(\varepsilon_1, \varepsilon_2)'$ *is given by a member of the single parameter copula family* $F_{\varepsilon_1,\varepsilon_2}(e_1, e_2) = C(F_{\varepsilon_1}(e_1), F_{\varepsilon_2}(e_2); \rho)$, *for* $(e_1, e_2) \in \mathbb{R}^2$, *where* $C(\cdot, \cdot; \rho)$ *satisfies the concordant ordering with respect to* ρ.

Assumption 3.3.1 defines a class of data generating processes that is sufficient for us to establish the relationship between endogeneity as captured by the dependence parameter ρ, and the widest SV bound width $\overline{\omega}(x)$. The derivation of the following proposition does not require the copula $C(\cdot, \cdot; \rho)$ nor the marginal distributions F_{ε_1} and F_{ε_2} to be specified.

Proposition 3.3.3 *Under Assumptions 3.2.1 and 3.3.1, the widest SV bound width* $\overline{\omega}(x)$ *is weakly increasing in* ρ *when* $ATE(x) > 0$, *and* $\overline{\omega}(x)$ *is weakly decreasing in* ρ *when* $ATE(x) < 0$.

Proposition 3.3.3 implies that the (widest) SV bound width could be significantly impacted by the degree of endogeneity, even if the extreme values of the CPS are fixed. In addition, Proposition 3.3.3 also reveals that the effect of endogeneity is asymmetric. To be more specific, with a positive treatment effect negative endogeneity helps narrow down the ATE bound width, while the opposite holds true for a negative treatment effect. Therefore, when measuring IVs' identification gains in an ATE partial identification framework, both the sign and the strength of endogeneity play an important role. A set of "seemingly weak" IVs, judged from the first-stage estimation alone, may actually achieve significant identification gains if in a problem with certain sign and level of endogeneity, thus considered as having enough identification *power*. Conversely, a set of "seemingly strong" IVs can be surprisingly *powerless* due to an undesirable sign or degree of endogeneity, resulting in wide ATE bounds. Thus, the conventional tests for detecting IV strength, such as F-statistic and pseudo R^2, or the associated weak IV tests

designed for linear models, can be misleading in measuring IV identification power. The result here shows that *IV strength* is a different concept from the *IV identification power* in this binary model.

3.3.3. Covariate Support and Variability

As we have seen from the construction of the SV bounds in Section 3.2, both IVs and covariates contribute to identifying the ATE under model (3.1). It is perhaps not surprising to find that there are situations where covariates fail to further tighten the SV bounds, a feature previously noted in Chiburis (2010). This happens when, conditional on D, the covariates in X have no additional effects on the outcome Y, leading to $\omega^{SV}(x) = \overline{\omega}(x)$. The following proposition formalises these statements.

Proposition 3.3.4 *Let Assumption 3.2.1 hold. If the random variable $\nu_1(D, X)|D$ is degenerate, then $\omega^{SV}(x) = \overline{\omega}(x)$.*

Proposition 3.3.4 implies that any further reduction in the SV bound width from $\overline{\omega}(x)$ to $\omega^{SV}(x)$ can be attributed to the additional identification information in the covariate X. In particular, if focusing on the second layer of the intersections over $\mathbf{X}_{0+}(x), \mathbf{X}_{0-}(x), \mathbf{X}_{1+}(x)$ and $\mathbf{X}_{1-}(x)$ in bounds (3.3) and (3.4), we can see that such identification gain is extracted from the matching pair $(x, z), (x', z') \in \Omega_{X,Z}$ such that $\Pr[D = 1|x, z] = \Pr[D = 1|x', z']$. Thus, broader support and greater variability in X increases the probability of finding a matching pair.

To sum up, from the discussion in Section 3.3, we know that the identification power for the ATE SV bounds is determined by the extreme values of the CPS, the sign and the degree of endogeneity, and the variability (or support) of the covariates in the outcome equation.

3.4. Decomposing Identification Gains

Based on the discussions above, in this section we introduce a novel decomposition of the identification gains of the SV bounds. It disentangles the identification gains into

components that are attributable to the gains obtained from the IVs and the exogenous covariates. To construct the decomposition let us first introduce the benchmark ATE bounds of Manski (1990) (Manski bounds), which are obtained without reference to IVs and are given by

$$\begin{aligned} L^M(x) &= -\Pr[Y=1, D=0|x] - \Pr[Y=0, D=1|x], \\ U^M(x) &= \Pr[Y=1, D=1|x] + \Pr[Y=0, D=0|x], \end{aligned} \tag{3.6}$$

where (with obvious notations) $L^M(X)$ and $U^M(x)$ are the lower bound and upper bound respectively. From (3.6), it is apparent that the width of the Manski bounds, defined as $\omega^M(x) = U^M(x) - L^M(x)$, is one for any given $x \in \Omega_X$, with the lower bound and upper bound falling on either side of zero. Thus, $\left[L^M(x), U^M(x)\right]$ is uninformative as to the sign or location of the treatment effect, and it is often referred to in the literature as "the worst case scenario" (see Bhattacharya et al., 2012; Chiburis, 2010; Tamer, 2010, for example).

Our proposed decomposition of identification gains is inspired by the implications of the theoretical results in Section 3.3. For any given $x \in \Omega_X$, the decomposition consists of four components, denoted by $C_1(x)$ to $C_4(x)$ respectively. Each component corresponds to the identification gains made by the SV bounds over the benchmark Manski bounds.

(i) $C_1(x)$: **Contribution of IV Validity**. The first component of the identification gains is the reduction in the SV bound width relative to the benchmark Manski bound width, due to the identification of the ATE(x) sign. This contribution is accredited to IV validity, since by (3.2) we can identify the sign of the ATE(x) if the IVs are independent of the error term $(\varepsilon_1, \varepsilon_2)$ and $\nu_2(X, Z)|X$ is nondegenerate (or equivalently, if the IVs are valid) regardless of the IV strength.[1] For $\forall x \in \Omega_X$,

$$C_1(x) = 1[\text{ATE}(x) \leq 0]U^M(x) - 1[\text{ATE}(x) \geq 0]L^M(x),$$

which is equivalent to the width of the negative (positive) part of Manski bounds

[1] If ATE(x)=0 is identified by (3.2), i.e. $\Pr[Y=1|x,p] = \Pr[Y=1|x,p']$ for any $p > p'$, then it is obvious that the first contribution of SV bounds already leads to the point identification of the ATE(x), and the IV identification power $IIP(x)$, which will be introduced in Section 3.5, achieves its maximum value one.

if ATE(x) is identified to be positive (negative).

(ii) $C_2(x)$: **Contribution of IV Strength**. Conditional on the first component, IV validity, the second component captures to the further reduction achieved by the SV bound width via intersecting over all possible values of Z. This is reflected in the dependence of the SV bounds in (3.3) and (3.4) on the two extreme values of the CPS, and the closer the extreme values to $[0, 1]$ are, the greater is $C_2(x)$. Therefore, identification gains attributed to IV strength can be measured as

$$C_2(x) = \omega^M(x) - \overline{\omega}(x) - C_1(x).$$

(iii) $C_3(x)$: **Contribution of Covariates**. The third component is the incremental reduction in the SV bound width brought about by intersecting over all possible values of the exogenous covariates X that fall into the areas described by the sets $\mathbf{X}_{0+}(x), \mathbf{X}_{0-}(x), \mathbf{X}_{1+}(x)$ and $\mathbf{X}_{1-}(x)$ via matching for the same propensity score values. As implied by Proposition 3.3.4, this component is attributed to the variation of exogenous covariates:

$$C_3(x) = \overline{\omega}(x) - \omega^{SV}(x).$$

(iv) $C_4(x)$: **Remaining SV Bound Width**. The last component is due to the unobservable error terms, and relates to the remaining SV bound width that cannot be further reduced by the observable data under the SV modelling assumptions. This component can be thought of as the signal-to-noise ratio of the error terms. By construction, we have $C_4(x) = \omega^{SV}(x)$.

It is easy to see that $C_1(x) + C_2(x) + C_3(x) + C_4(x) = \omega^M(x) = 1$. If $\nu_2(X, Z)|X$ is degenerate and the IVs have no explanatory power for the treatment, then $C_1(x) = C_2(x) = C_3(x) = 0$ and the SV bounds reduce to Manski bounds. It is worth to note that although we do not decompose the identification gains based on the sign and the degree of endogeneity, the magnitude of all the four components varies with them. According to Proposition 3.3.3, the sign and the endogeneity degree affects $\overline{\omega}(x)$, which enters all four

components either directly or indirectly due that the summation of the four components is a fixed value one. In addition, $C_1(x)$ to $C_4(x)$ can always be identified and estimated from the data. In practice, once the model has been estimated (parametrically or non-parametrically), the estimates can be used to construct the decomposition. Detailed numerical illustrations and simulations of the decomposition are presented in Sections 3.6 and 3.7.

3.5. IV Identification Power (IIP)

By construction, the identification gains decomposition satisfies $C_1(x)+C_2(x)+C_3(x)+C_4(x) = \omega^M(x) = 1$, $\forall x \in \Omega_X$, with each $C_j(x)$ representing the proportion of total identification gains that can be attributed to the corresponding component. Based on the decomposition, we can then construct a quantitative measurement of IV identification power in the partial identification setting. Suppose Assumption 3.2.1 holds, bar condition (c). For $\forall x \in \Omega_X$, define the IV identification power $IIP(x)$ as

$$IIP(x) := \begin{cases} \omega^M(x) - \overline{\omega}(x), & \text{if } \nu_2(X,Z)|X=x \text{ is nondegenerate} \\ 0, & \text{if } \nu_2(X,Z)|X=x \text{ is degenerate} \end{cases} \tag{3.7}$$

where $\overline{\omega}(x)$ is the widest width of the SV bounds defined in Proposition 3.3.3. Setting $IIP(x) = 0$ when $\nu_2(X,Z)|X=x$ is degenerate is equivalent to setting $\overline{\omega}(x) = \omega^M(x) = 1$, meaning that the widest width of the SV bounds equates to the width of the benchmark Manski bounds because the IVs are irrelevant.[2] From the decomposition, we have $IIP(x) = C_1(x) + C_2(x)$ when the IVs are valid and relevant. Thus $IIP(x)$ represents the proportion of the identification gains that is due to the IVs alone and it can be viewed as an index of the IV identification power. The overall IV identification power can be obtained by taking the expectation of $IIP(x)$ over Ω_X, i.e. $\mathbb{E}_X[IIP(X)]$.

The following proposition formalises some important properties of $IIP(x)$ as an indicator of the IV identification power.

[2]The definition allows $IIP(x)$ to be discontinuous at $\Omega_{P|x} = p_x$ for some constant $p_x \in [0,1]$, i.e. when $\Omega_{P|x}$ is a singleton.

Proposition 3.5.1 *The index $IIP(x)$ lies in the unit interval $[0,1]$, and under Assumption 3.2.1 $IIP(x)$ has the following properties:*

(a) $IIP(x)$ always lies in $[0,1]$ and can identify whether at least one of the IVs used to achieve the SV bounds is relevant;

(b) $IIP(x) = 0$ if none of the IVs in Z are relevant, then the SV bounds reduce to the benchmark Manski bounds;

(c) $IIP(x) = 1$ if the IVs in Z have perfect predictive power for the treatment D (identification at infinity holds), in the sense that there exists a p^* and p^{**} in $\Omega_{P|x}$ such that $Pr[D=0|x,p^*] = 1$ and $Pr[D=1|x,p^{**}] = 1$. Moreover, the ATE(x) is point identified when $IIP(x) = 1$.

Proposition 3.5.1 indicates that $IIP(x)$ is a meaningful measure of IV usefulness for improving the ATE partial identification. Therefore, values of $IIP(x)$ can be compared, across different sets of IVs, or across different values of x given the same set of IVs, since they are standardised relative to the same baseline benchmark. $IIP(x)$ or $\mathbb{E}_X[IIP(X)]$ can also be compared across various studies if necessary. For example, $IIP(x) = 0.4$ can be interpreted as that the Manski bound width can be reduced by 0.4 by using instruments alone. In this sense, the measure of $IIP(x)$ is a meaningful measure independent of the specific SV bounds. Theoretically, the value of $IIP(x)$ should lie in $[0,1]$ and the width of Manski bounds is always one. Then $IIP(x)$ can be interpreted as the percentage points of the identification gains brought by the IVs. In finite sample settings where the estimated Manski bound width may on longer be exact one, the sample explanation can be obtained by computing the ratio $IIP(x)/\omega^M(x)$ using their associated estimates. In addition, the values of $IIP(x)$ at its end points are intuitively interpretable; $IIP(x) = 0$ identifies situations where the IVs are completely irrelevant, and, when the IVs are able to perfectly predict the treatment status (when identification at infinity holds,) $IIP(x) = 1$ and point identification of the ATE(x) is achieved.

Numerical analysis is used in Section 3.6 to illustrate the behaviour of $IIP(x)$ in a class of representative models. At this point we note that $IIP(x)$ ignores the component

of identification gains attributable to the exogenous covariates, namely $C_3(x)$. In view of the additivity of the identification gains decomposition, this neglect seems entirely reasonable since we know, from Section 3.3, that for a given degree of endogeneity and extremes of the CPS, the value of $\overline{\omega}(x)$ does not vary with the identification information contained by the covariates. This indicates that $IIP(x)$ is a measure of identification gains due to IVs alone, without the contribution of the additional identification power provided by the exogenous covariates. It measures the smallest identification gains relative to the benchmark Manski bound that can be achieved by a given set of IVs. More importantly, focusing on $IIP(x)$ introduces considerable computational simplification when comparing sets of IVs, as it avoids the second layer of the intersection bounds required to compute the SV bounds.

3.6. Numerical Illustration

In this section we illustrate numerically the theoretical results on the decomposition of SV bounds studied in Section 3.2, and how each component affects the SV bounds. We consider as our data generating process (DGP) a version of the model in (3.1) with a linear additive latent structure, which is similar to that studied in Li et al. (2019):

$$\begin{aligned} Y &= 1[\alpha D + \beta X + \varepsilon_1 > 0], \\ D &= 1[\gamma Z + \pi X + \varepsilon_2 > 0], \end{aligned} \tag{3.8}$$

where the exogenous regressor X and the IV Z are assumed mutually independent, without loss of generality, $X \sim \mathbb{N}(0,1)$ and $Z \in \{-1,1\}$ with $\Pr(Z=1) = 1/2$. In addition, $(X,Z)' \perp (\varepsilon_1, \varepsilon_2)$ where the error term $(\varepsilon_1, \varepsilon_2)$ is zero mean bivariate normal with unit variances and correlation ρ. For this specification, given the distribution of Z, there is a monotonic one-to-one mapping from the coefficient of the IV, γ, to the range of the conditional propensity score. We capture changes in the extreme values of the CPS using the parameter grid $\gamma = -4 : 0.2 : 4$. Different levels of endogeneity were explored using the grid $\rho = -0.99 : 0.05 : 0.99$. We set $\alpha = 1$ and $\pi = 0$ across all parameter settings. Under this DGP, the SV bound width is affected by α, β and the variation of the exogenous covariates. Since α and the distribution of X are held fixed, we select β from

the set $\{0.05, 0.25, 0.45\}$, so that changes in β capture the variation of the exogenous covariates given the distribution of X. Using the DGP as characterised by (3.8) we compute the SV bounds $[L^{SV}(x), U^{SV}(x)]$ and the Manski bounds $[L^{M}(x), U^{M}(x)]$ and implement the identification gains decomposition according to the true DGP. In what follows we present the outcomes obtained when $x = \mathbb{E}[X]$.

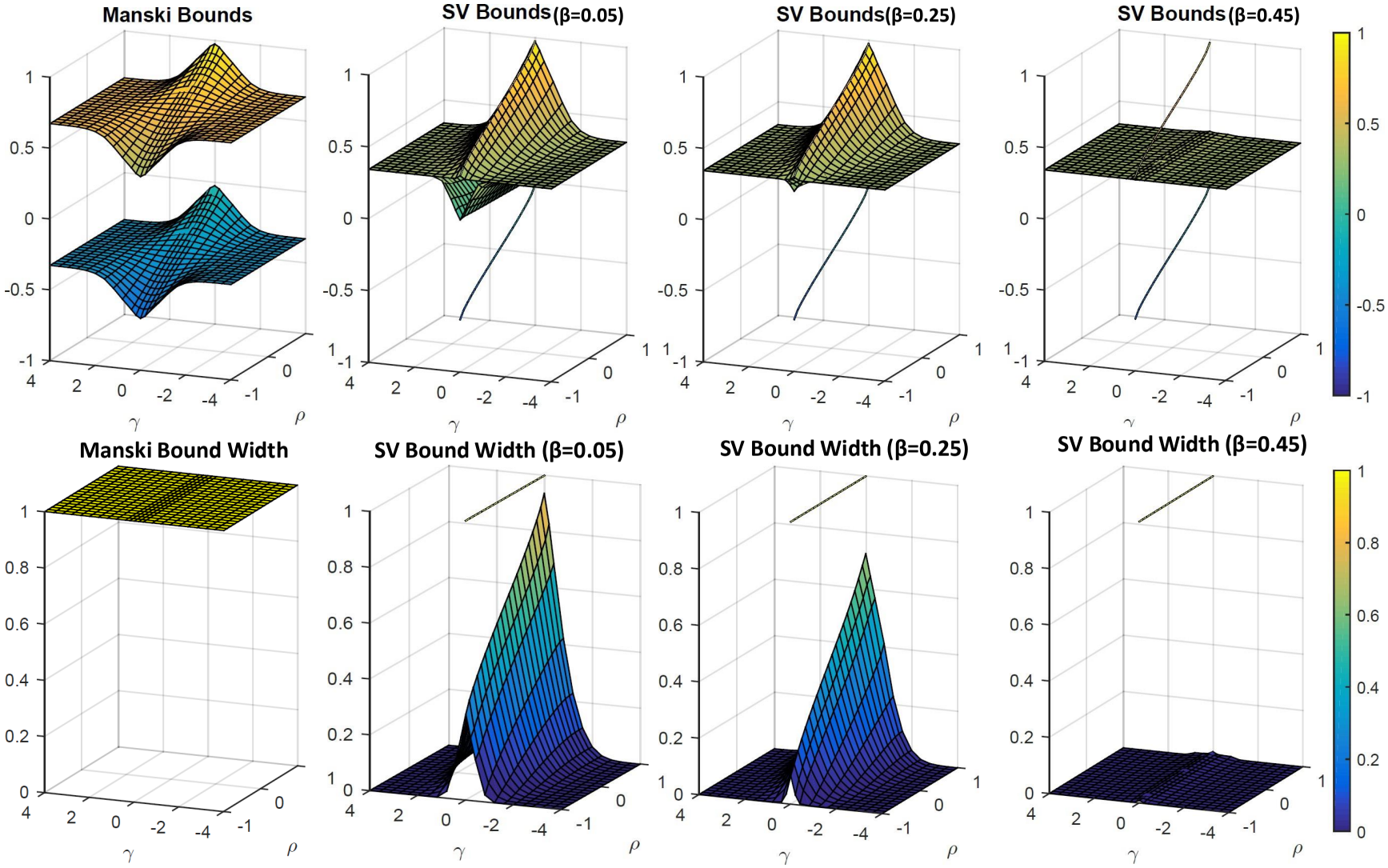
3.6.1. Determination of ATE Bounds

In Figure 3.6.1, the subplots in the first row display the upper and lower bounds of the ATE(x), and the subplots in the second row present the corresponding bound width. For the Manski bounds we can see that the width is always one, and the upper and lower bounds stand on either side of zero, as previously noted. The SV bounds reduce to the Manski bounds when the IVs are irrelevant with $\gamma = 0$ (the separate lines in the graphs at $\gamma = 0$). When γ moves away from $\gamma = 0$, the SV bound width has a significant drop. Then, as the magnitude of γ increases, i.e. as the ending points of the CPS expand, the SV bound width decreases. In addition, since $\alpha > 0$ and the ATE(x) is positive, the SV bound width increases as ρ increases. Moreover, comparison of the plots for different values of β reveals that β plays a critical role in determining the SV bounds in the sense that larger β produces significantly narrower bound width. When $\beta = 0.05$ the SV bound width is non-negligible when the absolute value of γ is small, while when $\beta = 0.45$, point identification of the ATE(x) is achieved for most of the (γ, ρ) pairs. These indicate that for a given IV strength, as measured by γ or the associated range of CPS, the lower the value of ρ in the $(-1, +1)$ range or the bigger the impact of x, the narrower the SV bounds that can be achieved. In other words, for given IV strength, a larger identification gain can be achieved if the error correlation ρ is large in magnitude and also has an opposite sign from the sign of the ATE(x).

3.6.2. Identification Gains Decomposition

The decomposition of identification gains obtained when $\gamma \in \{1, 2\}$, $\rho \in \{-0.8, -0.5, 0.5, 0.8\}$ and $\beta \in \{0.05, 0.25, 0.45\}$ is displayed for $x = \mathbb{E}[X]$ in Figure 3.6.2. We can see that when the ATE(x) is positive, the contribution of IV validity, as measured by $C_1(x)$, is determined by the Manski lower bound, and decreases as ρ increases (conversely the nu-

Figure 3.6.1: Manski and SV Bounds for ATE ($x = \mathbb{E}[X]$)



Note: Three dimensional plots of the ATE bounds as function of (γ, ρ). When $\gamma = 0$, SV bounds reduce to Manski bounds with bound width one.

merical results not reported here show that when the ATE(x) is negative $C_1(x)$ increases as ρ increases), while $C_1(x)$ is invariant to β. By way of contrast, the contribution of the component $C_2(x)$ also does not change by β, but it increases significantly as the magnitude of γ increases due to the impact of the IVs on the range of the CPS. The component of identification gains due to the exogenous covariates, $C_3(x)$, also contributes significantly to the identification gains. When β is relatively large (e.g. $\beta = 0.45$), the SV bound width is close to zero and point identification is virtually achieved.

3.6.3. IV Identification Power

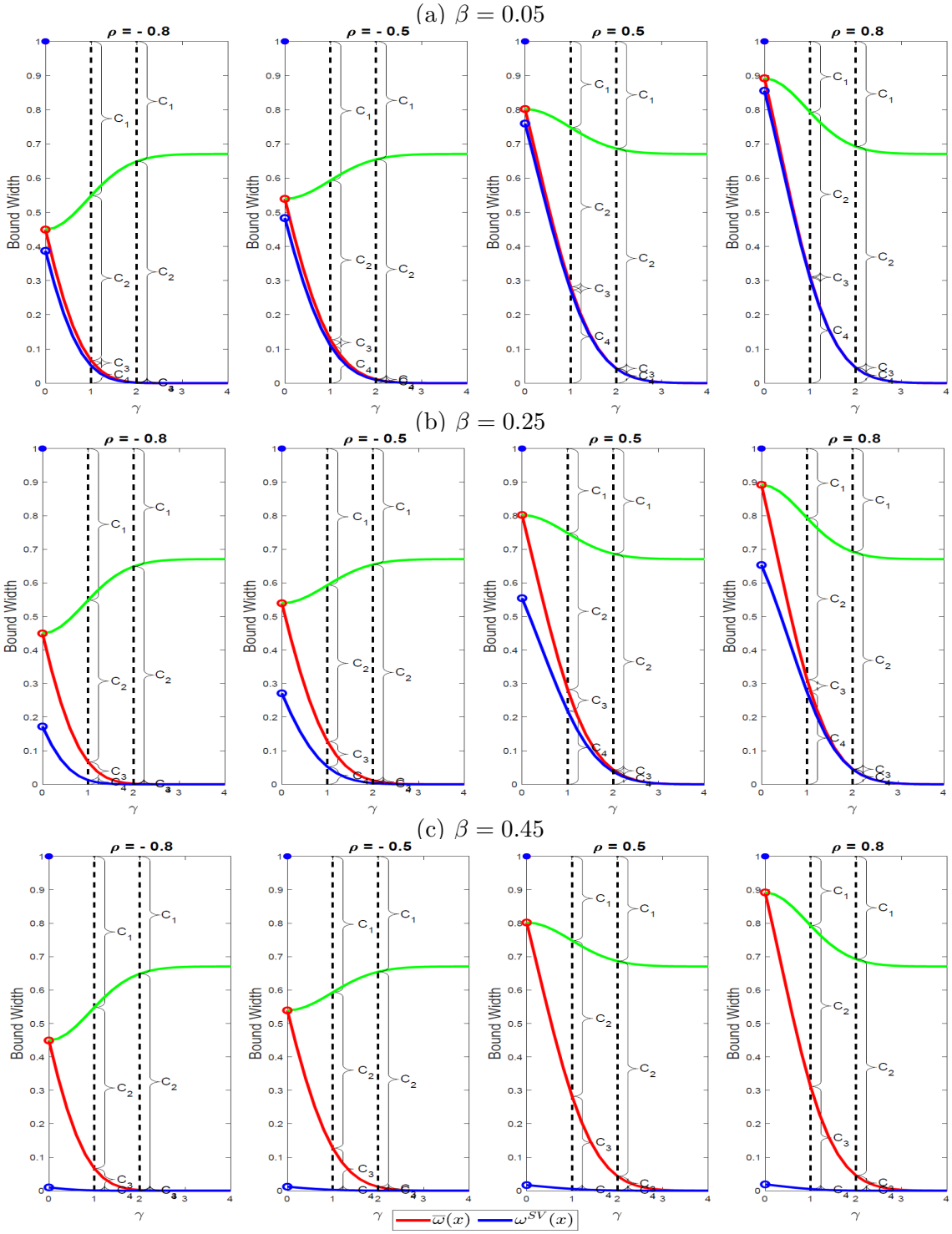
Figure 3.6.3 depicts the index $IIP(x)$ as a function of (γ, ρ) on the lattice $\{-4 : 0.2 : 4\} \times \{-0.99 : 0.05 : 0.99\}$. The plot confirms that, when the ATE(x) is positive, the IV identification power $IIP(x)$ increases as the IV strength ($|\gamma|$) increases, but for the same IV strength, the $IIP(x)$ is higher the lower the value of ρ. We also found, based on the results not reported here, that, when the ATE(x) is negative, a rising level of positive endogeneity drives up $IIP(x)$ and reduces the width of SV bounds.

By way of summary, the theoretical results presented in Sections 3.3, 3.4 and 3.5 are clearly reflected in the features observed in the numerical outcomes reported here. Firstly, $IIP(x)$ is bigger when IVs are stronger ($|\gamma|$ higher). In addition, for a given IV strength in the first-stage treatment equation, higher $IIP(x)$ can be achieved if the endogeneity ρ has an opposite sign from the ATE(x) and is of high magnitude ($|\rho|$). And if the endogeneity is of the same sign as the ATE(x), then the lower the degree of endogeneity the better the identification power. Of course adding the additional identification gain $C_3(x)$ to $IIP(x)$ leads to the SV bound width $\omega^{SV}(x)$, and the $C_3(x)$ depends on the properties of the covariates.

3.7. Finite Sample Evaluation of IV Strength and Relevance

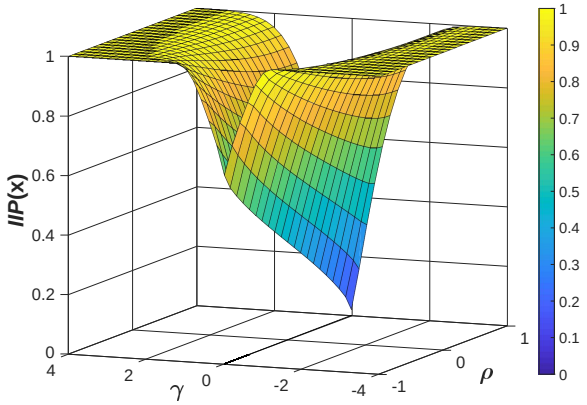
Next, we study the empirical performance of our decomposition analysis for alternative sets of IVs. We present finite sample results to show how $IIP(x)$ can be used to rank

Figure 3.6.2: Decomposition of Identification Gains ($x = \mathbb{E}[X]$)



Note: The green line depicts the amount of IV validity contribution $C_1(x)$. To aid legibility $C_1(x), \ldots, C_4(x)$ have been rendered as $C_1, \ldots, C_4$ in each of the subplots in this figure. x-axis displays the values of γ. For space limitation, we only represent the figure for nonnegative values of γ.

Figure 3.6.3: Instrument Identification Power ($x = \mathbb{E}[X]$)



Note: Three dimensional plot of $IIP(x)$ as function of (γ, ρ). The value of β does not affect the $IIP(x)$ in this case because $\pi = 0$ and no matches of $\Pr[D = 1|x, z]$ =$\Pr[D = 1|x', z']$ exist for $x = \mathbb{E}[X]$ and $z, z' \in \{-1, 1\}$. When $\gamma = 0$, the $IIP(x) = 0$ because IV is irrelevant.

the identification power of different sets of IVs and to potentially detect irrelevant IVs, when determining which set of IVs should be used to construct the ATE bounds. The advantage of this strategy over conventional IV *strength* evaluations (such as those akin to the first-stage IV F-statistic or the CPS) is that $IIP(x)$ captures the IV identification power in terms of their ability to shrink the width of the ATE bounds, incorporating the IV strength and their interaction with the direction and magnitude of endogeneity in the nonlinear model. The identification power $IIP(x)$ can provide testable implication of IV relevance, but a formal test is out of the scope of this chapter. Consider i.i.d. samples generated from a similar DGP to (3.8) with two IVs:

$$\begin{aligned} Y &= 1[\alpha D + \beta X + \varepsilon_1 > 0], \\ D &= 1[\pi X + \gamma_1 Z_1 + \gamma_2 Z_2 + \varepsilon_2 > 0] \end{aligned} \tag{3.9}$$

where two IVs in $Z = (Z_1, Z_2)'$ are $Z_1 \sim Bernoulli(1/2)$ and $Z_2 \in \{-3, -2, -1, 0, 1, 2, 3\}$ with probabilities $(0.1, 0.1, 0.2, 0.2, 0.2, 0.1, 0.1)$. Set $\alpha = 1$, $\beta = 1$, $\pi = -1$, $(\gamma_1, \gamma_2) = (0.5, 0.2)$, and assume the error term $(\varepsilon_1, \varepsilon_2)$ is jointly normal with mean zero, variance one and correlation $\rho \in \{0.5, 0.8\}$. In addition, Z_1, Z_2 and X are mutually independent, and also independent to $(\varepsilon_1, \varepsilon_2)$. Consider two cases of covariate variability: **case** 1, continuous $X \sim \mathbb{N}(0, 1)$; **case** 2, binary $X \sim Bernoulli(1/2)$. We conduct the analysis in this section at $x = 0$. The value of the ATE$(x) = \mathbb{E}[Y_1 - Y_0|X = 0]$ under the DGP (3.9) is 0.341.

In order to evaluate the finite sample performance of $IIP(x)$ as an index for measuring IV identification power, we consider five alternative sets of IV options. In addition to the two valid IVs of Z_1 and Z_2 in the DGP, we introduce two "pseudo" IVs: $\widetilde{Z}_2 = 1[Z_2 > 0]$, which is a misspecified binary IV that only partially reflects Z_2, and an irrelevant IV $Z_3 \in \{0, 1\}$ such that $\Pr[Z_3 = 1] = 2/3$, and $Z_3 \perp (\varepsilon_1, \varepsilon_2, Z_1, Z_2, X)$. To illustrate the behaviour of the $IIP(x)$ estimation, we use sample data for (Y, D, X) generated from the DGP in (3.9) to estimate models with five alternative IV sets: (1) only one valid IV Z_1 (omitting Z_2); (2) only one valid IV Z_2 (omitting Z_1); (3) one valid Z_1 and one misspecified $\widetilde{Z}_2$; (4) two valid IVs Z_1 and Z_2; and (5) two valid Z_1 and Z_2 plus one irrelevant Z_3.

Table 3.7.1: Population CPS Range and $IIP(x)$ ($x = 0$, cases 1 and 2)

Sets IVs	CPS definition	CPS Range	$IIP(x)$ ($\rho = 0.5$)	$IIP(x)$ ($\rho = 0.8$)
(1) only Z_1	$\Pr[D = 1\|x, Z_1]$	$[0.500, 0.682]$	0.305	0.232
(2) only Z_2	$\Pr[D = 1\|x, Z_2]$	$[0.367, 0.795]$	0.493	0.443
(3) $Z_1, \widetilde{Z}_2$	$\Pr[D = 1\|x, Z_1, \widetilde{Z}_2]$	$[0.410, 0.799]$	0.456	0.403
(4) Z_1, Z_2	$\Pr[D = 1\|x, Z_1, Z_2]$	$[0.274, 0.864]$	0.625	0.594
(5) Z_1, Z_2, Z_3	$\Pr[D = 1\|x, Z_1, Z_2, Z_3]$	$[0.274, 0.864]$	0.625	0.594

Note: The population CPS and $IIP(x)$ are the same for case 1 and case 2.

Table 3.7.1 presents the theoretical CPS range and $IIP(x)$ for the cases 1 and 2, at $x = 0$. Note that the covariate variability does not impact the population CPS nor $IIP(x)$, so that the values of CPS range and $IIP(x)$ for case 1 are the same to those for case 2. Looking at the CPS range as a measure of IV strength, we can see that the CPS range is the widest when both valid and relevant IVs Z_1 and Z_2 are used as in (4). Adding an irrelevant IV Z_3 does not change the theoretical CPS range, so theoretically (5) has the same IV strength as (4). The CPS range decreases when only one of the two valid IVs are used as in (1) and (2), with Z_2 being stronger with wider CPS range than Z_1. As expected, when a valid IV is incorrectly specified as a proxy dummy $\widetilde{Z}_2$ in (3), the CPS range is narrower than that of the best set in (4), but wider than that in (1) with Z_1 alone. Interestingly, comparing IV set (3) with (2), set (2) with only one valid IV actually results in wider CPS range than that for the two IVs in set (3) with Z_2 misspecified, though the CPS interval for (3) is not completely nested within the interval for (2).

Whilst the CPS range indicates the IV strength, it is the $IIP(x)$ that captures the identification power of each IV set, measuring the reduction of SV bound width relative to the benchmark Manski bound width due to the contribution of IVs. As seen from the two $IIP(x)$ columns in Table 3.7.1, the same IV strength can achieve bigger identification gains for $\rho = 0.5$ than that with $\rho = 0.8$. This is consistent with the results in Section 3.6: as ρ and ATE(x) are both positive in this case, the lower absolute value of ρ, the higher the $IIP(x)$ is. For example for IV set (4), the Manski bound width can be reduced by 0.594 (or 59.4%) by the two IVs when $\rho = 0.8$, and it increases to 0.625 (or 62.5%) if $\rho = 0.5$. The equally most powerful IV sets are (4) and (5), and the least powerful set is (1).

We next present the finite sample estimation of the Manski and SV bounds, and conduct the decomposition analysis based on the estimates of the bounds. Sample size is set to be $n = 500, 5000, 10000$ and replicate $M = 1000$ times. Tables 3.7.2 to 3.7.5 present the sample average (over M replications) of the estimated bounds, estimated $C_1(x)$ to $C_4(x)$ and $IIP(x)$ of the five IV sets at $x = 0$. We use the "half-median-unbiased estimator" (HMUE) of the intersecting bounds proposed by Chernozhukov, Lee, and Rosen (2013) (hereafter CLR) to estimate the benchmark Manski bounds and the SV bounds. In particular, we employ maximum likelihood estimation (MLE) to estimate the bounding functions and to select the critical values for bias correction according to the simulation-based methodology of CLR.[3]

The results of Tables 3.7.2 to 3.7.5 relate to the two different covariate distributions (**case** 1, $X \sim \mathbb{N}(0,1)$; **case** 2, $X \sim Bernoulli(1/2)$) and two ρ values ($\rho = 0.5, 0.8$). Let us look firstly the first row in each table, which lists the ATE bounds and decomposition components under the true DGP. We can see that in **case 1** (Tables 3.7.2 and 3.7.3),

[3]The CLR half-median-unbiased estimator produces a upper bound estimator that exceeds its true value and a lower bound estimator that falls below its true value, each with probability at least a half asymptotically. We report the HMUE of the Manski bounds, for comparison purpose. Other estimation methods for Manski bounds are also available; see for example Imbens and Manski (2004). Theoretically, the construction of the SV bounds requires the matching of pairs (x, z) and (x', z') such that $\Pr[D = 1|x, z] = \Pr[D = 1|x', z']$. In practice, it is hard to find such pairs with equal CPS especially when the variation of covariates is limited. In the simulations, the SV bounds are computed by matching (x, z) and (x', z') such that $|\Pr[D = 1|x, z] - \Pr[D = 1|x', z']| < c$ and $c = 1\%$. Although the estimated SV bounds depend on c, the estimated $IIP(x)$ does not. Therefore the choice of c has no impacts on the performance of the $IIP(x)$.

where the covariate possesses sufficient variation, the true SV bounds point identify the ATE(x) for both $\rho = 0.5$ and $\rho = 0.8$. In **case 2** (Tables 3.7.4 and 3.7.5), the true SV bounds fail to point identify the ATE(x) due to the limited variation in X.

Next, we focus on the left part of each table, which displays the HMUEs of the ATE bounds, and the Hausdorff distance between the true bounds and the estimated bounds, evaluated at $x = 0$. Simulation results of bounds at different values of x display similar patterns to those at $x = 0$, therefore are not reported due to the space limitation. The Hausdorff distance between sets A and B is defined as $\max\{\sup_{a\in A} d(a, B), \ \sup_{b\in B} d(b, A)\}$ where $d(b, A) := \inf_{a\in A} \|b - a\|$ and ∞ if either A or B is empty. Hausdorff distance is a natural generalisation of Euclidean distance and has been employed to study convergence properties when a set rather than a point is the parameter of interest; see e.g. Chernozhukov et al. (2007), Hansen et al. (1995) and Manski and Tamer (2002). For all four tables, we can see that the estimated Manski bounds are the same across all five IV sets, always include zero, and have a width a little over one. The estimated SV bounds identify the sign of ATE(x) for all five IV sets. Moreover, the IV sets with greater identification power lead to narrower estimated SV bounds and also improve the estimation accuracy in most of the scenarios. More precisely, the Hausdorff distance of the estimated SV bounds to the true bounds decreases as the IV identification power increases. Moving to the right part of each of table, first, we note that for each given IV set, all the estimated $C_1(x)$ to $C_4(x)$ and $IIP(x)$ converges to their true values as sample size n increases, indicating that the estimated identification gain is more accurate for larger sample size.[4] We also note that the estimated $C_1(x)$ which is determined by the Manski bounds, is the same for different IV sets. This result is quite intuitive because the identification gains brought by the IV validity should not vary with the IV strength. Comparison of Tables 3.7.2 and 3.7.3 or Tables 3.7.4 and 3.7.5 also reveals that the impacts of endogeneity degree on IV identification power can be captured by the estimated $IIP(x)$. Importantly, the true ranking of $IIP(x)$ as in Table 3.7.1 can be correctly revealed by finite sample estimates of $IIP(x)$.

[4]Because $C_1(x)$ to $C_4(x)$ are functions of $L^M(x)$, $U^M(x)$, $\overline{\omega}(x)$ and $\omega^{SV}(x)$, the estimates of $C_1(x)$ to $C_4(x)$ are computed using the HMUE of the bounds or their widths. We compute $\overline{\omega}(x)$ as the width of the estimated bounds (by HMUE of CLR) $[\underline{L}^{SV}(x), \overline{U}^{SV}(x)]$ in (3.14) if ATE(x) > 0 is identified, or (3.16) if otherwise.

It is interesting to analyse the effect of adding an additional but completely irrelevant IV on the finite sample performance of ATE partial identification, by comparing the results obtained using IV sets (4) and (5). Adding Z_3 to (Z_1, Z_2) actually produces a small *decrease* of the estimated $IIP(x)$, on average, for almost all different DGP designs considered in this section. The Cramer-Von Mises test and the Kolmogorov–Smirnov test confirm that the average values of the estimates of $IIP(x)$ under scenario (4) are significantly different from those obtained under scenario (5), when sample size is $n = 500$ and $n = 5000$ for both endogeneity degrees and for both case 1 and case 2. While when sample size is sufficiently large $n = 10000$, the estimates of $IIP(x)$ under scenario (4) and (5) are no longer significantly different, except for case 2 with $\rho = 0.8$. This suggests that in practice, the loss of information (efficiency) that arises from using irrelevant IV can have a statistically significant practical effect on the IV identification power, which can be captured by our proposed index $IIP(x)$. Such an information loss could lead to wider ATE bounds, especially when the covariate possesses limited variation. Particularly, from Table 3.7.4 and Table 3.7.5 we can see that when the covariate X is a binary variable (case 2), on average, the estimated SV bounds using (Z_1, Z_2) are significantly narrower than those estimated by the IV set including the irrelevant IV (Z_1, Z_2, Z_3), especially for small sample size. Analysing the results across the replications, we find that about 78% (for both endogeneity degrees) of the replications give narrower estimated SV bounds with IV set (Z_1, Z_2) than those with (Z_1, Z_2, Z_3), for sample size $n = 500$; and this rate becomes to 53% ($\rho = 0.5$) and 64% ($\rho = 0.8$) for sufficiently large sample size $n = 10000$.

On the other hand, the IV irrelevancy cannot always be detected by simply comparing the estimated SV bound width under different IV sets. That is, adding an irrelevant IV in (5) could further shrink the SV bound width when the covariate X is continuous, although the improvement happens at the third decimal and the degree of the improvement decreases as sample size increases.The shrinkage of the estimated SV bounds using the irrelevant Z_3 is due to the finite sample estimation error. In particular, because the estimates of the coefficient of the irrelevant Z_3 will be nonzero with probability one, it results in more matched pairs of (x, z) and (x', z') such that $|\Pr[D = 1|x, z] - \Pr[D = 1|x', z']| < c$ (see footnote 3) especially when covariate is continuous. For case 1 in Table 3.7.2 and Table 3.7.3, we find that when sample size is $n = 500$, (i) there are 22% ($\rho = 0.5$) and

17% ($\rho = 0.8$) of the 1000 replications where at least one (either lower or upper) estimated SV bound using (Z_1, Z_2) is closer to its true value, compared to that obtained by using the irrelevant IV; and (ii) 12% of the replications yield wider estimated SV bounds when using the irrelevant IV, for both endogeneity degrees. These outcomes reinforce *a-fortiori* the warning that simply adding extra IVs without assessing their identification power is unlikely to be a good practical modelling strategy, but the finite sample estimates of our proposed $IIP(x)$ is more reliable in detecting the loss of efficiency of IV irrelevancy.

Table 3.7.2: **Case 1**. True and Estimated Bounds, and Decomposition of Identification Gains ($\rho = 0.5,\ X \sim \mathbb{N}(0,1),\ x = 0$)

		Bounds				Decomposition				
		Manski		SV						
		$[L^M(x), U^M(x)]$	$d_H(x)$	$[L^{SV}(x), U^{SV}(x)]$	$d_H(x)$	$C_1(x)$	$C_2(x)$	$C_3(x)$	$C_4(x)$	IIP(x)
True DGP	Z_1, Z_2	$[-0.179, 0.821]$		$[0.341, 0.341]$		0.179	0.446	0.375	0.000	0.625
$n = 500$	(1) only Z_1	[-0.246,0.899]	0.092	[0.117,0.775]	0.434	0.246	0.186	0.056	0.658	0.432
	(2) only Z_2			[0.246,0.562]	0.227		0.342	0.241	0.316	0.587
	(3) $Z_1, \widetilde{Z}_2$			[0.193,0.759]	0.418		0.218	0.116	0.565	0.464
	(4) Z_1, Z_2			[0.290,0.455]	0.121		0.436	0.298	0.165	0.682
	(5) Z_1, Z_2, Z_3			[0.300,0.451]	0.116		0.424	0.324	0.151	0.670
$n = 5000$	(1) only Z_1	[-0.202,0.846]	0.030	[0.121,0.768]	0.427	0.202	0.145	0.053	0.648	0.347
	(2) only Z_2			[0.266,0.372]	0.078		0.334	0.406	0.106	0.536
	(3) $Z_1, \widetilde{Z}_2$			[0.221,0.757]	0.416		0.194	0.116	0.536	0.395
	(4) Z_1, Z_2			[0.312,0.377]	0.043		0.446	0.335	0.066	0.648
	(5) Z_1, Z_2, Z_3			[0.316,0.373]	0.038		0.442	0.347	0.057	0.644
$n = 10000$	(1) only Z_1	[-0.198,0.838]	0.022	[0.123,0.768]	0.427	0.198	0.139	0.054	0.645	0.337
	(2) only Z_2			[0.263,0.363]	0.080		0.331	0.407	0.101	0.528
	(3) $Z_1, \widetilde{Z}_2$			[0.225,0.756]	0.414		0.189	0.118	0.531	0.387
	(4) Z_1, Z_2			[0.317,0.365]	0.031		0.444	0.346	0.048	0.642
	(5) Z_1, Z_2, Z_3			[0.320,0.362]	0.027		0.443	0.353	0.042	0.641

Note: The estimated bounds, the Hausdorff distance $d_H(x)$ and the decompositions are the averages over 1000 replications.

Table 3.7.3: **Case 1**. True and Estimated Bounds, and Decomposition of Identification Gains ($\rho = 0.8,\ X \sim \mathbb{N}(0,1),\ x = 0$)

		Bounds				Decomposition				
		Manski		SV						
		$[L^M(x), U^M(x)]$	$d_H(x)$	$[L^{SV}(x), U^{SV}(x)]$	$d_H(x)$	$C_1(x)$	$C_2(x)$	$C_3(x)$	$C_4(x)$	IIP(x)
True DGP	Z_1, Z_2	$[-0.096, 0.904]$		$[0.341, 0.341]$		0.096	0.498	0.406	0.000	0.594
$n = 500$	(1) only Z_1	[-0.157,0.996]	0.098	[0.124,0.873]	0.532	0.157	0.205	0.041	0.750	0.362
	(2) only Z_2			[0.233,0.559]	0.229		0.382	0.288	0.326	0.539
	(3) $Z_1, \widetilde{Z}_2$			[0.191,0.848]	0.507		0.246	0.093	0.657	0.403
	(4) Z_1, Z_2			[0.291,0.437]	0.107		0.495	0.355	0.146	0.652
	(5) Z_1, Z_2, Z_3			[0.298,0.431]	0.100		0.482	0.382	0.133	0.639
$n = 5000$	(1) only Z_1	[-0.121,0.924]	0.028	[0.128,0.860]	0.519	0.121	0.149	0.042	0.732	0.271
	(2) only Z_2			[0.254,0.357]	0.088		0.346	0.475	0.103	0.467
	(3) $Z_1, \widetilde{Z}_2$			[0.208,0.853]	0.512		0.210	0.068	0.645	0.332
	(4) Z_1, Z_2			[0.312,0.378]	0.043		0.489	0.369	0.066	0.610
	(5) Z_1, Z_2, Z_3			[0.315,0.373]	0.038		0.486	0.380	0.058	0.607
$n = 10000$	(1) only Z_1	[-0.117,0.918]	0.022	[0.129,0.860]	0.519	0.117	0.146	0.042	0.731	0.263
	(2) only Z_2			[0.258,0.357]	0.083		0.346	0.473	0.099	0.463
	(3) $Z_1, \widetilde{Z}_2$			[0.212,0.851]	0.510		0.209	0.071	0.639	0.326
	(4) Z_1, Z_2			[0.316,0.369]	0.034		0.491	0.374	0.053	0.607
	(5) Z_1, Z_2, Z_3			[0.319,0.365]	0.030		0.491	0.381	0.046	0.607

Note: The estimated bounds, the Hausdorff distance $d_H(x)$ and the decompositions are the averages over 1000 replications.

Table 3.7.4: **Case 2**. True and Estimated Bounds, and Decomposition of Identification Gains ($\rho = 0.5,\ X \sim Bernoulli(1/2),\ x = 0$)

		Bounds				Decomposition				
		Manski		SV						
		$[L^M(x), U^M(x)]$	$d_H(x)$	$[L^{SV}(x), U^{SV}(x)]$	$d_H(x)$	$C_1(x)$	$C_2(x)$	$C_3(x)$	$C_4(x)$	IIP(x)
True DGP	Z_1, Z_2	$[-0.179, 0.821]$		$[0.283, 0.547]$		0.179	0.446	0.111	0.264	0.625
$n = 500$	(1) only Z_1	[-0.263,0.904]	0.102	[0.060,0.776]	0.237	0.263	0.185	0.002	0.716	0.448
	(2) only Z_2			[0.110,0.669]	0.179		0.359	-0.014	0.559	0.621
	(3) $Z_1, \widetilde{Z}_2$			[0.098,0.769]	0.224		0.237	-0.004	0.671	0.499
	(4) Z_1, Z_2			[0.166,0.647]	0.131		0.439	-0.017	0.481	0.701
	(5) Z_1, Z_2, Z_3			[0.160,0.656]	0.140		0.433	-0.025	0.496	0.695
$n = 5000$	(1) only Z_1	[-0.206,0.849]	0.034	[0.068,0.769]	0.223	0.206	0.148	0.000	0.701	0.354
	(2) only Z_2			[0.135,0.640]	0.148		0.337	0.007	0.506	0.543
	(3) $Z_1, \widetilde{Z}_2$			[0.115,0.754]	0.207		0.211	0.000	0.639	0.417
	(4) Z_1, Z_2			[0.210,0.619]	0.079		0.446	-0.006	0.409	0.653
	(5) Z_1, Z_2, Z_3			[0.208,0.620]	0.081		0.444	-0.007	0.412	0.650
$n = 10000$	(1) only Z_1	[-0.198,0.841]	0.024	[0.069,0.768]	0.221	0.198	0.141	0.001	0.699	0.339
	(2) only Z_2			[0.138,0.640]	0.145		0.333	0.005	0.502	0.531
	(3) $Z_1, \widetilde{Z}_2$			[0.118,0.751]	0.204		0.207	0.000	0.633	0.406
	(4) Z_1, Z_2			[0.216,0.612]	0.070		0.447	-0.006	0.396	0.645
	(5) Z_1, Z_2, Z_3			[0.217,0.613]	0.071		0.447	-0.003	0.396	0.645

Note: The estimated bounds, the Hausdorff distance $d_H(x)$ and the decompositions are the averages over 1000 replications.

Table 3.7.5: **Case 2**. True and Estimated Bounds, and Decomposition of Identification Gains ($\rho = 0.8,\ X \sim Bernoulli(1/2),\ x = 0$)

		Bounds				Decomposition				
		Manski		SV						
		$[L^M(x), U^M(x)]$	$d_H(x)$	$[L^{SV}(x), U^{SV}(x)]$	$d_H(x)$	$C_1(x)$	$C_2(x)$	$C_3(x)$	$C_4(x)$	IIP(x)
True DGP	Z_1, Z_2	$[-0.096, 0.904]$		$[0.319, 0.593]$		0.096	0.498	0.132	0.274	0.594
$n = 500$	(1) only Z_1	[-0.165,0.972]	0.084	[0.077,0.868]	0.276	0.165	0.183	-0.001	0.790	0.348
	(2) only Z_2			[0.114,0.751]	0.212		0.330	0.006	0.637	0.495
	(3) $Z_1, \widetilde{Z}_2$			[0.133,0.863]	0.270		0.243	-0.001	0.730	0.408
	(4) Z_1, Z_2			[0.209,0.732]	0.154		0.458	-0.008	0.523	0.623
	(5) Z_1, Z_2, Z_3			[0.200,0.738]	0.164		0.441	-0.007	0.538	0.606
$n = 5000$	(1) only Z_1	[-0.117,0.925]	0.026	[0.086,0.861]	0.268	0.117	0.149	0.001	0.776	0.266
	(2) only Z_2			[0.144,0.720]	0.175		0.340	0.010	0.576	0.457
	(3) $Z_1, \widetilde{Z}_2$			[0.154,0.848]	0.256		0.232	-0.001	0.694	0.349
	(4) Z_1, Z_2			[0.255,0.694]	0.102		0.486	0.001	0.439	0.603
	(5) Z_1, Z_2, Z_3			[0.255,0.696]	0.105		0.483	0.001	0.440	0.600
$n = 10000$	(1) only Z_1	[-0.111,0.919]	0.019	[0.087,0.860]	0.267	0.111	0.146	0.000	0.773	0.257
	(2) only Z_2			[0.148,0.713]	0.171		0.338	0.015	0.565	0.450
	(3) $Z_1, \widetilde{Z}_2$			[0.158,0.846]	0.253		0.230	0.000	0.688	0.342
	(4) Z_1, Z_2			[0.263,0.693]	0.100		0.491	-0.002	0.430	0.603
	(5) Z_1, Z_2, Z_3			[0.263,0.692]	0.100		0.489	0.001	0.429	0.601

Note: The estimated bounds, the Hausdorff distance $d_H(x)$ and the decompositions are the averages over 1000 replications.

3.8. Empirical Application: Women LFP and Childbearing

In this section, we apply our novel decomposition and IV evaluation method to study the effects of childbearing on women's labour supply. The dataset analysed here is from the 1980 Census Public Use Micro Samples (PUMS), available at Angrist and Evans (2009). We follow the data construction in Angrist and Evans (1998), where the sample consists of married women aged 21-35 with two or more children. The dateset contains 254,652 observations; see Table 2 in Angrist and Evans (1998) for more details and descriptive statistics. The binary outcome Y indicates if a individual was paid for work in the year prior to the census ($Y = 1$), or otherwise ($Y = 0$). The treatment effect of interest is the impact of having more than two children on the labour force participation Y. Thus, the binary treatment is $D \in \{0, 1\}$, with $D = 1$ denoting having more than two children.

Following Angrist and Evans (1998, Table 11) we use as continuous regressors woman's age, woman's age at first birth, and ages of the first two children (quarters), and binary regressors for first child being a boy, second child being a boy, black, hispanic, and other race, as well as the intersections of the above mentioned continuous and indicator variables. For computational simplicity, we reduce dimension of covariates by utilising the conditional propensity score $X_P := \widehat{\Pr}[D = 1|X]$ as a covariate, where $\widehat{\Pr}[D = 1|X]$ is estimated via a probit model and X includes all of the regressors mentioned above. Three sets of IVs are considered in this section: (1) the binary indicator that the first two children are the same sex ("*Samesex*"), (2) the binary indicator that the second birth was a twin ("*Twins*"), and (3) both indicators ("*Both*=$\{$*Samesex*, *Twins*$\}$"). To provide a basis for comparison of SV bounds with other ATE bounding analyses, we also compute the ATE bounds in Heckman and Vytlacil (2001) (hereafter HV bounds) and Chesher (2010) (hereafter Chesher bounds). To be consistent with our previous numerical analyses in Section 3.7, we use the method of CLR to compute all the four bounds of interest, via MLE for estimating bounding functions and the simulation-based method for correcting the bias of the intersecting bounds.

Table 3.8.1 reports the weighted average of the HMUE and of the CLR two-sided confi-

dence intervals (at 90%, 95% and 99% significant level) of the four bounds of ATE(X_P), with weights given by the estimated kernel density of X_P. Panels (a), (b) and (c) display the results using IV *Samesex*, *Twins* and *Both*, respectively. The estimated average of the Manski bounds in all three panels are essentially identical, since the Manski bounds do not depend on IVs. In all panels, the HV bounds make an improvement over the benchmark Manski bounds, with the HV bound width using *Twins* being narrower than that using *Samesex*, and the HV bound width using *Both* being the narrowest. The Chesher bounds using *Samesex* fail to identify the sign of the ATE(X_P), as it is a union of both negative and positive intervals. When the IV *Twins* or *Both* is used instead, the weighted average of 95% confidence interval of the Chesher bounds is $[-0.349, -0.019]$ (using *Twins*) or $[-0.335, -0.026]$ (using *Both*), revealing negative effects of having a third child on women's labour force participation. For the SV bounds, the results using the IV *Twins* or *Both* dramatically outperform those using *Samesex*. The 95% confidence interval using *Samesex*, *Twins* and *Both* are $[-0.548, -0.022]$, $[-0.272, -0.031]$ and $[-0.269, -0.042]$, respectively. The SV bounds estimates confirm the negative effect of a third child on women's labour force participation. The two-stage least square (2SLS) estimates of Angrist and Evans (1998, Table 11) give an ATE estimate of -0.123 with 95% confidence interval of $[-0.178, -0.068]$ using IV *Samesex*, and an estimate of -0.087 with 95% confidence interval of $[-0.120, -0.054]$ using IV *Twins*. As would be expected, the 95% two-sided confidence intervals of all four bounds cover the 2SLS estimates and their associated 95% confidence intervals for both IVs. To summarise the results above, we can see that for ATE bounds in which the IV plays a key role in extracting identifying information, i.e. HV, Chesher and SV bounds, the IV *Both* gives us the narrowest bounds (on average).

The ranking of the IV identification power of the three available IVs revealed by the discussion above is confirmed and explained by the identification gains decomposition and the IIP reported in Table 3.8.2. The results based on the 95% confidence interval show that given the same contribution of IV validity for the three IVs, which is 44.6% on average, the identification power of *Twins* (68.2%) is significantly larger than that of *Samesex* (47.1%). Closer inspection of the data reveals that the contribution of *Twins* to the identification gains exceeds that of *Samesex*, because whenever *Twins*$= 1$

the treatment $D = 1$, i.e. *Twins* is a perfect predictor of being treated, whereas this is not the case for *Samesex*. It is this feature, of course, that explains the superior performance when the HV, Chesher and SV bounds are evaluated using *Twins* rather than *Samesex*. Moreover, when both IVs *Samesex* and *Twins* are used, the identification power of *Both* (70.3%) also exceeds that of either one of the single IV *Samesex* or *Twins*. It indicates that although the identification power of *Samesex* is dominated by *Twins*, *Samesex* can still make extra contributions when identifying the ATE. It is intuitive because the mechanisms of the two IVs driving the probability of having a third child are different. One remark on the above analysis is that, for other ATE bounds that exploits the identification information of IVs, for example the HV and Chesher bounds, IVs with higher IIP clearly leads to narrower bounds for the ATE. It indicates that although the IIP is constructed to measure the IV's contribution to the SV bounds, it is also a meaningful measure for the IV identification power and can be utilised to indicate the IV relevance in other ATE bounds.

Table 3.8.1: Average of the Estimated Bounds

(a) IV: Samesex

	Manski	HV	Chesher	SV
HMUE	[-0.560,0.439]	[-0.537,0.401]	[-0.537,-0.011] ∪ [0.011,0.401]	[-0.538,-0.030]
90% CI	[-0.566,0.445]	[-0.546,0.411]	[-0.546,-0.005] ∪ [0.005,0.411]	[-0.546,-0.023]
95% CI	[-0.567,0.446]	[-0.548,0.412]	[-0.548,-0.004] ∪ [0.004,0.412]	[-0.548,-0.022]
99% CI	[-0.569,0.448]	[-0.551,0.416]	[-0.551,-0.001] ∪ [0.001,0.416]	[-0.551,-0.020]

(b) IV: Twins

	Manski	HV	Chesher	SV
HMUE	[-0.560,0.439]	[-0.304,0.113]	[-0.305,-0.061]	[-0.185,-0.101]
90% CI	[-0.566,0.445]	[-0.341,0.151]	[-0.342,-0.026]	[-0.259,-0.042]
95% CI	[-0.567,0.446]	[-0.349,0.158]	[-0.349,-0.019]	[-0.272,-0.031]
99% CI	[-0.569,0.448]	[-0.364,0.172]	[-0.365,-0.004]	[-0.299,-0.012]

(c) IV: Both={Samesex,Twins}

	Manski	HV	Chesher	SV
HMUE	[-0.560,0.439]	[-0.295,0.097]	[-0.295,-0.065]	[-0.200,-0.105]
90% CI	[-0.566,0.445]	[-0.329,0.131]	[-0.329,-0.032]	[-0.259,-0.051]
95% CI	[-0.567,0.446]	[-0.336,0.137]	[-0.335,-0.026]	[-0.269,-0.042]
99% CI	[-0.569,0.448]	[-0.349,0.151]	[-0.349,-0.011]	[-0.289,-0.027]

Note: The first row of panels (a)-(c) reports the weighted average of the HMUE of the four ATE bounds, and the second to fourth rows report the weighted average of the CLR two-sided confidence interval at different significant levels.

Table 3.8.2: Decomposition of Identification Gains and Instrument Identification Power

(a) IV: Samesex

	C_1	C_2	C_3	C_4	IIP
Based on HMUE	0.439	0.034	0.019	0.508	0.473
Based on 90% CI	0.445	0.026	0.018	0.523	0.472
Based on 95% CI	0.446	0.024	0.018	0.526	0.471
Based on 99% CI	0.448	0.021	0.019	0.532	0.471

(b) IV: Twins

	C_1	C_2	C_3	C_4	IIP
Based on HMUE	0.439	0.317	0.163	0.081	0.756
Based on 90% CI	0.445	0.250	0.100	0.216	0.695
Based on 95% CI	0.446	0.236	0.090	0.242	0.682
Based on 99% CI	0.448	0.209	0.075	0.286	0.657

(c) IV: Both={Samesex,Twins}

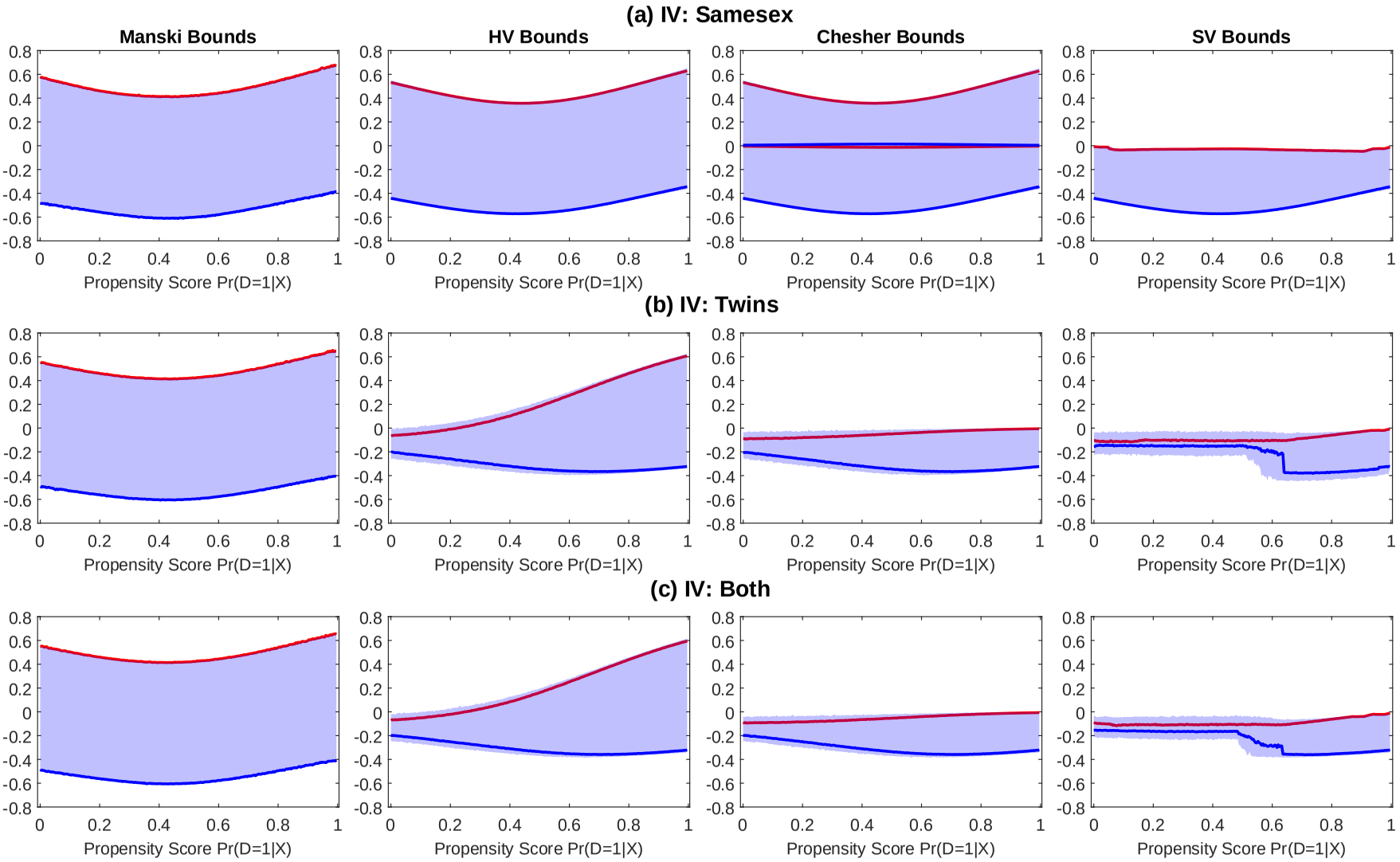
	C_1	C_2	C_3	C_4	IIP
Based on HMUE	0.439	0.330	0.134	0.096	0.769
Based on 90% CI	0.445	0.270	0.090	0.206	0.715
Based on 95% CI	0.446	0.257	0.085	0.226	0.703
Based on 99% CI	0.448	0.232	0.078	0.260	0.681

Note: C_1-C_4 and IIP are the weighted average of their associated conditional estimates given X_P, with the kernel density of X_P as weights. For both panels (a) to (c), C_1 to C_4 are computed as described in the footnote 4, and the estimates in each row correspond to different significance levels of the CLR estimation.

To explore the heterogeneity of the treatment effects, Figure 3.8.1 graphs the four bounds of interest against X_P. From Figure 3.8.1, we can see that when the more powerful of the three IVs are employed, namely *Twins* or *Both*, the HV bounds narrow down the possible range of the ATE(X_P) relative to the benchmark Manski bounds, especially for individuals with a small probability of having a third child. In addition, they can even identify the negative effect for individuals with a propensity score X_P close to zero. Similar properties are exhibited by the Chesher bounds. The SV bounds indicate that for women who are less likely to have more than two children, it is more probable that there will be a negative effect on their labour force participation once they have a third child, roughly in the region of -10% to -15%. For individuals who are more likely to have more than two children, the effect of having a third child is still negative but with larger possible range, roughly from -10% to -40% when their propensity score is about 0.6, and roughly from 0% to -30% when their propensity score is close to one.

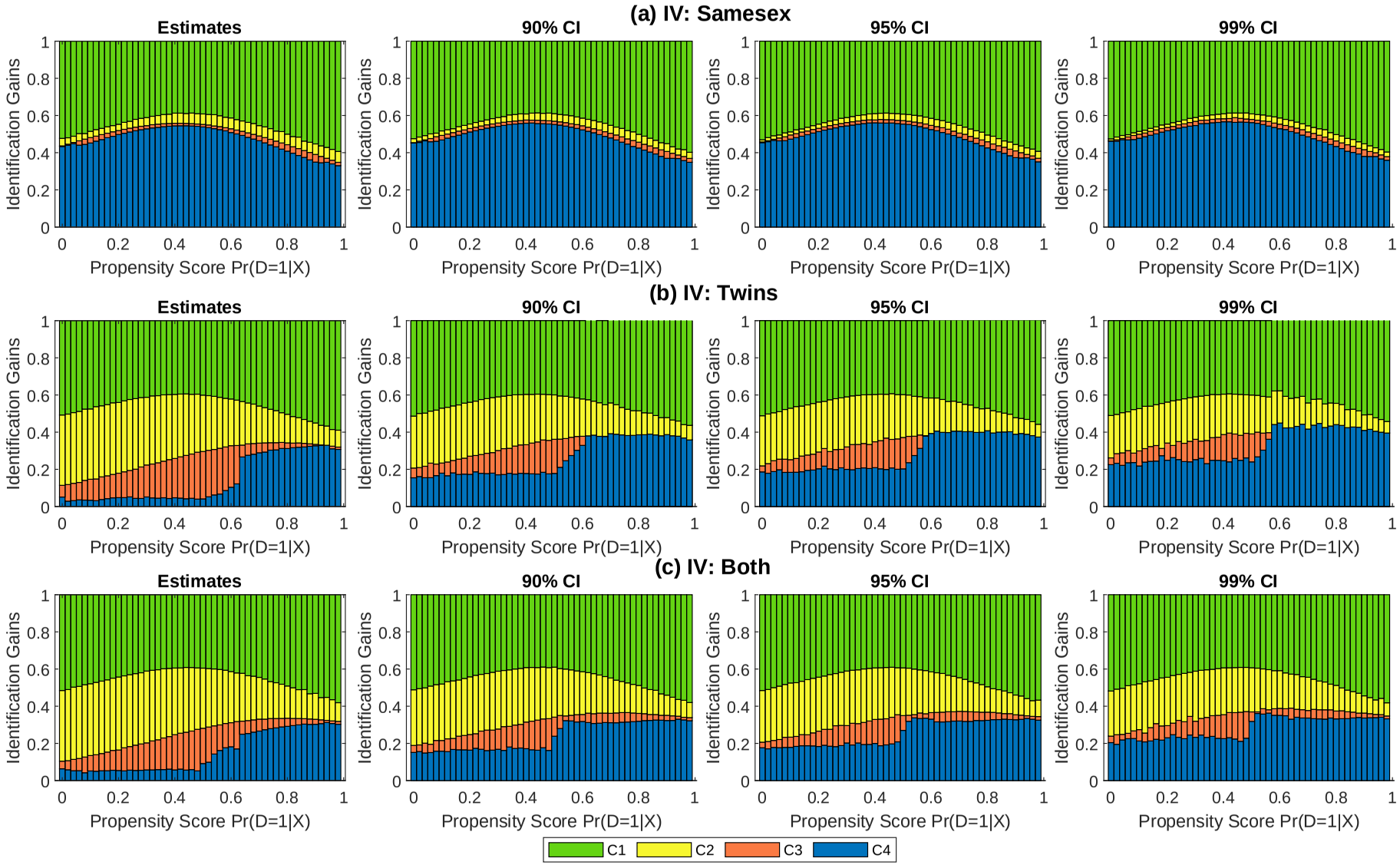
To check the heterogeneity of the IV identification power, Figure 3.8.2 displays the decompositions plotted against X_P. It is obvious that the IV identification power of *Twins* and *Both* are significantly larger than that of *Samesex*, across all possible values of X_P. Furthermore, the contribution of the covariate appears to be amplified when *Twins* is involved in deriving the bounds, leading to a further reduction in the width of the unexplained part relative to the benchmark.

Figure 3.8.1: Estimated Bounds of ATE(x)



Note: Panels (a)-(c) plot the estimates ATE(x) as functions of the propensity score X_P. The red lines are the upper bounds and blue lines are the lower bounds. The blue shaded area represents the 95% confidence regions.

Figure 3.8.2: Decomposition of Identification Gains



Note: Panels (a)-(c) depict the estimated decomposition of identification gains over Manski bound width $(C_j(x)/\omega^M(x)$ with $j = 1, 2, 3, 4)$ against the conditional probability of being treated $X_P = \widehat{\Pr}(D = 1|X)$.

3.9. Conclusion

In this chapter we explore the factors that determine the identification gains for the ATE in models with binary endogenous variables. We use the reduction in the size of the ATE identified set as a measure for identification power, and conduct our analysis with the identification gains achieved by the SV bounds (Shaikh and Vytlacil, 2011) against the benchmark Manski bounds (Manski, 1990). We decompose the identification gains into the impacts of the IV validity, the IV strength and the variability of the exogenous covariates. More importantly, we construct an index of "IIP" as a measure for the IV identification power.

We have developed theoretical results to show the complex mechanism through which IVs affect the identification of the ATE. We find that the IV identification power in a non-parametric and partially identified model is fundamentally different from the traditional understanding of the IV strength as in a parametric linear model, which is measured, for instance, by the pseudo R^2 or F-statistic from the reduced form treatment equation. We have shown that in partially identified non-linear models it is not only the traditional *IV strength* that determines the identification gains obtained when bounding the ATE, but also the interplay of the IVs with the degree of endogeneity and the variability of exogenous covariates. The conventional notion of IV strength or weakness no longer provides a full picture of the IV identification power, and is not the sole arbiter of IV usefulness. More specifically, we demonstrate that for the same IV strength given by the first-stage treatment equation, having the endogeneity with an opposite sign from that of the ATE can produce greater IV identification power, relative to the case when the endogeneity has the same sign as the ATE. That the endogeneity plays a similar role when testing IV weakness in binary outcome models with continuous endogenous regressors has been noted previously in Frazier, Renault, Zhang, and Zhao (June 28, 2019).

Our proposed index IIP provides a more appropriate measure of IV identification power, namely, the contribution made by the IVs in shrinking the ATE identified set. Importantly, we illustrate how the range of the conditional propensity score and the IIP relate to the ATE bounds for different levels of endogeneity, finite sample sizes and covariate variabilities. The results show that the IIP works well in finite sample settings as a

tool for measuring the IV identification power and for providing guidance on detecting irrelevant IVs. We find that missing IVs, or misspecification of relevant IVs can result in wider ATE identified sets and identification power loss. We also find that the loss of efficiency in finite sampl from adding an irrelevant IV can be more reliably detected by the estimated $IIP(x)$, even irrelevant IV could sometimes result in narrower SV bound width. The empirical application also demonstrates the practical usefulness of our novel decomposition of the identification gains and of the IIP index.

The study of IIP in this chapter sheds new light on IV relevancy in partial identification frameworks, and offers a potential criterion for IV selection in high dimension settings. It also raises new questions as to what constitutes an adequate definition of weak IVs in conjunction with ATE bounding analyses. Explorations of these issues are left for future research.

3.10. Appendix

Throughout the proof, let $P = \Pr[D = 1|X, Z]$ with support Ω_P and let $p(x, z) = \Pr[D = 1|x, z]$.

3.10.1. Lemmas

Lemma 3.10.1 *Under Assumption 3.2.1 (a) and (b), for any $p, p' \in \Omega_{P|x}$ such that $p > p'$, we have*

$$Pr[D = 0|x, p] + Pr[Y = y, D = 1|x, p] - \{Pr[D = 0|x, p'] + Pr[Y = y, D = 1|x, p']\} \leq 0,$$
$$Pr[D = 1|x, p] + Pr[Y = y, D = 0|x, p] - \{Pr[D = 1|x, p'] + Pr[Y = y, D = 0|x, p']\} \geq 0,$$

for $y \in \{0, 1\}$. In addition,

$$Pr[Y = y, D = 1|x, p] - Pr[Y = y, D = 1|x, p'] \geq 0,$$
$$Pr[Y = y, D = 0|x, p] - Pr[Y = y, D = 0|x, p'] \leq 0,$$

for $y \in \{0,1\}$. Lastly, if $\nu_1(1,x) > \nu_1(0,x)$ given $x \in \Omega_X$, then $Pr[Y=1|x,p] - Pr[Y=1|x,p'] \geq 0$. If $\nu_1(1,x) \leq \nu_1(0,x)$ given $x \in \Omega_X$, then $Pr[Y=1|x,p] - Pr[Y=1|x,p'] \leq 0$. Strict inequalities hold if Assumption 3.2.1 (c) is imposed on the DGP.

Proof of Lemma 3.10.1. Under Assumption 3.2.1 (a) and (b), for $p, p' \in \Omega_{P|x}$ with $p > p'$, we have

$$\begin{aligned}
&\Pr[D=0|x,p] + \Pr[Y=1,D=1|x,p] - \{\Pr[D=0|x,p'] + \Pr[Y=1,D=1|x,p']\} \\
=&\Pr[\varepsilon_1 < \nu_1(1,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] - \Pr[p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] \\
=&-\Pr[\varepsilon_1 \geq \nu_1(1,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] \\
\leq&0.
\end{aligned}$$

Similar manipulations show that

$$\begin{aligned}
&\Pr[D=0|x,p] + \Pr[Y=0,D=1|x,p] - \{\Pr[D=0|x,p'] + \Pr[Y=0,D=1|x,p']\} \leq 0, \\
&\Pr[D=1|x,p] + \Pr[Y=1,D=0|x,p] - \{\Pr[D=1|x,p'] + \Pr[Y=1,D=0|x,p']\} \geq 0, \text{ and} \\
&\Pr[D=1|x,p] + \Pr[Y=0,D=0|x,p] - \{\Pr[D=1|x,p'] + \Pr[Y=0,D=0|x,p']\} \geq 0.
\end{aligned}$$

In addition, using relatively straightforward if somewhat tedious algebra, we can obtain the following inequalities

$$\begin{aligned}
&\Pr[Y=0,D=1|x,p] - \Pr[Y=0,D=1|x,p'] = \Pr[\varepsilon_1 \geq \nu_1(1,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] \geq 0, \\
&\Pr[Y=1,D=1|x,p] - \Pr[Y=1,D=1|x,p'] = \Pr[\varepsilon_1 < \nu_1(1,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] \geq 0, \\
&\Pr[Y=0,D=0|x,p] - \Pr[Y=0,D=0|x,p'] = -\Pr[\varepsilon_1 \geq \nu_1(0,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] \leq 0, \text{ and} \\
&\Pr[Y=1,D=0|x,p] - \Pr[Y=1,D=0|x,p'] = -\Pr[\varepsilon_1 < \nu_1(0,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] \leq 0.
\end{aligned}$$

Now suppose that $\nu_1(1,x) > \nu_1(0,x)$ given $x \in \Omega_X$. Then it follows that

$$\begin{aligned}
&\Pr[Y=1|x,p] - \Pr[Y=1|x,p'] \\
=&\Pr[Y=1,D=1|x,p] + \Pr[Y=1,D=0|x,p] \\
&\quad - \Pr[Y=1,D=1|x,p'] - \Pr[Y=1,D=0|x,p'] \\
=&\Pr[\varepsilon_1 < \nu_1(1,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] - \Pr[\varepsilon_1 < \nu_1(0,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p]
\end{aligned}$$

$$
\begin{aligned}
&= \Pr[\nu_1(0,x) \leq \varepsilon_1 < \nu_1(1,x), p' \leq F_{\varepsilon_2}(\varepsilon_2) < p] \\
&\geq 0.
\end{aligned}
$$

Finally, using a parallel argument in the case where $\nu_1(1,x) \leq \nu_1(0,x)$ given $x \in \Omega_X$, we can conclude that the inequalities stated in the lemma hold. ■

Lemma 3.10.2 *Under Assumptions 3.2.1 and 3.3.1, the following results hold. Joint probabilities $Pr[Y = y, D = d|x, p]$ for $y, d \in \{0, 1\}$ are functions of the dependence parameter ρ. In addition,*

(a) $Pr[Y = 1, D = 1|x, p]$ and $Pr[Y = 0, D = 0|x, p]$ are weakly increasing in ρ;

(b) $Pr[Y = 1, D = 0|x, p]$ and $Pr[Y = 0, D = 1|x, p]$ are weakly decreasing in ρ.

Proof of Lemma 3.10.2. For any given $p \in \Omega_P$,

$$
\begin{aligned}
\Pr[Y = 1, D = 1|x, p] &= \Pr[\varepsilon_1 < \nu_1(1,x), F_{\varepsilon_2}(\varepsilon_2) < p|x, p] \\
&= \Pr[\varepsilon_1 < \nu_1(1,x), F_{\varepsilon_2}(\varepsilon_2) < p] \\
&= C(F_{\varepsilon_1}(\nu_1(1,x)), p; \rho). \qquad (3.10)
\end{aligned}
$$

Because the copula $C(\cdot, \cdot; \rho)$ satisfies the concordant ordering with respect to ρ, we know that $\Pr[Y = 1, D = 1|x, p]$ is weakly increasing in ρ. Since

$$
\Pr[Y = 0, D = 1|x, p] = \Pr[D = 1|x, p] - \Pr[Y = 1, D = 1|x, p] = p - C(F_{\varepsilon_1}(\nu_1(1,x)), p; \rho),
$$

$\Pr[Y = 0, D = 1|x, p]$ is decreasing in ρ. In addition,

$$
\begin{aligned}
\Pr[Y = 0, D = 0|x, p] &= \Pr[\varepsilon_1 \geq \nu_1(0,x), F_{\varepsilon_2}(\varepsilon_2) \geq p|x, p] \\
&= \Pr[\varepsilon_1 \geq \nu_1(0,x), F_{\varepsilon_2}(\varepsilon_2) \geq p] \\
&= \Pr[\varepsilon_1 \geq \nu_1(0,x)] - \Pr[\varepsilon_1 \geq \nu_1(0,x), F_{\varepsilon_2}(\varepsilon_2) < p] \\
&= \Pr[\varepsilon_1 \geq \nu_1(0,x)] - \Pr[F_{\varepsilon_2}(\varepsilon_2) < p] + \Pr[\varepsilon_1 < \nu_1(0,x), F_{\varepsilon_2}(\varepsilon_2) < p] \\
&= 1 - F_{\varepsilon_1}(\nu_1(0,x)) - p + C(F_{\varepsilon_1}(\nu_1(0,x)), p; \rho). \qquad (3.11)
\end{aligned}
$$

From (3.11) we can see that $\Pr[Y = 0, D = 0|x, p]$ is weakly increasing in ρ, which immediately implies that $\Pr[Y = 1, D = 0|x, p]$ is weakly decreasing in ρ. ■

3.10.2. Proofs

Proof of Proposition 3.3.1. To begin, let us first introduce the following notation:

$$
\begin{aligned}
L_0(x,p) &= \Pr[Y=1, D=0|x,p] + \sup_{x' \in \mathbf{X}_{0-}(x)} \Pr[Y=1, D=1|x', p], \\
L_1(x,p) &= \Pr[Y=1, D=1|x,p] + \sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[Y=1, D=0|x', p], \\
U_0(x,p) &= \Pr[Y=1, D=0|x,p] + p \inf_{x' \in \mathbf{X}_{0+}(x)} \Pr[Y=1|x', p, D=1], \\
U_1(x,p) &= \Pr[Y=1, D=1|x,p] + (1-p) \inf_{x' \in \mathbf{X}_{1-}(x)} \Pr[Y=1|x', p, D=0].
\end{aligned}
$$

Then the SV bounds become

$$
L^{SV}(x) = L_1(x, \overline{p}) - U_0(x, \underline{p}) \text{ and } U^{SV}(x) = U_1(x, \overline{p}) - L_0(x, \underline{p}), \tag{3.12}
$$

and under Assumption 3.2.1 the SV bounds are sharp if $\Omega_{X,P} = \Omega_X \times \Omega_P$ (Shaikh and Vytlacil, 2011, Theorem 2.1).

Next we show that $L_0(x, p)$ is weakly decreasing in p (*ceteris paribus*). Under Assumption 3.2.1 and $\Omega_{X,P} = \Omega_X \times \Omega_P$, for $\forall x \in \Omega_X$ there exists $x_0^l \in \mathbf{X}_{0-}(x)$, such that $\nu_1(1, x_0^l) = \sup_{x \in \mathbf{X}_{0-}(x)} \nu_1(1, x)$ and

$$
L_0(x,p) = \Pr[Y=1, D=0|x,p] + \Pr[Y=1, D=1|x_0^l, p],
$$

(For detailed particulars see the proof of Shaikh and Vytlacil, 2011, Theorem 2.1 (ii)[5]). For $p, p' \in \Omega_P$ and $p' < p$, we have now have

$$
\begin{aligned}
L_0(x,p) - L_0(x,p') =& \Pr[Y=1, D=0|x,p] + \Pr[Y=1, D=1|x_0^l, p] \\
&- \Pr[Y=1, D=0|x,p'] - \Pr[Y=1, D=1|x_0^l, p'] \\
=& \Pr[\varepsilon_1 \leq \nu_1(1, x_0^l), p' < \varepsilon_2 \leq p) - \Pr[\varepsilon_1 \leq \nu_1(0, x), p' < \varepsilon_2 \leq p)
\end{aligned}
$$

[5]The proof is contained in the supplementary material of Shaikh and Vytlacil (2011).

$$
\begin{aligned}
&= \Pr[\nu_1(0,x) < \varepsilon_1 \leq \nu_1(1,x_0^l), p' < \varepsilon_2 \leq p) \\
&\leq 0, \qquad (3.13)
\end{aligned}
$$

where the last inequality follows because $x_0^l \in \mathbf{X}_{0-}(x)$, and the Lemma 2 in Shaikh and Vytlacil (2011) shows that $x_0^l \in \mathbf{X}_{0-}(x)$ implies $\nu_1(1,x_0^l) \geq \nu_1(0,x)$. Thus, from (3.13), $L_0(x,p)$ is weakly decreasing in p.

Similar arguments show that $L_1(x,p)$ is weakly increasing in p, $U_0(x,p)$ is weakly increasing in p, and $U_1(x,p)$ is weakly decreasing in p. Hence $L^{SV}(x)$ is weakly increasing in $\overline{p}$ and $U^{SV}(x)$ is weakly decreasing in $\overline{p}$. On the other hand, $L^{SV}(x)$ is weakly decreasing in $\underline{p}$ and $U^{SV}(x)$ is weakly increasing in $\underline{p}$. This completes the proof of the proposition. ■

Proof of Proposition 3.3.2. Suppose that $\text{ATE}(x) > 0$ for $x \in \Omega_X$. Under Assumption 3.2.1, from the definitions of $\mathbf{X}_{0+}(x)$, $\mathbf{X}_{0-}(x)$, $\mathbf{X}_{1+}(x)$ and $\mathbf{X}_{1-}(x)$, we know that $\mathbf{X}_{0+}(x)$ and $\mathbf{X}_{1+}(x)$ are nonempty for $\forall x \in \Omega_X$, since x itself belongs to these two sets. While, $\mathbf{X}_{0-}(x)$ and $\mathbf{X}_{1-}(x)$ may be empty for some $x \in \Omega_X$. Recall that the supremum and infimum are defined as zero and one over an empty set, respectively. Thus, for the four functions defined in the proof of Proposition 3.3.1 we have

$$
\begin{aligned}
&L_0(x,p) \geq \Pr[Y=1, D=0|x,p], \\
&L_1(x,p) \geq \Pr[Y=1|x,p], \\
&U_0(x,p) \leq \Pr[Y=1|x,p], \text{ and} \\
&U_1(x,p) \leq \Pr[Y=1, D=1|x,p] + \Pr[D=0|x,p].
\end{aligned}
$$

The ATE SV bounds are therefore bounded by $[L^{SV}(x), U^{SV}(x)] \subset [\underline{L}^{SV}(x), \overline{U}^{SV}(x)]$, where

$$
\begin{aligned}
\underline{L}^{SV}(x) &= \sup_{p \in \Omega_{P|x}} \Pr[Y=1|x,p] - \inf_{p \in \Omega_{P|x}} \Pr[Y=1|x,p], \text{ and} \\
\overline{U}^{SV}(x) &= \inf_{p \in \Omega_{P|x}} \{\Pr[Y=1, D=1|x,p] + \Pr[D=0|x,p]\} - \sup_{p \in \Omega_{P|x}} \Pr[Y=1, D=0|x,p], \qquad (3.14)
\end{aligned}
$$

and the widest possible width $\overline{\omega}(x) := \overline{U}^{SV}(x) - \underline{L}^{SV}(x)$ is

$$\overline{\omega}(x) := \inf_{p\in\Omega_{P|x}} \{\Pr[Y=1, D=1|x,p] + \Pr[D=0|x,p]\} - \sup_{p\in\Omega_{P|x}} \Pr[Y=1, D=0|x,p] \\ - \sup_{p\in\Omega_{P|x}} \Pr[Y=1|x,p] + \inf_{p\in\Omega_{P|x}} \Pr[Y=1|x,p]\,.$$

From Lemma 3.10.1 it follows that

$$\begin{aligned}\overline{\omega}(x) =& \Pr[Y=1, D=1|x,\overline{p}(x)] + \Pr[D=0|x,\overline{p}(x)] - \Pr[Y=1, D=0|x,\underline{p}(x)] \\ & - \Pr[Y=1|x,\overline{p}(x)] + \Pr[Y=1|x,\underline{p}(x)] \\ =& \Pr[Y=1, D=1|x,\underline{p}(x)] + \Pr[Y=0, D=0|x,\overline{p}(x)]. \end{aligned} \tag{3.15}$$

Now consider the case where $\text{ATE}(x) < 0$. In contrast to the positive $\text{ATE}(x)$ case, $\mathbf{X}_{0-}(x)$ and $\mathbf{X}_{1-}(x)$ are nonempty for $\forall x \in \Omega_X$ since x itself belongs to these two sets, while $\mathbf{X}_{0+}(x)$ and $\mathbf{X}_{1+}(x)$ may be empty for some $x \in \Omega_X$. Thus, the following inequalities hold

$$\begin{aligned} &L_0(x,p) \geq \Pr[Y=1|x,p], \\ &L_1(x,p) \geq \Pr[Y=1, D=1|x,p], \\ &U_0(x,p) \leq \Pr[Y=1, D=0|x,p] + \Pr[D=1|x,p], \text{ and} \\ &U_1(x,p) \leq \Pr[Y=1|x,p], \end{aligned}$$

based on which we can bound the SV bounds as $[L^{SV}(x), U^{SV}(x)] \subset [\underline{L}^{SV}(x), \overline{U}^{SV}(x)]$, where

$$\begin{aligned} \overline{U}^{SV}(x) &= \inf_{p\in\Omega_{P|x}} \Pr[Y=1|x,p] - \sup_{p\in\Omega_{P|x}} \Pr[Y=1|x,p], \text{ and} \\ \underline{L}^{SV}(x) &= \sup_{p\in\Omega_{P|x}} \Pr[Y=1, D=1|x,p] - \inf_{p\in\Omega_{P|x}} \{\Pr[Y=1, D=0|x,p] + \Pr[D=1|x,p]\}\,. \end{aligned} \tag{3.16}$$

The widest possible width of the SV bounds is now therefore

$$\overline{\omega}(x) = \inf_{p\in\Omega_{P|x}} \Pr[Y=1|x,p] - \sup_{p\in\Omega_{P|x}} \Pr[Y=1|x,p] - \sup_{p\in\Omega_{P|x}} \Pr[Y=1, D=1|x,p]$$

$$+ \inf_{p \in \Omega_{P|x}} \{\Pr[Y=1, D=0|x,p] + \Pr[D=1|x,p]\},$$

and from Lemma 3.10.1 we have that

$$\begin{aligned}\overline{\omega}(x) =& \Pr[Y=1|x,\overline{p}(x)] - \Pr[Y=1|x,\underline{p}(x)] - \Pr[Y=1, D=1|x,\overline{p}(x)] \\ &+ \Pr[Y=1, D=0|x,\underline{p}(x)] + \Pr[D=1|x,\underline{p}(x)] \\ =& \Pr[Y=1, D=0|x,\overline{p}(x)] + \Pr[Y=0, D=1|x,\underline{p}(x)]. \end{aligned} \tag{3.17}$$

The nature of the relationship between $\overline{\omega}(x)$ and $\underline{p}(x)$ and $\overline{p}(x)$ follows directly from the expressions in (3.15) and (3.17) upon application of Lemma 3.10.1. ■

Proof of Proposition 3.3.3. The proof follows directly from the expression for $\overline{\omega}(x)$ in Proposition 3.3.2 and Lemma 3.10.2. ■

Proof of Proposition 3.3.4. Without loss of generality, assume that the distribution of ε_2 has been "normalised" to be uniform over $[0,1]$. Degeneracy of $\nu_1(D,X)|D$ indicates that there exists a function $m_1 : \{0,1\} \mapsto \mathbb{R}$ such that $\nu_1(d,x) = m_1(d)$ for all $(d,x) \in \{0,1\} \times \Omega_X$. Take ATE$(x)$ to be positive. When $H(x,x')$ is well defined and $\nu_1(D,X) = m_1(D)$, $\mathbf{X}_{0+}(x) = \mathbf{X}_{1+}(x) = \Omega_X$, and $\mathbf{X}_{0-}(x) = \mathbf{X}_{1-}(x) = \emptyset$. Since ε_2 is continuously distributed we can conclude that $\forall (x,z), (z',x') \in \Omega_{X,Z}$ such that $\Pr[D=1|z',x'] = \Pr[D=1|x,z]$ we must have $\nu_2(x,z) = \nu_2(z',x')$.

For $L^{SV}(x)$, consider $\sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[Y=1, D=0|x',p]$. If $\mathbf{X}_{1+}(x)$ is empty, or if there does not exist a z' such that $\Pr[D=1|x',z'] = p$, then $\sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[Y=1, D=0|x',p]$ is set to zero. Since $\mathbf{X}_{1+}(x)$ equals Ω_X because $\nu_1(D,X) = m_1(D)$, we have $\Pr[D=1|x',z'] = p$ for at least $(z',x') = (x,z)$, and thus $\sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[Y=1, D=0|x',p]$ is well-defined. It follows that

$$\begin{aligned}\sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[Y=1, D=0|x',p] &= \sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[\nu_1(0,x') > \varepsilon_1, \nu_2(x',z') \leq \varepsilon_2 | x', p] \\ &= \sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[m_1(0) > \varepsilon_1, \nu_2(x,z) \leq \varepsilon_2 | x', p] \\ &= \sup_{x' \in \mathbf{X}_{1+}(x)} \Pr[m_1(0) > \varepsilon_1, \nu_2(x,z) \leq \varepsilon_2 | x, p]\end{aligned}$$

$$=\Pr[Y=1,D=0|x,p], \tag{3.18}$$

where the second equality arises because the CDF of ε_2 is the strictly positive and $\nu_1(0,x')=m_1(0)$ is degenerate. The third equality is due to the assumed independence of (X,Z). Similarly,

$$\begin{aligned} p \inf_{x'\in\mathbf{X}_{0+}(x)} \Pr[Y=1|x',p,D=1] &= \inf_{x'\in\mathbf{X}_{0+}(x)} \Pr[Y=1,D=1|x',p] \\ &= \inf_{x'\in\mathbf{X}_{0+}(x)} \Pr[\nu_1(1,x')>\varepsilon_1,\nu_2(x',z')>\varepsilon_2|x',p] \\ &= \inf_{x'\in\mathbf{X}_{0+}(x)} \Pr[m_1(1)>\varepsilon_1,\nu_2(x,z)>\varepsilon_2|x,p] \\ &= \Pr[Y=1,D=1|x,p]. \end{aligned} \tag{3.19}$$

By virtue of equations (3.18) and (3.19), and Lemma 3.10.1, $L^{SV}(x)$ can be rewritten as

$$\begin{aligned} L^{SV}(x) &= \sup_{p\in\Omega_{P|x}} \{\Pr[Y=1,D=1|x,p]+\Pr[Y=1,D=0|x,p]\} \\ &\qquad - \inf_{p\in\Omega_{P|x}} \{\Pr[Y=1,D=0|x,p]+\Pr[Y=1,D=1|x,p]\} \\ &= \sup_{p\in\Omega_{P|x}} \Pr[Y=1|x,p] - \inf_{p\in\Omega_{P|x}} \Pr[Y=1|x,p] \\ &= \Pr[Y=1|x,\overline{p}(x)] - \Pr[Y=1|x,\underline{p}(x)]. \end{aligned} \tag{3.20}$$

For $U^{SV}(x)$, because $\mathbf{X}_{0-}(x)$ and $\mathbf{X}_{1-}(x)$ are empty, from Lemma 3.10.1 we get

$$\begin{aligned} U^{SV}(x) &= \inf_{p\in\Omega_{P|x}} \{\Pr[Y=1,D=1|x,p]+(1-p)\} - \sup_{p\in\Omega_{P|x}} \Pr[Y=1,D=0|x,p] \\ &= \Pr[Y=1,D=1|x,\overline{p}(x)]+(1-\overline{p}(x))-\Pr[Y=1,D=0|x,\underline{p}(x)]. \end{aligned} \tag{3.21}$$

THe expressions in (3.20) and (3.21) now yield the result that

$$\begin{aligned} \omega^{SV} =& \Pr[Y=1,D=1|x,\overline{p}(x)]+(1-\overline{p}(x))-\Pr[Y=1,D=0|x,\underline{p}(x)] \\ &- \Pr[Y=1|x,\overline{p}(x)]+\Pr[Y=1|x,\underline{p}(x)] \\ =& \Pr[Y=0,D=0|x,\overline{p}(x)]+\Pr[Y=1,D=1|x,\underline{p}(x)], \end{aligned}$$

which is equal to $\overline{\omega}(x)$. The proof for the negative ATE(x) case is completely analogous, the details are omitted. ■

Proof of Proposition 3.5.1. (a) We first show that $IIP(x)$ is well-defined in the sense that we are able to identify whether Z is relevant or not. If, for a given $x \in \Omega_X$, there exists a z and z' in $\Omega_{Z|x}$ such that $z \neq z'$ and $\Pr[D = 1|x, z] \neq \Pr[D = 1|x, z']$, then the IV Z is relevant. If Z is relevant then $IIP(x) = 1 - \overline{\omega}(x)$ where $\overline{\omega}(x)$ is the widest possible width defined in Proposition 3.3.2. Otherwise, Z is irrelevant, and by Proposition 3.3.4, if Z is irrelevant the SV bounds reduce to the benchmark Manski bounds and we have $IIP(x) = 0$.

Next, we prove that $IIP(x) \in [0, 1]$. Since $\overline{\omega}(x)$ is a summation of some conditional probabilities $\forall x \in \Omega_X$, it follows that $\overline{\omega}(x) \geq 0$ and $IIP(x) \leq 1$. Whenever Z is relevant the sign of ATE(x) is identified, and from Lemma 3.10.1 it follows that if ATE$(x) > 0$ then

$$\begin{aligned}\overline{\omega}(x) &= \Pr[Y = 1, D = 1|x, \underline{p}(x)] + \Pr[Y = 0, D = 0|x, \overline{p}(x)] \\ &\leq \Pr[Y = 1, D = 1|x] + \Pr[Y = 0, D = 0|x], \end{aligned} \tag{3.22}$$

which is less than one, and if ATE$(x) < 0$ then

$$\begin{aligned}\overline{\omega}(x) &= \Pr[Y = 1, D = 0|x, \overline{p}(x)] + \Pr[Y = 0, D = 1|x, \underline{p}(x)] \\ &\leq \Pr[Y = 1, D = 0|x] + \Pr[Y = 0, D = 1|x], \end{aligned} \tag{3.23}$$

which is also less than one. Thus, $IIP(x) = 1 - \overline{\omega}(x) \geq 0$, $\forall x \in \Omega_X$, and $IIP(x) \in [0, 1]$.

(b) If Z is irrelevant, by definition we have $IIP(x) = 0$ and the SV bounds reduce to the benchmark Manski bounds by Proposition 3.3.4. To establish necessity we will show that the presumption that the events Z is relevant and $IIP(x) = 0$ occur simultaneously leads to a contradiction. If Z is relevant, then the index $IIP(x) = 1 - \overline{\omega}(x)$. The goal, therefore, is to show that relevant Z leads to strictly less one $\overline{\omega}(x)$, by verifying the inequalities (3.22) and (3.23) are strict. Take (3.22) as an example and the result for

(3.23) can be verified analogously. Since

$$\begin{aligned}
&\Pr[Y=1,D=1|x]-\Pr[Y=1,D=1|x,\underline{p}(x)]\\
&=\int_{p\in\Omega_{P|x}}\left[\Pr[Y=1,D=1|x,p]-\Pr[Y=1,D=1|x,\underline{p}(x)]\right]d\Pr[P=p|X=x]\\
&=\int_{p\in\Omega_{P|x}}\Pr\left[\varepsilon_1<\mu_1(1,x),\underline{p}(x)\leq\varepsilon_2<p\right]d\Pr[P=p|X=x], \qquad (3.24)
\end{aligned}$$

the relevance of Z guarantees that there exists a $p\in\Omega_{P|x}$ such that $p\neq\underline{p}(x)$ and $\Pr[P=p|X=x]>0$. Then, the continuity of the joint distribution of the $(\varepsilon_1,\varepsilon_2)$ with support $\mathbb{R}^2$ implies that (3.24) is strictly positive. Similar arguments can be applied to show that $\Pr[Y=0,D=0|x]-\Pr[Y=0,D=0|x,\overline{p}(x)]>0$. Therefore, $\overline{\omega}(x)<\Pr[Y=1,D=1|x]+\Pr[Y=0,D=0|x]\leq 1$, leading to $IIP(x)>0$.

(c) If Z is a perfect predictor of the treatment D in the sense that there exist a z^* and a z^{**} in $\Omega_{Z|x}$ such that $\Pr(D=0|x,z^*)=1$ and $\Pr(D=1|x,z^{**})=1$, this obviously implies Z is relevant and $IIP(x)=1-\overline{\omega}(x)$. Furthermore, $\underline{p}(x)=p(x,z^*)$ and $\overline{p}(x)=p(x,z^{**})$. Hence, it can be easily shown from the expressions for $\overline{\omega}(x)$ that perfect prediction by Z leads to the equality $\overline{\omega}(x)=0$ for both $\text{ATE}(x)>0$ and $\text{ATE}(x)<0$. Thus $IIP(x)=1-\overline{\omega}(x)=1$.

Moreover, since $\overline{\omega}(x)$ is the widest possible width for the SV bounds, we have $0\leq\omega^{SV}(x)\leq\overline{\omega}(x)$, and when $\overline{\omega}(x)=0$ it follows that $\omega^{SV}(x)=0$. The $\text{ATE}(x)$ is point identified if $IIP(x)=1$. ■

Chapter 4

Spillovers of Program Benefits with Mismeasured Networks

4.1. Introduction

In the literature on treatment effects, the SUTVA (Rubin, 1990) is widely adopted for causal inference. It states that the treatment of one unit does not affect others' outcomes. However, the spillover effects of the treatment via network interactions have been documented in many applications, for example, cash transfer programs (Barrera-Osorio et al., 2011), health programs (Dupas, 2014), public policy programs (Kremer and Miguel, 2007), education programs (Opper, 2019), and information diffusion (Banerjee et al., 2013). In these studies, spillovers are conceived as a mechanism through which a treatment could propagate and affect many others' socioeconomic behaviour. Therefore, correctly measuring the spillovers of a program intervention is incredibly relevant for understanding whether and how the treatment influences individuals' outcome through their social interactions, and providing meaningful policy advice for effective treatment allocation (Angelucci and Di Maro, 2016; Viviano, 2019).

Existing methods of studying spillover effects typically assume that accurate information about the network connections of all the sampled units is available (e.g. Leung, 2020b; Ma et al., 2020; Vazquez-Bare, 2019; Viviano, 2019). However, this assumption

is difficult to verify in practice and questionable in many settings (Sävje, 2019). Increasing evidence has shown that social connections are substantially misreported. For example, Comola and Fafchamps (2017) document a massive discrepancy (about 73%) between the responses of inter-household transfers reported from givers and receivers in the village of Nyakatoke, Tanzania. The ratio 73% is computed as the number of reported transfers coming from only giver or only receiver (1250) over the number of total reported transfers (1721), see Comola and Fafchamps (2017) page 560-561. These transfers are oft-used to construct risk-sharing networks; therefore, the nonnegligible portion of network links are nonreciprocal due to the discrepancies. Ignoring or mis-connecting either side of the responses may lead to misclassified networks. Similar non-reciprocal problem has also been found in other survey data, e.g. 40% of risk-sharing network links from rural Philippines (Fafchamps and Lund, 2003) and more than 10% of the friendship among adolescents in Add-Health dataset (Calvó-Armengol et al., 2009; Patacchini et al., 2017) are non-reciprocal. When constructing generational family network links in the PROGRESA data via the respondents' surnames, Angelucci, De Giorgi, Rangel, and Rasul (2010) find various forms of measurement errors. They include poor recalling of and typos in the surnames, false connections of two genuinely unrelated families sharing the same surnames, and misspecification of network boundary by restricting the network within the same village. In the study of technology diffusion among pineapple farmers in Ghana, Conley and Udry (2010) also notice the potential misclassification of information neighbours, due to the lack of precise definitions of the information neighbours and the existence of multi-contextual social connections.

This chapter investigates the identification and estimation of the treatment and spillover effects of a program intervention with mismeasured network data. There are several attractive features of the proposed method. First, it allows flexible forms of heterogeneity in the treatment and spillover effects, which is important to inform how the treatment response varies across population (Manski, 2001). Second, the analysis can be applied to settings with a large network that is not block-diagonal and that contains missing or misreported links. Moreover, modelling the network formation or its misclassification probability is not required to implement the proposed method.

We focus on a randomised program intervention and a superpopulation model studied by Leung (2020b). If a network is correctly measured, the direct treatment effect can be identified from the variation of the ego unit's own treatment status, and the spillover effect can be identified via the variation of a statistic summarising the exposure to the treated peers. However, the network measurement errors introduced in this chapter sophisticate the identification by contaminating the true channels of the network interference. Ignoring those errors will lead to biased estimation. The measurement errors considered in this chapter are nonclassical; that is, they depend on the network interactions. In addition, the measurement errors are assumed to be independent of the potential outcomes and the treatment, conditional on the statistic of the network and exogenous covariates. This independence assumption is referred to as "nondifferential", and is often invoked in the literature studying measurement error models (e.g. Bound, Brown, and Mathiowetz, 2001).

In this chapter, we propose a novel strategy to nonparametrically point-identify the treatment and spillover effects with a mismeasured network proxy, when an instrumental variable for the latent network (or equivalently, an additional network proxy) is available. The identification consists of two steps. Firstly, we adopt the matrix diagonalisation method proposed by Hu (2008) to identify several distributions of the true *number of network neighbours* (hereafter degree), under the help of the instrument network proxy. Secondly, the distribution involving the true *number of treated network neighbours*, which measures the exposure to the treated peers, is identified. The identification in the second step relies on the observation that, network proxies in some studies might satisfy one assumption: there is only one type of measurement error. It means that the network proxy either includes *no* false links while allowing missing ones ("no false positive"), or includes *no* missing links but allowing false ones ("no false negative"). This one type of measurement error assumption dramatically simplifies the interdependence of the observed network-based variables with their latent counterparts, which is the main difficulty of identification. Testable implication of such an assumption is also available.

Inference in network settings is nonstandard due to the data correlation induced by the network interaction. In particular, outcomes of two units are correlated if they are

connected or share common network neighbours (Leung, 2020b). In this chapter, the mismeasured network adds to the complication by introducing an extra source of correlation, through the spillover of the measurement errors. Such spillover occurs, because a false network connection of two units will alter both their observable exposures to the treated neighbours. In addition to the above network-induced correlation, this chapter also considers data dependency due to general forms of heteroscedasticity, autocorrelation and clustering, so that units that are not friends and do not share common friends may also correlate with each other. Such correlation may be caused by, for instance, family background, school culture, or community diversity. All sources of data correlation described above generate distinct technical issues for the causal inference.

We propose a semiparametric estimation approach, which overcomes the difficulty caused by the spillover of measurement errors, and the resulting estimator is shown to be consistent and asymptotically normal. To derive limit theorems, we extend the univariate central limit theorem (CLT) of Chandrasekhar and Jackson (2016) to multivariate settings. The estimation approach in this chapter possesses several advantages: (i) it fits situations where there may be no clear spatial or ordered structure; (ii) it does not require a large number of independent subnetworks and allows general forms of data dependence; and (iii) it allows a sufficiently large number of units have nonzero correlation with an increasing number of other units.

In the simulation exercises, we verify the advantages of the proposed methodology over the naive estimation that ignores the network measurement errors. The bias reduction provided by the semiparametric approach is substantial and its causal inference is more reliable than that of the naive estimation. Moreover, it is confirmed that the semiparametric method still outperforms the naive estimation, even if its key identification assumption, for example, the one type of measurement error, is mildly violated.

4.2. Literature Review

This chapter is among a few papers that have studied the spillover effect of a program intervention with mismeasured networks. Hardy, Heath, Lee, and McCormick (2019)

consider a parametric model for the potential outcomes and for the network misclassifications, and use a likelihood-based approach to estimate the spillover effect. In a nonparametric setting, when only a network proxy is available, He and Song (2018) provide a lower bound for the spillover effect under the restriction that the spillover is nonnegative. This chapter is substantially different from the papers above, because it does not rely on modelling the network misclassifications, and more importantly, it provides a formal solution for the nonparametric point-identification of the spillover effect, when two network proxies are available.[1]

One study of the structural model of social interactions with mismeasured networks is related to this chapter. Gao and Li (2019) explore the endogenous and exogenous peer effects via the linear-in-means model with two mismeasured network proxies. Their identification result depends on three key assumptions. First, there exist two different latent network structures for the same group of individuals. Second, the error contaminated network-based variables are assumed to be independent conditional on their latent counterparts, which implicitly requires the networks to be non-stochastic. At last, a copula is used to capture the dependence between the mismeasured network effects. Like the study in this chapter, Gao and Li (2019) exploit the matrix diagonalisation method, however our analysis focuses on the reduced-form treatment response function that is modeled nonparametrically, enabling flexible forms of heterogeneous treatment and spillover effects. See Hardy et al. (2019), Leung (2019a) and Manski (2013) which also emphasise the difference between the structural model of social interactions and the reduced-form model focusing on the treatment response function. In addition, the identification strategy in this chapter does not require different network structures for the same set of individuals, the non-stochastic network, or a copula structure for the network-based variables. Instead, identification in this chapter is achieved by restricting the network measurement errors. Other related papers are, for example, Advani and Malde (2018), Chandrasekhar and Lewis (2011), Goldsmith-Pinkham and Imbens (2013), and Lewbel, Qu, and Tang (2019).

[1] Sävje, Aronow, and Hudgens (2017) find that when there is limited or moderate degree of network interactions, ignoring the network interference would not impact the asymptotic properties of the average treatment effect estimators. Chin (2018) studies the average treatment effects under unmodeled network interference. However, neither of them explore the spillover effect.

Consequences and solutions of misclassified networks on estimating network statistics or network formation are discussed by, for example Breza, Chandrasekhar, McCormick, and Pan (2020), Candelaria and Ura (2020), Comola and Fafchamps (2017), Kossinets (2006), Liu (2013) and Thirkettle (2019). However, it is not clear how to apply these methods to identify treatment and spillover effects in a causal setting.

The literature exploring limit theorems using network-dependent data is developing rapidly. Some papers assume that the social network can be partitioned into a large number of disjoint and independent subnetworks (e.g. Lewbel et al., 2019; Vazquez-Bare, 2019). However, this independence assumption may not be plausible in practice, because it ignores the links across subnetworks. Chandrasekhar and Lewis (2011) adopt mixing conditions to restrict the dependence of network connections, while, in many contexts, there is no underlying metric space to define the standard "mixing" forms of dependence. Leung (2020b) introduces the notion of "dependence graph" to capture the network-correlated effects, and derives limit theorems under the conditional local dependence; that is, the outcomes of two units are independent if they are not network neighbours nor share common network neighbours. However, in the setting considered in this chapter, the measurement errors disrupt the true network dependence structure, so that some seemingly uncorrelated units may actually correlate with each other due to the latent network connections, and vice versa. Therefore, an alternative data dependence structure is needed. We adopt the "dependence neighbourhoods" structure proposed by Chandrasekhar and Jackson (2016) to control the data correlation, which does not require the correct network links to be observed and employs less restrictions on the dependence structure. The dependency neighbourhood used in this chapter is similar to the dependency graph of Leung (2020b) in the sense that they both aim to control the data dependence. Nonetheless, they are different, since the dependency neighbourhoods can capture more general forms of correlation induced by network measurement errors and unobservables. Other papers study limit theorems of network dependent data include Chin (2018), Kojevnikov, Marmer, and Song (2019), Kuersteiner (2019), Lee and Ogburn (2020), Leung and Moon (2019), Leung (2019b, 2020a), Liu and Hudgens (2014), Song (2018), van der Laan (2014) and references therein.

4.3. Model Setup

Let $\mathbf{D} = \{D_i\}_{i\in\mathcal{P}}$ and $\mathbf{Z} = \{Z_i\}_{i\in\mathcal{P}}$ denote vectors consisting of units' (or individuals, nodes, agents) treatment status and observable characteristics for a super-population $\mathcal{P}$, respectively. Denote $\mathbf{A}^*$ as the true, latent and binary adjacency matrix, corresponding to an unweighted and undirected random network over the super-population $\mathcal{P}$. Each row of $\mathbf{A}^*$, denoted by $\mathbf{A}_i^*$, represents unit i's network connection with unit j.[2] Let $A_{ij}^* = 1$ if i and j are linked (or equivalently, network neighbours[3]), otherwise $A_{ij}^* = 0$. As a convention, self links are ruled out, i.e. $A_{ii}^* = 0$ for $\forall i \in \mathcal{P}$. Given the adjacency matrix $\mathbf{A}^*$, we define the set of unit i's first-degree network neighbours by $\mathcal{N}_i^* = \{j \in \mathcal{P} : A_{ij}^* = 1\}$. Denote $\mathcal{F}_i^* = \sum_{j\in\mathcal{P}} A_{ij}^*$ as the cardinality of $\mathcal{N}_i^*$, and $\mathcal{F}_i^*$ is usually referred to as the "network degree" of unit i. For each $i \in \mathcal{P}$, the outcome Y_i is defined as

$$Y_i = \tilde{r}(i, \mathbf{D}, \mathbf{A}^*, \mathbf{Z}, \varepsilon_i), \tag{4.1}$$

where $\tilde{r}$ is a unknown real-valued function and ε_i is an unobservable error term. The Y_i in (4.1) acknowledges that one unit's outcome depends on not only his or her own treatment status, but also the treatments assigned to other units, i.e., the spillover effect. We impose the assumption below to restrict the dependence of the outcome Y_i on $(i, \mathbf{D}, \mathbf{A}^*, \mathbf{Z}, \varepsilon_i)$.

Assumption 4.3.1 (Network Interference) *For* $\forall i, k \in \mathcal{P}$, $\forall(\mathbf{D}, \mathbf{A}^*, \mathbf{Z})$ *and* $\forall(\tilde{\mathbf{D}}, \tilde{\mathbf{A}}^*, \tilde{\mathbf{Z}})$,

$$\tilde{r}(i, \mathbf{D}, \mathbf{A}^*, \mathbf{Z}, e) = \tilde{r}(k, \tilde{\mathbf{D}}, \tilde{\mathbf{A}}^*, \tilde{\mathbf{Z}}, e),$$

for all $e \in \Omega_{\varepsilon_i} \cup \Omega_{\varepsilon_k}$, *if the following conditions hold simultaneously: (i)* $D_i = \tilde{D}_k$; *(ii)* $\sum_{j\in\mathcal{P}} A_{ij}^* = \sum_{j\in\mathcal{P}} \tilde{A}_{kj}^*$; *(iii)* $\sum_{j\in\mathcal{P}} A_{ij}^* D_j = \sum_{j\in\mathcal{P}} \tilde{A}_{kj}^* \tilde{D}_j$; *(iv)* $Z_i = \tilde{Z}_k$.

Assumption 4.3.1 is the equivalent to model (2) in Leung (2020b). It states that the outcome is fully determined by (i) unit's own treatment status; (ii) the network degree;

[2] The vectors of treatment status and observable characteristics, and the adjacency matrix are infinite-dimensional. We follow Leung (2020b) and obviate further details to ease the illustration.

[3] It is worthy to notice that there are two different definitions of neighbours utilised in this chapter. The first one, which is referred to as "network neighbours", is defined by the network links $\mathbf{D}$. The second one, which is referred to as "dependent neighbours", is defined via the dependency neighbourhoods and is used to characterise correlations of random variables of interest (see Section 4.5.1 for more details).

(iii) the number of the first-order treated network neighbours $S_i^* := \sum_{j\in\mathcal{P}} A_{ij}^* D_j$; and (iv) unit's own covariates.

Assumption 4.3.1 substantially reduces the dimensionality of the outcome and reveals two crucial features of the network interactions. First, the interference occurs locally, only among the first-order network neighbours. Thus, (D_i, S_i^*) can be viewed as the "effective treatment" (Manski, 2013). Second, the outcome is invariant to any permutations of the treatments received by the first-order network neighbours, meaning that the interactions are anonymous. The anonymous interaction is also referred to as "stratified interference", see Baird et al. (2018), Basse and Feller (2018) and Hudgens and Halloran (2008) among others. Aronow and Samii (2017), Leung (2019a) and Sävje (2019) consider the possible mis-specification of models similarly defined by Assumption 4.3.1, and tests for Assumption 4.3.1 are feasible in Athey et al. (2018). Under Assumption 4.3.1, equation (4.1) can be simplified to

$$Y_i = r(D_i, S_i^*, Z_i, \mathcal{F}_i^*, \varepsilon_i), \quad \text{for } \forall i \in \mathcal{P} \tag{4.2}$$

where r represents a real-valued unknown function. Such an outcome structure permits adequate controls for the observable and unobservable heterogeneity of the treatment response. Given (4.2), it is easy to see that unit i's outcome Y_i is directly affected by his or her own treatment status D_i (treatment effect), and is also affected by S_i^* because of the exposure to the treated peers (spillover effect). The network $\mathcal{N}_i^*$ affects the outcome via two pathways: the network degree $\mathcal{F}_i^*$ and the treated network neighbours incorporated in S_i^*. The network degree is a critical attribute because it quantifies the influence of each unit in the social network and controlling $\mathcal{F}_i^*$ in (4.2) enables us to target subpopulation based on different levels of influence. Besides, it acts as a control variable for the degree heterogeneity to allow the correlation between the network formation and the potential outcomes. The notion of "degree heterogeneity" is proposed by Graham (2017). Similar control variable method is used in e.g. Johnsson and Moon (2015).

The following notations will be used throughout the chapter. For any generic random variables X and Y, denote f_X and $f_{X|Y}$ as the probability distribution function of X and the conditional probability distribution function of X given Y, respectively. Ω_X denotes

the support of the random variable X. By notation abuse, $|B|$ denotes the cardinality of any set B, or the absolute value for any scalar B. For any vector $a \in \mathbb{R}^p$, let $\|a\|_1 = \sum_{i=1}^p |a_i|$ be its L^1 norm, $\|a\| = (a'a)^{1/2}$ be its Euclidean norm and $\|a\|_\infty = \max_{1\le i\le p} |a_i|$. Given a matrix $A = (a_{ij})$, we set $\|A\| = [tr(A'A)]^{1/2}$ and $\|A\|_\infty = \max_{1\le i,j\le p} |a_{ij}|$. More generally, for an array (or a vector) of functions, say $a = \{a_i\}$ with $a_i : \Omega_X \mapsto \mathbb{R}$, denote $\|a\|_\infty = \sup_{x\in\Omega_X} \sup_i |a_i(x)|$, where i could stand for a multiple index. For an arbitrary parameter β, denote $d_\beta = dim(\beta)$. $\perp$ means statistical independence.

4.3.1. Treatment and Spillover Effects

To motivate the potential identification issues, let us begin by defining key concepts and introducing basic assumptions.

Definition 4.3.1 (CASF) *For $\forall(d, s, z, n) \in \{0, 1\} \times \Omega_{S^*,Z,\mathcal{F}^*}$, the conditional average structural function (CASF) is defined as*

$$m^*(d, s, z, n) = \mathbb{E}\left[r(d, s, Z_i, \mathcal{F}_i^*, \varepsilon_i)\middle| Z_i = z, \mathcal{F}_i^* = n\right].$$

In this chapter, we focus on treatment and spillover effects, measured the average change in potential outcomes in response to the counterfactual manipulation of the treatment assigned to the ego unit and network peers, respectively. Similar definitions measuring the direct effect and the spillover effect of treatment are also introduced in Hudgens and Halloran (2008) and Sobel (2006) to name a few. See Tchetgen and VanderWeele (2012) for a discussion about relationships between various notions of causal effects in the presence of network interference. The analysis in this chapter can be straightforwardly extended to studies dealing with other notions of treatment effect estimands.

Definition 4.3.2 (Treatment and Spillover Effects) *For $\forall(s, z, n) \in \Omega_{S^*,Z,\mathcal{F}^*}$, define*

$$\textit{treatment effect: } \tau_d(s, z, n) = m^*(1, s, z, n) - m^*(0, s, z, n),$$
$$\textit{spillover effect: } \tau_s(s, z, n) = m^*(0, s, z, n) - m^*(0, 0, z, n).$$

The assumption below introduces the ignorability conditions that account for network interference, based on which the causal effects of interest can be recovered if the actual network data is available.

Assumption 4.3.2

(a) (**Randomised Treatment**) $\{D_i\}_{i\in\mathcal{P}}$ *are i.i.d. and* $\{D_i\}_{i\in\mathcal{P}} \perp \{\varepsilon_j, Z_j, \mathcal{N}_j^*\}_{j\in\mathcal{P}}$.

(b) (**Unconfounded Network**) *For* $\forall i \in \mathcal{P}$, $\varepsilon_i \perp \left(\mathcal{N}_i^*, \{D_j\}_{j\in\mathcal{N}_i^*}\right) \big| Z_i, \mathcal{F}_i^*$.

Assumption 4.3.2 (a) states that the treatment is randomly assigned and independent of the potential outcomes, and does not affect the network. Randomised intervention has been used in a wide range of experimental contexts, including Miguel and Kremer (2004), Aral and Walker (2012), Oster and Thornton (2012), Cai et al. (2015b) to name a few, and see Athey and Imbens (2017) for a review. Therefore, Assumption 4.3.2 (a) is a straightforward starting point for the analysis. Assumption 4.3.2 (b) requires the unconfounded network, which is weaker than the fully exogenous network, by allowing the correlation between the degree $\mathcal{F}_i^*$ and the unobservable characteristics, for example, through the spillovers of unobservables. See Leung (2020b) for a similar assumption and supportive examples. The network unconfoundedness to the treatment and the potential outcomes is likely to hold in randomised experiments where the network data is collected before the intervention.

Assumption 4.3.3 (Distribution)

(a) $\{Z_i\}_{i\in\mathcal{P}}$ *are i.i.d. and* $\mathcal{F}_i^*$ *given* Z_i *is identically distributed across* $i \in \mathcal{P}$.

(b) For $\forall i \in \mathcal{P}$, ε_i *given* $(Z_i, \mathcal{F}_i^*)$ *is identically distributed.*

Assumption 4.3.3 (a) implies that the covariate Z_i is of randomly drawn samples, which is standard in the literature on network effect models, e.g., Johnsson and Moon (2015) and Auerbach (2019). In the analysis of this chapter, it is feasible to relax the i.i.d. of Z_i and allow it to possess dependent structure under the framework described in Section 4.5.1. We maintain such an i.i.d. assumption only for illustration simplicity. It also

requires the conditional distribution of the network degree to be invariant across units. An example of the dyadic network formation in Appendix 4.9.1 can be used to verify the existence of such an identical distribution. Also see a strategic network formation model in Leung (2020b) that satisfies (a). Moreover, the identical distribution of the error term ε_i given $(Z_i, \mathcal{F}_i^*)$ in condition (b) permits that the expressions of the CASF, the treatment effect τ_d and the spillover effect τ_s are all identical for any unit $i \in \mathcal{P}$.

If the actual network $\mathcal{N}_i^*$ is correctly observed, under the assumptions introduced so far, the CASF can be identified by[4]

$$m^*(d, s, z, n) = \mathbb{E}\left[Y_i \middle| D_i = d, S_i^* = s, Z_i = z, \mathcal{F}_i^* = n\right],$$

which ensures that the treatment and spillover effects are also identifiable. However, it appears that, in many applications, we fail to obtain fully accurate network information. Ignoring the missing or misclassified network links may lead to biased estimation and misleading causal implications.

4.3.2. Bias of CASF with Mismeasured Network

This subsection presents the potential bias of the CASF identified from the mismeasured network data. Suppose that researchers randomly draw N units from the population $\mathcal{P}$, and collect their outcomes of interest, treatment status, covariates, network information and treatment assignments of their network neighbours. Thus, researchers can observe:

$$(Y_i, D_i, Z_i, \mathcal{N}_i, \{D_j\}_{j \in \mathcal{N}_i}), \text{ for } i = 1, 2, ..., N,$$

where $\mathcal{N}_i$ denotes the observed identities of unit i's network neighbours with cardinality $\mathcal{F}_i$, and the convention of no self connections is maintained, i.e. $i \notin \mathcal{N}_i$. Note that

[4]For $\forall (d, s, z, n) \in \{0, 1\} \times \Omega_{S^*, Z, \mathcal{F}^*}$, it can be shown that

$$\begin{aligned}\mathbb{E}\left[Y_i \middle| D_i = d, S_i^* = s, Z_i = z, \mathcal{F}_i^* = n\right] =& \mathbb{E}\left[r(D_i, S_i^*, Z_i, \mathcal{F}_i^*, \varepsilon_i) \middle| D_i = d, S_i^* = s, Z_i = z, \mathcal{F}_i^* = n\right] \\ =& \mathbb{E}\left[r(d, s, z, n, \varepsilon_i) \middle| Z_i = z, \mathcal{F}_i^* = n\right] \\ =& m^*(d, s, z, n),\end{aligned}$$

where the second equality is due to the unconfoundedness of (D_i, S_i^*) in Lemma 4.9.12 and the last equality is by Definition 4.3.1.

there are no restrictions on the sampling scheme of the network data. Namely, $\mathcal{N}_i$ can be obtained from a single and fully observed network, or from a (possibly partially observed) sampled network. In addition, $\mathcal{N}_i$ can be either self-reported, acquired from the administrative data, or constructed by researchers based on specific rules. Throughout the chapter, $\mathcal{N}_i$ is referred to as the "network proxy". Given $\mathcal{N}_i$, the number of observed treated network neighbours is denoted by $S_i = \sum_{j \in \mathcal{N}_i} D_j$.

The assumption below extends Assumption 4.3.2 to accommodate the observable network proxy by restricting the misclassification of the network links.

Assumption 4.3.4 (Nondifferential Misclassification)

(a) $\{D_i\}_{i \in \mathcal{P}} \perp \{\varepsilon_j, Z_j, \mathcal{N}_j^*, \mathcal{N}_j\}_{j \in \mathcal{P}}$;

(b) For $\forall i \in \mathcal{P}$, $\varepsilon_i \perp \left(\mathcal{N}_i^*, \{D_j\}_{j \in \mathcal{N}_i^*}, \mathcal{N}_i, \{D_j\}_{j \in \mathcal{N}_i}\right) \big| Z_i, \mathcal{F}_i^*$.

(c) For $\forall i \in \mathcal{P}$, $\mathcal{F}_i$ *given* $(Z_i, \mathcal{F}_i^*)$ *is identically distributed.*

Assumption 4.3.4 (a) and (b) indicate that, given the actual network information and individual's characteristics, the observed proxy $\mathcal{N}_i$ does not contain relevant information to predict the outcome. This is often referred to as "nondifferential misclassification" in the measurement error models literature, e.g. Battistin and Sianesi (2011), Hu (2008) and Lewbel (2007). In addition, Assumption 4.3.4 (c) holds in many contexts, for example, when units fail to respond with probability proportional to their actual degrees ("the load effect"), or inversely proportional to their actual degrees ("the periphery effect") (Kossinets, 2006). A set of sufficient conditions for Assumption 4.3.4 (c) is provided in Appendix 4.9.1.

Now, denote the conditional mean of the outcome given the observables as

$$m_i(d, s, z, n) = \mathbb{E}[Y_i | D_i = d, S_i = s, Z_i = z, \mathcal{F}_i = n],$$

where the subscript i of m_i represents the possibly non-identical conditional mean of the outcome given the observables, which is caused by the unknown dependence between

the error contaminated network-variables $(S_i, \mathcal{F}_i)$ and their latent counterparts $(S_i^*, \mathcal{F}_i^*)$. The relationship between m_i and m^* can be obtained by the proposition below.

Proposition 4.3.3 *Under Assumptions 4.3.1-4.3.4, for* $\forall i \in \mathcal{P}$ *and* $\forall (d, s, z, n) \in \{0,1\} \times \Omega_{S,Z,\mathcal{F}}$,

$$m_i(d, s, z, n) = \sum_{(s^*, n^*) \in \Omega_{S^*, \mathcal{F}^*}} m^*(d, s^*, z, n^*) f_{S_i^*, \mathcal{F}_i^* | D_i = d, S_i = s, \mathcal{F}_i = n, Z_i = z}(s^*, n^*).$$

Proposition 4.3.3 characterises the bias in the CASF estimand if ignoring the measurement errors of the network links. The expression of m_i makes it clear that the bias of m_i is governed by the latent distribution of the actual network-based variable $(S_i^*, \mathcal{F}_i^*)$ given its observed counterpart $(S_i, \mathcal{F}_i)$. This bias will be larger, if the misclassification probability of $(S_i, \mathcal{F}_i)$ is higher. Importantly, due to the nonparametric setting of m^*, simply differencing $m_i(1, s, z, n)$ and $m_i(0, s, z, n)$ in general cannot give the treatment effect $\tau_d(s, z, n)$, even though the treatment is randomised and correctly-observed. However, it will be true if the response to the variation of the ego unit's treatment status is homogeneity in both the observables and unobservables, relying on strong structural restriction. Similar weighted average expressions of the identifiable parameter are presented by Gao and Li (2019) for the endogenous peer effects and by Hardy et al. (2019) for the treatment spillover effects.

4.4. Identification

Let us first introduce a key lemma in decomposing the latent distribution $f_{S_i^*, \mathcal{F}_i^* | D_i, S_i, \mathcal{F}_i, Z_i}$ into identifiable components.

Lemma 4.4.1 *Under Assumption 4.3.2 (a) and 4.3.4 (a),*

(a) $\mathcal{N}_i^* \perp S_i^* \big| Z_i, \mathcal{F}_i^*$ *and* $\mathcal{F}_i \perp S_i^* \big| Z_i, \mathcal{F}_i^*$;

(b) $\mathcal{N}_i \perp S_i \big| Z_i, \mathcal{F}_i$ *and* $\mathcal{F}_i^* \perp S_i \big| Z_i, \mathcal{F}_i$;

(c) for $\forall (s, n) \in \Omega_{S,\mathcal{F}}$, $f_{S_i^* | \mathcal{F}_i^* = n, Z_i}(s) = f_{S_i | \mathcal{F}_i = n, Z_i}(s) = C_n^s f_D(1)^s f_D(0)^{n-s}$, *for* $f_D(d) :=$

$Pr(D_i = d)$ *with* $d \in \{0, 1\}$.

Lemma 4.4.1 (a) delivers two implications. First, the distribution of the number of treated network neighbours, S_i^*, is fully determined by the true network degree and exogenous covariates, instead of the identity of network neighbours or the observable network degree. It further restricts the anonymous interactions. Lemma 4.4.1 (b) states the similar properties of S_i. The identifiability of $f_{S_i^*|\mathcal{F}_i^*,Z_i}$ in (c) is intuitive, because the treatments are randomly assigned and the summation of any given n i.i.d. treatment status follows a binomial distribution.

Given Proposition 4.3.3 and Lemma 4.4.1, to identify the CASF m^*, we can first decompose the latent distribution function $f_{S_i^*,\mathcal{F}_i^*|D_i,S_i,\mathcal{F}_i,Z_i}$ as follows.

Proposition 4.4.2 *Under Assumptions 4.3.2 and 4.3.4*

$$f_{S_i^*,\mathcal{F}_i^*|D_i,S_i,\mathcal{F}_i,Z_i} = \frac{f_{S_i|S_i^*,\mathcal{F}_i^*,\mathcal{F}_i,Z_i} \times f_{S_i^*|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i^*|Z_i}}{f_{S_i|\mathcal{F}_i,Z_i} \times f_{\mathcal{F}_i|Z_i}}. \tag{4.3}$$

It is clear that $f_{\mathcal{F}_i|Z_i}$ and $f_{S_i|\mathcal{F}_i,Z_i}$ can be identified directly from the observables under the assumptions exploited in the previous sections, and $f_{S_i^*|\mathcal{F}_i^*,Z_i}$ is identifiable based on Lemma 4.4.1.

In what follows, we will deal with the identification of the remaining distributions in the decomposition in two steps. First, suppose that two network proxies are available for each sampled individual $i \in \{1, 2, ..., N\}$, denoted by $\mathcal{N}_i$ and $\tilde{\mathcal{N}}_i$. They may come from repeated observations of a sampled network over time, different dimensions of connections (e.g., kinship and borrowing-lending), multi-contextual interactions (e.g., various social events or afflictions), or self-reported and administrative network data. Intuitively, the additional network proxy $\tilde{\mathcal{N}}_i$ can be understood as an instrument for the true latent network (Hu, 2008; Hu and Schennach, 2008), that is conceptually similar to the ones utilised in conventional instrumental variable methods. Following the same construction, for $\tilde{\mathcal{N}}_i$, denote its cardinality as $\tilde{\mathcal{F}}_i$ and the number of treated network neighbours as $\tilde{S}_i = \sum_{j \in \tilde{\mathcal{N}}_i} D_j$. Given the two observed network proxies $\mathcal{N}_i$ and $\tilde{\mathcal{N}}_i$, apply the method of matrix diagonalisation of Hu (2008) to achieve the identification of $f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}$

and $f_{\mathcal{F}_i^*|Z_i}$. Due to the complex and unconstrained interdependence between the observed $(S_i, \mathcal{F}_i)$, $(\tilde{S}_i, \tilde{\mathcal{F}}_i)$ and their latent counterpart $(S_i^*, \mathcal{F}_i^*)$ through the underlying network $\mathcal{N}_i^*$, it is not feasible to identify the latent distribution $f_{S_i|S_i^*,\mathcal{F}_i^*,\mathcal{F}_i,Z_i}$ by simply repeating the matrix diagonalisation approach. Therefore, in the second step, we introduce a crucial assumption on the network measurement errors, which dramatically simplifies the interdependence and ensures the identification of $f_{S_i|S_i^*,\mathcal{F}_i^*,\mathcal{F}_i,Z_i}$.

4.4.1. Identification via Matrix Diagonalisation

Assumptions 4.4.1 to 4.4.4 below are crucial when establishing the identification results via the matrix diagonalisation technique similar to that used by Hu (2008). Modifications to the assumptions and method are made, to fit the network setting considered in this chapter.

Assumption 4.4.1 (Exclusion Restriction) $\mathcal{F}_i \perp \tilde{\mathcal{F}}_i \big| Z_i, \mathcal{F}_i^*$.

Assumption 4.4.1 can be interpreted as a standard exclusion restriction that $\tilde{\mathcal{F}}_i$ does not provide extra information about $\mathcal{F}_i$ than the actual degree $\mathcal{F}_i^*$ already provides. It can also be understood as that the instrumental variable $\tilde{\mathcal{N}}_i$ is conditionally independent of the measurement errors in $\mathcal{N}_i$. It rules out the situations where both network proxies are mismeasured due to random omission of the same group of units when constructing the networks. A set of sufficient conditions for Assumption 4.4.1 is given in Appendix 4.9.1. The exclusion restriction is the key to implementing the matrix diagonalisation method.

Assumption 4.4.2 (Sparsity) $\Omega_{\tilde{\mathcal{F}}} = \Omega_{\mathcal{F}} = \Omega_{\mathcal{F}^*}$ *with finite cardinality* $K_{\mathcal{F}}$.

Assumption 4.4.2 requires that the network is sparse, i.e. each individual has finite friends, and that the number of friends does not increase with the sample size. Sparse networks are commonly observed in empirical applications (Chandrasekhar, 2016), and are a standard assumption in the literature on network effects, e.g. De Paula, Richards-Shubik, and Tamer (2018), Qu and Lee (2015) and Viviano (2019). By the i.i.d. of the treatment assignment, it is clear that $\Omega_{\widetilde{S}} = \Omega_S = \Omega_{S^*}$.

To illustrate the basic idea of the matrix diagonalisation technique, let us introduce the following notations. Without loss of generality, set $\Omega_{\mathcal{F}^*} = \Omega_{\mathcal{F}} = \Omega_{\tilde{\mathcal{F}}} = \{0, 1, ..., K_{\mathcal{F}} - 1\}$. Denote the $K_{\mathcal{F}} \times K_{\mathcal{F}}$ matrix $F_{\mathcal{F}|\mathcal{F}^*,Z}$ as

$$F_{\mathcal{F}|\mathcal{F}^*,Z} = \begin{bmatrix} f_{\mathcal{F}_i|\mathcal{F}_i^*=0,Z_i}(0) & \cdots & f_{\mathcal{F}_i|\mathcal{F}_i^*=K_{\mathcal{F}}-1,Z_i}(0) \\ \vdots & \ddots & \vdots \\ f_{\mathcal{F}_i|\mathcal{F}_i^*=0,Z_i}(K_{\mathcal{F}}-1) & \cdots & f_{\mathcal{F}_i|\mathcal{F}_i^*=K_{\mathcal{F}}-1,Z_i}(K_{\mathcal{F}}-1) \end{bmatrix}. \tag{4.4}$$

Similarly, define $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$ via replacing $f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}$ in (4.4) by $f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*,Z_i}$. In addition, define two observable $K_{\mathcal{F}} \times K_{\mathcal{F}}$ matrices

$$F_{\tilde{\mathcal{F}},\mathcal{F}|Z} = \{f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|Z_i}(i,j)\}, \text{ and } \quad E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} = \left\{\int_{y\in\Omega_Y} y f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|Z_i}(i,j,y)dy\right\},$$

with $i, j = 0, 1, ..., K_{\mathcal{F}} - 1$, and define a $K_{\mathcal{F}} \times K_{\mathcal{F}}$ diagonal matrix

$$T_{Y|\mathcal{F}^*,Z} = diag\left(\mathbb{E}[Y_i|\mathcal{F}_i^* = 0, Z_i], \mathbb{E}[Y_i|\mathcal{F}_i^* = 1, Z_i], \cdots, \mathbb{E}[Y_i|\mathcal{F}_i^* = K_{\mathcal{F}} - 1, Z_i]\right).$$

The main idea of the matrix diagonalisation method is to identify the latent distributions of interest via diagonalising the matrix of directly observable distributions $E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z}$ as

$$T_{Y|\mathcal{F}^*,Z} = F^{-1}_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z} \times \left(E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z}\right) \times F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}.$$

Then, recover the latent distributions in $F_{\mathcal{F}|\mathcal{F}^*,Z}$ and $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$ via the eigen-decomposition approach: columns of $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$ are the eigenvectors of the matrix $E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z}$, and diagonal elements of $T_{Y|\mathcal{F}^*,Z}$ are the corresponding eigenvalues. Note that the discussion above is based on the preassumption about the invertibility of $F_{\mathcal{F}|\mathcal{F}^*,Z}$ and $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$, which is formalised by Assumption 4.4.3 below.

Assumption 4.4.3 (Rank Condition) *The ranks of $F_{\mathcal{F}|\mathcal{F}^*,Z}$ and $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$ are both $K_{\mathcal{F}}$.*

The next assumption is the key to identifying latent probabilities via eigen-decomposition.

Assumption 4.4.4 (Eigen-decomposition)

(a) For $\forall n, n' \in \Omega_{\mathcal{F}^*}$ *such that* $n \neq n'$*, we have* $\mathbb{E}[Y_i|\mathcal{F}_i^* = n, Z_i] \neq \mathbb{E}[Y_i|\mathcal{F}_i^* = n', Z_i]$.

(b) For $\forall n^* \in \Omega_{\mathcal{F}^*}$ *and any* $n \neq n^*$*, we have*

$$f_{\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(n^*) > f_{\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(n), \quad f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*=n^*,Z_i}(n^*) > f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*=n^*,Z_i}(n).$$

Assumption 4.4.4 (a) is a sufficient condition to avoid duplicate eigenvalues so that the eigen-decomposition is unique. It is automatically satisfied if $\mathbb{E}[Y_i|\mathcal{F}_i^*, Z_i]$ is monotone in $\mathcal{F}_i^*$ and it also holds for more general scenarios. Noticing that the condition (a) is a special case of a more general condition $\mathbb{E}[\varpi(Y_i)|\mathcal{F}_i^* = n, Z_i] \neq \mathbb{E}[\varpi(Y_i)|\mathcal{F}_i^* = n', Z_i]$, where the transformation function $\varpi(\cdot)$ can be user-specified, such as $\varpi(y) = (y - \mathbb{E}[Y_i])^2$ (variance) or $\varpi(y) = 1[y \leq y_0]$ (quantile) for some given y_0. Assumption 4.4.4 (b) permits that the order of the eigenvectors is identifiable. It indicates that the observable network degrees are informative proxies for the latent degree, which implicitly assumes that the probability of correctly reporting is higher than that of misreporting. Similar restrictions are widely invoked in the literature on measurement error models. See, for example, Battistin and Sianesi (2011), Battistin, De Nadai, and Sianesi (2014), Chen, Hong, and Nekipelov (2011), Hu and Schennach (2008), Lewbel (2007) and Mahajan (2006).

Theorem 4.4.3 *Suppose Assumption 4.3.4 is satisfied by* $\tilde{\mathcal{N}}_i$ *and* $\mathcal{N}_i$*. Under Assumptions 4.3.1-4.3.3 and 4.4.1,*

(a) $f_{\mathcal{F}_i|Z_i}$*,* $f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|Z_i}$ *and* $f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|Z_i}$ *are identical across* $i \in \mathcal{P}$*.*

(b) If further assume Assumptions 4.4.2-4.4.4 hold, then $f_{\mathcal{F}_i^*|Z_i}$*,* $f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}$ *and* $f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*,Z_i}$ *are nonparametrically identified.*

4.4.2. Identification via One Type of Measurement Error

Next, let us proceed with the identification of $f_{S_i|S_i^*,\mathcal{F}_i,\mathcal{F}_i^*,Z_i}$. The matrix diagonalisation method is infeasible in this step, because of the violation of the exclusion restriction analogue to Assumption 4.4.1. In other words, the conditional independence $S_i \perp \tilde{S}_i$

given $(S_i^*, \mathcal{Z}_i)$ with $\mathcal{Z}_i = (Z_i, \mathcal{F}_i^*)$ does not hold. To be more specific, consider the expression of S_i in terms of S_i^* below

$$S_i = S_i^* - \sum_{j \in \mathcal{N}_i^* / \mathcal{N}_i} D_j + \sum_{j \in \mathcal{N}_i / \mathcal{N}_i^*} D_j, \tag{4.5}$$

where for any sets A and B, let $A/B := A \bigcap B^c$ with B^c being the complement of B. The set $\mathcal{N}_i^*/\mathcal{N}_i$ contains all the missing network links of i (false negative), and the set $\mathcal{N}_i/\mathcal{N}_i^*$ includes all the false network links (false positive). Similarly, $\tilde{S}_i = S_i^* - \sum_{j \in \mathcal{N}_i^*/\tilde{\mathcal{N}}_i} D_j + \sum_{j \in \tilde{\mathcal{N}}_i/\mathcal{N}_i^*} D_j$. Given $(S_i^*, \mathcal{Z}_i)$, the remaining parts in $S_i - S_i^*$ and $\tilde{S}_i - S_i^*$ contain two sources of randomness: (i) the network measurement errors and (ii) the treatment status of the missing and falsely connected network neighbours. Although the measurement errors of the two network proxies are conditional independent as implicitly implied by the exclusion restriction, without further restrictions, however, we cannot rule out the dependence arising from those mismeasured network neighbours' treatment status, which may appear in both S_i and $\tilde{S}_i$. Therefore, S_i and $\tilde{S}_i$ are correlated conditionally on $(S_i^*, \mathcal{Z}_i)$, contradicting the exclusion restriction.

Based on the discussion above, the main issue in identifying $f_{S_i|S_i^*,\mathcal{F}_i,\mathcal{F}_i^*,Z_i}$ arises from the dependence between $(S_i, \mathcal{F}_i)$ and $(S_i^*, \mathcal{F}_i^*)$. Their dependence is not easy to characterise, because $(S_i, \mathcal{F}_i)$ and $(S_i^*, \mathcal{F}_i^*)$ relate to each other via the underlying network $\mathcal{N}_i^*$ which is unobservable, and the arbitrary measurement error further complicates their relationship. The latter occurs because, without imposing any constraint on the measurement errors, given $(S_i^* = s^*, \mathcal{F}_i^* = n^*, \mathcal{F}_i = n)$, there will be various realisations of $\mathcal{N}_i$ and $\mathcal{N}_i^*$, each of which may lead to a substantially different S_i. For example, when $n = n^*$, all network links may be classified correctly, therefore $\mathcal{N}_i = \mathcal{N}_i^*$. If so, S_i would be entirely determined by its latent counterpart S_i^*. However, it is also possible that not a single element in $\mathcal{N}_i$ and $\mathcal{N}_i^*$ will be the same, although they have the same cardinality. If that is the case, then S_i would be solely governed by the treatment status of the misreported false network neighbours $\sum_{j \in \mathcal{N}_i} D_j$, and would no longer depend on $(S_i^*, \mathcal{F}_i^*)$. Therefore, without further restricting the measurement error, there will be too little information and too much uncertainty to pin down $f_{S_i|S_i^*,\mathcal{F}_i,\mathcal{F}_i^*,Z_i}$.

For any given $n \in \Omega_{\mathcal{F}}$ and $n^* \in \Omega_{\mathcal{F}^*}$, the $(n+1) \times (n^*+1)$ unknown conditional probabilities of S_i which characterise the dependence structure between $(S_i, \mathcal{F}_i)$ and $(S_i^*, \mathcal{F}_i^*)$, can be formalised by the following $(n+1) \times (n^*+1)$ matrix:

$$F_{S|S^*,\mathcal{F},\mathcal{F}^*,Z} = \begin{bmatrix} f_{S_i|S_i^*=0,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(0) & \cdots & f_{S_i|S_i^*=n^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(0) \\ \vdots & \ddots & \vdots \\ f_{S_i|S_i^*=0,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(n) & \cdots & f_{S_i|S_i^*=n^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(n) \end{bmatrix}. \tag{4.6}$$

Denote a $(n+1) \times 1$ vector $F_{S|\mathcal{F},Z}$ and a $(n^*+1) \times 1$ vector $F_{S^*|\mathcal{F}^*,Z}$ by

$$\begin{aligned} F_{S|\mathcal{F},Z} =& [f_{S_i|\mathcal{F}_i=n,Z_i}(0), \cdots, f_{S_i|\mathcal{F}_i=n,Z_i}(n)]', \\ F_{S^*|\mathcal{F}^*,Z} =& [f_{S_i^*|\mathcal{F}_i^*=n^*,Z_i}(0), \cdots, f_{S_i^*|\mathcal{F}_i^*=n^*,Z_i}(n^*)]', \end{aligned}$$

where both the vectors are identifiable. This yields a system of $(n+1)$ linear equations with $(n+1) \times (n^*+1)$ unknowns from Lemma 4.4.1 (b) and the law of total probability:[5]

$$F_{S|\mathcal{F},Z} = F_{S|S^*,\mathcal{F},\mathcal{F}^*,Z} \times F_{S^*|\mathcal{F}^*,Z}, \tag{4.7}$$

which, however, is underdetermined because there are fewer equations than unknowns. Therefore, it is necessary to impose restrictions to reduce the number of unknown parameters to get a unique solution for the system (4.7). Fortunately, this goal is achieved, if the possibility of either false negative or false positive can be ruled out.

Without loss of generality, suppose that no false negative holds (i.e. $\mathcal{N}_i^* \subset \mathcal{N}_i$), which essentially requires the observed network to be larger than the true network. Firstly, $\mathcal{N}_i^* \subset \mathcal{N}_i$ enforces a sparsity constraint on the unknowns: given $S_i^* = s^*$, the probability of $S_i = s$ with $s < s^*$ should be zero, as the only source of misclassification in S_i is from those false connections. Therefore, the elements above the main diagonal of the matrix

[5] Equation (4.7) is because

$$\begin{aligned} f_{S_i|\mathcal{F}_i=n,Z_i}(s) =& f_{S_i|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s) \\ =& \sum_{s^* \in \Omega_{S^*}} f_{S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s) \times f_{S_i^*|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s^*) \\ =& \sum_{s^* \in \Omega_{S^*}} f_{S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s) \times f_{S_i^*|\mathcal{F}_i^*=n^*,Z_i}(s^*). \end{aligned}$$

$F_{S|S^*,\mathcal{F},\mathcal{F}^*,Z}$ are all zero. Secondly, $\mathcal{N}_i^* \subset \mathcal{N}_i$ also dramatically simplifies the dependence structure between $(S_i, \mathcal{F}_i)$ and $(S_i^*, \mathcal{F}_i^*)$ via limiting the possible realisations of $\mathcal{N}_i$ and $\mathcal{N}_i^*$, so that the elements in each k-diagonal ($k = 0, -1, ..., -n$) of the matrix $F_{S|S^*,\mathcal{F},\mathcal{F}^*,Z}$ will be the same. It is because, under the no false negative assumption, the treated true network neighbours are all observed as treated network neighbours, and the untreated true network neighbours are all observed as untreated. Hence, the $S_i - S_i^*$ extra observed treated neighbours can only come from the $\mathcal{F}_i - \mathcal{F}_i^*$ falsely connected neighbours. Due that the treatment assignment is randomised, intuitively,

$$f_{S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s) = f_{S_i-S_i^*|\mathcal{F}_i-\mathcal{F}_i^*=n-n^*,Z_i}(s-s^*),$$

indicating that the distribution of interest reduces to the probability of randomly choosing $s - s^*$ units out of $n - n^*$ units, which does not vary with the realisations of S_i^* as long as the difference $S_i - S_i^*$ is the same. Denote $\Delta S_i = S_i - S_i^*$ and $\Delta\mathcal{F}_i = \mathcal{F}_i - \mathcal{F}_i^*$. Now, under the no false negative assumption, for any $n^* \leq n$, the matrix $F_{S|S^*,\mathcal{F},\mathcal{F}^*,Z}$ can be simplified to

$$F_{S|S^*,\mathcal{F},\mathcal{F}^*,Z} = \begin{bmatrix} f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(0) & 0 & \cdots & 0 \\ f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(1) & f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(0) & \cdots & 0 \\ \vdots & f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(1) & \ddots & 0 \\ \vdots & \vdots & \ddots & f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(0) \\ f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(\Delta n) & \vdots & \vdots & f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(1) \\ \vdots & f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(\Delta n) & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(n) & f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(n-1) & \cdots & f_{\Delta S_i|\Delta\mathcal{F}_i=\Delta n,Z_i}(\Delta n) \end{bmatrix}, \tag{4.8}$$

with $(n + 1)$ unknowns, the same as the number of equations, which ensures a unique solution for (4.7) and the identification of $f_{S_i|S_i^*,\mathcal{F}_i,\mathcal{F}_i^*,Z_i}$. Under no false negative, we do not consider the case $n < n^*$, because $\mathcal{N}_i^* \subset \mathcal{N}_i$ implies that the event $(\mathcal{F}_i^*, \mathcal{F}_i) = (n^*, n)$ with $n < n^*$ is a zero probability even, and a conditional probability conditional on a zero probability even is undefined. Similarly, under no false positive, we do not consider the case $n > n^*$. It is worth to note the equivalence between $f_{S_i|S_i^*,\mathcal{F}_i,\mathcal{F}_i^*,Z_i}$ and $f_{S_i^*|S_i,\mathcal{F}_i,\mathcal{F}_i^*,Z_i}$

via re-scaling:

$$f_{S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s) = f_{S_i^*|S_i=s,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s^*) f_{S_i|\mathcal{F}_i=n,Z_i}(s) / f_{S_i^*|\mathcal{F}_i^*=n^*,Z_i}(s^*),$$

where the equality is based on Lemma 4.4.1. Therefore, similar arguments can be applied when no false positive assumption holds. The discussion above only requires one of the two network proxies satisfying the desired property, and does not impose any restriction on the measurement errors of the other proxy, except for those assumed previously. Without loss of generality, hereafter we use $\mathcal{N}_i$ to denote the one that satisfies the requirement.

Assumption 4.4.5 (One Type of Measurement Error) *For each unit $i \in \mathcal{P}$, the proxy $\mathcal{N}_i$ satisfies either no false positive, i.e. $\mathcal{N}_i \subset \mathcal{N}_i^*$, or no false negative, i.e. $\mathcal{N}_i^* \in \mathcal{N}_i$.*

Borrowing the terminology from Calvi et al. (2018), Assumption 4.4.5 is referred to as "one type of measurement error". As can be seen from the next lemma, exploiting Assumption 4.4.5 benefits us the significant simplicity of the interdependence between the observable $(S_i, \mathcal{F}_i)$ and the latent $(S_i^*, \mathcal{F}_i^*)$, which dramatically reduces the number of unknown probabilities.

Lemma 4.4.4 *Suppose Assumptions 4.3.2, 4.3.4 and 4.4.5 hold. Let $\Delta s = |s - s^*|$ and $\Delta n = |n - n^*|$. For $\forall (s^*, n^*) \in \Omega_{S^*,\mathcal{F}^*}$ and $\forall (s, n) \in \Omega_{S,\mathcal{F}}$, we have that $f_{S_i^*|S_i,\mathcal{F}_i^*,\mathcal{F}_i,Z_i}$ is identical across $i \in \mathcal{P}$.*[6]

(a) If no false negative $\mathcal{N}_i^* \subset \mathcal{N}_i$ holds, then for $n^* \leq n$,

$$f_{S_i|S_i^*=s^*,\mathcal{F}_i^*=n^*,\mathcal{F}_i=n,Z_i}(s) = \begin{cases} C_{\Delta n}^{\Delta s} f_D(1)^{\Delta s} f_D(0)^{\Delta n - \Delta s}, & \text{if } s^* \leq s \text{ and } \Delta s \leq \Delta n \\ 0, & \text{otherwise.} \end{cases}$$

[6] The conditions $(s^*, n^*) \in \Omega_{S^*,\mathcal{F}^*}$ and $(s, n) \in \Omega_{S,\mathcal{F}}$ implicitly imply that $0 \leq s \leq n$ and $0 \leq s^* \leq n^*$.

(b) If no false positive $\mathcal{N}_i \subset \mathcal{N}_i^*$ *holds, then for* $n \leq n^*$

$$f_{S_i^*|S_i=s,\mathcal{F}_i^*=n^*,\mathcal{F}_i=n,Z_i}(s^*) = \begin{cases} C_{\Delta n}^{\Delta s} f_D(1)^{\Delta s} f_D(0)^{\Delta n - \Delta s}, & \text{if } s \leq s^* \text{ and } \Delta s \leq \Delta n \\ 0, & \text{otherwise.} \end{cases}$$

It is perhaps not surprising that S_i conditional on $(S_i^*, \mathcal{F}_i^*, \mathcal{F}_i, Z_i)$ follows a binomial distribution, given the equivalence of $f_{S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s^*)$ to the probability of randomly assigning treatment to Δs out of Δn units. The result in Lemma 4.4.4 enables a faster and easier way to compute $f_{S_i^*|S_i,\mathcal{F}_i^*,\mathcal{F}_i,Z_i}$ without solving the linear system. Nevertheless, the linear system greatly facilitates the identification analysis and determines the identification status of $f_{S_i^*|S_i,\mathcal{F}_i^*,\mathcal{F}_i,Z_i}$, and the solution of the system produces the same result as that obtained by simply exploiting the binomial distribution.

Theorem 4.4.5 *Under Assumptions 4.3.2-4.3.4 and 4.4.5,* $f_{S_i^*,\mathcal{F}_i^*|D_i,S_i,\mathcal{F}_i,Z_i}$ *is identical across* $i \in \mathcal{P}$ *and is nonparametrically identified.*

"No false positive" assumption is satisfied in many situations, for instance, when the mismeasurement is caused by sampling-induced errors, such as missing links ("induced subgraph" in Kossinets, 2006); restricting the network within a village (Angelucci et al., 2010); or limiting the maximum number of nominated friends (Cai et al., 2015b). It is also satisfied when non-sampling-induced errors arise, for example, when a survey respondent becomes uninterested in naming the full list of the friends due to survey fatigue; when there exists a lack of measurability of abstract but meaningful connections (e.g. esteem or authority); when constructing networks by intersecting repeated network observations, assuming that the overlap includes those effectual interactions; when keeping only the reciprocated network links while non-reciprocated or undirected network links exist (Comola and Fafchamps, 2017); when collecting data in certain contexts where participants are unwilling to cooperate, like criminals' connections or adolescents' sexual network (Kossinets, 2006); or when constructing a network measure based on a particular dimension of social connections, while ignoring other relevant interactions (Conley and Udry, 2010).

"No false negative" is also a reasonable assumption. It may be the case when observing a large network that includes ineffectual interactions, such as social media friends, email connections, and virtual communities; when simply assuming all units within a certain geographical boundary are linked; when the observed network is formed as a union of multi-dimensional networks (e.g. kinship, borrower-lender relationships, and advice-giving in Banerjee et al., 2013); when assuming a link exists if either side of the two nodes reports an interaction; or when constructing a network based on participation in multiple social events or affiliations ("multicontextual approach" in Kossinets, 2006).

If the network proxy $\mathcal{N}_i$ satisfies the one type of measurement error assumption, the matrix $F_{\mathcal{F}|\mathcal{F}^*,Z}$ in (4.4) should be upper triangular if there is no false positive, and lower triangular if there is no false negative. Based on Theorem 4.4.3, since $F_{\mathcal{F}|\mathcal{F}^*,Z}$ is identifiable, it is possible to test the one type of measurement error assumption via testing the null hypothesis that all elements in either the upper or the lower triangular of matrix $F_{\mathcal{F}|\mathcal{F}^*,Z}$ are zero. One possible testing approach is the subsampling or bootstrap method proposed by Romano and Shaikh (2012) with proper adjustments to accommodate the network data. Other possible testing approaches may be established following Leung (2020a) if the $\sqrt{N}$ convergence rate of estimator for $F_{\mathcal{F}|\mathcal{F}^*,Z}$ is satisfied. It might be the case if the outcome Y_i and covariate Z_i are discrete, then a smooth kernel estimation is not needed and the $\sqrt{N}$ convergence rate can be achieved based on the proof of Theorem 4.5.2 in Section 4.5. A formal test if left for future research.

Given the results in Theorem 4.4.3 and Theorem 4.4.5, the identification of the CASF, the treatment and spillover effects can be achieved.

Theorem 4.4.6 (Identification) *Suppose Assumption 4.3.4 is satisfied by $\tilde{\mathcal{N}}_i$ and $\mathcal{N}_i$. Let Assumptions 4.3.1-4.3.3, and 4.4.1-4.4.5 hold.*

(a) For $\forall(d,s,z,n) \in \{0,1\}\times\Omega_{S,Z,\mathcal{F}}$ such that $f_{\mathcal{F}_i|Z_i=z}(n) > 0$, $m_i(d,s,z,n) = \mathbb{E}[Y_i|D_i = d, S_i = s, \mathcal{F}_i = n, Z_i = z]$ is identical for all $i \in \mathcal{P}$.

(b) The CASF m^*, the treatment effect τ_d and the spillover effect τ_s are nonparametrically identified wherever they are well-defined.

4.4.3. Discussion and Extension

Anonymous Interactions

As implied by Lemma 4.4.1, the anonymous interactions $S_i^* \perp \mathcal{N}_i^* | Z_i, \mathcal{F}_i^*$ is critical for the identification of m^*. The key factor to ensure the anonymous interactions is that, for any given unit i, the treatment assignments to units other than i (i.e. $\{D_j\}_{j\in\mathcal{P}, j\neq i}$) conditional on $(Z_i, \mathcal{F}_i^*)$ are i.i.d. across j. It might be violated if some covariates not only enter the network formation process, but also influence the treatment assignment. It is because, if one would like to believe that the homophily exists in the network formation, i.e. individuals are more likely to establish a link if they are similar, then unit i's characteristics and the peers' identity will reveal relevant information about the characteristics of the peers and non-peers. In this case, conditioning on the covariate Z_i, the i.i.d. of $\{D_j\}_{j\in\mathcal{P}, j\neq i}$ would fail to hold.

Unconfounded Treatment

Given the discussion in Section 4.4.3, it is apparent that there exist two settings where a fully randomised treatment assumption can be relaxed to allow stratified randomisation based on individuals' characteristics. The first setting accounts for homophily and requires that there exists a subset of individual's characteristics $Z_{1,i} \subset Z_i$ such that $Z_{1,i}$ does not affect the network formation. Then, the treatments can be randomly assigned based on $Z_{1,i}$. For example, in the microfinance program, interventions can be allocated randomly given the participants' social status, e.g. occupation; it is unlikely that the network measured by "go to pray together" will be affected by the occupation, because people with whom a individual goes to pray would rely on their religion, gender and caste, rather than the social status. The second setting suits situations where it is reasonable to believe that the network is formed following the random graph model of Erdös and Rényi (1959); that is, each link is formed independently with the same probability. In this case, the treatments can be randomly assigned based on Z_i. It would be interesting to study the consequences of further relaxing this condition and adopting more general unconfounded treatment assignments.

Directed or Weighed Links

The analysis so far does not require the network $\mathcal{N}_i^*$ to be undirected, and the generalisation to the directed network is straightforward. If the unweighted restriction is also relaxed, then the spillover effects can be captured by $S_i^* = \sum_{j \in \mathcal{N}_i^*} \pi(Z_{1,j}) D_j$ where $Z_{1,j}$ is a subset of Z_i and $\pi(\cdot)$ is a known weighting function. For the same reason discussed in Section 4.4.3, it is required that $Z_{1,j}$ does not impact the network formation in the presence of homophily among units. For example, in the microfinance program, a unit with a higher degree of financial literacy might be assigned a higher weight. However, financial literacy is unlikely to directly affect the network connections of women from South India, because the network data is collected before the microfinance program is implemented.

4.5. Asymptotic Properties

Section 4.5 is organised as follows. Section 4.5.1 introduces the concept of dependency neighbourhood, which helps to define the distance of correlated samples and to establish the asymptotic properties of the estimation approach. Section 4.5.2 presents the nonparametric kernel estimation and Section 4.5.3 discusses the semiparametric estimation procedure.

4.5.1. Dependency Neighbourhoods

Let W_i be an observable random variable or vector. For sample size N, the dependency neighbourhood of unit i, denoted by $\Delta(i, N)$, satisfies $\Delta(i, N) \subset \{1, 2, ..., N\}$, $i \in \Delta(i, N)$ and conditions in Assumption 4.5.1. Any unit j such that $j \in \Delta(i, N)$ is referred to as unit i's dependent neighbour (hereafter DN), while the dependent neighbour is not necessarily a network neighbour. Following Chandrasekhar and Jackson (2016), we define the dependency neighbourhood by restricting the relative correlation of $\{W_i\}_{i=1}^N$ inside and outside $\{\Delta(i, N)\}_{i=1}^N$. For any integrable function b, denote the sum of covariance of all pairs of units in each others' dependency neighbourhoods as

$$\Sigma_N^b = \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \text{Cov}\left(b(W_i), b(W_j)\right), \tag{4.9}$$

which captures the variation of $b(W_i)$ for all the N units and the dependence across all pairs $(b(W_i), b(W_j))$, where j is the dependent neighbour of i. The assumption below characterises the two principal properties of the dependency neighbourhood.

Assumption 4.5.1 (Dependency Neighbourhood) *For any integrable function* $b : \Omega_W \mapsto \mathbb{R}^{d_b}$,

(a) $\Sigma_N^b \to \infty$ *as* $N \to \infty$;

(b) $\sum_{i=1}^N \sum_{j \notin \Delta(i,N)} Cov(b(W_i), b(W_j)) = o(\Sigma_N^b)$.

Condition (a) ensures that the dependence of units in each others' dependent neighbourhoods contains sufficient information that is necessary for deriving asymptotic properties using these dependent variables. Intuitively, condition (b) requires that $\Delta(i, N)$ is a collection of units with a relatively high correlation with unit i, compared to those in its complement. The set $\Delta(i, N)$ may not be unique, because it is defined asymptotically. In addition, the size of $\Delta(i, N)$ may change (generally expand) as the sample size increases.

As mentioned in Chandrasekhar and Jackson (2016), there is substantial freedom in constructing these sets in different studies. For example, the dependency neighbourhoods can be defined based on individuals' participation in common actions, affiliations, and social events, regardless of their network interactions; individuals' identities that lead to strong social norms and clear barriers across groups, such as caste, tribe or race (Currarini et al., 2009, 2010); or social or geographical locations, such as occupation, class, school, village or community. Essentially, the dependency neighbourhoods $\{\Delta(i, N)\}_{i=1}^N$ can be understood as defined by individuals' exogenous attributes and the analysis in this chapter is conducted conditional on these attributes: that is, the dependent neighbourhoods are treated as non-stochastic.

4.5.2. First Step Kernel Estimation

The nonparametric kernel estimation of density function has been extensively studied; see Newey and McFadden (1994), Newey (1994) and Li and Racine (2007) among others. To ease illustration, denote the observable variable by $W_i = (W_i^{c'}, W_i^{d'})'$ where W_i^c

represents the vector containing continuous variables and W_i^d is the vector containing discrete variables. Recall that the supports of W_i^c and W_i^d are Ω_{W^c} and Ω_{W^d}, respectively. Note that W_i may be used to denote different observable variables at different places. For a bandwidth $h > 0$ and $\forall w = (w^{c'}, w^{d'})' \in \Omega_{W^c, W^d}$, denote

$$K(W_i^c, w^c) = \frac{1}{h^Q} \prod_{q=1}^{Q} \kappa \left(\frac{W_{i,q}^c - w_q^c}{h} \right),$$

where $\kappa(\cdot)$ is the univariate kernel function and Q is the dimension of vector W_i^c. Denote the nonparametric kernel estimator of f_{W_i} as

$$\hat{f}_{W_i}(w) = \frac{1}{N} \sum_{i=1}^{N} K(W_i^c, w^c) 1\left[W_i^d = w^d\right]. \tag{4.10}$$

For expositional simplicity, we restrict the bandwidth for all continuous variables to be the same. In practice, our method also allows for different bandwidths, while a data-driven method for bandwidth selection is not the focus of this chapter. Given (4.10), the estimators for the nuisance parameter γ^0 is:

$$\hat{\gamma}_N = \left[\hat{f}_{\tilde{\mathcal{F}}_i, \mathcal{F}_i, Y_i, Z_i},\ \hat{f}_{\tilde{\mathcal{F}}_i, \mathcal{F}_i, Z_i},\ \hat{f}_{S_i, \mathcal{F}_i, Z_i},\ \hat{f}_{\mathcal{F}_i, Z_i},\ \hat{f}_{Z_i} \right]'.$$

Assumption below is employed for deriving the uniform convergence of the nonparametric kernel estimator.

Assumption 4.5.2 *Let $W_i^c = (Y_i, Z_i^{c'})'$ and $W_i^d = (D_i, Z_i^{d'}, S_i, \mathcal{F}_i, \tilde{S}_i, \tilde{\mathcal{F}}_i)'$.*

(a) $\Omega_{W^c} \subset \mathbb{R}^Q$ is a compact and convex set and the cardinality of Ω_{W^d} is finite.

(b) Each element in γ^0 is bounded and continuously differentiable in w^c to order two with bounded derivatives on an open set containing Ω_{W^c}.

(c) $\kappa(\cdot)$ is nonnegative kernel function and is differentiable with uniformly bounded first derivative. In addition, for some constant $K_1, K_2 > 0$

$$\int \kappa(v)dv = 1, \quad \kappa(v) = \kappa(-v), \quad \int v^2 \kappa(v)dv = K_1, \quad \int \kappa(v)^2 dv = K_2.$$

(d) $h \to 0, \quad Nh^Q \to \infty, \quad \ln(N)/(Nh^Q) \to 0, \ \textit{as}\ N \to \infty.$

(e) *Let* $\bar{r}_N = \sup_{1\le i\le N} |\Delta(i,N)|$. *The cardinalities of dependency neighbourhoods satisfy*

$$\bar{r}_N \left[\ln(N)/(Nh^Q)\right]^{1/2} = O(1), \quad \frac{1}{N}\sum_{i=1}^{N} |\Delta(i,N)|^2 = O(1).$$

Conditions (a) and (b) state the regularity conditions of the support and data distribution. Conditions (c) and (d) describe features of the kernel function and the bandwidth, which are standard for nonparametric kernel estimation. In addition, to accommodate the dependence across units, we need to impose restrictions on the size of dependency neighbourhood. Condition (e) allows the situation where a sufficiently large number of units possess an increasing number of DNs, say $O([\ln(N)N/h^Q]^{1/2})$ units with $O([Nh^Q/\ln(N)]^{1/4})$ DNs, and the rest with a bounded number of DNs. Although we require a sparse network, the number of DNs may increase with the sample size.

To address issues arising from the dependence between observations and to derive the uniform convergence rate of the first-step kernel estimation, we adopt the method of Masry (1996), which is based on the approximation theorems developed by Bradley et al. (1983) to approximate dependent random variables by independent ones. Let us first introduce a partition of samples, based on which a notion of "distance" can be developed. Intuitively, the dependence strength among the units' observable variables can be used to describe their relative distance: units are far away from each other when they are less correlated. Therefore, the dependency neighbourhood would be a useful tool to construct the distance measure. For any given sample size N, partition the index set $\{1,2,...,N\}$ into q_N mutually exclusive subsets $\mathbb{S}_1,...,\mathbb{S}_{q_N}$ with $\bigcup_{1\le l\le q_N}\mathbb{S}_l = \{1,2,...,N\}$. The subscript N of q_N means that q_N may go to infinity as $N\to\infty$. Without loss of generality, suppose that i_0 is an arbitrary unit from the N observed samples and $\mathbb{I}_0 := \{i_0\}$. For $k = 1,2,...,q_N$, define

$$\mathbb{I}_k = \underbrace{\bigcup_{i\in\mathbb{I}_{k-1}} \Delta(i,N)}_{\text{DNs of } \mathbb{I}_{k-1}} \bigcup \underbrace{\Big\{ i\in\{1,2,...,N\} :\ i \notin \bigcup_{j\in\mathbb{I}_{k-1}} \Delta(j,N) \text{ and } \bigcup_{j\in\mathbb{I}_{k-1}} \Delta(j,N) \bigcap \Delta(i,N) \neq \emptyset \Big\}}_{\text{DNs of DNs of } \mathbb{I}_{k-1}}, \tag{4.11}$$

where $\emptyset$ denotes an empty set. Apparently, $\mathbb{I}_k$ includes all the DNs of the units in $\mathbb{I}_{k-1}$, including $\mathbb{I}_{k-1}$ itself and all the units who are not DNs of the units in $\mathbb{I}_{k-1}$ but share common DNs with them (hereafter DNs of DNs). In other words, $\mathbb{I}_k$ consists of units that are closely correlated to those in $\mathbb{I}_{k-1}$. Because $\mathbb{I}_{k-1} \subseteq \bigcup_{i\in\mathbb{I}_{k-1}} \Delta(i,N)$ by definition of the dependency neighbourhood, $\mathbb{I}_1 \subseteq \cdots \mathbb{I}_{k-1} \subseteq \mathbb{I}_k \cdots \subseteq \mathbb{I}_{q_N}$ is an increasing sequence. Given $\mathbb{I}_1, ..., \mathbb{I}_{q_N}$, let $\mathbb{S}_1 = \mathbb{I}_1$ and $\mathbb{S}_k = \mathbb{I}_k/\mathbb{I}_{k-1}$ for $k = 2, ..., q_N$.

Importantly, $\mathbb{I}_k/\mathbb{I}_{k-1} = \emptyset$ may occur in two scenarios. First, when each of the N units are included in at least one of the sets $\mathbb{I}_1, ..., \mathbb{I}_{k-1}$, so that there are no DNs of $\mathbb{I}_{k-1}$ or DNs of DNs of $\mathbb{I}_{k-1}$, that are not included in $\mathbb{I}_{k-1}$. In this scenario, we have obtained the desirable partition with $q_N = k-1$. Second, it may also occur when the units in $\mathbb{I}_{k-1}$ form an isolated cluster that is disjointed from other units. That is, when none of the units in $\mathbb{I}_{k-1}$ have DNs outside $\mathbb{I}_{k-1}$ or share common DNs with units outside $\mathbb{I}_{k-1}$. If so, we can pick an arbitrary unit $i_k \in \{1, 2, ..., N\}$ such that $i_k \notin \mathbb{I}_{k-1}$. Then, define $\mathbb{I}_k$ as (4.11) via replacing $\mathbb{I}_{k-1}$ with $\{i_k\}$, and repeat the above process until all the observed units are exhausted. For any given sample size N, the partition exists and every unit is included in exactly one set of $\mathbb{S}_1, ..., \mathbb{S}_{q_N}$. The largest possible value of q_N can be N, for example, when all observations are i.i.d. and each $\Delta(i,N)$ is set to be a singleton $\{i\}$.[7]

The above partition helps to order the units so that their dependence strength becomes weaker (or equivalently, their distance becomes larger), when they belong to far apart sets in $\mathbb{S}_1, ..., \mathbb{S}_{q_N}$. Given such an ordering, we can then introduce the dependence coefficient that is an analogue of the strong mixing coefficient of a stochastic process:

$$\alpha_k = \sup_{A\in\Gamma_1^{k-2}, B\in\Gamma_k^k} \left|\Pr(A,B) - \Pr(A)\Pr(B)\right|,$$

where $\Gamma_1^{k-2} = \sigma\left(\{W_i,\ i \in \bigcup_{1\le l\le k-2}\mathbb{S}_l\}\right)$ and $\Gamma_k^k = \sigma\left(\{W_i,\ i\in\mathbb{S}_k\}\right)$ for $k = 1, 2, ..., q_N$ are the σ-fields on sets of random variables of units in $\bigcup_{1\le l\le k-2}\mathbb{S}_l$ and $\mathbb{S}_k$, respectively.

[7]The partition is constructed in a similar vein with the "dependency graph" in Leung (2020b), which is introduced to capture data correlation and is built upon the true network connections. The key idea of the dependency graph is that potential outcomes of two units are independent if they are neither network neighbours nor share common network neighbours. In this chapter, we use the dependency neighbourhoods to define distance, because the true network is not available and the network measurement errors induce extra data correlation.

Assumption 4.5.3 (Local Dependence) *Let* $L_N = [N/(\ln(N)h^{Q+2})]^{Q/2}$. *The dependence coefficient* α_k *satisfies*

$$\Psi_N := L_N \left(\frac{N}{\ln(N)}\right)^{1/2} \sum_{k=1}^{q_N} \alpha_k, \quad \sum_{N=1}^{\infty} \Psi_N < \infty.$$

Assumption 4.5.3 requires that the dependence coefficient α_k converges to zero, meaning that the observable variables of the units in $\bigcup_{1\leq l\leq k-2} \mathbb{S}_l$ and in $\mathbb{S}_k$ tend toward being independent as the sample size increases. It is the case, for example, when the units are only locally dependent upon their DNs and those with whom they share common DNs, while their dependence with others who are not their DNs nor DNs of DNs is negligible and goes to zero. It is also the case when there are many independent clusters and only finite dependent ones. This assumption controls the asymptotic dependence among observables and is akin to the mixing coefficient decaying condition but in a setting with network-induced data dependence. It ensures that the uniform convergence of the first-step kernel estimation holds, even when a relatively large scale of local dependency among units exists. A similar assumption is exploited in Masry (1996) to restrict the time series data, and in Sävje (2019) to control the dependence of network measurement errors.

Lemma 4.5.1 provides two sufficient conditions under which Assumption 4.5.3 holds

Lemma 4.5.1 *Assumption 4.5.3 is satisfied, if either of the following conditions hold.*

(a) $\{W_i,\ \forall\ i \in \mathbb{S}_k\} \perp \{W_j,\ \forall\ i \in \mathbb{S}_{k'}\}$ *for any* $k \neq k'$ *and* $k, k' = 1, 2, ..., q_N$;

(b) $\{W_i,\ \forall\ i \in \mathbb{S}_k\} \perp \{W_j,\ \forall\ i \in \mathbb{S}_{k+2}\}$ *and* $k = 1, 2, ..., q_N - 2$.

The proof of Lemma 4.5.1 is trivial therefore omitted. Condition (a) indicates that Assumption 4.5.3 holds if the population consists of many disjoint and independent clusters, and each $\mathbb{S}_k$ represents one of those clusters. In this case, we rule out the possibility of the network interference or more general forms of data dependence across clusters. Condition (b) requires that the units are independent, if they are not DNs and do not share common DNs. Condition (b) is weaker than (a), because it allows the

possibility of data dependence across clusters, although the dependence is limited to the "nearby" clusters connected by the DNs or the DNs of DNs. It also allows a single large network where no clear boundaries can be drawn to divide the population into clusters.

Given the local dependence assumption, the uniform convergence result of the kernel estimation can be established.

Theorem 4.5.2 *Let Assumptions 4.5.2 and 4.5.3 hold, then*

$$\left\|\hat{\gamma}_N - \gamma^0\right\|_\infty = O_p\left(\left[\ln(N)/(Nh^Q)\right]^{1/2} + h^2\right).$$

The uniform convergence rate of the kernel estimation in Theorem 4.5.2 is consistent with that of the conventional kernel estimation under i.i.d. or strong mixing settings (e.g., Newey, 1994; Li and Racine, 2007; Masry, 1996).

Let $\hat{\phi}_N := \phi(\hat{\gamma}_N)$ represent the estimator of the latent distribution function $f_{S^*,\mathcal{F}_i^*|D_i,S_i,Z_i,\mathcal{F}_i}$. According to Proposition 4.4.2, we can obtain a plug-in estimator $\hat{\phi}_N$ via replacing the distributions on the right hand side of (4.3) by their kernel estimators based on $\hat{\gamma}_N$ in (4.10). Denote $\phi^0 = \phi(\gamma^0)$ as the true latent distribution function. Given the uniform convergence of $\hat{\gamma}_N$ in Theorem 4.5.2, we only need to consider the convergence of $\hat{\phi}_N$ in a small neighbourhood of γ^0.

Corollary 1 *Let Assumption 4.3.1-4.3.4 and 4.4.1-4.4.5 hold. Under assumptions in Theorem 4.5.2, suppose that there exists a constant $\epsilon > 0$ such that $f_{\mathcal{F}_i|Z_i} > \epsilon$. Then, for $\eta \to 0$ as $N \to \infty$,*

$$\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty \leq \eta} \left\|\hat{\phi}_N - \phi^0\right\|_\infty = O_p(\left\|\hat{\gamma}_N - \gamma^0\right\|_\infty).$$

4.5.3. Semiparametric Estimation

In this subsection, we study the estimation of CASF m^* by simplifying $m^* = m^*(\cdot;\theta)$ as a known function up to the unknown parameter $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$. Consequently, $m_i(\cdot) = m(\cdot) = m(\cdot;\theta,\phi)$ is also known up to (θ,ϕ). Based on Theorem 4.4.5, we know that

$m_i(\cdot)$ is identical across all i. Thus, we can suppress the subscript i, i.e. $m_i(\cdot) = m(\cdot)$. In addition, $m(\cdot) = m(\cdot;\theta,\phi)$ is because of $m(\cdot)$ being a function of the CASF $m^*(\cdot;\theta)$ and nuisance parameter ϕ. Note that the identification of m^* in Section 4.4 does not rely on such an simplification. More importantly, imposing such a parametric structure on m^* still allows flexible heterogeneity of the treatment and spillover effects, which can be captured by interactions of D_i and S_i, with covariate Z_i and network degree $\mathcal{F}_i$, as well as their polynomials.

Consistency

For notational simplicity, let $X_i^* := (D_i, S_i^*, Z_i, \mathcal{F}_i^*)'$ and $X_i := (D_i, S_i, Z_i, \mathcal{F}_i)'$ with support Ω_{X^*} and Ω_X, respectively. In addition, denote $T_i^* = (S_i^*, \mathcal{F}_i^*)'$. Let $x_j^* := (d, s_j^*, z, n_j^*)$ with $t_j^* = (s_j^*, n_j^*) \in \Omega_{S^*,\mathcal{F}^*}$, and $j \in \{1, 2, ..., K_T\}$ represents the lexicographical ordering of the possible values of T_i^* as described in Appendix (4.9.79). Similarly, let $x_j := (d, s_j, z, n_j)$ with $t_j := (s_j, n_j) \in \Omega_{S,\mathcal{F}}$. By definition of $m(\cdot;\theta,\phi)$, the following moment condition holds:

$$\mathbb{E}\left[Y_i - m(X_i;\theta,\phi)\middle|X_i\right] = 0.$$

From Proposition 4.3.3, $m(\cdot;\theta,\phi)$ and the CASF $m^*(\cdot;\theta)$ are linked through the formula $m(x;\theta,\phi) = \sum_{j=1}^{K_T} m^*(x_j^*;\theta) f_{T_i^*|X_i=x}(t_j^*)$. Recall that X_i is identically distributed for all i under the assumptions in Section 4.3. Denote the objective function and its sample analogue as

$$\mathcal{L}(\theta,\phi) = \mathbb{E}\left\{\tau_i \left[Y_i - m\left(X_i;\theta,\phi\right)\right]^2\right\}, \quad \text{and} \quad \mathcal{L}_N(\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N} \tau_i \left[Y_i - m\left(X_i;\theta,\phi\right)\right]^2,$$

where $\tau_i := \tau(X_i)$ is the non-negative weight. Following Newey (1994), we use the weight function τ to focus the optimisation problem on regions where the kernel estimation is relatively reliable. Hu (2008) also adopts the weight function and sets it as a fixed trimming $\tau(x) = 1[x \in \mathbf{X}]$ with $\mathbf{X} \subset \Omega_X$ a fixed set. Other types of weight functions such as data-driven weight functions or methods for selection of weight functions are out of the scope of this chapter. Then, θ is estimated by minimising $\mathcal{L}_N(\theta, \hat{\phi}_N)$ given the

estimator $\hat{\phi}_N$ from Theorem 4.5.2:

$$\hat{\theta}_N = \arg\min_{\theta\in\Theta} \mathcal{L}_N(\theta, \hat{\phi}_N). \tag{4.12}$$

Let $W_i = (Y_i, X_i')'$ be the vector containing all the observed variables and $w = (y, x')' \in \Omega_W$.

Assumption 4.5.4

(a) $\Theta \subset \mathbb{R}^{d_\theta}$ *is compact,* $\theta^0 \in int(\Theta)$ *and* θ^0 *is identifiable from the weighted conditional moment function* $\mathcal{L}(\theta, \phi^0) = 0$.

(b) $\tau(\cdot)$ *is nonnegative and* $\sup_{x\in\Omega_X} |\tau(x)| < C$ *for some constant* $C > 0$.

(c) $m^*(x;\theta)$ *is continuous in* θ *for all* $x \in \Omega_X$*, and is an integrable function of* X_i *for all* $\theta \in \Theta$.

(d) Denote the random variable $x^*_{i,j} = (D_i, s^*_j, Z_i, n^*_j)$ *with* $t^*_j = (s^*_j, n^*_j) \in \Omega_{T^*}$ *and* $j = 1, 2, ..., K_T$*. There exists a function* $h_1(x)$ *such that* $|m^*(x;\theta)|^2 \leq h_1(x)$ *for all* $\theta \in \Theta$*, and* $\mathbb{E}[h_1(x^*_{i,j})] < \infty$ *for all* $j = 1, 2, ..., K_T$.

(e) Let $e(w, \theta) := \tau(x)[y - m(x;\theta,\phi^0)]^2$ *and* $e_i(\theta) := e(W_i, \theta)$*. For any given constant* $\eta > 0$*, denote* $U_i(\theta, \eta) = \sup_{\theta'\in\Theta,\ \|\theta'-\theta\|<\eta} |e_i(\theta') - e_i(\theta)|$*. There exists a function* $h_2(w)$ *such that* $|e(w, \theta)| \leq h_2(w)$ *for all* $\theta \in \Theta$ *and* $\mathbb{E}[h_2(W_i)] < \infty$*. In addition,* $\sup_{\theta\in\Theta} \mathbb{E}[|e_i(\theta)|^{2+\delta}] < C$ *for some constants* $\delta > 0$ *and* $C > 0$.

Theorem 4.5.3 (Consistency) *Let assumptions in Theorem 4.4.6 hold. Under Assumptions 4.5.1- 4.5.4, we have* $\|\hat{\theta}_N - \theta^0\| = o_p(1)$.

Asymptotic Normality

To show asymptotic normality of the estimator $\hat{\theta}_N$, we need to account for the presence of the nuisance parameter ϕ and the data dependence arising from the mismeasured network, which requires a significant generalisation of the classical CLT. In particular, the often used CLT developed for mixing processes does not work for our purpose, as it

relies on some ordering structure to measure the "distance" between units. Therefore, we adopt and extend the univariate CLT for the network data proposed by Chandrasekhar and Jackson (2016) to a multivariate setting, which will be applied in this section to derive the asymptotic normality for $\hat{\theta}_N$. See Lemma 4.9.7 in the Appendix.

Let $g(W_i;\theta,\phi) = \tau_i[Y_i - m(X_i;\theta,\phi)]\frac{\partial m(X_i;\theta,\phi)}{\partial\theta}$. From the first order condition of the optimisation problem (4.12), $\hat{\theta}_N$ solves $\frac{1}{N}\sum_{i=1}^N g(W_i;\hat{\theta}_N,\hat{\phi}_N) = 0$. Then, by the mean value theorem we can obtain

$$0 = \frac{1}{N}\sum_{i=1}^N g(W_i;\hat{\theta}_N,\hat{\phi}_N) = \frac{1}{N}\sum_{i=1}^N g(W_i;\theta^0,\hat{\phi}_N) + \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial\theta'}(\hat{\theta}_N - \theta^0), \tag{4.13}$$

where $\tilde{\theta}_N$ is between $\hat{\theta}_N$ and θ^0. If $\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial\theta'}$ is invertible, rearranging (4.13) leads to

$$\sqrt{N}(\hat{\theta}_N - \theta^0) = \left[\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial\theta'}\right]^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^N g(W_i;\theta^0,\hat{\phi}_N).$$

Let us introduce some useful notations. Recall that $\phi(\cdot) = \phi(\cdot;\gamma)$. We set $\mathbf{t} := (t_1,...,t_{K_T})'$ and $\phi(\mathbf{t};\gamma) = [f_{T_i^*|X_i}(t_1),...,f_{T_i^*|X_i}(t_{K_T})]'$. Let $\mathbf{1}_{d_\gamma}$ be a $d_\gamma \times 1$ vector of ones. Denote $\nu(w;\theta,\gamma) = \mathbb{E}\left[\tau(X_i)\frac{\partial}{\partial\theta}\mathcal{R}(W_i;\theta,\phi)\frac{\partial\phi(\mathbf{t};\gamma)}{\partial\gamma'}\Big|_{\gamma=\gamma(w)}\mathbf{1}_{d_\gamma}\Big|w\right]$ and $\delta(W_i;\theta,\gamma) := \nu(W_i;\theta,\gamma) - \mathbb{E}[\nu(W_i;\theta,\gamma)]$, where

$$\mathcal{R}(W_i;\theta,\phi) = \begin{bmatrix} [Y_i - m(X_i;\theta,\phi)]\, m^*(x^*_{i,1};\theta) \\ \vdots \\ [Y_i - m(X_i;\theta,\phi)]\, m^*(x^*_{i,K_T};\theta) \end{bmatrix}'.$$

Assumption 4.5.5

(a) $m^*(x;\theta)$ is continuously differentiable in θ up to order three with bounded third

order derivative uniformly in x, i.e. for any $r, q = 1, 2, ..., d_\theta$,

$$\sup_{x \in \Omega_X} \left| \frac{\partial}{\partial \theta} \left(\frac{\partial^2 m^*(x;\theta)}{\partial \theta_r \partial \theta_q} \right) \right| < C, \text{ for all } \theta \in \Theta.$$

(b) *There exist functions $H_1(x)$ and $H_2(x)$ such that $\left\| \frac{d^2 m^*(x;\theta)}{d\theta d\theta'} \right\|^2 \leq H_1(x)$, $\left\| \frac{dm^*(x;\theta)}{d\theta} \right\|^2 \leq H_2(x)$ for all $\theta \in \Theta$ and $\mathbb{E}[H_1(x_{i,j})] < \infty$, $\mathbb{E}[H_2(x_{i,j})] < \infty$ for all $j = 1, 2, ..., K_T$.*

(c) $\mathbb{E}\left[\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'} \right]$ *exists and is nonsingular. In addition,* $\mathbb{E}\left[\left\| \frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'} \right\|^2 \right] < \infty$.

Assumption 4.5.5 (a) and (b) introduce regularity conditions on the smoothness of the CASF $m^*(\cdot, \theta)$. Condition (c) ensures that the limit of the Hessian matrix exists and is invertible. To simplify notation, denote $\nu(W_i) := \nu(W_i; \theta^0, \gamma^0)$ and $\delta(W_i) := \delta(W_i; \theta^0, \gamma^0)$.

Assumption 4.5.6

(a) $N^{1/2}[\ln(N)/(Nh^Q)] \to 0$ *and* $Nh^4 \to 0$ *as* $N \to \infty$.

(b) $\nu(w; \theta, \gamma) = \nu(w^c, w^d; \theta, \gamma)$ *is continuously differentiable in w^c almost everywhere and satisfies $\sum_{w^d \in \Omega_{W^d}} \int \|\nu(w)\| dw^c < \infty$. In addition, $\| Var[\nu(W_i)] \| < \infty$.*

Assumption 4.5.6 implies that the convergence rate of $\hat{\gamma}_N$ is faster than $N^{1/4}$. It is a typical restriction on the bandwidth to guarantee the asymptotic normality for semi-parametric two-step estimators that depend on kernel density, for example Newey and McFadden (1994).

We first show that the $d_\theta \times d_\theta$ Hessian matrix $\frac{1}{N} \sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'}$ converges in probability uniformly.

Lemma 4.5.4 *Let the assumptions in Theorem 4.5.3 hold.*

(a) *Under Assumption 4.5.5, for a small enough $\eta \to 0$ as $N \to \infty$, we have*

$$\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \frac{1}{N} \sum_{i=1}^N \frac{\partial g(W_i; \tilde{\theta}_N, \hat{\phi}_N)}{\partial \theta'} - \mathbb{E}\left[\frac{\partial g(W_i; \theta^0, \phi^0)}{\partial \theta'} \right] \right\| = o_p(1).$$

(b) Under Assumption 4.5.6, we can get

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g(W_i;\theta^0,\hat{\phi}_N) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\big[g(W_i;\theta^0,\phi^0)+\delta(W_i)\big] + o_p(1).$$

Denote the dependence neighbourhoods covariance matrix

$$\Sigma_N^{\tilde{g}} = \sum_{i=1}^{N}\sum_{j\in\Delta(i,N)} \mathbb{E}\Big\{ \big[g(W_i;\theta^0,\phi^0)+\delta(W_i)\big]\big[g(W_j;\theta^0,\phi^0)+\delta(W_j)\big]' \Big\}.$$

Denote the $d_\theta \times 1$ vector $\tilde{g}_i = g(W_i;\theta^0,\phi^0)+\delta(W_i)$ with $\tilde{g}_i = (\tilde{g}_{i,1},...,\tilde{g}_{i,d_\theta})'$. Then, $\Sigma_N^{\tilde{g}} = \sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\mathbb{E}[\tilde{g}_i\tilde{g}_j']$. In addition, by notation abuse, let $\mathbf{S}_i^c = \sum_{j\notin\Delta(i,N)}\tilde{g}_j$. For any vector a, let $a \geq 0$ mean that each of its entries are nonnegative. For any matrix $A = \{a_{ij}\}$, $vec(A)$ denotes the vectorisation of A and $|A| = \{|a_{ij}|\}$.

Assumption 4.5.7

(a) For all $i \in \mathcal{P}$, $\Delta(i,N)$ is symmetric such that $j \in \Delta(i,N)$ if and only if $i \in \Delta(j,N)$.

(b) There exists a finite, strictly positive-definite and symmetric matrix $\Omega \in \mathbb{R}^{d_\theta}\times\mathbb{R}^{d_\theta}$ such that $\|\frac{1}{N}\Sigma_N^{\tilde{g}} - \Omega\| \to 0$ as $N\to\infty$.

(c) The following conditions hold for $\{\tilde{g}_i\}_{i=1}^N$.

(c1) $\left\|\sum_{i=1}^{N}\sum_{j,k\in\Delta(i,N)}\mathbb{E}\left[\left|vec(\tilde{g}_i\tilde{g}_j')\tilde{g}_k'\right|\right]\right\|_\infty = o\left(\left\|[\Sigma_N^{\tilde{g}}]^{3/2}\right\|_\infty\right)$;

(c2) $\left\|\sum_{i,k=1}^{N}\sum_{j\in\Delta(i,N)}\sum_{l\in\Delta(k,N)}\mathbb{E}\Big[\left(\tilde{g}_i\tilde{g}_j' - \mathbb{E}[\tilde{g}_i\tilde{g}_j']\right)'\left(\tilde{g}_k\tilde{g}_l' - \mathbb{E}[\tilde{g}_k\tilde{g}_l']\right)\Big]\right\|_\infty = o\left(\left\|[\Sigma_N^{\tilde{g}}]^2\right\|_\infty\right)$;

(c3) $\left\|\sum_{i=1}^{N}\sum_{j\notin\Delta(i,N)} Cov(\tilde{g}_i,\tilde{g}_j)\right\|_\infty = o\left(\left\|\Sigma_N^{\tilde{g}}\right\|_\infty\right)$;

(c4) $\mathbb{E}\left[\tilde{g}_i\mathbf{S}_i^c\middle|\mathbf{S}_i^c\right] \geq 0$ *for all* $i\in\mathcal{P}$.

Assumption 4.5.7 (a) guarantees that the covariance matrix $\Sigma_N^{\tilde{g}}$ is symmetric. Condition (b) ensures that the samples possess sufficiently large variation so that the CLT holds.

Meanwhile, it requires the limit of $\Sigma_N^{\tilde{g}}/N$ being a constant matrix Ω, instead of varying with the sample size, which imposes restriction on the allowable divergence rate of $\Sigma_N^{\tilde{g}}$ to some degree. Similar assumptions are used to study the asymptotic properties of covariance matrix estimator by White and Domowitz (1984).

Moreover, Assumption 4.5.7 (c) is crucial for the multivariate CLT under the dependency neighbourhood structure. Similar assumption is used in Chandrasekhar and Jackson (2016) to establish the asymptotic normality for univariate random variable with support $[0, 1]$. We extend the assumption to accommodate general multivariate random vectors without imposing any restrictions on their support. In particular, conditions (c1) and (c2) restrict the rate of dependency between the dependency neighbourhoods, while (c3) limits the rate of dependency outside the dependence neighbourhoods. Besides, condition (c4) states that on average, units outside each others' dependency neighbourhoods do not tend to interact negatively.[8]

Theorem 4.5.5 (Asymptotic Normality) *Suppose assumptions in Theorem 4.5.3, Assumptions 4.5.5-4.5.7 hold. Then*

$$\sqrt{N}(\hat{\theta}_N - \theta^0) \xrightarrow{d} \mathbb{N}(0, H^{-1}\Omega H^{-1}),$$

where $H = \mathbb{E}\left[\partial g(W_i; \theta^0, \phi^0)/\partial \theta'\right]$ *and* $\mathbb{N}$ *represents the normal distribution.*

Given that the function form of $\delta(w)$ is known, following Newey and McFadden (1994), we construct the estimator of $\delta(W_i)$ by substituting $(\hat{\theta}_N, \hat{\gamma}_N)$ for (θ^0, ϕ^0), i.e. $\hat{\delta}(W_i) := \delta(W_i; \hat{\theta}_N, \hat{\gamma}_N)$. Notably, the consistency and asymptotic normality of $\hat{\theta}_N$ only require the existence of dependency neighbourhoods. If how the dependency neighbourhoods $\{\Delta(i, N)\}_{i=1}^N$ are defined in a given study is known, it suffices a consistent variance estimator. The corollary below provides a consistent estimator of the variance-covariance matrix $H^{-1}\Omega H^{-1}$, which is essential when constructing asymptotic confidence intervals and conducting hypothesis tests.

[8] Chandrasekhar and Jackson (2016) also use condition that is similar to Assumption 4.5.7 (c4) to ease their proof. We note that the condition (c4) is not necessary for the asymptotic normality in this chapter and can be replaced by more primitive assumptions.

Corollary 2 (Variance Estimator) *Under assumptions in Theorem 4.5.5, we can get*

$$\left\|\hat{H}_N^{-1}\hat{\Omega}_N\hat{H}_N^{-1} - H^{-1}\Omega H^{-1}\right\| \xrightarrow{p} 0 \text{ as } N \to \infty,$$

where

$$\begin{aligned}\hat{H}_N =& \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i;\hat{\theta}_N,\hat{\phi}_N)}{\partial \theta'},\\ \hat{\Omega}_N =& \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\left[g(W_i;\hat{\theta}_N,\hat{\phi}_N)+\hat{\delta}(W_i)\right]\left[g(W_j;\hat{\theta}_N,\hat{\phi}_N)+\hat{\delta}(W_j)\right]'.\end{aligned}$$

Note that the consistency of the variance estimator $\hat{H}_N^{-1}\hat{\Omega}_N\hat{H}_N^{-1}$ is robust to a mild degree of misspecification of the dependency neighbourhoods. For example, if there are only finite units whose dependency neighbourhoods are misspecified, the variance estimator is still consistent due to the consistency of $(\hat{\theta}_N, \hat{\phi}_N, \hat{\gamma}_N)$ and Assumption 4.5.7 (c3). Moreover, if the knowledge of the dependency neighbourhoods is not available, one may resort to the resampling method proposed by Leung (2020a) to conduct inference for the parameter of interest. Rigourous study is left for future research.

4.6. Simulation

In this section, we illustrate the finite-sample behavior of the proposed estimation procedure via Monte Carlo. The data generating process (DGP) and network formation design for this Monte Carlo is similar to Leung (2020b). However, in contrast to Leung (2020b), the observed network proxies are contaminated by measurement errors. Consider the following DGP for the outcome Y_i:

$$Y_i = \theta_0 + \theta_1 D_i + \theta_2 D_i \mathcal{F}_i^* Z_i + \theta_3 S_i^* + \theta_4 S_i^{*2} + \theta_5 S_i^* Z_i + \theta_6 S_i^* \mathcal{F}_i^* + \varepsilon_i, \tag{4.14}$$

where $D_i \overset{i.i.d.}{\sim} Bernoulli(0.3)$ and $Z_i \overset{i.i.d.}{\sim} Bernoulli(0.5)$ are generated independently. In addition, the error term $\varepsilon_i = \varepsilon_i^{idio} + \varepsilon_i^{peer}$ where ε_i^{idio} denotes the idiosyncratic disturbance, and $\varepsilon_i^{peer} = \sum_{j\in\mathcal{P}} A_{ij}^* v_j$ captures the unobservable peer effects where $v_j \overset{i.i.d.}{\sim} \mathbb{N}(0, 0.5)$ is a random error. Set $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)' = (0, 1, 1/3, 1, -1, -1/2, 1)'$. We aim

to estimate the treatment effects $\tau_d(0,0,3)$ and $\tau_d(0,1,3)$ (with true value 1 and 2, respectively) and the spillover effects $\tau_s(1,0,3)$ and $\tau_s(1,1,3)$ (with true values 3 and 2.5, respectively), as in Definition 4.3.2.

We allocate units on a $[0,1] \times [0,1]$ space according to their exogenous geographic locations $\rho_i = (\rho_{i1}, \rho_{i2}) \overset{i.i.d.}{\sim} Uniform[0,1] \times Uniform[0,1]$. The actual network links are generated as follows:

$$A^*_{ij} = 1[\beta_1 + \beta_2(Z_i + Z_j) + \beta_3 d(\rho_i, \rho_j) + \zeta_{ij} > 0] \times 1[i \neq j],$$

where $\zeta_{ij} = \zeta_{ji}$ is a random shock that is i.i.d. $\mathbb{N}(0,1)$ (across dyads) and is independent of (Z_i, ρ_i) for all i and j. In addition, $d(\rho_i, \rho_j)$ indicates the distance between two units

$$d(\rho_i, \rho_j) = \begin{cases} 0, & \text{if } r^{-1}\|\rho_i - \rho_j\|_1 \leq 1 \\ \infty, & \text{otherwise} \end{cases},$$

where the scaling constant $r = (r_{deg}/N)^{1/2}$ guarantees the network sparsity and the parameter r_{deg} controls the average degree: $\mathbb{E}[\mathcal{F}^*_i]$ is an increasing function of r_{deg}. Consider two levels of sparsity $r_{deg} = 5$ and 8. Set $\beta = (\beta_1, \beta_2, \beta_3)' = (-0.25, 0.5, -1)$. Statistics of the latent network $\mathbf{A}^* = \{A^*_{ij}\}_{i,j=1}^N$ are summarised in Table 4.6.1. Suppose two self-reported and mismeasured network proxies are available for all units: for $i, j = 1, 2, ..., N$,

$$A_{ij} = \omega_i\big[U_{ij}A^*_{ij} + V_{ij}(1 - A^*_{ij})\big] + (1 - \omega_i)A^*_{ij},$$
$$\tilde{A}_{ij} = \tilde{\omega}_i\big[\tilde{U}_{ij}A^*_{ij} + \tilde{V}_{ij}(1 - A^*_{ij})\big] + (1 - \tilde{\omega}_i)A^*_{ij},$$

where ω_i, U_{ij}, V_{ij}, $\tilde{\omega}_i$, $\tilde{U}_{ij}$ and $\tilde{V}_{ij}$ are mutually independent and randomly generated binary indicators, taking value one with probabilities p^ω, p^U, p^V, $p^{\tilde{\omega}}$, $p^{\tilde{U}}$, $p^{\tilde{V}}$, respectively. In particular, taking A_{ij} as an example, ω_i indicates whether unit i ever misreports his or her links, and p^ω captures the overall level of misreporting. If unit i misreports, there are two types of classification errors: $U_{ij} = 0$ indicates that units i and j will be misclassified as unlinked if they are actually linked with $A^*_{ij} = 1$ (false negative); $V_{ij} = 1$ indicates that units i and j will be misclassified as linked if they are in fact unlinked with $A^*_{ij} = 0$

(false positive). Therefore, $1-p^U$ and p^V are the probability of false negative and false positive, respectively.

Following the design of Leung (2020b), assume the full network is collected for both proxies, meaning that $\mathcal{P}=\{1,2,...,N\}$. Given the DGP design, the dependency neighbourhood of each unit i can be set as a collection of units that are located close to unit i with distance less than r, i.e., $\Delta(i,N)=\{j\in\{1,2,...,N\},\ \|\rho_i-\rho_j\|_1\leq r\}$.

We generate data using sample size $N\in\{1000,2000,5000\}$ with replications $M=1000$. In the first-step kernel estimation, we set the bandwidth to be $h=N^{-3/8}$.

Table 4.6.1: Statistics of Latent Links

	$r_{deg}=5$					$r_{deg}=8$				
	$\mathcal{F}_i^*$		S_i^*		total	$\mathcal{F}_i^*$		S_i^*		total
N	avg.	max	avg.	max		avg.	max	avg.	max	
1k	5.65	15.52	1.69	7.31	5649	8.92	21.45	2.68	9.53	8919
2k	5.73	16.36	1.72	7.90	11458	9.08	22.54	2.72	10.26	18167
5k	5.80	17.39	1.74	8.55	29018	9.23	23.78	2.77	11.07	46165

Note: statistics reported in this table are the average over 1000 replications.

4.6.1. Semiparametric Estimation with Two Network Proxies

The overall misclassification rates are set as $p^\omega=p^{\tilde{\omega}}=0.6$. For the first proxy, let $1-p^U\in\{0.2,0.4\}$ and $p^V=\delta^V/N$ with $\delta^V\in\{0.1,0.5\}$ to ensure the network sparsity. For the second proxy, set $1-p^{\tilde{U}}\in\{0.2,0.4\}$ and $p^{\tilde{V}}=0$. Then, the first proxy possesses both the false negative and false positive classification errors, and the second one contains no false positive errors. Table 4.6.2 reports the statistics of the two mismeasured network proxies for different misclassification rates. We can see that when p^U or $p^{\tilde{U}}$ is 0.2, the misclassification rates are relatively low, varying from 12% to 17%. While when p^U or $p^{\tilde{U}}$ is set to be 0.4, the misclassification rates become quite high, varying between 24% to 29%. In what follows, we compare three estimation procedures:

(1) **SPE**: the semiparametric estimation studied in Section 4.5.3 using two proxies;

as well as two naive estimation procedures (ordinary least square (OLS)) that ignore potential misclassification errors:

(2) **Naive 1**: OLS of Y_i on $(1, D_i, D_i\mathcal{F}_i Z_i, S_i, S_i^2, S_i Z_i, S_i\mathcal{F}_i)$;

(3) **Naive 2**: OLS of Y_i on $(1, D_i, D_i\tilde{\mathcal{F}}_i Z_i, \tilde{S}_i, \tilde{S}_i^2, \tilde{S}_i Z_i, \tilde{S}_i\tilde{\mathcal{F}}_i)$.

Tables 4.6.3 to 4.6.6 display the estimation results for the treatment and the spillover effects obtained using the above three approaches. The bias, the standard deviation (sd), the mean squared error (mse), and the coverage rate (cr) of the 95% confidence interval for the true value of the causal parameter are reported.

For the treatment effect $\tau_d(0,0,3)$ (Table 4.6.3), the three estimation methods are roughly comparable in terms of the mse and cr. This finding is not surprising given that the treatment status of each ego unit is correctly observed and the network measurement errors do not impact the naive estimation of $\tau_d(0,0,3)$ for the units with $Z_i = 0$.

Analysing the results for the treatment effect $\tau_d(0,1,3)$ (Table 4.6.4), and the spillover effects $\tau_s(1,0,3)$ (Table 4.6.5) and $\tau_s(1,1,3)$ (Table 4.6.6), several interesting patters emerge. First and most importantly, the bias of the SPE is significantly lower than the bias of the two naive estimations in most cases. This is especially true if the network degree is relatively small ($r_{deg} = 5$), the misclassification rate is relatively low ($1 - p^U = 1 - p^{\tilde{U}} = 0.2$), or the sample size is sufficiently large ($N = 5000$).

In addition, as expected, the bias of the SPE decreases as the sample size increases for most cases. While, the two naive estimations are biased in all settings, and the bias is quite severe when the misclassification rate is relatively high ($1 - p^U = 1 - p^{\tilde{U}} = 0.4$) or the network degree is relatively large ($r_{deg} = 8$). Increasing the sample size fails to mitigate the bias of the two naive estimations. For instance, consider the estimation of the spillover $\tau_s(1,0,3)$ under $r_{deg} = 8$ in panel (b) of Table 4.6.5. Under the low misclassification rate $1 - p^U = 1 - p^{\tilde{U}} = 0.2$, $\delta^V = 0.1$ and $N = 1000$, the bias of SPE (0.076) is 11.6% of the bias of Naive 1 (0.653), and is 9.7% of the bias of Naive 2 (0.780). When sample size increases to $N = 5000$, the bias of SPE (-0.034) decreases to 5.2% of the bias of Naive 1 (0.650) and 4.5% of the bias of Naive 2 (0.753). While, in the case of a high misclassification rate $1 - p^U = 1 - p^{\tilde{U}} = 0.4$ and $\delta^V = 0.1$, the naive estimators have even larger bias: the biases of Naive 1 and 2 are roughly double the biases in the

Table 4.6.2: Statistics of Misclassified Links ($p^{\omega} = 0.6,\ p^V = \delta^V/N$)

(a) $r_{deg} = 5$

δ^V	$1-p^U$ (%)	p^V (%)	N	$\mathcal{F}_i$ avg.	$\mathcal{F}_i$ max	S_i avg.	S_i max	Misclassified links 1 to 0	0 to 1	total	ratio (%)
		0	1k	4.97	14.84	1.49	6.95	677.5	0	677.5	12.01
0	20	0	2k	5.04	15.63	1.51	7.51	1371	0	1371	12.00
		0	5k	5.11	16.66	1.53	8.12	3482	0	3482	12.01
		0.010	1k	5.03	14.87	1.51	6.97	677.5	59.64	737.2	13.05
0.1	20	0.005	2k	5.10	15.67	1.53	7.56	1371	119.1	1490	13.01
		0.002	5k	5.17	16.66	1.55	8.17	3482	299.5	3782	13.03
		0.050	1k	5.27	15.02	1.58	7.08	677.5	298.2	975.7	17.28
0.5	20	0.025	2k	5.34	15.81	1.60	7.65	1371	596.3	1968	17.18
		0.010	5k	5.41	16.81	1.62	8.28	3482	1501	4983	17.17
		0	1k	4.29	14.70	1.29	6.81	1356	0	1356	24.03
0	40	0	2k	4.35	15.53	1.31	7.37	2746	0	2746	24.00
		0	5k	4.41	16.56	1.32	7.99	6961	0	6961	24.02
		0.010	1k	4.35	14.72	1.31	6.79	1356	59.64	1416	25.07
0.1	40	0.005	2k	4.42	15.55	1.32	7.38	2746	119.1	2865	25.01
		0.002	5k	4.47	16.52	1.34	7.99	6961	299.5	7260	25.02
		0.050	1k	4.59	14.74	1.38	6.84	1356	298.2	1654	29.29
0.5	40	0.025	2k	4.65	15.56	1.40	7.42	2746	596.3	3343	29.18
		0.010	5k	4.71	16.54	1.41	8.03	6961	1501	8462	29.16

(b) $r_{deg} = 8$

δ^V	$1-p^U$ (%)	p^V (%)	N	$\mathcal{F}_i$ avg.	$\mathcal{F}_i$ max	S_i avg.	S_i max	Misclassified links 1 to 0	0 to 1	total	ratio (%)
		0	1k	7.85	20.50	2.36	9.08	1069	0	1069	12.02
0	20	0	2k	7.99	21.55	2.40	9.75	2177	0	2177	12.01
		0	5k	8.12	22.83	2.44	10.58	5540	0	5540	12.01
		0.010	1k	7.91	20.52	2.37	9.07	1069	59.45	1128	12.65
0.1	20	0.005	2k	8.05	21.55	2.42	9.81	2177	118.9	2296	12.64
		0.002	5k	8.18	22.82	2.45	10.55	5540	299.3	5839	12.65
		0.050	1k	8.15	20.60	2.45	9.15	1069	297.3	1366	15.32
0.5	20	0.025	2k	8.29	21.64	2.49	9.88	2177	595.3	2772	15.26
		0.010	5k	8.43	22.90	2.53	10.62	5540	1500	7039	15.25
		0	1k	6.78	20.40	2.03	8.92	2139	0	2139	24.03
0	40	0	2k	6.90	21.48	2.07	9.62	4356	0	4356	24.01
		0	5k	7.02	22.75	2.10	10.45	11078	0	11078	24.02
		0.010	1k	6.84	20.43	2.05	8.87	2139	59.45	2199	24.65
0.1	40	0.005	2k	6.96	21.48	2.09	9.66	4356	118.9	4475	24.63
		0.002	5k	7.08	22.77	2.12	10.41	11078	299.3	11377	24.64
		0.050	1k	7.08	20.43	2.12	8.90	2139	297.3	2437	27.32
0.5	40	0.025	2k	7.20	21.48	2.16	9.67	4356	595.3	4951	27.25
		0.010	5k	7.32	22.77	2.19	10.43	11078	1500	12577	27.24

Note: The results in this table can be applied to both $(\mathcal{F}_i, S_i)$ and $(\tilde{\mathcal{F}}_i, \tilde{S}_i)$.

case of a low misclassification rate. Although the bias of SPE also increases in cases with high misclassification rate compared to that in cases with low misclassification rate, it diminishes as the sample size increases. Hence, the simulations verify that ignoring the network classification errors results in non-negligible bias that does not abate as the sample size increases.

In addition, we can see that the sd and the mse of SPE decrease as the sample size increases. The mse of SPE outperforms those of Naive 1 and Naive 2 in most cases when the sample size is relatively large. Moreover, the coverage rate of the SPE is closer to the nominal level than either naive method and approaches the nominal level as the sample size increases. In contrast, the coverage rates of the naive approaches drop rapidly as the sample size increases or as the misclassification worsens. For example, when the misclassification rate is low $1 - p^U = 1 - p^{\tilde{U}} = 0.2$ and $\delta^V = 0.1$, for the spillover effect $\tau_s(1, 1, 3)$ under $r_{deg} = 8$ (panel (b) in Table 4.6.6), the cr is 11.9% for Naive 1 and 6.2% for Naive 2, while it is 93.1% for SPE. When $N = 5000$, the cr is 0% for both Naive 1 and 2, but is 93.8% for SPE.

However, it is important to note that, relatively speaking, the accuracy of the SPE decreases as r_{deg} increases, or as the misclassification rate increases. To sum up, the SPE works significantly better than the naive estimators that neglect network misclassifications, especially if the sample size is relatively large.

Table 4.6.3: Estimation of Treatment Effect $\tau_d(0,0,3)$ ($p^{\omega} = p^{\tilde{\omega}} = 0.6,\ p^V = \delta^V/N$)

(a) $r_{deg} = 5$

δ^V	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)	N	SPE				Naive 1				Naive 2			
				bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
	(20, 0.010)		1k	-0.060	0.349	0.125	0.931	-0.073	0.292	0.091	0.943	-0.063	0.294	0.090	0.935
0.1	(20, 0.005)	(20,0)	2k	-0.027	0.245	0.061	0.941	-0.077	0.206	0.048	0.933	-0.071	0.208	0.048	0.931
	(20, 0.002)		5k	-0.016	0.133	0.018	0.937	-0.060	0.132	0.021	0.924	-0.063	0.130	0.021	0.916
	(20, 0.050)		1k	-0.053	0.354	0.128	0.941	-0.076	0.319	0.108	0.942	-0.061	0.284	0.084	0.946
0.5	(20, 0.025)	(20,0)	2k	-0.032	0.243	0.060	0.941	-0.097	0.219	0.057	0.925	-0.061	0.205	0.046	0.939
	(20, 0.010)		5k	-0.028	0.133	0.019	0.942	-0.083	0.139	0.026	0.909	-0.062	0.133	0.022	0.922
	(40, 0.010)		1k	0.075	0.538	0.296	0.950	-0.035	0.405	0.165	0.952	-0.016	0.390	0.153	0.948
0.1	(40, 0.005)	(40,0)	2k	0.051	0.384	0.150	0.942	-0.018	0.276	0.076	0.948	-0.019	0.273	0.075	0.955
	(40, 0.002)		5k	0.038	0.236	0.057	0.938	-0.013	0.173	0.030	0.948	0.004	0.182	0.033	0.945
	(40, 0.050)		1k	0.059	0.547	0.303	0.940	-0.040	0.398	0.160	0.958	-0.012	0.399	0.160	0.950
0.5	(40, 0.025)	(40,0)	2k	0.040	0.368	0.137	0.941	-0.047	0.280	0.081	0.954	-0.015	0.283	0.080	0.953
	(40, 0.010)		5k	0.022	0.219	0.048	0.940	-0.052	0.189	0.038	0.944	0.012	0.175	0.031	0.952

(b) $r_{deg} = 8$

δ^V	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)	N	SPE				Naive 1				Naive 2			
				bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
	(20, 0.010)		1k	0.060	0.574	0.334	0.953	-0.161	0.527	0.304	0.941	-0.142	0.507	0.277	0.949
0.1	(20, 0.005)	(20,0)	2k	-0.048	0.284	0.083	0.940	-0.140	0.359	0.148	0.929	-0.130	0.392	0.170	0.928
	(20, 0.002)		5k	-0.020	0.180	0.033	0.954	-0.141	0.237	0.076	0.910	-0.139	0.233	0.074	0.908
	(20, 0.050)		1k	-0.016	0.535	0.287	0.930	-0.170	0.518	0.298	0.938	-0.170	0.522	0.302	0.938
0.5	(20, 0.025)	(20,0)	2k	0.019	0.399	0.160	0.954	-0.141	0.394	0.175	0.935	-0.155	0.361	0.154	0.934
	(20, 0.010)		5k	-0.019	0.169	0.029	0.963	-0.162	0.241	0.084	0.899	-0.144	0.243	0.080	0.902
	(40, 0.010)		1k	0.383	0.792	0.774	0.934	-0.119	0.776	0.617	0.946	-0.120	0.756	0.585	0.942
0.1	(40, 0.005)	(40,0)	2k	0.356	0.569	0.451	0.933	-0.118	0.574	0.343	0.946	-0.103	0.560	0.325	0.945
	(40, 0.002)		5k	0.280	0.343	0.196	0.897	-0.101	0.354	0.135	0.935	-0.086	0.354	0.133	0.938
	(40, 0.010)		1k	0.367	0.794	0.765	0.919	-0.184	0.757	0.607	0.948	-0.121	0.749	0.575	0.949
0.5	(40, 0.025)	(40,0)	2k	0.323	0.556	0.413	0.934	-0.148	0.552	0.326	0.937	-0.115	0.550	0.316	0.950
	(40, 0.010)		5k	0.211	0.342	0.162	0.910	-0.154	0.348	0.145	0.928	-0.103	0.362	0.141	0.945

Note: SPE lists the semiparametric estimation results proposed in Section 4.5.3. Estimates of Naive 1 are computed using OLS with $\{Y_i, D_i, S_i, Z_i, \mathcal{F}_i\}_{i=1}^N$; and estimates of Naive 2 are computed using OLS with $\{Y_i, D_i, \tilde{S}_i, Z_i, \tilde{\mathcal{F}}_i\}_{i=1}^N$. True value of the treatment effect $\tau_d(0,0,3) = 1$.

Table 4.6.4: Estimation of Treatment Effect $\tau_d(0,1,3)$ ($p^{\omega} = p^{\tilde{\omega}} = 0.6,\ p^V = \delta^V/N$)

(a) $r_{deg} = 5$

δ^V	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)	N	SPE bias	SPE sd	SPE mse	SPE cr	Naive 1 bias	Naive 1 sd	Naive 1 mse	Naive 1 cr	Naive 2 bias	Naive 2 sd	Naive 2 mse	Naive 2 cr
	(20, 0.010)		1k	-0.068	0.313	0.102	0.948	0.131	0.279	0.095	0.921	0.120	0.268	0.086	0.936
0.1	(20, 0.005)	(20,0)	2k	-0.059	0.215	0.050	0.939	0.121	0.195	0.053	0.897	0.140	0.190	0.056	0.883
	(20, 0.002)		5k	-0.052	0.126	0.019	0.930	0.132	0.122	0.032	0.815	0.133	0.126	0.034	0.818
	(20, 0.050)		1k	-0.066	0.323	0.108	0.944	0.078	0.283	0.086	0.941	0.133	0.270	0.090	0.920
0.5	(20, 0.025)	(20,0)	2k	-0.059	0.209	0.047	0.946	0.075	0.201	0.046	0.943	0.136	0.195	0.057	0.892
	(20, 0.010)		5k	-0.057	0.114	0.016	0.931	0.081	0.124	0.022	0.907	0.135	0.115	0.031	0.778
	(40, 0.010)		1k	0.040	0.528	0.281	0.953	0.287	0.405	0.247	0.885	0.318	0.408	0.268	0.882
0.1	(40, 0.005)	(40,0)	2k	0.007	0.350	0.123	0.949	0.299	0.293	0.175	0.834	0.303	0.291	0.176	0.825
	(40, 0.002)		5k	0.001	0.209	0.044	0.957	0.305	0.181	0.126	0.600	0.316	0.183	0.133	0.581
	(40, 0.050)		1k	0.054	0.522	0.276	0.946	0.255	0.393	0.219	0.898	0.303	0.411	0.261	0.892
0.5	(40, 0.025)	(40,0)	2k	0.027	0.325	0.106	0.953	0.252	0.286	0.145	0.863	0.325	0.275	0.181	0.788
	(40, 0.010)		5k	0.003	0.196	0.039	0.952	0.248	0.181	0.094	0.730	0.322	0.185	0.138	0.590

(b) $r_{deg} = 8$

δ^V	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)	N	SPE bias	SPE sd	SPE mse	SPE cr	Naive 1 bias	Naive 1 sd	Naive 1 mse	Naive 1 cr	Naive 2 bias	Naive 2 sd	Naive 2 mse	Naive 2 cr
	(20, 0.010)		1k	-0.024	0.516	0.267	0.953	0.071	0.445	0.203	0.941	0.083	0.446	0.205	0.944
0.1	(20, 0.005)	(20,0)	2k	-0.061	0.323	0.108	0.951	0.084	0.319	0.109	0.937	0.078	0.335	0.118	0.941
	(20, 0.002)		5k	-0.089	0.190	0.044	0.939	0.077	0.200	0.046	0.936	0.087	0.202	0.048	0.938
	(20, 0.050)		1k	0.062	0.629	0.400	0.966	0.062	0.448	0.204	0.955	0.075	0.454	0.212	0.943
0.5	(20, 0.025)	(20,0)	2k	-0.054	0.373	0.142	0.960	0.058	0.330	0.112	0.945	0.082	0.305	0.100	0.943
	(20, 0.010)		5k	-0.096	0.177	0.041	0.937	0.044	0.208	0.045	0.942	0.078	0.210	0.050	0.927
	(40, 0.010)		1k	0.329	0.813	0.768	0.932	0.267	0.703	0.565	0.938	0.279	0.709	0.581	0.933
0.1	(40, 0.005)	(40,0)	2k	0.299	0.571	0.416	0.932	0.300	0.511	0.351	0.908	0.306	0.502	0.346	0.901
	(40, 0.002)		5k	0.173	0.336	0.143	0.916	0.272	0.318	0.175	0.877	0.285	0.322	0.185	0.851
	(40, 0.010)		1k	0.300	0.814	0.752	0.934	0.244	0.655	0.488	0.939	0.298	0.700	0.579	0.933
0.5	(40, 0.025)	(40,0)	2k	0.256	0.538	0.355	0.925	0.240	0.474	0.282	0.912	0.298	0.500	0.339	0.903
	(40, 0.010)		5k	0.119	0.327	0.121	0.937	0.218	0.316	0.147	0.888	0.284	0.330	0.190	0.863

Note: SPE lists the semiparametric estimation results proposed in Section 4.5.3. Estimates of Naive 1 are computed using OLS with $\{Y_i, D_i, S_i, Z_i, \mathcal{F}_i\}_{i=1}^N$; and estimates of Naive 2 are computed using OLS with $\{Y_i, D_i, \tilde{S}_i, Z_i, \tilde{\mathcal{F}}_i\}_{i=1}^N$. True value of the treatment effect $\tau_d(0,1,3) = 2$.

Table 4.6.5: Estimation of Spillover Effect $\tau_s(1,0,3)$ ($p^{\omega} = p^{\tilde{\omega}} = 0.6,\ p^V = \delta^V/N$)

(a) $r_{deg} = 5$

δ^V	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)	N	SPE				Naive 1				Naive 2			
				bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
	(20, 0.010)		1k	0.035	0.488	0.240	0.957	0.270	0.214	0.119	0.749	0.366	0.215	0.180	0.582
0.1	(20, 0.005)	(20,0)	2k	0.040	0.373	0.141	0.961	0.254	0.160	0.090	0.628	0.351	0.155	0.147	0.365
	(20, 0.002)		5k	0.073	0.209	0.049	0.945	0.252	0.102	0.074	0.306	0.342	0.100	0.127	0.074
	(20, 0.050)		1k	0.050	0.543	0.297	0.947	-0.096	0.220	0.058	0.924	0.355	0.211	0.170	0.596
0.5	(20, 0.025)	(20,0)	2k	0.054	0.354	0.128	0.952	-0.091	0.159	0.033	0.919	0.352	0.151	0.147	0.360
	(20, 0.010)		5k	0.082	0.209	0.051	0.930	-0.104	0.105	0.022	0.841	0.346	0.102	0.130	0.089
	(40, 0.010)		1k	0.051	0.750	0.565	0.945	0.436	0.283	0.270	0.650	0.533	0.301	0.375	0.562
0.1	(40, 0.005)	(40,0)	2k	0.079	0.607	0.375	0.949	0.432	0.209	0.230	0.441	0.532	0.216	0.330	0.299
	(40, 0.002)		5k	0.165	0.351	0.150	0.923	0.415	0.138	0.192	0.144	0.507	0.138	0.276	0.046
	(40, 0.050)		1k	0.037	0.753	0.568	0.952	0.082	0.289	0.090	0.948	0.519	0.309	0.364	0.611
0.5	(40, 0.025)	(40,0)	2k	0.086	0.557	0.317	0.942	0.079	0.212	0.051	0.936	0.517	0.204	0.309	0.285
	(40, 0.010)		5k	0.175	0.369	0.167	0.927	0.057	0.138	0.022	0.932	0.505	0.138	0.274	0.044

(b) $r_{deg} = 8$

δ^V	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)	N	SPE				Naive 1				Naive 2			
				bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
	(20, 0.010)		1k	0.076	0.861	0.748	0.935	0.653	0.367	0.561	0.544	0.780	0.354	0.733	0.392
0.1	(20, 0.005)	(20,0)	2k	0.051	0.675	0.459	0.947	0.651	0.283	0.504	0.358	0.769	0.269	0.664	0.176
	(20, 0.002)		5k	-0.034	0.403	0.163	0.945	0.650	0.171	0.451	0.040	0.753	0.167	0.595	0.010
	(20, 0.050)		1k	0.119	0.891	0.809	0.943	0.265	0.368	0.205	0.885	0.774	0.372	0.737	0.459
0.5	(20, 0.025)	(20,0)	2k	0.029	0.746	0.557	0.941	0.263	0.263	0.138	0.833	0.764	0.267	0.655	0.181
	(20, 0.010)		5k	-0.013	0.375	0.141	0.951	0.245	0.176	0.091	0.730	0.745	0.174	0.586	0.014
	(40, 0.010)		1k	1.270	1.023	2.659	0.789	1.260	0.535	1.872	0.335	1.348	0.517	2.085	0.254
0.1	(40, 0.005)	(40,0)	2k	1.040	0.831	1.772	0.796	1.244	0.367	1.683	0.104	1.314	0.379	1.869	0.063
	(40, 0.002)		5k	0.743	0.602	0.915	0.787	1.179	0.253	1.454	0.008	1.269	0.255	1.676	0.001
	(40, 0.010)		1k	1.171	1.045	2.462	0.823	0.873	0.506	1.019	0.581	1.356	0.516	2.106	0.231
0.5	(40, 0.025)	(40,0)	2k	0.993	0.854	1.715	0.843	0.814	0.368	0.798	0.401	1.298	0.369	1.822	0.053
	(40, 0.010)		5k	0.653	0.620	0.811	0.845	0.803	0.241	0.702	0.090	1.270	0.253	1.678	0.002

Note: SPE lists the semiparametric estimation results proposed in Section 4.5.3. Estimates of Naive 1 are computed using OLS with $\{Y_i, D_i, S_i, Z_i, \mathcal{F}_i\}_{i=1}^N$; and estimates of Naive 2 are computed using OLS with $\{Y_i, D_i, \tilde{S}_i, Z_i, \tilde{\mathcal{F}}_i\}_{i=1}^N$. True value of the treatment effect $\tau_s(1,0,3) = 3$.

Table 4.6.6: Estimation of Spillover Effect $\tau_s(1,1,3)$ ($p^{\omega} = p^{\tilde{\omega}} = 0.6,\ p^V = \delta^V/N$)

(a) $r_{deg} = 5$

δ^V	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)	N	SPE				Naive 1				Naive 2			
				bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
	(20, 0.010)		1k	0.060	0.846	0.719	0.942	0.617	0.243	0.440	0.272	0.709	0.246	0.564	0.172
0.1	(20, 0.005)	(20,0)	2k	0.028	0.522	0.273	0.957	0.603	0.173	0.393	0.064	0.693	0.181	0.513	0.029
	(20, 0.002)		5k	0.021	0.284	0.081	0.954	0.605	0.112	0.379	0.001	0.695	0.112	0.496	0.000
	(20, 0.050)		1k	0.127	0.928	0.878	0.941	0.289	0.246	0.144	0.768	0.690	0.246	0.537	0.203
0.5	(20, 0.025)	(20,0)	2k	0.090	0.555	0.316	0.952	0.299	0.186	0.124	0.633	0.704	0.173	0.525	0.025
	(20, 0.010)		5k	0.047	0.306	0.096	0.954	0.294	0.117	0.100	0.275	0.703	0.114	0.507	0.000
	(40, 0.010)		1k	0.619	0.979	1.342	0.897	1.273	0.343	1.739	0.037	1.386	0.359	2.050	0.030
0.1	(40, 0.005)	(40,0)	2k	0.377	0.864	0.889	0.915	1.288	0.242	1.718	0.000	1.406	0.249	2.038	0.000
	(40, 0.002)		5k	0.232	0.591	0.403	0.937	1.277	0.167	1.658	0.000	1.385	0.164	1.944	0.000
	(40, 0.050)		1k	0.564	1.016	1.351	0.905	0.865	0.344	0.866	0.297	1.370	0.369	2.013	0.041
0.5	(40, 0.025)	(40,0)	2k	0.272	0.869	0.829	0.926	0.875	0.247	0.828	0.053	1.387	0.249	1.986	0.000
	(40, 0.010)		5k	0.207	0.588	0.388	0.949	0.868	0.156	0.777	0.000	1.389	0.167	1.957	0.000

(b) $r_{deg} = 8$

δ^V	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)	N	SPE				Naive 1				Naive 2			
				bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
	(20, 0.010)		1k	0.382	1.416	2.150	0.931	1.320	0.419	1.919	0.119	1.456	0.403	2.281	0.062
0.1	(20, 0.005)	(20,0)	2k	0.139	1.055	1.133	0.950	1.345	0.321	1.912	0.013	1.480	0.304	2.282	0.007
	(20, 0.002)		5k	-0.053	0.621	0.389	0.938	1.369	0.191	1.912	0.000	1.489	0.196	2.256	0.000
	(20, 0.050)		1k	0.498	1.411	2.239	0.940	0.921	0.412	1.019	0.396	1.474	0.424	2.353	0.066
0.5	(20, 0.025)	(20,0)	2k	0.029	1.003	1.007	0.944	0.928	0.293	0.947	0.111	1.470	0.304	2.255	0.004
	(20, 0.010)		5k	-0.019	0.627	0.393	0.950	0.931	0.201	0.908	0.002	1.476	0.201	2.220	0.000
	(40, 0.010)		1k	2.673	1.634	9.812	0.647	3.000	0.618	9.379	0.001	3.120	0.631	10.136	0.001
0.1	(40, 0.005)	(40,0)	2k	2.185	1.432	6.826	0.704	2.974	0.443	9.041	0.000	3.113	0.449	9.892	0.000
	(40, 0.002)		5k	1.437	1.152	3.391	0.779	2.997	0.301	9.072	0.000	3.125	0.294	9.851	0.000
	(40, 0.010)		1k	2.540	1.698	9.335	0.724	2.482	0.607	6.526	0.012	3.125	0.613	10.140	0.001
0.5	(40, 0.025)	(40,0)	2k	2.135	1.460	6.692	0.727	2.459	0.423	6.225	0.000	3.109	0.441	9.860	0.000
	(40, 0.010)		5k	1.262	1.175	2.974	0.837	2.489	0.279	6.274	0.000	3.120	0.295	9.821	0.000

Note: SPE lists the semiparametric estimation results proposed in Section 4.5.3. Estimates of Naive 1 are computed using OLS with $\{Y_i, D_i, S_i, Z_i, \mathcal{F}_i\}_{i=1}^N$; and estimates of Naive 2 are computed using OLS with $\{Y_i, D_i, \tilde{S}_i, Z_i, \tilde{\mathcal{F}}_i\}_{i=1}^N$. True value of the treatment effect $\tau_s(1,1,3) = 2.5$.

4.6.2. Robustness of the Semiparametric Estimation

Two key identification assumptions, i.e. the exclusion restriction and the one type of measurement error, may be violated in some applications. In this section, we assess the following empirical questions using additional Monte Carlo experiments: Is SPE robust to the violation of these two assumptions? Does SPE still perform better than the naive estimation if any violation is present? To answer the above questions, we consider a Monte Carlo exercise that analyses the behaviour of the SPE when the observable networks are generated according to one of the following empirical relevant departures from the above identification conditions:

(i) violation of the "exclusion restriction": generate random errors $(U^*_{ij}, V^*_{ij}, \tilde{U}^*_{ij}, \tilde{V}^*_{ij})'$ from a joint normal distribution for all $i, j = 1, 2, ..., N$,

$$\begin{pmatrix} U^*_{ij} \\ V^*_{ij} \\ \tilde{U}^*_{ij} \\ \tilde{V}^*_{ij} \end{pmatrix} = \mathbb{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \varrho & 0 \\ 0 & 1 & 0 & \varrho \\ \varrho & 0 & 1 & 0 \\ 0 & \varrho & 0 & 1 \end{pmatrix} \right), \quad \begin{aligned} U_{ij} &= 1[\Phi(U^*_{ij}) < 1 - p^U],\ V_{ij} = 1[\Phi(V^*_{ij}) < p^V] \\ \tilde{U}_{ij} &= 1[\Phi(\tilde{U}^*_{ij}) < 1 - p^{\tilde{U}}],\ \tilde{V}_{ij} = 1[\Phi(\tilde{V}^*_{ij}) < p^{\tilde{V}}]. \end{aligned}$$

where $\varrho \in \{0.05, 0.1\}$ controls the correlation between the misclassification errors;

(ii) violation of the "one type of measurement error": generate $\tilde{V}_{ij}$ via $p^{\tilde{V}} = \delta^{\tilde{V}}/N$ with $\delta^{\tilde{V}} \in \{0.05, 0.1\}$;

while all remaining elements of the Monte Carlo are precisely as in Section 4.6.1. Results for the three approaches are reported in Table 4.6.7 and 4.6.8. Table 4.6.8 displays the results for cases with relatively large sample size ($N = 5000$), which is sufficient to illustrate the asymptotic performance of the SPE relative to the naive approaches.

To check the robustness of the SPE method, compare the results in Tables 4.6.3 to 4.6.6 with their counterparts in Tables 4.6.7 and 4.6.8. We can see that the violation of either assumptions deteriorates the performance of SPE, but only at a limited degree.

Take the spillover $\tau_s(1, 0, 3)$ as an example. When $r_{deg} = 5$, $N = 5000$ and misclassification rate is relatively low ($1 - p^U = 1 - p^{\tilde{U}} = 0.2$, $\delta^V = 0.1$), the bias and the mse

of SPE under the point identification condition are 0.073 and 0.049, respectively, with the coverage rate 94.5%. When the exclusion restriction fails to hold ($\varrho = 0.1$), the bias, mse and coverage rate are 0.101, 0.097 and 92.0%, respectively. When the one type of measure error is violated ($\delta^{\tilde{V}} = 0.05$), the bias, mse and coverage rate are 0.104, 0.103 and 93.3%, respectively.

The question whether SPE still outperforms the naive estimation can be answered by comparing the results in Tables 4.6.7 and 4.6.8. For the treatment effect $\tau_d(0,0,3)$, the bias and the mse of SPE are smaller than those of the two naive methods when the misclassification rate is relatively low ($1 - p^U = 1 - p^{\tilde{U}} = 0.2$); while the SPE produces a slightly larger bias compared to that of the two naive methods when the misclassification rate is relatively high ($1 - p^U = 1 - p^{\tilde{U}} = 0.4$). For the treatment effect $\tau_d(0,1,3)$, the spillover effects $\tau_s(1,0,3)$ and $\tau_s(1,1,3)$, the bias and the mse of the SPE are better than those of the two naive estimators across almost all designs. Notably, the coverage rate of the SPE is much closer to the nominal level than either of the naive estimators. For example, consider the case where $r_{deg} = 8$ with low misclassification rate ($1 - p^U = 1 - p^{\tilde{U}} = 0.2$). If the exclusion restriction is violated, the coverage rate of the spillover effects $\tau_s(1,0,3)$ and $\tau_s(1,1,3)$ obtained by the SPE method lies in the range of 93.0% to 94.3%, while for the native estimators, the coverage rate is less than 6% for $\tau_s(1,0,3)$ and can even be 0% for $\tau_s(1,1,3)$. If the one type of measurement error assumption fails, the SPE's coverage rate of $\tau_s(1,0,3)$ and $\tau_s(1,1,3)$ varies from 94.9% to 95.5%, while it varies from 0% to less than 4% for the naive estimations.

The results in this section show that (i) the SPE approach is robust to mild violations of the one type of measurement error assumption; and (ii) the SPE is superior to the naive methods except in rare cases, in the sense that the bias reduction provided by the SPE is substantial and its causal inference is much more reliable.

Table 4.6.7: Robustness Check for Exclusion Restriction ($p^{\omega} = p^{\tilde{\omega}} = 0.6$, $p^{V} = \delta^{V}/N$, $p^{\tilde{V}} = 0$, $\delta^{V} = 0.1$, $N = 5k$)

(a) $r_{deg} = 5$

ϱ	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)		SPE				Naive 1				Naive 2			
				bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
0.05	(20,0.002)	(20,0)	$\tau_d(0,0,3)$	-0.020	0.118	0.014	0.915	-0.067	0.128	0.021	0.910	-0.054	0.135	0.021	0.929
			$\tau_d(0,1,3)$	-0.053	0.118	0.017	0.931	0.123	0.118	0.029	0.834	0.140	0.124	0.035	0.800
			$\tau_s(1,0,3)$	0.145	0.184	0.055	0.865	0.249	0.097	0.072	0.273	0.347	0.099	0.130	0.063
			$\tau_s(1,1,3)$	0.095	0.301	0.100	0.935	0.604	0.111	0.377	0.000	0.701	0.114	0.504	0.000
0.1	(20,0.002)	(20,0)	$\tau_d(0,0,3)$	-0.021	0.110	0.013	0.934	-0.065	0.129	0.021	0.917	-0.062	0.131	0.021	0.923
			$\tau_d(0,1,3)$	-0.059	0.112	0.016	0.910	0.131	0.120	0.032	0.802	0.132	0.120	0.032	0.815
			$\tau_s(1,0,3)$	0.157	0.200	0.065	0.869	0.247	0.098	0.070	0.296	0.346	0.099	0.129	0.060
			$\tau_s(1,1,3)$	0.101	0.295	0.097	0.920	0.596	0.109	0.368	0.000	0.696	0.115	0.497	0.000
0.05	(40,0.002)	(40,0)	$\tau_d(0,0,3)$	0.049	0.219	0.051	0.917	-0.019	0.175	0.031	0.950	-0.001	0.180	0.032	0.949
			$\tau_d(0,1,3)$	0.005	0.210	0.044	0.945	0.306	0.186	0.128	0.620	0.318	0.184	0.135	0.606
			$\tau_s(1,0,3)$	0.245	0.329	0.168	0.877	0.424	0.130	0.196	0.101	0.513	0.136	0.281	0.032
			$\tau_s(1,1,3)$	0.347	0.545	0.417	0.910	1.290	0.155	1.687	0.000	1.397	0.164	1.979	0.000
0.1	(40,0.002)	(40,0)	$\tau_d(0,0,3)$	0.070	0.225	0.055	0.930	0.005	0.182	0.033	0.956	0.014	0.174	0.030	0.951
			$\tau_d(0,1,3)$	0.010	0.208	0.043	0.954	0.307	0.187	0.129	0.613	0.313	0.182	0.131	0.586
			$\tau_s(1,0,3)$	0.236	0.330	0.165	0.882	0.413	0.139	0.190	0.153	0.499	0.134	0.267	0.053
			$\tau_s(1,1,3)$	0.344	0.509	0.377	0.886	1.282	0.161	1.670	0.000	1.388	0.157	1.951	0.000

(b) $r_{deg} = 8$

ϱ	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)		SPE				Naive 1				Naive 2			
				bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
0.05	(20, 0.002)	(20, 0)	$\tau_d(0,0,3)$	-0.049	0.192	0.039	0.959	-0.143	0.243	0.080	0.908	-0.141	0.241	0.078	0.923
			$\tau_d(0,1,3)$	-0.093	0.193	0.046	0.943	0.078	0.211	0.050	0.930	0.089	0.208	0.051	0.922
			$\tau_s(1,0,3)$	-0.026	0.394	0.156	0.939	0.647	0.183	0.452	0.059	0.751	0.173	0.594	0.007
			$\tau_s(1,1,3)$	-0.050	0.612	0.377	0.930	1.364	0.206	1.904	0.000	1.482	0.203	2.236	0.000
0.1	(20, 0.002)	(20, 0)	$\tau_d(0,0,3)$	-0.048	0.177	0.034	0.960	-0.144	0.241	0.079	0.914	-0.143	0.243	0.079	0.911
			$\tau_d(0,1,3)$	-0.102	0.191	0.047	0.929	0.080	0.213	0.052	0.929	0.082	0.207	0.050	0.926
			$\tau_s(1,0,3)$	-0.030	0.356	0.128	0.943	0.634	0.180	0.434	0.069	0.754	0.179	0.601	0.015
			$\tau_s(1,1,3)$	-0.046	0.591	0.351	0.937	1.352	0.201	1.868	0.000	1.488	0.203	2.256	0.000
0.05	(40, 0.002)	(40, 0)	$\tau_d(0,0,3)$	0.284	0.341	0.197	0.888	-0.098	0.349	0.132	0.943	-0.073	0.354	0.131	0.952
			$\tau_d(0,1,3)$	0.184	0.343	0.151	0.933	0.274	0.323	0.180	0.863	0.300	0.322	0.193	0.851
			$\tau_s(1,0,3)$	0.786	0.597	0.975	0.774	1.180	0.250	1.454	0.002	1.269	0.255	1.675	0.003
			$\tau_s(1,1,3)$	1.581	1.114	3.741	0.749	2.997	0.298	9.071	0.000	3.121	0.299	9.829	0.000
0.1	(40, 0.002)	(40, 0)	$\tau_d(0,0,3)$	0.310	0.312	0.193	0.855	-0.086	0.346	0.127	0.944	-0.065	0.349	0.126	0.942
			$\tau_d(0,1,3)$	0.195	0.325	0.144	0.917	0.285	0.333	0.192	0.861	0.306	0.325	0.199	0.851
			$\tau_s(1,0,3)$	0.781	0.618	0.992	0.796	1.178	0.249	1.449	0.004	1.270	0.249	1.674	0.000
			$\tau_s(1,1,3)$	1.498	1.131	3.522	0.776	2.995	0.290	9.057	0.000	3.119	0.293	9.813	0.000

Table 4.6.8: Robustness Check for One Type of Measurement Error ($p^{\omega} = p^{\tilde{\omega}} = 0.6,\ p^{V} = \delta^{V}/N,\ p^{\tilde{V}} = \delta^{\tilde{V}}/N,\ N = 5k$)

(a) $r_{deg} = 5$

δ^V	$\delta^{\tilde{V}}$	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)		SPE				Naive 1				Naive 2			
					bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
0.1	0.05	(20,0.002)	(20,0.001)	$\tau_d(0,0,3)$	-0.027	0.124	0.016	0.930	-0.069	0.136	0.023	0.908	-0.063	0.132	0.021	0.928
				$\tau_d(0,1,3)$	-0.055	0.115	0.016	0.923	0.125	0.122	0.031	0.818	0.133	0.119	0.032	0.803
				$\tau_s(1,0,3)$	0.125	0.204	0.057	0.880	0.248	0.102	0.072	0.312	0.298	0.107	0.100	0.200
				$\tau_s(1,1,3)$	0.104	0.303	0.103	0.933	0.600	0.116	0.373	0.001	0.646	0.119	0.432	0.002
0.1	0.1	(20,0.002)	(20,0.002)	$\tau_d(0,0,3)$	-0.027	0.129	0.017	0.934	-0.062	0.134	0.022	0.925	-0.059	0.132	0.021	0.920
				$\tau_d(0,1,3)$	-0.053	0.116	0.016	0.927	0.124	0.123	0.031	0.825	0.131	0.122	0.032	0.809
				$\tau_s(1,0,3)$	0.098	0.221	0.058	0.904	0.247	0.099	0.071	0.289	0.251	0.101	0.073	0.303
				$\tau_s(1,1,3)$	0.104	0.306	0.104	0.942	0.600	0.108	0.372	0.000	0.602	0.111	0.374	0.001
0.1	0.05	(40,0.002)	(40,0.001)	$\tau_d(0,0,3)$	0.049	0.231	0.056	0.941	-0.018	0.180	0.033	0.958	-0.003	0.178	0.032	0.945
				$\tau_d(0,1,3)$	-0.007	0.208	0.043	0.959	0.295	0.177	0.119	0.634	0.306	0.186	0.128	0.617
				$\tau_s(1,0,3)$	0.183	0.342	0.150	0.910	0.415	0.136	0.191	0.157	0.466	0.135	0.235	0.066
				$\tau_s(1,1,3)$	0.320	0.540	0.394	0.909	1.278	0.156	1.656	0.000	1.342	0.159	1.826	0.000
0.1	0.1	(40,0.002)	(40,0.002)	$\tau_d(0,0,3)$	0.026	0.232	0.054	0.939	-0.013	0.184	0.034	0.947	-0.019	0.175	0.031	0.945
				$\tau_d(0,1,3)$	-0.012	0.211	0.045	0.954	0.293	0.186	0.120	0.634	0.290	0.184	0.118	0.663
				$\tau_s(1,0,3)$	0.159	0.348	0.146	0.927	0.424	0.134	0.198	0.121	0.412	0.135	0.188	0.132
				$\tau_s(1,1,3)$	0.282	0.524	0.354	0.926	1.293	0.159	1.698	0.000	1.281	0.156	1.666	0.000

(b) $r_{deg} = 8$

δ^V	$\delta^{\tilde{V}}$	$(1-p^U, p^V)$ (%)	$(1-p^{\tilde{U}}, p^{\tilde{V}})$ (%)		SPE				Naive 1				Naive 2			
					bias	sd	mse	cr	bias	sd	mse	cr	bias	sd	mse	cr
0.1	0.05	(20, 0.002)	(20, 0.001)	$\tau_d(0,0,3)$	-0.052	0.202	0.044	0.956	-0.141	0.239	0.077	0.911	-0.129	0.244	0.076	0.922
				$\tau_d(0,1,3)$	-0.097	0.201	0.050	0.943	0.074	0.205	0.047	0.934	0.090	0.216	0.055	0.926
				$\tau_s(1,0,3)$	-0.040	0.419	0.177	0.955	0.641	0.178	0.443	0.058	0.697	0.178	0.517	0.035
				$\tau_s(1,1,3)$	-0.027	0.669	0.448	0.949	1.355	0.206	1.878	0.000	1.417	0.202	2.048	0.000
0.1	0.1	(20, 0.002)	(20, 0.002)	$\tau_d(0,0,3)$	-0.061	0.180	0.036	0.959	-0.147	0.225	0.072	0.910	-0.153	0.237	0.080	0.899
				$\tau_d(0,1,3)$	-0.105	0.186	0.046	0.934	0.066	0.194	0.042	0.936	0.069	0.206	0.047	0.934
				$\tau_s(1,0,3)$	-0.049	0.379	0.146	0.955	0.642	0.172	0.442	0.041	0.644	0.168	0.443	0.039
				$\tau_s(1,1,3)$	-0.026	0.644	0.415	0.949	1.358	0.196	1.883	0.000	1.363	0.194	1.896	0.000
0.1	0.05	(40, 0.002)	(40, 0.001)	$\tau_d(0,0,3)$	0.264	0.338	0.184	0.903	-0.077	0.347	0.126	0.941	-0.068	0.335	0.117	0.943
				$\tau_d(0,1,3)$	0.169	0.318	0.130	0.917	0.305	0.320	0.195	0.843	0.311	0.316	0.196	0.825
				$\tau_s(1,0,3)$	0.728	0.646	0.947	0.817	1.178	0.243	1.447	0.001	1.227	0.251	1.568	0.002
				$\tau_s(1,1,3)$	1.359	1.150	3.168	0.802	2.988	0.291	9.011	0.000	3.047	0.292	9.368	0.000
0.1	0.1	(40, 0.002)	(40, 0.002)	$\tau_d(0,0,3)$	0.254	0.349	0.187	0.909	-0.094	0.351	0.132	0.943	-0.108	0.360	0.141	0.932
				$\tau_d(0,1,3)$	0.165	0.330	0.136	0.916	0.286	0.322	0.185	0.852	0.284	0.325	0.187	0.868
				$\tau_s(1,0,3)$	0.737	0.636	0.948	0.831	1.176	0.246	1.443	0.004	1.184	0.250	1.465	0.002
				$\tau_s(1,1,3)$	1.446	1.134	3.377	0.787	2.987	0.292	9.009	0.000	2.992	0.295	9.038	0.000

4.7. Empirical Application: Diffusion of Insurance Information among Rice Farmers

This section applies the proposed SPE method to data on social network of rice farmers from 185 villages in rural China. The data was collected by Cai, De Janvry, and Sadoulet (2015b) to investigate the take-up decisions of a weather insurance, which typically has low rates of adoption even when the government provides heavy subsidies. The primary interest of Cai et al. (2015b) is to study whether and how the diffusion of weather insurance knowledge through social networks affects the insurance take-up rate.[9] Thus, two rounds of sessions are offered with a three-days gap to allow information sharing by the first round participants. In each round, there are two types of sessions held simultaneously: the 20-minute simple session where only the contract is discussed, and the 45-minute intensive session where the details of how the insurance operates and the expected benefits are explained. About 5000 rice-producing households from 185 villages are randomly assigned to one of the two information sessions aiming at generating household-level variation in insurance knowledge. The authors are particularly interested in the spillover effects: whether the second round participants' take-up decisions are affected by their friends who are invited to the first round intensive session. Hence, the baseline model for the treatment and spillover effects is:

$$Takeup_{ig} = \theta_0 + \theta_1 Intensive_{ig} + \theta_2 Network_{ig} + \theta_3 Cov_{ig} + \theta_4 NetSize_{ig} + \eta_g + \varepsilon_{ig}, \tag{4.15}$$

where $Takeup_{ig}$ is a binary indicator of whether household i in village g decides to buy the insurance, $Intensive_{ig}$ is a dummy variable taking value one if the household is invited to an intensive session, $Network_{ig}$ is the fraction of household i's friends who have been invited to the first round intensive session, $NetSize_{ig}$ is a set of dummies indicating network degree, Cov_{ig} includes household characteristics and η_g represents village fixed effect. If household i nominates zero friends, then $Network_{ig}$ is set to be zero. Household characteristics in Cov_{ig} include gender, age and education of household head, rice production area, risk aversion and perceived probability of future disasters.

[9] Data is available at Cai, De Janvry, and Sadoulet (2015a) https://doi.org/10.3886/E113593V1.

Dummies in $NetSize_{ig}$ are indicators of the number of nominated friends, where the dummy of zero nominated friends is dropped to avoid collinearity. Instead of the baseline model (4.15), we also consider an alternative model specification where the interaction term $Intensive_{ig} * Network_{ig}$ is included. In the same spirit of Cai et al. (2015b), because $Intensive_{ig}$ is whether household is invited to an intensive session, the treatment and spillover effects are studied from an intention-to-treat perspective. Nevertheless, almost 90% of households who are invited to one of the sessions actually attend. Therefore, the dropout is not a main concern.

Data from the social network survey is used to construct the household-level network measures. The social network survey requires the sampled household heads to nominate five friends with whom they discuss rice production or financial issues, while not all the respondents list up to five friends. No geographical restriction is imposed, which means the nominated friends can either live in the same village with the respondent or outside the village. This network measure is nonreciprocal and is referred to as the "general measure" in Cai et al. (2015b). The general measure may contain two types of measurement error: those with less than five friends are likely to report false friends (false positive) and those with more than five friends may censor the number of network links (false negative). Another household-level network measure used in Cai et al. (2015b), referred to as the "strong measure", is defined as the bilaterally linked friends (reciprocal) using the same information from the social network survey. The social network survey is conducted before the experiment, therefore the network formation should not be affected by the treatment assignments or the take-up decisions.

The analysis in this section utilises both these two measures, and assumes that the strong measure includes only false negative links. It is worth noting that although the two network measures are probably correlated even conditional on the true network information, according to the simulation results in Section 4.6.2, the SPE can be viewed as a bias-reduction method in the presence of network measurement error. Estimation is implemented as described in Section 4.5.[10] In this application, the dependence neigh-

[10]To mitigate estimation error arising from small sample size, the first step estimation uses samples from both the first and the second rounds and their network data based on the social network survey, with sample size 4588.

bourhoods can be set as villages, meaning that the DNs of a respondent i are those from the same village with i.

Two further remarks are worth noticing. First, as verified by Cai et al. (2015b), the second round participants should not be affected by the take-up decisions made by the first round participants if this information is not revealed to them (see Table 6 column 7 and Table 7 column 6 of Cai et al., 2015b). According to the survey, only 9% of the households who are not informed of any first round take-up information know at least one of their friends' decision. Thus, the endogenous peer effects (i.e. the spillovers of friends' take-up decisions) should not be a major concern in this application. Secondly, the first round simple session also exhibits no significant spillover effects on the second round participants (see Table 2 column 3 of Cai et al., 2015b).

Table 4.7.1: Effect of Social Networks on Insurance Take-up

	Naive		SPE	Naive		SPE
	General	Strong		General	Strong	
	(1)	(2)	(3)	(4)	(5)	(6)
Intensive	0.0298	0.0228	0.0265	0.0809**	0.0409	0.0556
	(0.0332)	(0.0334)	(0.0462)	(0.0397)	(0.0341)	(0.0735)
Network	0.291***	0.113*	0.196	0.444***	0.231***	0.244
	(0.0820)	(0.0606)	(0.2492)	(0.1089)	(0.0859)	(0.2472)
Intensive*Network				−0.329**	−0.221**	-0.106
				(0.161)	(0.111)	(0.189)
η_g	Yes	Yes	Yes	Yes	Yes	Yes
Cov_{ig}	Yes	Yes	Yes	Yes	Yes	Yes

Note: Samples are from the second round sessions "Simple2-NoInfo" and "Intensive2-NoInfo" as defined and used by Cai et al. (2015b). Number of observations is 1255. Standard error (se) is reported in the parenthesis. For the naive method, column "General" shows the result using the general measure of the network and column "Strong" displays the result using the strong measure of the network. The SPE method is implemented by assuming the network classification error is correlated to literacy. The se of the naive method is computed using clustered se with villages as clusters. The se of the SPE method is calculated based on Corollary 2 with villages as dependency neighbourhoods.

Estimation results are summarised in Table 4.7.1. The baseline model (columns (1) to (3)) and the alternative model with an interaction term of the treatment and network exposure (columns (4) to (6)) are estimated using the household-level samples from the second round sessions, where no overall attendance/take-up rate or individual insurance purchase resulting from the first round sessions in their village are revealed to the participants. The results for the naive method using the general measure of the network data

in columns (1) and (4) in Table 4.7.1 are the same to those in Table 2 columns (2) and (4) of Cai et al. (2015b), based on which they draw two conclusions. First, the spillover effect on insurance take-up is significantly positive. For example, column (1) (or column (2)) in Table 4.7.1 reveals that a 20% increase in the ratio of friends attending the first round intensive session will lead to a 29.1% × 20%=5.82% (or 11.3% × 20%=2.26%) increase in farmer's own take-up probability. Second, people are less likely to be affected by their friends if they attend the intensive session themselves. Column (4) (or column (5)) in Table 4.7.1 reveals that for farmers who have been directly educated about the insurance details, if the ratio of friends attending the first round intensive session increases by 20%, their own take-up probability will increase by (44.4%-32.9%) × 20%=2.3% (or (23.1%-22.1%) × 20%=0.2%); while this probability increases by 44.4% × 20%=8.88% (or 23.1% × 20%=4.62%) for farmers who have not attended the intensive session.

If the general measure and the strong measure possess network misclassification, then the estimates of the naive approach are biased. The SPE method can then be used to provide some guidance of the degree and direction of the potential bias. The SPE estimates in Table 4.7.1 are obtained by assuming that the measurement errors of the two network measures (both general and strong) are dependent on the household-head's literacy. By comparing the results in columns (1) and (2) to those in column (3), we can see that the SPE estimate of the spillover effect induced by a 20% increase in the ratio of treated friends is 19.6% × 20%=3.92%. Thus, the naive method using the general measure may overestimate the spillover effect, while the naive method using the strong measure is likely to underestimate the spillover effect. In addition, based on the SPE results in column (6), people who attend the intensive session themselves have a (24.4%-10.6%) × 20%=2.76% increase in their take-up probability when extra 20% of their friends are exposed to the intensive insurance-information education. While this change increases to 24.4% × 20%=4.88% for people who do not attend the intensive session. Hence, the comparison between columns (5) and (6) indicates that the results for the naive method using the general measure underestimate the spillover effect for the treated individuals, and overestimate the spillover effect for the untreated ones. In addition, the naive method using the strong measure dramatically underestimates the spillover effect for the treated individuals, but only slightly underestimates the spillover

effect for the controlled individuals.

4.8. Conclusion

Motivated by applications of program evaluation under network interference, this chapter studies the identification and estimation of treatment and spillover effects when the network is mismeasured. The novel identification strategy proposed in this chapter utilises two network proxies, where one of them is used as an instrumental variable for the latent network and the other is assumed to contain only one type of measurement error. A semiparametric estimation approach is proposed to estimate the causal effects of interest. Simulation results confirm that the proposed estimation approach (i) outperforms the naive estimators that neglect the network misclassification, and (ii) is preferred to the naive approaches, even if its key assumption is mildly violated, at least in terms of bias, mse and coverage rate. Therefore, the proposed estimation approach constitutes an effective method to reduce the bias caused by network measurement errors, and provides reliable causal inferences.

The proposed semiparametric estimation approach exploits a parametric structural assumption of the outcome variable to avoid the curse of dimensionality, which opens new questions on the trade-off between the potential model misspecification and the network mismeasurement-robust estimation. It is also meaningful and feasible to investigate the estimation in a more flexible semiparametric setup, including partially linear models, index models and random-coefficient models.

This chapter is particularly suitable for studies where the treatment is randomly assigned with perfect compliance. While, for some empirical studies, it is reasonable to allow for non-compliance (Vazquez-Bare, 2020). Future research will explore the impacts of relaxing the perfect compliance assumption, and develop methods for the identification and estimation of spillover effects that can accommodate the non-compliance.

Finally, this chapter assumes that the exposure to the treated peers that affects the outcome is correctly specified, meaning that the spillover effect is local through the first-order network neighbours. The literature on network effects often stresses the existence

of higher-order interference, i.e. the interference with friends of friends. The inclusion of higher-order interference complicates the analysis in this chapter by introducing higher-order spillovers of the treatment and of the measurement errors. It also further complicates the dependence structure among the observable and latent network-based variables. Given these complications, it is a nontrivial exercise to extend the analysis in this chapter to deal with higher-order interference. However, for studies where the treatment response is primarily governed by the first-order spillover, it is possible to apply the analysis of this chapter via assuming the higher-order spillover effects can be relegated to the unobservable error terms. The rationale is that, based on the studies of Leung (2019a) and Sävje (2019), the exposure misspecification that results from ignoring the higher-order interference does not alter the estimation results, if the specification errors are well counterbalanced by the deceasing data correlation as the order of the interference increases. Rigorous exploration along this direction is left for future research.

4.9. Appendix

We first introduce notations used in the Appendix. $\mathbf{I}_K$ is the $K \times K$ identity matrix. $\lambda_{max}(A)$ and $\lambda_{min}(A)$ denote the largest and the smallest eigenvalues of a square matrix A, respectively. We use C to represent some positive constant and its value may be different at different uses. *s.o.* denotes the terms of smaller order.

4.9.1. Examples

This section provides sufficient conditions or examples for the assumptions in the main text.

Example 1 (Assumption 4.3.3 (a)) *Suppose the network links follow the dyadic formation below:*

$$A^*_{ij} = 1[\omega(Z_i, Z_j) > \eta_{ij}] \cdot 1[i \neq j], \text{ with } i, j \in \mathcal{P}$$

where $\omega : \Omega^2_Z \mapsto \mathbb{R}$ *and the unobserved link specific error term* η_{ij} *is independent to* $\{Z_i\}_{i\in\mathcal{P}}$ *and is i.i.d. across* (i, j)*. Then,* A^*_{ij} *given* Z_i *is a function of* (Z_j, η_{ij})*, which is i.i.d. across* j *and* $\mathcal{F}^*_i = \sum_{j\in\mathcal{P}} A^*_{ij}$ *is identically distributed following the binomial distribution. Such a network formation is considered in for example Johnsson and Moon (2015).*

Example 2 (Assumption 4.3.4 (c)) *For any given latent* A^*_{ij}*, consider the following data generating process of the observable* A_{ij}

$$A_{ij} = U_{ij}A^*_{ij} + V_{ij}(1 - A^*_{ij}), \quad \text{with } i, j \in \mathcal{P} \tag{4.9.16}$$

where $\mathcal{N}_i = \{j \in \mathcal{P} : A_{ij} = 1\}$ *and the classification errors* (U_{ij}, V_{ij}) *are random indicators taking values from* $\{0,1\}$*. From (4.9.16), we can obtain that*

$$\mathcal{F}_i = \sum_{j\in\mathcal{P}} A_{ij} = \sum_{j\in\mathcal{P}} (U_{ij} - V_{ij})A^*_{ij} + \sum_{j\in\mathcal{P}} V_{ij} = \sum_{j\in\mathcal{N}^*_i} (U_{ij} - V_{ij}) + \sum_{j\in\mathcal{P}} V_{ij}.$$

Let two vectors $\mathbf{U}_i = \{U_{ij}\}_{j\in\mathcal{P}}$ *and* $\mathbf{V}_i = \{V_{ij}\}_{j\in\mathcal{P}}$*. If the random vector* $(\mathbf{U}_i, \mathbf{V}_i)$ *is conditionally independent to* $\mathcal{N}^*_i$ *and identically distributed across* $i \in \mathcal{P}$ *given* $(Z_i, \mathcal{F}^*_i)$*,*

then the identical distribution of $\mathcal{F}_i$ given $(Z_i, \mathcal{F}_i^*)$ holds.

Example 3 (Assumption 4.4.1) *For each $i \in \mathcal{P}$ and any given latent A_{ij}^*, suppose the observable links are generated as*

$$\begin{aligned} A_{ij} &= \omega_j \big[U_{ij} A_{ij}^* + V_{ij}(1 - A_{ij}^*)\big], \\ \tilde{A}_{ij} &= \tilde{\omega}_j \big[\tilde{U}_{ij} A_{ij}^* + \tilde{V}_{ij}(1 - A_{ij}^*)\big], \end{aligned} \quad \text{with } j \in \mathcal{P} \tag{4.9.17}$$

with U_{ij}, V_{ij}, $\tilde{U}_{ij}$, $\tilde{V}_{ij}$, ω_j and $\tilde{\omega}_j$ are all binary random variables taking values from $\{0,1\}$. ω_j and $\tilde{\omega}_j$ can be understood as indicators of sampling-induced errors, e.g. $\omega_j = 0$ means unit j is not sampled when constructing $\mathcal{N}_i$, while only links among pairs of sampled units are accounted for. (U_{ij}, V_{ij}) and $(\tilde{U}_{ij}, \tilde{V}_{ij})$ can be understood as indicators of non-sampling-induced errors, e.g. $\tilde{U}_{ij} = 0$ represents unit i's misreporting of her link with unit j when constructing $\tilde{\mathcal{N}}_i$. Then, the observed sets of links are $\mathcal{N}_i = \{j \in \mathcal{P} : \ A_{ij} = 1\}$ and $\tilde{\mathcal{N}}_i = \{j \in \mathcal{P} : \ \tilde{A}_{ij} = 1\}$. Therefore,

$$\begin{aligned} \mathcal{F}_i &= \sum_{j\in\mathcal{P}} A_{ij} = \sum_{j\in\mathcal{P}} (U_{ij} - V_{ij})\omega_j A_{ij}^* + \sum_{j\in\mathcal{P}} \omega_j V_{ij} = \sum_{j\in\mathcal{N}_i^*} (U_{ij} - V_{ij})\omega_j + \sum_{j\in\mathcal{P}} V_{ij}\omega_j, \\ \tilde{\mathcal{F}}_i &= \sum_{j\in\mathcal{P}} \tilde{A}_{ij} = \sum_{j\in\mathcal{P}} (\tilde{U}_{ij} - \tilde{V}_{ij})\tilde{\omega}_j A_{ij}^* + \sum_{j\in\mathcal{P}} \tilde{\omega}_j \tilde{V}_{ij} = \sum_{j\in\mathcal{N}_i^*} (\tilde{U}_{ij} - \tilde{V}_{ij})\tilde{\omega}_j + \sum_{j\in\mathcal{P}} \tilde{V}_{ij}\tilde{\omega}_j. \end{aligned} \tag{4.9.18}$$

Then, one set of sufficient conditions for Assumption 4.4.1 is provided by the lemma below.

Lemma 4.9.1 *Let Assumption 4.3.4 (a) holds for both $\mathcal{N}_i$ and $\tilde{\mathcal{N}}_i$. Suppose the random vector $(U_{ij}, V_{ij}, \tilde{U}_{ij}, \tilde{V}_{ij}, \omega_j, \tilde{\omega}_j)$ given $(Z_i, \mathcal{F}_i^*)$ is i.i.d. across j for all $i \in \mathcal{P}$. If*

(a) $\{U_{ij}, V_{ij}, \omega_j\}_{j\in\mathcal{P}} \perp \{\tilde{U}_{ik}, \tilde{V}_{ik}, \tilde{\omega}_k\}_{k\in\mathcal{P}} \big| Z_i, \mathcal{N}_i^*$;

(b) $\{U_{ij}, V_{ij}, \tilde{U}_{ij}, \tilde{V}_{ij}, \omega_j, \tilde{\omega}_j\}_{j\in\mathcal{P}} \perp \mathcal{N}_i^* \big| Z_i, \mathcal{F}_i^*$;

then Assumption 4.4.1 is satisfied by $\mathcal{F}_i$ and $\tilde{\mathcal{F}}_i$ given in (4.9.18).

Proof of Lemma 4.9.1. (i) From condition (a) that $\{U_{ij}, V_{ij}, \omega_j\}_{j\in\mathcal{P}} \perp \{\tilde{U}_{ik}, \tilde{V}_{ik}, \tilde{\omega}_k\}_{k\in\mathcal{P}} \big| Z_i, \mathcal{N}_i^*$, we have $\mathcal{N}_i \perp \tilde{\mathcal{N}}_i \big| Z_i, \mathcal{N}_i^*$, which implies $\mathcal{F}_i \perp \tilde{\mathcal{F}}_i \big| Z_i, \mathcal{N}_i^*$. If we can further show that

$\mathcal{F}_i \perp \mathcal{N}_i^* \big| Z_i, \mathcal{F}_i^*$ and $\tilde{\mathcal{F}}_i \perp \mathcal{N}_i^* \big| Z_i, \mathcal{F}_i^*$ hold, then the desired result follows, because

$$\begin{aligned} f_{\mathcal{F}_i|\mathcal{F}_i^*,\tilde{\mathcal{F}}_i,Z_i}(n) &= \sum_{\mathcal{J}\in\Omega_{\mathcal{N}^*}} f_{\mathcal{F}_i|\mathcal{F}_i^*,\tilde{\mathcal{F}}_i,\mathcal{N}_i^*=\mathcal{J},Z_i}(n) \times f_{\mathcal{N}_i^*|\mathcal{F}_i^*,\tilde{\mathcal{F}}_i,Z_i}(\mathcal{J}) \\ &= \sum_{\mathcal{J}\in\Omega_{\mathcal{N}^*}} f_{\mathcal{F}_i|\mathcal{F}_i^*,\mathcal{N}_i^*=\mathcal{J},Z_i}(n) \times f_{\mathcal{N}_i^*|\mathcal{F}_i^*,Z_i}(\mathcal{J}) \\ &= f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}(n), \end{aligned}$$

which indicates $\mathcal{F}_i \perp \tilde{\mathcal{F}}_i \big| Z_i, \mathcal{F}_i^*$.

Given the expressions in (4.9.18), based on the i.i.d. of $(U_{ij}, V_{ij}, \omega_j)$ across j, applying the same arguments used to prove Lemma 4.4.1 (a), we can show that given $(Z_i, \mathcal{F}_i^*)$, the distribution of $\sum_{j\in\mathcal{N}_i^*}(U_{ij} - V_{ij})\omega_j$ does not depend on $\mathcal{N}_i^*$, i.e. $\sum_{j\in\mathcal{N}_i^*}(U_{ij} - V_{ij})\omega_j \perp \mathcal{N}_i^* \big| Z_i, \mathcal{F}_i^*$. In addition, from condition (b) we can obtain the independence of $\sum_{j\in\mathcal{P}} V_{ij}\omega_j$ to $\mathcal{N}_i^*$ given $Z_i, \mathcal{F}_i^*$. Thus, it follows from the above results and (4.9.18) that $\mathcal{F}_i \perp \mathcal{N}_i^* \big| Z_i, \mathcal{F}_i^*$. Similarly, $\tilde{\mathcal{F}}_i \perp \mathcal{N}_i^* \big| Z_i, \mathcal{F}_i^*$ also holds. ■

4.9.2. Lemmas

This section introduces some useful lemmas which are used in the proofs of Appendix Section 4.9.3.

Lemma 4.9.2 *Denote $\mathcal{H}$ as a set of measurable functions such that $|h| \leq 1$ for $\forall h \in \mathcal{H}$, and denote $sign(x) = 1[x \geq 0] - 1[x < 0]$ for any real value x. For any random variables X and Z, a solution to $\max_{h\in\mathcal{H}} |\mathbb{E}[Xh(Z)]|$ is $h(Z) = sign(\mathbb{E}[X|Z])$, and $\max_{h\in\mathcal{H}} |\mathbb{E}[Xh(Z)]| = \mathbb{E}[X sign(X|Z)]$.*

Proof of Lemma 4.9.2. By the law of iterated expectation

$$|\mathbb{E}[Xh(Z)]| = \left| \int \mathbb{E}[X|Z]h(Z) d\Pr(Z) \right| \leq \int |\mathbb{E}[X|Z]h(Z)| \, d\Pr(Z) \leq \int |\mathbb{E}[X|Z]| \, d\Pr(Z).$$

Then, by $|\mathbb{E}[X|Z]| = \mathbb{E}[X|Z] sign(\mathbb{E}[X|Z])$, it is clear that $h(Z) = sign(\mathbb{E}[X|Z])$. ■

Lemma 4.9.3 (Uniform Law of Large Number under Dependency Neighbourhood) *For any function $b : \Omega_W \times \Theta \mapsto \mathbb{R}^p$, if the following conditions hold*

(i) Θ *is compact;*

(ii) $b(w;\theta)$ *is continuous in* θ *over* Θ*;*

(iii) there exists a function $h(w)$ *with* $\|b(w;\theta)\| \leq h(w)$ *for all* $\theta \in \Theta$ *and* $\mathbb{E}[h(W_i)] < \infty$*;*

(iv) for some constant $\eta > 0$*, define*

$$u(w;\theta,\eta) = \sup_{\theta' \in \Theta,\ \|\theta'-\theta\|<\eta} \|b(w;\theta') - b(w;\theta)\|,$$
$$\Sigma_N^b(\theta) = \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} Cov(b(W_i;\theta), b(W_j;\theta)),$$
$$\Sigma_N^u(\theta,\eta) = \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} Cov(u(W_i;\theta,\eta), u(W_j;\theta,\eta)).$$

(a) for all $\theta \in \Theta$ *and any fixed* η*,*

$$\left\| \sum_{i=1}^{N} \sum_{j \notin \Delta(i,N)} Cov(b(W_i;\theta), b(W_j;\theta)) \right\| = o\left(\|\Sigma_N^b(\theta)\|\right),$$
$$\sum_{i=1}^{N} \sum_{j \notin \Delta(i,N)} Cov(u(W_i;\theta,\eta), u(W_j;\theta,\eta)) = o\Big(\Sigma_N^u(\theta,\eta)\Big).$$

(b) $1/N \sum_{i=1}^{N} |\Delta(i,N)| = O(1)$*; (c)* $\sup_{\theta\in\Theta} \mathbb{E}[\|b(W_i;\theta)\|^{2+\delta}] < C$ *for some constants* $\delta > 0$ *and* $C > 0$*, and all* i*;*

then $\mathbb{E}[b(W_i;\theta)]$ *is continuous in* θ *and* $\sup_{\theta\in\Theta} \left\| \frac{1}{N} \sum_{i=1}^{N} \{b(W_i;\theta) - \mathbb{E}[b(W_i;\theta)]\} \right\| \xrightarrow{p} 0$.

Proof of Lemma 4.9.3. This proof is based on the proof of Lemma 1 in Tauchen (1985). Let $b_r(W_i;\theta)$ be the r-th element in vector $b(W_i;\theta)$, $r = 1, 2, ..., p$. Define a matrix $\Lambda_{ij}(\theta)$ such that its rq-th entry is $corr(b_r(W_i;\theta), b_q(W_j;\theta))$, $r, q = 1, 2, ..., p$. Denote a diagonal matrix $V_i(\theta) = diag(\text{Var}[b_1(W_i;\theta)], ..., \text{Var}[b_p(W_i;\theta)])$.

By condition (iv) (c), for all i and given η, there exist constants $C_1, C_2 > 0$ such that

$\sup_{\theta\in\Theta}\text{Var}[b_r(W_i;\theta)] < C_1$ for all $r = 1, ..., p$, and $\sup_{\theta\in\Theta}\text{Var}[u(W_i;\theta,\eta)] < C_2$. Then,

$$\begin{aligned}\left\|\Sigma_N^b(\theta)\right\| \leq & \sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\left\|\text{Cov}(b(W_i;\theta), b(W_j;\theta))\right\| \\ \leq & \sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\left\|V_i(\theta)^{1/2}\Lambda_{ij}(\theta)V_j(\theta)^{1/2}\right\| \\ \leq & C_1 p\sum_{i=1}^{N}|\Delta(i,N)| = O(N),\end{aligned}$$

where the last line follows from $1/N\sum_{i=1}^{N}|\Delta(i,N)| = O(1)$ in condition (iv) (b). Similarly, $\Sigma_N^u(\theta,\eta) = O(N)$. Applying Chebyshev's inequality, we have that for any $\epsilon > 0$

$$\begin{aligned}&\Pr\left(\left\|\frac{1}{N}\sum_{i=1}^{N}\{b(W_i;\theta) - \mathbb{E}[b(W_i;\theta)]\}\right\| > \epsilon\right) \\ \leq & \frac{1}{\varepsilon^2 N^2}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\{b(W_i;\theta) - \mathbb{E}[b(W_i;\theta)]\}\right\|^2\right] \\ = & \frac{1}{\varepsilon^2 N^2} tr\left(\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\text{Cov}\left(b(W_i;\theta), b(W_j;\theta)\right) + \sum_{i=1}^{N}\sum_{j\notin\Delta(i,N)}\text{Cov}\left(b(W_i;\theta), b(W_j;\theta)\right)\right) \\ = & \frac{p}{\varepsilon^2 N^2}\left(\left\|\Sigma_N^b(\theta)\right\| + s.o.\right) \\ = & O\left(\frac{1}{\varepsilon^2 N}\right),\end{aligned}$$

where the second equality comes from that $tr(A) \leq p\|A\|_\infty \leq p\|A\|$ for any $p \times p$ square matrix A, and the third equality is due to condition (iv) (a). By choosing ϵ such that $\epsilon \to 0$ and $\epsilon^2 N \to \infty$ as $N \to \infty$, we can get

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\{b(W_i;\theta) - \mathbb{E}[b(W_i;\theta)]\}\right\| = o_p(1).$$

Similar arguments can be used to show that $\frac{1}{N}\sum_{i=1}^{N}\{u(W_i;\theta,\eta) - \mathbb{E}[u(W_i;\theta,\eta)]\} = o_p(1)$. By condition (ii) the continuity of $b(w;\theta)$ in θ, we have that with fixed θ, $\lim u(w;\theta,\eta) = 0$ as $\eta \to 0$. Thus, by dominated convergence theorem, for any $\epsilon > 0$, there exists a $\bar{\eta}(\theta)$

such that

$$\mathbb{E}[u(W_i;\theta,\eta)] \leq \epsilon, \text{ whenever } \eta \leq \bar{\eta}(\theta). \tag{4.9.19}$$

Let $B(\theta)$ be an open ball of radius $\bar{\eta}(\theta)$ about θ. Due to the compactness of Θ, there exist a finite sequence of open balls $B_k := B(\theta_k)$ with $k = 1, 2, ..., K$ such that $\Theta \subset \bigcup_{k=1}^{K} B_k$. Let $\eta_k = \bar{\eta}(\theta_k)$ and $u_k = \mathbb{E}[u(W_i;\theta_k,\eta_k]$. By (4.9.19) and dominated convergence theorem, if $\theta \in B_k$ then $u_k \leq \epsilon$ and $\|\mathbb{E}[b(W_i;\theta)] - \mathbb{E}[b(W_i;\theta')]\| \leq \epsilon$. Next, for $\forall\theta \in \Theta$, there exists a k such that $\theta \in B_k$, then

$$\begin{aligned}
&\left\| \frac{1}{N}\sum_{i=1}^{N} b(W_i;\theta) - \mathbb{E}[b(W_i;\theta)] \right\| \\
\leq &\frac{1}{N}\sum_{i=1}^{N} \|b(W_i;\theta) - b(W_i;\theta_k)\| + \left\| \frac{1}{N}\sum_{i=1}^{N} b(W_i;\theta_k) - \mathbb{E}[b(W_i;\theta_k)] \right\| \\
&+ \|\mathbb{E}\left[b(W_i;\theta_k)\right] - \mathbb{E}[b(W_i;\theta)]\| \\
\leq &\frac{1}{N}\sum_{i=1}^{N} u(W_i;\theta_k,\eta_k) - u_k + u_k + \left\| \frac{1}{N}\sum_{i=1}^{N} b(W_i;\theta_k) - \mathbb{E}[b(W_i;\theta_k)] \right\| + \epsilon \\
\leq &4\epsilon
\end{aligned} \tag{4.9.20}$$

whenever $N \geq \bar{N}_k(\epsilon)$, by $u_k \leq \epsilon$. Thus, whenever $N \geq \max_k \bar{N}_k(\epsilon)$, we have that

$$\sup_{\theta\in\Theta} \left\| \frac{1}{N}\sum_{i=1}^{N} \{b(W_i;\theta) - \mathbb{E}[b(W_i;\theta)]\} \right\| \leq 4\epsilon.$$

■

Lemma 4.9.4 (Theorem 3 of Bradley et al. (1983)) *Suppose X and Y are random variables taking their values on a Borel space Γ and $\mathbb{R}$, respectively. Suppose U is a uniform $[0,1]$ random variable independent of (X,Y). Suppose μ and γ are positive numbers such that $\mu \leq \|Y\|_\gamma < \infty$. Let $\|Y\|_\gamma = (\mathbb{E}[|Y|^\gamma])^{1/\gamma}$. Then there exists a real-valued random variable $Y^* = g(X,Y,U)$ where g is a measurable function from $\Gamma \times \mathbb{R} \times [0,1]$ into $\mathbb{R}$, such that*

(i) Y^* is independent of X;

(ii) the probability distributions of Y^* *and* Y *are identical;*

(iii) $Pr(|Y^* - Y| \geq \mu) \leq 18(\|Y\|_\gamma/\mu)^{\gamma/(2\gamma+1)}[\alpha(\mathfrak{B}(X), \mathfrak{B}(Y))]^{2\gamma/(2\gamma+1)}$,

where for any two σ*-fields* $\mathfrak{B}_1, \mathfrak{B}_2$, $\alpha(\mathfrak{B}_1, \mathfrak{B}_2) = \sup |Pr(\mathfrak{B}_1 \bigcap \mathfrak{B}_2) - Pr(\mathfrak{B}_1)Pr(\mathfrak{B}_2)|$.

The following lemmas are pioneered by Stein (1986) and utilised in for example Chen et al. (2010), Ross (2011) and Goldstein and Rinott (1996) among others, to derive central limit theorems for dependency graphs. We re-state them here such that the proofs are self-contained.

Lemma 4.9.5 (Meckes et al. (2009) Lemma 1) *Let* $Z \in \mathbb{R}^p$ *be a standard normal random vector with mean zero and covariance matrix* $\mathbf{I}_d$.

(i) If a function $f : \mathbb{R}^p \mapsto \mathbb{R}$ *is twice continuously differentiable with compact support, then*

$$\mathbb{E}\left[tr\left(\frac{d^2 f(Z)}{dzdz'}\right) - Z'\frac{df(Z)}{dz}\right] = 0.$$

(ii) If a random vector $X \in \mathbb{R}^p$ *is such that*

$$\mathbb{E}\left[tr\left(\frac{d^2 f(X)}{dxdx'}\right) - X'\frac{df(X)}{dx}\right] = 0$$

for every $f \in C^2(\mathbb{R}^p)$ *that is twice continuously differentiable with finite absolute mean value* $\mathbb{E}\left[|tr\left(d^2 f(X)/dxdx'\right) - X'df(X)/dx|\right] < \infty$, *then* X *is a standard normal random vector.*

Lemma 4.9.6 (Goldstein and Rinott (1996) Lemma 3.1) *Let* $Z \in \mathbb{R}^p$ *be a standard norm random vector and let* $h : \mathbb{R}^p \mapsto \mathbb{R}$ *have three bounded derivatives. Define* $(T_u h)(x) = \mathbb{E}[h(xe^{-u} + \sqrt{1 - e^{-2u}}Z)]$ *for* $x \in \mathbb{R}^p$. *Then* $f(x) = -\int_0^\infty [T_u h(x) - \mathbb{E}[h(Z)]]du$ *solves*

$$tr\left(\frac{d^2 f(x)}{dxdx'}\right) - x'\frac{df(x)}{dx} = h(x) - \mathbb{E}[h(Z)].$$

In addition, for any k-th partial derivative we have that

$$\left|\frac{\partial^k f(x)}{\prod_{j=1}^k \partial x_j}\right| \leq \frac{1}{k} \sup_{x\in\Omega_X} \left\|\frac{d^2 h(x)}{dxdx'}\right\|_\infty .$$

Further, for any $\lambda \in \mathbb{R}^p$ and positive definite $p \times p$ matrix Σ, then f^*, denoted by the change of variable $f^*(x) := f(\Sigma^{-1/2}(x-\lambda))$ solves

$$tr\left(\Sigma\nabla^2 f^*(x)\right) - (x-\lambda)'\nabla f^*(x) = h(\Sigma^{-1/2}(x-\lambda)) - \mathbb{E}[h(Z)],$$

and

$$\left|\frac{\partial^k f^*(x)}{\prod_{j=1}^k \partial x_j}\right| \leq \frac{p^k}{k}\|\Sigma^{-1/2}\|_\infty^k \|\nabla^k h\|_\infty.$$

The lemma below is based on Theorem 1.4 of Goldstein and Rinott (1996) which aims at providing a bound on the distance to normality for any sum of dependent random vectors whose dependence structure is formed via dependency neighbourhoods.

Lemma 4.9.7 (Multivariate CLT under Dependency Neighbourhood) *Let $\{W_i\}_{i=1}^N$ be random vectors in $\mathbb{R}^p$ with $\mathbb{E}[W_i] = 0$ and $Z \in \mathbb{R}^p$ be a standard normal random vector. Denote*

$$\mathbf{S}_N = \sum_{i=1}^N W_i \quad \text{and} \quad \Sigma_N = \sum_{i=1}^N \sum_{j\in\Delta(i,N)} \mathbb{E}[W_i W_j'].$$

In addition, denote $\mathbf{S}_i^c = \sum_{j\notin\Delta(i,N)} W_j$. Assume Σ_N is symmetric positive definite. If the following conditions hold,

(i) there exists a finite, strictly positive-definite and symmetric $p \times p$ matrix Ω such that $\|\frac{1}{N}\Sigma_N - \Omega\| \to 0$ as $N \to \infty$;

(ii) *(a)* $$\left\|\sum_{i=1}^N \sum_{j,k\in\Delta(i,N)} \mathbb{E}\left[\left|vec(W_i W_j')W_k'\right|\right]\right\|_\infty = o\left(\left\|\Sigma_N^{3/2}\right\|_\infty\right);$$

(b) $$\left\|\sum_{i,k=1}^N \sum_{j\in\Delta(i,N)} \sum_{l\in\Delta(k,N)} \mathbb{E}\Big[\left(W_i W_j' - \mathbb{E}[W_i W_j']\right)'\left(W_k W_l' - \mathbb{E}[W_k W_l']\right)\Big]\right\|_\infty = o\left(\|\Sigma_N^2\|_\infty\right);$$

(c) $$\left\|\sum_{i=1}^N \sum_{j\notin\Delta(i,N)} Cov(W_i, W_j)\right\|_\infty = o\left(\|\Sigma_N\|_\infty\right);$$

(d) $\mathbb{E}\left[W_i\mathbf{S}_i^c \middle| \mathbf{S}_i^c\right] \geq 0$ *for all* $i \in \mathcal{P}$*;*

then $\Sigma_N^{-1/2}\mathbf{S}_N \xrightarrow{d} \mathbb{N}(0, \mathbf{I}_p)$.

Proof of Lemma 4.9.7. Denote $\mathbf{S}_{i,q}^c$ as the q-th element of $\mathbf{S}_i^c$. Let $h : \mathbb{R}^p \mapsto \mathbb{R}$ be a function with bounded mixed partial derivatives up to order three. Denote $\nabla^k h$ the k-th derivative of h. Let $\nabla_r f^*(x) = \partial f^*(x)/\partial x_r$ and $\nabla_{rq}^2 f^*(x) = \partial^2 f^*(x)/\partial x_r \partial x_q$. It follows directly from the proof of Theorem 1.4 in Goldstein and Rinott (1996) that

$$\begin{aligned}
&\left|\mathbb{E}\left[h\left(\Sigma_N^{-1/2}\mathbf{S}_N\right) - \mathbb{E}[h(Z)]\right]\right| \\
\leq &\frac{p^2}{2}\left\|\Sigma_N^{-1/2}\right\|_\infty^2 \|\nabla^2 h\|_\infty \sum_{r,q=1}^{p} \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\left(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}]\right)\right)^2\right]} \\
&+ \left|\sum_{r=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[W_{i,r}\nabla_r f^*(\mathbf{S}_i^c)\right]\right| \\
&+ \frac{p^3}{6}\left\|\Sigma_N^{-1/2}\right\|_\infty^3 \|\nabla^3 h\|_\infty \sum_{r,q,u=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[\left|W_{i,r}\sum_{j\in\Delta(i,N)}W_{j,q}\sum_{k\in\Delta(i,N)}W_{k,u}\right|\right], \qquad (4.9.21)
\end{aligned}$$

where f^* is defined as in Lemma 4.9.6. Consider the second term on the right hand side of (4.9.21)

$$\begin{aligned}
&\left|\sum_{r=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[W_{i,r}\nabla_r f^*(\mathbf{S}_i^c)\right]\right| \\
\leq &\left|\sum_{r=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left\{W_{i,r}\left[\nabla_r f^*(\mathbf{S}_i^c) - \nabla_r f^*(0)\right]\right\}\right| + \left|\sum_{r=1}^{p}\nabla_r f^*(0)\sum_{i=1}^{N}\mathbb{E}\left[W_{i,r}\right]\right| \\
= &\left|\sum_{r,q=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[W_{i,r}\mathbf{S}_{i,q}^c\nabla_{rq}^2 f^*(\widetilde{\mathbf{S}}_i^c)\right]\right|, \qquad (4.9.22)
\end{aligned}$$

where $\widetilde{\mathbf{S}}_i^c$ is between $\mathbf{S}_i^c$ and 0 and the last equality comes from the mean value theorem and the fact that $\mathbb{E}[W_{i,r}] = 0$. Without loss of generality, suppose there exists a function $\tilde{f}$ such that $\widetilde{\mathbf{S}}_i^c = \tilde{f}(\mathbf{S}_i^c)$. Then, we can further bound (4.9.22) as below:

$$\left|\sum_{r,q=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[W_{i,r}\mathbf{S}_{i,q}^c\nabla_{rq}^2 f^*(\widetilde{\mathbf{S}}_i^c)\right]\right| = \left|\sum_{r,q=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[W_{i,r}\mathbf{S}_{i,q}^c\nabla_{rq}^2 f^*\left(\tilde{f}(\mathbf{S}_i^c)\right)\right]\right|$$

$$\leq \frac{p^2}{2}\|\Sigma_N^{-1/2}\|_\infty^2\|\nabla^2 h\|_\infty \sum_{r,q=1}^{p}\sum_{i=1}^{N}\mathbb{E}\Big\{W_{i,r}\mathbf{S}_{i,q}^c sign\left(\mathbb{E}\left[W_{i,r}\middle|\mathbf{S}_i^c\right]\mathbf{S}_{i,q}^c\right)\Big\}, \tag{4.9.23}$$

where the inequality is because of Lemma 4.9.2 and $|\nabla_{rq}^2 f^* \circ \tilde{f}| \leq \frac{p^2}{2}\|\Sigma_N^{-1/2}\|_\infty^2\|\nabla^2 h\|_\infty$ by Lemma 4.9.6. Therefore, we have that

$$\begin{aligned}
&\left|\mathbb{E}\left[h\left(\Sigma_N^{-1/2}\mathbf{S}_N\right) - \mathbb{E}[h(Z)]\right]\right| \\
\leq &\frac{p^2}{2}\left\|\Sigma_N^{-1/2}\right\|_\infty^2\|\nabla^2 h\|_\infty \sum_{r,q=1}^{p}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}])\right)^2\right]} \\
&+ \frac{p^2}{2}\left\|\Sigma_N^{-1/2}\right\|_\infty^2\|\nabla^2 h\|_\infty \sum_{r,q=1}^{p}\sum_{i=1}^{N}\mathbb{E}\Big\{W_{i,r}\mathbf{S}_{i,q}^c sign\left(\mathbb{E}\left[W_{i,r}\middle|\mathbf{S}_i^c\right]\mathbf{S}_{i,q}^c\right)\Big\} \\
&+ \frac{p^3}{6}\left\|\Sigma_N^{-1/2}\right\|_\infty^3\|\nabla^3 h\|_\infty \sum_{r,q,u=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[\left|W_{i,r}\sum_{j\in\Delta(i,N)}W_{j,q}\sum_{k\in\Delta(i,N)}W_{k,u}\right|\right],
\end{aligned} \tag{4.9.24}$$

for some constant $p > 0$. Let us start from the first term. By the Cauchy–Schwarz inequality

$$\begin{aligned}
&\sum_{r,q=1}^{p}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}])\right)^2\right]} \\
\leq &\left(\sum_{r,q=1}^{p}\mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}])\right)^2\right]\right)^{1/2}\left(\sum_{r,q=1}^{p}1\right)^{1/2} \\
= &p\mathbb{E}\left[\left\|\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}])\right\|^2\right]^{1/2},
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}])\right\|^2\right] \\
= &\mathbb{E}\left[tr\left(\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}])'\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}])\right)\right] \\
= &tr\left(\sum_{i,k=1}^{N}\sum_{j\in\Delta(i,N)}\sum_{l\in\Delta(k,N)}\mathbb{E}\Big[(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}])'(W_{k,r}W_{l,q} - \mathbb{E}[W_{k,r}W_{l,q}])\Big]\right)
\end{aligned}$$

$$\leq p \left\| \sum_{i,k=1}^{N} \sum_{j \in \Delta(i,N)} \sum_{l \in \Delta(k,N)} \mathbb{E}\Big[\left(W_{i,r}W_{j,q} - \mathbb{E}[W_{i,r}W_{j,q}]\right)' \left(W_{k,r}W_{l,q} - \mathbb{E}[W_{k,r}W_{l,q}]\right) \Big] \right\|_{\infty}. \tag{4.9.25}$$

Besides, since $\mathbb{E}\left[W_{i,r}\middle|\mathbf{S}_i^c\right]\mathbf{S}_i^c \geq 0$ for all $i = 1, ..., N$ and $r = 1, ..., p$, the second term becomes to

$$\begin{aligned} \sum_{r,q=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[W_{i,r}\mathbf{S}_{i,q}^c\right] &= \sum_{r,q=1}^{p}\sum_{i=1}^{N}\sum_{j \notin \Delta(i,N)} \mathrm{Cov}\left(W_{i,r}, W_{j,q}\right) \\ &\leq p^2 \left\| \sum_{i=1}^{N}\sum_{j \notin \Delta(i,N)} \mathrm{Cov}\left(W_i, W_j\right) \right\|_{\infty}. \end{aligned} \tag{4.9.26}$$

For the last term, we can obtain

$$\begin{aligned} \sum_{r,q,u=1}^{p}\sum_{i=1}^{N}\mathbb{E}\left[\left|W_{i,r}\sum_{j \in \Delta(i,N)} W_{j,q}\sum_{k \in \Delta(i,N)} W_{k,u}\right|\right] &\leq \sum_{r,q,u=1}^{p}\sum_{i=1}^{N}\sum_{j,k \in \Delta(i,N)} \mathbb{E}\left[\left|W_{i,r}W_{j,q}W_{k,u}\right|\right] \\ &\leq p^3 \left\| \sum_{i=1}^{N}\sum_{j,k \in \Delta(i,N)} \mathbb{E}\left[\left|vec(W_iW_j')W_k'\right|\right] \right\|_{\infty}. \end{aligned} \tag{4.9.27}$$

Moreover, since $\|N^{-1}\Sigma_N - \Omega\| \to 0$, implying that there exist $\underline{\epsilon}, \bar{\epsilon}$ such that

$$0 < \underline{\epsilon} \leq \frac{1}{N}\lambda_{min}(\Sigma_N) \leq \frac{1}{N}\lambda_{max}(\Sigma_N) < \bar{\epsilon} < \infty.$$

In addition, by the property of norm and the symmetry of Σ_N, we have that

$$\left\|\Sigma_N^{-1/2}\right\|_{\infty}^2 \leq \left\|\Sigma_N^{-1/2}\right\|^2 = tr\left(\Sigma_N^{-1}\right) = \sum_{r=1}^{p}\lambda_r^{-1}(\Sigma_N) \leq p\lambda_{min}^{-1}(\Sigma_N) = O(N^{-1}),$$

where $\lambda_r(\Sigma_N)$ means the r-th largest eigenvalue of matrix Σ_N. Similarly,

$$\|\Sigma_N\|_{\infty}^2 = O(N^2), \quad \left\|\Sigma_N^{3/2}\right\|_{\infty}^2 = O(N^3), \quad \left\|\Sigma_N^2\right\|_{\infty}^2 = O(N^4). \tag{4.9.28}$$

Now, plugging (4.9.25), (4.9.26) and (4.9.27) into (4.9.24) gives us

$$\left|\mathbb{E}\left[h\left(\Sigma_N^{-1/2}\mathbf{S}_N\right) - \mathbb{E}[h(Z)]\right]\right|$$

$$
\begin{aligned}
\leq & C\left\|\Sigma_N^{-1/2}\right\|_\infty^2 \left\| \sum_{i,k=1}^N \sum_{j\in\Delta(i,N)} \sum_{l\in\Delta(k,N)} \mathbb{E}\Big[\left(W_iW_j' - \mathbb{E}[W_iW_j']\right)' \left(W_kW_l' - \mathbb{E}[W_kW_l']\right) \Big] \right\|_\infty^{1/2} \\
& + C\left\|\Sigma_N^{-1/2}\right\|_\infty^2 \left\| \sum_{i=1}^N \sum_{j\notin\Delta(i,N)} \mathrm{Cov}\left(W_i, W_j\right) \right\|_\infty \\
& + C\left\|\Sigma_N^{-1/2}\right\|_\infty^3 \left\| \sum_{i=1}^N \sum_{j,k\in\Delta(i,N)} \mathbb{E}\left[\left|vec(W_iW_j')W_k\right|\right] \right\|_\infty \\
= & \left\|\Sigma_N^{-1/2}\right\|_\infty^2 o\left(\left\|\Sigma_N^2\right\|_\infty^{1/2} + \left\|\Sigma_N\right\|_\infty\right) + \left\|\Sigma_N^{-1/2}\right\|_\infty^3 o\left(\left\|\Sigma_N^{3/2}\right\|_\infty\right) \\
= & o(1),
\end{aligned}
$$

implying that $\Sigma_N^{-1/2}\mathbf{S}_N \xrightarrow{d} \mathbb{N}(0, \mathbf{I}_p)$. ■

In what follows, we first present several lemmas that will be used to show the asymptotic properties of the jacobian and hessian matrix of the objective function.

Lemma 4.9.8 *Under Assumptions 4.5.4, 4.5.5 and the i.i.d. of $x_{i,j}$ across i for any given $j = 1, 2, ..., K_T$, we have that*

$$
\sup_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^N \left\| \frac{d^2 m^*(x_{i,j};\theta)}{d\theta d\theta'} \right\|^2 = O_p(1); \qquad \sup_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^N \left\| \frac{dm^*(x_{i,j};\theta)}{d\theta} \right\|^2 = O_p(1);
$$

and for $\tilde{\theta}_N \xrightarrow{p} \theta^0$,

$$
\begin{aligned}
& \frac{1}{N}\sum_{i=1}^N \left| m^*(x_{i,j};\tilde{\theta}_N) - m^*(x_{i,j};\theta^0) \right|^2 = o_p(1); \\
& \frac{1}{N}\sum_{i=1}^N \left\| \frac{dm^*(x_{i,j};\tilde{\theta}_N)}{d\theta} - \frac{dm^*(x_{i,j};\theta^0)}{d\theta} \right\|^2 = o_p(1); \\
& \frac{1}{N}\sum_{i=1}^N \left\| \frac{d^2 m^*(x_{i,j};\tilde{\theta}_N)}{d\theta d\theta} - \frac{d^2 m^*(x_{i,j};\theta^0)}{d\theta d\theta} \right\|^2 = o_p(1).
\end{aligned}
$$

Proof of Lemma 4.9.8. By Assumption 4.5.5 and the uniform convergence of i.i.d. samples (Lemma 2.4 of Newey and McFadden (1994))

$$
\sup_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^N \left\| \frac{d^2 m^*(x_{i,j};\theta)}{d\theta d\theta'} \right\|^2
$$

$$\leq \sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{d^2m^*(x_{i,j};\theta)}{d\theta d\theta'}\right\|^2-\mathbb{E}\left[\left\|\frac{d^2m^*(x_{i,j};\theta)}{d\theta d\theta'}\right\|^2\right]\right|+\sup_{\theta\in\Theta}\left|\mathbb{E}\left[\left\|\frac{d^2m^*(x_{i,j};\theta)}{d\theta d\theta'}\right\|^2\right]\right|$$
$$=o_p(1)+\sup_{\theta\in\Theta}\left|\mathbb{E}\left[\left\|\frac{d^2m^*(x_{i,j};\theta)}{d\theta d\theta'}\right\|^2\right]\right|. \tag{4.9.29}$$

Because $\sup_{\theta\in\Theta}\mathbb{E}\left[\left\|\frac{d^2m^*(x_{i,j};\theta)}{d\theta d\theta'}\right\|^2\right] \leq \mathbb{E}[H_1(x_{i,j})] < \infty$ by Assumption 4.5.5, (4.9.29) becomes to

$$\sup_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{d^2m^*(x_{i,j};\theta)}{d\theta d\theta'}\right\|^2=O_p(1). \tag{4.9.30}$$

Similar arguments can be used to show that $\sup_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{dm^*(x_{i,j};\theta)}{d\theta}\right\|^2=O_p(1)$. Besides, the mean value theorem gives

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left|m^*(x_{i,j};\tilde{\theta}_N)-m^*(x_{i,j};\theta^0)\right|^2=&\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\partial m^*(x_{i,j};\bar{\theta}_N)}{\partial\theta'}(\tilde{\theta}_N-\theta^0)\right|^2\\ \leq&\sup_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial m^*(x_{i,j};\theta)}{\partial\theta}\right\|^2\left\|\tilde{\theta}_N-\theta^0\right\|^2\\ =&o_p(1),\end{aligned} \tag{4.9.31}$$

for $\bar{\theta}_N$ between $\tilde{\theta}_N$ and θ^0. Similarly, we can also obtain that

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial m^*(x_{i,j};\tilde{\theta}_N)}{\partial\theta}-\frac{\partial m^*(x_{i,j};\theta^0)}{\partial\theta}\right\|^2\leq\sup_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial m^*(x_{i,j};\theta)}{\partial\theta'}\right\|^2\left\|\tilde{\theta}_N-\theta^0\right\|^2=o_p(1). \tag{4.9.32}$$

Moreover, since

$$\frac{d^2m^*(x_{i,j};\tilde{\theta}_N)}{d\theta_r d\theta_q}-\frac{d^2m^*(x_{i,j};\theta^0)}{d\theta_r d\theta_q}=\frac{\partial}{\partial\theta'}\left(\frac{d^2m^*(x_{i,j};\bar{\theta}_N)}{d\theta_r d\theta_q}\right)(\tilde{\theta}_N-\theta^0),$$

by the uniformly bounded third derivative of $m^*(x;\theta)$,

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{d^2m^*(x_{i,j};\tilde{\theta}_N)}{d\theta d\theta'}-\frac{d^2m^*(x_{i,j};\theta^0)}{d\theta d\theta'}\right\|^2$$

$$
\begin{aligned}
\leq & \frac{1}{N} \sum_{r,q=1}^{d_\theta} \sum_{i=1}^{N} \left| \frac{d^2 m^*(x_{i,j}; \tilde{\theta}_N)}{d\theta_r d\theta_q} - \frac{d^2 m^*(x_{i,j}; \theta^0)}{d\theta_r d\theta_q} \right|^2 \\
\leq & \frac{1}{N} \sum_{r,q=1}^{d_\theta} \sum_{i=1}^{N} \left\| \frac{\partial}{\partial \theta'} \left(\frac{d^2 m^*(x_{i,j}; \bar{\theta}_N)}{d\theta_r d\theta_q} \right) \right\|^2 \left\| \tilde{\theta}_N - \theta^0 \right\|^2 \\
= & O_p \left(\left\| \tilde{\theta}_N - \theta^0 \right\|^2 \right) = o_p(1). \qquad (4.9.33)
\end{aligned}
$$

■

Lemmas 4.9.9 to 4.9.11 show the key steps for establishing the asymptotics for the jacobian of the objective function. The proofs are based on Section 8 of Newey and McFadden (1994) and extended to adopt data under dependency-neighbourhoods structure.

Lemma 4.9.9 (Linearisation) *Under assumptions in Lemma 4.5.4 (b), there exists a function $G(\cdot;\gamma) : \Omega_W \mapsto \mathbb{R}^{d_\theta}$ which is linear in γ and satisfies*

$$
\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[g(W_i; \theta^0, \hat{\phi}_N) - g(W_i; \theta^0, \phi^0) - G(W_i; \hat{\gamma}_N - \gamma^0) \right] \right\| = o_p(1).
$$

Proof of Lemma 4.9.9. Recall that $g(W_i; \theta, \phi) = \tau_i [Y_i - m(X_i; \theta, \phi)] \frac{\partial m(X_i; \theta, \phi)}{\partial \theta}$. Then,

$$
\begin{aligned}
& \frac{1}{\sqrt{N}} \sum_{i=1}^{N} g(W_i; \theta^0, \hat{\phi}_N) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} g(W_i; \theta^0, \phi^0) \\
= & \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \tau_i \left[[Y_i - m(X_i; \theta^0, \hat{\phi}_N)] \frac{\partial m(X_i; \theta^0, \hat{\phi}_N)}{\partial \theta} - [Y_i - m(X_i; \theta^0, \phi^0)] \frac{\partial m(X_i; \theta^0, \phi^0)}{\partial \theta} \right], \qquad (4.9.34)
\end{aligned}
$$

where making use of the identity $\hat{a}\hat{b} - ab = (\hat{a} - a)b + a(\hat{b} - b) + (\hat{a} - a)(\hat{b} - b)$ leads to

$$
\begin{aligned}
& \frac{1}{\sqrt{N}} \sum_{i=1}^{N} g(W_i; \theta^0, \hat{\phi}_N) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} g(W_i; \theta^0, \phi^0) \\
= & - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \tau_i \left[m(X_i; \theta^0, \hat{\phi}_N) - m(X_i; \theta^0, \phi^0) \right] \frac{\partial m(X_i; \theta^0, \phi^0)}{\partial \theta} \\
& + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \tau_i \left[Y_i - m(X_i; \theta^0, \phi^0) \right] \left[\frac{\partial m(X_i; \theta^0, \hat{\phi}_N)}{\partial \theta} - \frac{\partial m(X_i; \theta^0, \phi^0)}{\partial \theta} \right]
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i\left[m(X_i;\theta^0,\hat{\phi}_N)-m(X_i;\theta^0,\phi^0)\right]\left[\frac{\partial m(X_i;\theta^0,\hat{\phi}_N)}{\partial\theta}-\frac{\partial m(X_i;\theta^0,\phi^0)}{\partial\theta}\right]\\
=&-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i\sum_{j=1}^{K_T}m^*(x_{i,j};\theta^0)\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\frac{\partial m(X_i;\theta^0,\phi^0)}{\partial\theta}\\
&+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i\left[Y_i-m(X_i;\theta^0,\phi^0)\right]\sum_{j=1}^{K_T}\frac{\partial m^*(x_{i,j};\theta^0)}{\partial\theta}\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\\
&-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i\sum_{j=1}^{K_T}m^*(x_{i,j};\theta^0)\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\sum_{j=1}^{K_T}\frac{\partial m^*(x_{i,j};\theta^0)}{\partial\theta}\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\\
:=&\mathcal{G}_1+\mathcal{G}_2+\mathcal{G}_3. \qquad (4.9.35)
\end{aligned}
$$

Firstly, consider $\mathcal{G}_3$. By the Cauchy–Schwarz inequality, (4.9.116) and Lemma 4.9.8,

$$
\begin{aligned}
&\|N^{-1/2}\mathcal{G}_3\|\\
\leq&\frac{C}{N}\sum_{i=1}^{N}\sum_{j,l=1}^{K_T}\left\|m^*(x_{i,j};\theta^0)\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\frac{\partial m^*(x_{i,l};\theta^0)}{\partial\theta}\left[\hat{f}_{T_i^*|X_i}(t_l)-f_{T_i^*|X_i}(t_l)\right]\right\|\\
\leq&\left(\sup_{\|\hat{\gamma}_N-\gamma^0\|_\infty<\eta}\left\|\hat{\phi}_N-\phi^0\right\|_\infty\right)^2\frac{1}{N}\sum_{j,l=1}^{K_T}\sum_{i=1}^{N}\left|m^*(x_{i,j};\theta^0)\right|\left\|\frac{\partial m^*(x_{i,l};\theta^0)}{\partial\theta}\right\|\\
\leq&\left(\sup_{\|\hat{\gamma}_N-\gamma^0\|_\infty<\eta}\left\|\hat{\phi}_N-\phi^0\right\|_\infty\right)^2\sum_{j,l=1}^{K_T}\left[\frac{1}{N}\sum_{i=1}^{N}m^*(x_{i,j};\theta^0)^2\right]^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial m^*(x_{i,l};\theta^0)}{\partial\theta}\right\|^2\right]^{1/2}\\
=&O_p\left(\|\hat{\gamma}_N-\gamma^0\|_\infty^2\right). \qquad (4.9.36)
\end{aligned}
$$

Thus, given (4.9.36) we can get that $\|\mathcal{G}_3\|=O_p(N^{1/2}\|\hat{\gamma}_N-\gamma^0\|_\infty^2)=o_p(1)$ by Assumption 4.5.6.

Next, let us consider $\mathcal{G}_1+\mathcal{G}_2$. Recall that the $1\times K_T$ row vector $\mathcal{R}(W_i;\theta,\phi)$ is defined as

$$
\mathcal{R}(W_i;\theta,\phi)=\begin{bmatrix}\left[Y_i-m(X_i;\theta,\phi)\right]m^*(x_{i,1};\theta)\\ \vdots\\ \left[Y_i-m(X_i;\theta,\phi)\right]m^*(x_{i,K_T};\theta)\end{bmatrix}'.
$$

Denote $\phi(\mathbf{t};\hat{\gamma}_N)=[\hat{f}_{T_i^*|X_i}(t_1),...,\hat{f}_{T_i^*|X_i}(t_{K_T})]'$ and $\phi(\mathbf{t};\gamma^0)=[f_{T_i^*|X_i}(t_1),...,f_{T_i^*|X_i}(t_{K_T})]'$.

Then, simple calculations yield that

$$\begin{aligned}\mathcal{G}_1+\mathcal{G}_2=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i\Bigg[\sum_{j=1}^{K_T}\Bigg(\left[Y_i-m(X_i;\theta^0,\phi^0)\right]\frac{\partial m^*(x_{i,j};\theta^0)}{\partial\theta}\\&-m^*(x_{i,j};\theta^0)\frac{\partial m(X_i;\theta^0,\phi^0)}{\partial\theta}\Bigg)\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\Bigg]\\=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i\left[\frac{\partial}{\partial\theta}\mathcal{R}(W_i;\theta^0,\phi^0)\left(\phi(\mathbf{t};\hat{\gamma}_N)-\phi(\mathbf{t};\gamma^0)\right)\right]\\=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i\left[\frac{\partial}{\partial\theta}\mathcal{R}(W_i;\theta^0,\phi^0)\frac{\partial\phi(\mathbf{t};\gamma^0)}{\partial\gamma'}\left(\hat{\gamma}_N-\gamma^0\right)\right]+\mathcal{G}_R,\end{aligned}\tag{4.9.37}$$

where the reminder term

$$\begin{aligned}\mathcal{G}_R:=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i\frac{\partial}{\partial\theta}\mathcal{R}(W_i;\theta^0,\phi^0)\left[\phi(\mathbf{t};\hat{\gamma}_N)-\phi(\mathbf{t};\gamma^0)-\frac{\partial\phi(\mathbf{t};\gamma^0)}{\partial\gamma'}\left(\hat{\gamma}_N-\gamma^0\right)\right]\\=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tau_i[\nabla\mathcal{R}_1+\nabla\mathcal{R}_2]\left[\phi(\mathbf{t};\hat{\gamma}_N)-\phi(\mathbf{t};\gamma^0)-\frac{\partial\phi(\mathbf{t};\gamma^0)}{\partial\gamma'}\left(\hat{\gamma}_N-\gamma^0\right)\right],\end{aligned}$$

with $\frac{\partial}{\partial\theta}\mathcal{R}(W_i;\theta^0,\phi^0):=\nabla\mathcal{R}_1+\nabla\mathcal{R}_2$ and

$$\begin{aligned}\nabla\mathcal{R}_1&=\left[Y_i-m(X_i;\theta^0,\phi^0)\right]\left[\tfrac{\partial m^*(x_{i,1};\theta^0)}{\partial\theta}\quad\dots\quad\tfrac{\partial m^*(x_{i,K_T};\theta^0)}{\partial\theta}\right],\\\nabla\mathcal{R}_2&=-\frac{\partial m(X_i;\theta^0,\phi^0)}{\partial\theta}\left[m^*(x_{i,1};\theta^0)\quad\cdots\quad m^*(x_{i,K_T};\theta^0)\right].\end{aligned}$$

Next, we show that $\mathcal{G}_R=o_p(1)$. Due to Theorem 4.5.2, we can focus on a small neighbourhood of γ^0 and bound the reminder term as follows:

$$\begin{aligned}\|N^{-1/2}\mathcal{G}_R\|\leq&\sup_{\|\hat{\gamma}_N-\gamma^0\|_\infty<\eta}\left\|\left(\hat{\phi}_N-\phi^0\right)-\frac{\partial\phi(\gamma^0)}{\partial\gamma'}\left(\hat{\gamma}_N-\gamma^0\right)\right\|_\infty\frac{1}{N}\sum_{i=1}^{N}\tau_i\left\|\nabla\mathcal{R}_1+\nabla\mathcal{R}_2\right\|\\\leq&O_p\left(\left\|\hat{\gamma}_N-\gamma^0\right\|_\infty^2\right)\left[\frac{1}{N}\sum_{i=1}^{N}\tau_i\left\|\nabla\mathcal{R}_1\right\|+\frac{1}{N}\sum_{i=1}^{N}\tau_i\left\|\nabla\mathcal{R}_2\right\|\right],\end{aligned}$$

where the $O_p(\|\hat{\gamma}_N-\gamma^0\|_\infty^2)$ is due to (4.9.113), and applying the Cauchy–Schwarz inequality to each of the term inside the bracket leads to

$$\frac{1}{N}\sum_{i=1}^{N}\tau_i\left\|\nabla\mathcal{R}_1\right\|\leq\frac{1}{N}\sum_{i=1}^{N}\tau_i\left|Y_i-m(X_i;\theta^0,\phi^0)\right|\left\|\left[\tfrac{\partial m^*(x_{i,1};\theta^0)}{\partial\theta}\quad\dots\quad\tfrac{\partial m^*(x_{i,K_T};\theta^0)}{\partial\theta}\right]\right\|$$

$$
\begin{aligned}
\leq & C\left[\frac{1}{N}\sum_{i=1}^{N}\tau_i\left[Y_i - m(X_i;\theta^0,\phi^0)\right]^2\right]^{1/2}\left[\frac{1}{N}\sum_{j=1}^{K_T}\sum_{i=1}^{N}\left\|\frac{\partial m^*(x_{i,j};\theta^0)}{\partial\theta}\right\|^2\right]^{1/2} \\
= & O_p(1),
\end{aligned}
$$

where the last line follows from (4.9.117) and Lemma 4.9.8. Similarly, from the Cauchy–Schwarz inequality and Lemma 4.9.8, we can also get that

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\tau_i\left\|\nabla\mathcal{R}_2\right\| \leq & \frac{C}{N}\sum_{i=1}^{N}\left\|\frac{\partial m(X_i;\theta^0,\phi^0)}{\partial\theta}\right\|\left\|\left[m^*(x_{i,1};\theta^0) \quad \cdots \quad m^*(x_{i,K_T};\theta^0)\right]\right\| \\
\leq & C\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial m(X_i;\theta^0,\phi^0)}{\partial\theta}\right\|^2\right]^{1/2}\left[\frac{1}{N}\sum_{j=1}^{K_T}\sum_{i=1}^{N}m^*(x_{i,j};\theta^0)^2\right]^{1/2} \\
= & O_p(1),
\end{aligned}
$$

Therefore, it yields from the above results and Assumption 4.5.6 that

$$
\left\|\mathcal{G}_R\right\| = O_p\left(N^{1/2}\|\hat{\gamma}_N - \gamma^0\|_\infty^2\right) = o_p(1). \tag{4.9.38}
$$

To fulfil this proof and find the function G, let $\tilde{\nu}(W_i) := \tau_i\left[\frac{\partial}{\partial\theta}\mathcal{R}(W_i;\theta^0,\phi^0)\frac{\partial\phi(\mathbf{t};\gamma^0)}{\partial\gamma'}\right]$ and $G(W_i;\gamma) = \tilde{\nu}(W_i)\gamma = \tau_i\left[\frac{\partial}{\partial\theta}\mathcal{R}(W_i;\theta^0,\phi^0)\frac{\partial\phi(\mathbf{t};\gamma^0)}{\partial\gamma'}\right]\gamma$, then by construction $G(W_i;\gamma)$ is linear in γ. Moreover, based on (4.9.36) and (4.9.38),

$$
\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[g(W_i;\theta^0,\hat{\phi}_N) - g(W_i;\theta^0,\phi^0) - G(W_i;\hat{\gamma}_N - \gamma^0)\right]\right\| \leq \|\mathcal{G}_3\| + \|\mathcal{G}_R\| = o_p(1). \tag{4.9.39}
$$

■

Lemma 4.9.10 (Stochastic Equicontinuity) *Let $F_W(w)$ be the true probability distribution function of W_i. Suppose assumptions in Lemma 4.5.4 (b) hold, then*

$$
\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[G(W_i;\hat{\gamma}_N - \gamma^0) - \int G(w;\hat{\gamma}_N - \gamma^0)dF_W(w)\right] = o_p(1).
$$

Proof of Lemma 4.9.10. By the linearity of $G(w;\gamma) = \tilde{\nu}(w)\gamma$ in γ, we can get

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[G(W_i; \hat{\gamma}_N - \gamma^0) - \int G(w; \hat{\gamma}_N - \gamma^0) dF_W(w) \right] \right\| \\
=&\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\tilde{\nu}(W_i) - \mathbb{E}[\tilde{\nu}(W_i)] \right] (\hat{\gamma}_N - \gamma^0) \right\| \\
\leq& C \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\tilde{\nu}(W_i) - \mathbb{E}[\tilde{\nu}(W_i)] \right] \right\| \left\| \hat{\gamma}_N - \gamma^0 \right\|_\infty . \qquad (4.9.40)
\end{aligned}$$

Denote $\tilde{\nu}_r(W_i)$ as the r-th entry of the vector $\tilde{\nu}(W_i)$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\tilde{\nu}(W_i) - \mathbb{E}[\tilde{\nu}(W_i)] \right] \right\|^2 \right] \\
=&\frac{1}{N} \mathbb{E}\left[\sum_{i=1}^{N} \Big(\tilde{\nu}(W_i) - \mathbb{E}[\tilde{\nu}(W_i)] \Big)' \sum_{i=1}^{N} \Big(\tilde{\nu}(W_i) - \mathbb{E}[\tilde{\nu}(W_i)] \Big) \right] \\
=&\frac{1}{N} \sum_{r=1}^{d_\theta} \left[\sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \mathrm{Cov}\Big(\tilde{\nu}_r(W_i), \tilde{\nu}_r(W_j) \Big) + \sum_{i=1}^{N} \sum_{j \notin \Delta(i,N)} \mathrm{Cov}\Big(\tilde{\nu}_r(W_i), \tilde{\nu}_r(W_j) \Big) \right] \\
=&\frac{1}{N} \sum_{r=1}^{d_\theta} \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \mathrm{Cov}\Big(\tilde{\nu}_r(W_i), \tilde{\nu}_r(W_j) \Big) + s.o., \qquad (4.9.41)
\end{aligned}$$

where the last line comes from Assumption 4.5.1. Note that due to $\mathrm{Var}[\tilde{\nu}_r(W_i)] < \infty$ as in Assumption 4.5.6 and $1/N \sum_{i=1}^{N} |\Delta(i,N)| = O(1)$,

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \mathrm{Cov}\Big(\tilde{\nu}_r(W_i), \tilde{\nu}_r(W_j) \Big) \leq \frac{C}{N} \sum_{i=1}^{N} |\Delta(i,N)| = O(1). \qquad (4.9.42)$$

Given (4.9.42), together with the consistency $\left\| \hat{\gamma}_N - \gamma^0 \right\|_\infty = o_p(1)$, we know that

$$\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[G(W_i; \hat{\gamma}_N - \gamma^0) - \int G(w; \hat{\gamma}_N - \gamma^0) dF_W(w) \right] \right\| = o_p(1). \qquad (4.9.43)$$

■

Lemma 4.9.11 (Mean-square Differentiability) *Under assumptions in Lemma 4.5.4*

(b), there exists a function $\delta : \Omega_W \mapsto \mathbb{R}^{d_\theta}$ *such that*

$$\int G(w; \tilde{\gamma}_N - \gamma) dF_W(w) = \int \delta(w) d\hat{F}_W(w),$$
$$\sqrt{N}\mathbb{E}\left[\left\| \int \delta(w) d\hat{F}_W(w) - \int \delta(w) d\tilde{F}_W(w) \right\|\right] = o(1),$$

where $\hat{F}_W(w)$ *is the kernel estimator of* $F_W(w)$ *and* $\tilde{F}_W(w) := 1/N \sum_{i=1}^N 1[W_i \leq w]$ *is the empirical distribution of* W_i*.*

Proof of Lemma 4.9.11. Following the derivations of Theorem 8.1 or (Theorem 8.11) in Newey and McFadden (1994), it is apparent from the linearity of $G(w;\gamma)$ in γ that and the law of iterated expectation,

$$\int G(w;\gamma) dF_W(w) = \int \nu(w)\gamma(w) dw,$$

where recall that the $d_\theta \times d_\gamma$ matrix $\nu(w)$ is defined as

$$\nu(w) = \mathbb{E}\left[\tau(X_i) \frac{\partial}{\partial \theta} \mathcal{R}(W_i; \theta^0, \phi^0) \frac{\partial \phi(\mathbf{t}; \gamma)}{\partial \gamma'}\Big|_{\gamma=\gamma(w)} \mathbf{1}_{d_\gamma} \Big| w\right].$$

In addition, let $\delta(w) := \nu(w) - \mathbb{E}[\nu(w)]$, we have

$$\int G(w; \tilde{\gamma}_N - \gamma) dF_W(w) = \int \delta(w) d\hat{F}_W(w),$$

with $\hat{F}_W(w)$ being the kernel estimator of the distribution of W_i.

At last, recall the empirical distribution $\tilde{F}_W(w) = 1/N \sum_{i=1}^N 1[W_i \leq w]$. By an abuse of notation, we denote $\kappa(\frac{w^c - \tilde{w}^c}{h}) := \prod_{q=1}^Q \kappa(\frac{w_q^c - \tilde{w}_q^c}{h})$. Consider the difference between the two integrals $\delta(F)$ defined as below, which can be interpreted as a smoothing bias term,

$$\begin{aligned}
\delta(F) :=& \int \delta(w) d\hat{F}_W(w) - \int \delta(w) d\tilde{F}_W(w) \\
=& \frac{1}{N} \sum_{i=1}^N \left[\sum_{w^d \in \Omega_{W^d}} \int \nu(w) \hat{f}_i^{ker}(w) dw^c - \nu(W_i) \right] \\
=& \frac{1}{N} \sum_{i=1}^N \left[\sum_{w^d \in \Omega_{W^d}} \int \nu(w) \frac{1}{h^Q} 1[W_i^d = w^d] \prod_{q=1}^Q \kappa\left(\frac{w_q^c - W_{iq}^c}{h}\right) dw^c - \nu(W_i) \right]
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{N}\sum_{i=1}^{N}\left[\int \nu(w^c, W_i^d)\frac{1}{h^Q}\prod_{q=1}^{Q}\kappa\left(\frac{w_q^c - W_{iq}^c}{h}\right)dw^c - \nu(W_i)\right]\\
=&\frac{1}{N}\sum_{i=1}^{N}\int\left[\nu(W_i^c + hv, W_i^d) - \nu(W_i)\right]\prod_{q=1}^{Q}\kappa\left(v_q\right)dv\\
:=&\frac{1}{N}\sum_{i=1}^{N}\delta_i(F). \qquad (4.9.44)
\end{aligned}
$$

Because the identical distribution of W_i across i, it follows from (4.9.44) that

$$
\begin{aligned}
&\sqrt{N}\mathbb{E}[\delta(F)] = \sqrt{N}\mathbb{E}[\delta_i(F)]\\
=&\sqrt{N}\mathbb{E}\left[\int \nu(W_i^c + hv, W_i^d)\prod_{q=1}^{Q}\kappa\left(v_q\right)dv - \nu(W_i)\right]\\
=&\sqrt{N}\iint \nu(\tilde{w}^c + hv, \tilde{w}^d)\prod_{q=1}^{Q}\kappa\left(v_q\right)dvdF_W(\tilde{w}) - \sqrt{N}\int \nu(w)dF_W(w)\\
=&\sqrt{N}\iint \nu(\tilde{w}^c, \tilde{w}^d)\prod_{q=1}^{Q}\kappa\left(v_q\right)dvdF_W(\tilde{w}^c - hv, \tilde{w}^d) - \sqrt{N}\int \nu(w)dF_W(w)\\
=&\sqrt{N}\left\{\iint \nu(\tilde{w}^c, \tilde{w}^d)\prod_{q=1}^{Q}\kappa\left(v_q\right)dvdF_W(\tilde{w}^c - hv, \tilde{w}^d) - \iint \nu(w)\prod_{q=1}^{Q}\kappa(v_q)dvdF_W(w)\right\}\\
=&\sqrt{N}\left\{\sum_{w^d\in\Omega_{W^d}}\iint \nu(w)\left[f_{W_i^c,W_i^d}(w^c - hv, w^d) - f_{W_i^c,W_i^d}(w^c, w^d)\right]\prod_{q=1}^{Q}\kappa\left(v_q\right)dvdw^c\right\},
\end{aligned}
\qquad (4.9.45)
$$

which together with (4.9.89) and Assumption 4.5.6, implies that

$$
\begin{aligned}
&\sqrt{N}\left\|\mathbb{E}[\delta(F)]\right\|\\
\leq&\sqrt{N}\sum_{w^d\in\Omega_{W^d}}\int\left\|\nu(w)\right\|\left\|\int\left[f_{W_i^c,W_i^d}(w^c - hv, w^d) - f_{W_i^c,W_i^d}(w^c, w^d)\right]\prod_{q=1}^{Q}\kappa\left(v_q\right)dv\right\|dw^c\\
\leq&C\sqrt{N}h^2\sum_{w^d\in\Omega_{W^d}}\int\left\|\nu(w)\right\|dw^c\\
=&o(1). \qquad (4.9.46)
\end{aligned}
$$

Next, let $\delta(F) = (\delta_1(F), ..., \delta_{d_\theta}(F))'$ with $\delta_r(F) = 1/N \sum_{i=1}^N \delta_{r,i}(F)$ and consider

$$\begin{aligned}\mathbb{E}\left[\left\|\sqrt{N}\delta(F) - \sqrt{N}\mathbb{E}[\delta(F)]\right\|^2\right] =& \sum_{r=1}^{d_\theta} \mathbb{E}\left[\left|\sqrt{N}\delta_r(F) - \sqrt{N}\mathbb{E}[\delta_r(F)]\right|^2\right] \\ =& N\sum_{r=1}^{d_\theta} \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N (\delta_{r,i}(F) - \mathbb{E}[\delta_{r,i}(F)])\right|^2\right] \\ =& \frac{1}{N}\sum_{r=1}^{d_\theta}\sum_{i=1}^N \sum_{j\in\Delta(i,N)} \text{Cov}\left(\delta_{r,i}(F), \delta_{r,j}(F)\right) + s.o.,\end{aligned} \tag{4.9.47}$$

where the last line follows from Assumption 4.5.1. Due to the identical distribution of W_i and (4.9.44), we can bound the covariance in (4.9.47) by

$$\begin{aligned}\left|\text{Cov}\left(\delta_{r,i}(F), \delta_{r,j}(F)\right)\right| \leq& \text{Var}\left[\delta_{r,i}(F)\right] \leq \mathbb{E}\left[\left|\delta_{r,i}(F)\right|^2\right] \\ =& \mathbb{E}\left[\left(\int \left[\nu_r(W_i^c + hv, W_i^d) - \nu_r(W_i)\right] \prod_{q=1}^Q \kappa\left(v_q\right) dv\right)^2\right].\end{aligned} \tag{4.9.48}$$

From Assumption 4.5.2 we know that $\int x\kappa(x)dx = 0$ and $\int x^2\kappa(x)dx = K_2$, and Assumption 4.5.6 that $\nu(w)$ is twice continuously differentiable in w^c. Expanding $\nu_r(W_i^c + hv, W_i^d)$ around W_i^c, then there exists a constant $C > 0$ such that

$$\left|\text{Cov}\left(\delta_{r,i}(F), \delta_{r,j}(F)\right)\right| \leq h^4\mathbb{E}\left[\left(\int v' \frac{\partial \nu_r(W_i^c + w^*, W_i^d)}{\partial w^c \partial (w^c)'} v \prod_{q=1}^Q \kappa\left(v_q\right) dv\right)^2\right] \leq Ch^4. \tag{4.9.49}$$

Substituting (4.9.49) into (4.9.47), since $1/N\sum_{i=1}^N |\Delta(i,N)| = O(1)$ as in Assumption 4.5.2,

$$\mathbb{E}\left[\left\|\sqrt{N}\delta(F) - \sqrt{N}\mathbb{E}[\delta(F)]\right\|^2\right] = O(h^4) = o(1). \tag{4.9.50}$$

Based on (4.9.46) and (4.9.50), since both the mean and variance of $\sqrt{N}\delta(F)$ are $o(1)$, by Chebyshev's inequality, it follows directly that $\mathbb{E}\left[\|\sqrt{N}\delta(F)\|\right] \xrightarrow{p} 0$. ■

4.9.3. Proofs

Proofs of Section 4.3

Lemma 4.9.12 *Under Assumption 4.3.2, we have that for* $\forall i \in \mathcal{P}$, $\varepsilon_i \perp (D_i, S_i^*) \big| Z_i, \mathcal{F}_i^*$.

Proof of Lemma 4.9.12. If we can show that $\Pr(\varepsilon_i < e | D_i, S_i^*, Z_i, \mathcal{F}_i^*) = \Pr(\varepsilon_i < e | Z_i, \mathcal{F}_i^*)$, then the stated result follows. By the law of total probability, we have for $\forall e \in \Omega_\varepsilon$,

$$\begin{aligned} &\Pr(\varepsilon_i < e | D_i, S_i^*, Z_i, \mathcal{F}_i^*) \\ =&\mathbb{E}\left[Pr\left(\varepsilon_i < e \middle| D_i, S_i^*, Z_i, \mathcal{F}_i^*, \mathcal{N}_i^*, \{D_j\}_{j\in\mathcal{N}_i^*}\right) \middle| D_i, S_i^*, Z_i, \mathcal{F}_i^*\right], \end{aligned} \tag{4.9.51}$$

where the expectation is with respect to $f_{\mathcal{N}_i^*, \{D_j\}_{j\in\mathcal{N}_i^*} | D_i, S_i^*, Z_i, \mathcal{F}_i^*}$. By definition, $S_i^* = \sum_{j\in\mathcal{N}_i^*} D_j$, therefore, S_i^* becomes fixed when given $(\mathcal{N}_i^*, \{D_j\}_{j\in\mathcal{N}_i^*})$. In addition, since Assumption 4.3.2 (a) implies that D_i is independent to $(\varepsilon_i, Z_i, \mathcal{F}_i^*, \mathcal{N}_i^*, \{D_j\}_{j\in\mathcal{N}_i^*})$. Then, we know that (D_i, S_i^*) can be eliminated from the conditional probability of $\varepsilon_i < e$ in (4.9.51), i.e.

$$\begin{aligned} \Pr(\varepsilon_i < e | D_i, S_i^*, Z_i, \mathcal{F}_i^*) =&\mathbb{E}\left[Pr\left(\varepsilon_i < e | Z_i, \mathcal{F}_i^*, \mathcal{N}_i^*, \{D_j\}_{j\in\mathcal{N}_i^*}\right) \middle| D_i, S_i^*, Z_i, \mathcal{F}_i^*\right] \\ =&\mathbb{E}\left[\Pr(\varepsilon_i < e | Z_i, \mathcal{F}_i^*) \middle| D_i, S_i^*, Z_i, \mathcal{F}_i^*\right] \\ =&\Pr(\varepsilon_i < e | Z_i, \mathcal{F}_i^*), \end{aligned} \tag{4.9.52}$$

where the second line is from Assumption 4.3.2 (b) and the last line implies $\varepsilon_i \perp (D_i, S_i^*) \big| Z_i, \mathcal{F}_i^*$. ■

Lemma 4.9.13 *Under Assumptions 4.3.2 and 4.3.4,* $\varepsilon_i \perp (D_i, S_i^*, S_i, \mathcal{F}_i) | Z_i, \mathcal{F}_i^*$ *for* $\forall i \in \mathcal{P}$.

Proof of Lemma 4.9.13. Denote $P_i^* = (\mathcal{N}_i^*, \{D_j\}_{j\in\mathcal{N}_i^*})$ and $P_i = (\mathcal{N}_i, \{D_j\}_{j\in\mathcal{N}_i})$. Then, we know from Assumptions 4.3.2 (a) and 4.3.4 (a) that $D_i \perp (P_i^*, P_i)$, because of the facts that $i \notin \mathcal{N}_i^*$, $i \notin \mathcal{N}_i$, $D_i \perp (\{D_j\}_{j\in\mathcal{N}_i^*}, \{D_j\}_{j\in\mathcal{N}_i}) \big| \mathcal{N}_i^*, \mathcal{N}_i$ and $\{D_i\}_{i\in\mathcal{P}} \perp (\mathcal{N}_i^*, \mathcal{N}_i)$. Moreover, since S_i^* and S_i are functions of P_i^* and P_i, respectively, we have

$D_i \perp (S_i^*, S_i)$. By the law of total probability,

$$\begin{aligned}
&Pr\left(\varepsilon_i < e \middle| D_i, S_i^*, S_i, \mathcal{F}_i, Z_i, \mathcal{F}_i^*\right) \\
=&Pr\left(\varepsilon_i < e \middle| S_i^*, S_i, \mathcal{F}_i, Z_i, \mathcal{F}_i^*\right) \\
=&\mathbb{E}\left[Pr\left(\varepsilon_i < e \middle| S_i^*, S_i, \mathcal{F}_i, Z_i, \mathcal{F}_i^*, P_i^*, P_i\right) \middle| S_i^*, S_i, \mathcal{F}_i, Z_i, \mathcal{F}_i^*\right], \qquad (4.9.53)
\end{aligned}$$

for $\forall e \in \Omega_\varepsilon$, where the expectation is with respect to $f_{P_i^*, P_i | S_i^*, S_i, \mathcal{F}_i, Z_i, \mathcal{F}_i^*}$. We know that $S_i^*, S_i, \mathcal{F}_i$ are fixed given (P_i^*, P_i). Thus, equation (4.9.53) becomes to

$$\begin{aligned}
Pr\left(\varepsilon_i < e \middle| D_i, S_i^*, S_i, \mathcal{F}_i, Z_i, \mathcal{F}_i^*\right) =&\mathbb{E}\left[Pr\left(\varepsilon_i < e \middle| Z_i, \mathcal{F}_i^*, P_i^*, P_i\right) \middle| S_i^*, S_i, \mathcal{F}_i, Z_i, \mathcal{F}_i^*\right] \\
=&\mathbb{E}\left[Pr\left(\varepsilon_i < e \middle| Z_i, \mathcal{F}_i^*\right) \middle| S_i^*, S_i, \mathcal{F}_i, Z_i, \mathcal{F}_i^*\right] \\
=&Pr\left(\varepsilon_i < e \middle| Z_i, \mathcal{F}_i^*\right),
\end{aligned}$$

where the second equality above is due to Assumption 4.3.4 (b). ■

Proof of Proposition 4.3.3. By Assumption 4.3.1 and the law of iterated expectation,

$$\begin{aligned}
&m_i(d, s, z, n) \\
=&\mathbb{E}\left[r(D_i, S_i^*, Z_i, \mathcal{F}_i^*, \varepsilon_i) \middle| D_i = d, S_i = s, Z_i = z, \mathcal{F}_i = n\right] \\
=&\sum_{(s^*, n^*) \in \Omega_{S^*, \mathcal{F}^*}} \mathbb{E}\left[r(d, s^*, z, n^*, \varepsilon_i) \middle| D_i = d, S_i = s, Z_i = z, \mathcal{F}_i = n, S_i^* = s^*, \mathcal{F}_i^* = n^*\right] \\
&\times f_{S_i^*, \mathcal{F}_i^* | D_i = d, S_i = s, Z_i = z, \mathcal{F}_i = n}(s^*, n^*). \qquad (4.9.54)
\end{aligned}$$

Based on Lemma 4.9.13 that $\varepsilon_i \perp (D_i, S_i^*, S_i, \mathcal{F}_i) | Z_i, \mathcal{F}_i^*$, we have that (4.9.54) becomes to

$$\begin{aligned}
&m_i(d, s, z, n) \\
=&\sum_{(s^*, n^*) \in \Omega_{S^*, \mathcal{F}^*}} \mathbb{E}\left[r(d, s^*, z, n^*, \varepsilon_i) \middle| Z_i = z, \mathcal{F}_i^* = n^*\right] f_{S_i^*, \mathcal{F}_i^* | D_i = d, S_i = s, \mathcal{F}_i = n, Z_i = z}(s^*, n^*) \\
=&\sum_{(s^*, n^*) \in \Omega_{S^*, \mathcal{F}^*}} m^*(d, s^*, z, n^*) f_{S_i^*, \mathcal{F}_i^* | D_i = d, S_i = s, \mathcal{F}_i = n, Z_i = z}(s^*, n^*), \qquad (4.9.55)
\end{aligned}$$

where the last equality is by Definition 4.3.1. ■

Proofs of Section 4.4

Lemma 4.9.14 *Under Assumption 4.3.2 (a), suppose Assumption 4.3.4 (a) and (b) are satisfied by both $\mathcal{N}_i$ and $\tilde{\mathcal{N}}_i$. Then, $Y_i \perp (\mathcal{F}_i, \tilde{\mathcal{F}}_i) \big| Z_i, \mathcal{F}_i^*$ holds.*

Proof of Lemma 4.9.14. First, same arguments used in the proof of Lemma 4.4.1 (a) can be applied to show that $S_i^* \perp (\mathcal{F}_i, \tilde{\mathcal{F}}_i) \big| Z_i, \mathcal{F}_i^*$. Second, rewrite Y_i in terms of the potential outcomes:

$$Y_i = \sum_{(d,s)\in\{0,1\}\times\Omega_{S^*}} 1[D_i = d, S_i^* = s] r\left(d, s, Z_i, \mathcal{F}_i^*, \varepsilon_i\right),$$

where by the randomness of the treatment assignment and $S_i^* \perp (\mathcal{F}_i, \tilde{\mathcal{F}}_i) \big| Z_i, \mathcal{F}_i^*$, we know that $1[D_i = d, S_i^* = s] \perp (\mathcal{F}_i, \tilde{\mathcal{F}}_i) \big| Z_i, \mathcal{F}_i^*$. Then, because Assumption 4.3.4 (b) implies that $(\mathcal{F}_i, \tilde{\mathcal{F}}_i)$ is independent to the potential outcome $r(d, s, Z_i, \mathcal{F}_i^*, \varepsilon_i)$ given $(Z_i, \mathcal{F}_i^*)$, we can conclude that $Y_i \perp (\mathcal{F}_i, \tilde{\mathcal{F}}_i) \big| Z_i, \mathcal{F}_i^*$. ■

Proof of Lemma 4.4.1. (a) If we can show that for any $(s, \mathcal{J}) \in \Omega_{S^*,\mathcal{N}^*}$ the equation below holds,

$$f_{S_i^*,\mathcal{N}_i^*|\mathcal{F}_i^*=n,Z_i}(s, \mathcal{J}) = f_{S_i^*|\mathcal{F}_i^*=n,Z_i}(s) \times f_{\mathcal{N}_i^*|\mathcal{F}_i^*=n,Z_i}(\mathcal{J}), \tag{4.9.56}$$

then the desired result follows. First of all, if either $s > n$ or $|\mathcal{J}| \neq n$, (4.9.56) holds trivially. Therefore, we consider $(s, \mathcal{J})$ such that $s \leq n$ and $|\mathcal{J}| = n$. Because for any fixed $\mathcal{J}$, $\{D_j\}_{j\in\mathcal{J}}$ is independent to $(Z_i, \mathcal{F}_i^*, \mathcal{N}_i^*)$ by Assumption 4.3.2 (a), then by i.i.d. of D_i

$$f_{S_i^*|\mathcal{F}_i^*=n,\mathcal{N}_i^*=\mathcal{J},Z_i}(s) = f_{\sum_{j\in\mathcal{J}} D_j|\mathcal{F}_i^*=n,\mathcal{N}_i^*=\mathcal{J},Z_i}(s) = f_{\sum_{j\in\mathcal{J}} D_j}(s) = C_n^s f_D^s(1) f_D(0)^{(n-s)}. \tag{4.9.57}$$

where $f_D(d) = \Pr(D_i = d)$ with $d \in \{0, 1\}$. On the other hand, by the law of total probability,

$$f_{S_i^*|\mathcal{F}_i^*=n,Z_i}(s) = \sum_{\mathcal{J}\in\Omega_{\mathcal{N}^*},\ s.t.\ |\mathcal{J}|=n} f_{\sum_{j\in\mathcal{J}} D_j|\mathcal{F}_i^*=n,\mathcal{N}_i^*=\mathcal{J},Z_i}(s) \times f_{\mathcal{N}_i^*|\mathcal{F}_i^*=n,Z_i}(\mathcal{J})$$

$$
\begin{aligned}
&= \sum_{\mathcal{J}\in\Omega_{\mathcal{N}^*},\ s.t.\ |\mathcal{J}|=n} f_{\sum_{j\in\mathcal{J}} D_j}(s) f_{\mathcal{N}_i^*|\mathcal{F}_i^*=n,Z_i}(\mathcal{J}) \\
&= C_n^s f_D^s(1) f_D(0)^{(n-s)}. \qquad (4.9.58)
\end{aligned}
$$

Therefore, (4.9.57) and (4.9.58) lead to

$$
\begin{aligned}
f_{S_i^*,\mathcal{N}_i^*|\mathcal{F}_i^*=n,Z_i}(s,\mathcal{J}) &= f_{S_i^*|\mathcal{F}_i^*=n,\mathcal{N}_i^*=\mathcal{J},Z_i}(s) \times f_{\mathcal{N}_i^*|\mathcal{F}_i^*=n,Z_i}(\mathcal{J}) \\
&= f_{S_i^*|\mathcal{F}_i^*=n,Z_i}(s) \times f_{\mathcal{N}_i^*|\mathcal{F}_i^*=n,Z_i}(\mathcal{J}).
\end{aligned}
$$

In addition, due to $S_i^* = \sum_{j\in\mathcal{N}_i^*} D_j$ and Assumption 4.3.4 (a), it is easy to see that $\mathcal{F}_i \perp S_i^*|Z_i,\mathcal{N}_i^*$. Thus, similar arguments used to show (4.9.60) give us

$$
\begin{aligned}
f_{S_i^*|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s) &= \sum_{\mathcal{J}\in\Omega_{\mathcal{N}^*},\ s.t.\ |\mathcal{J}|=n^*} f_{S_i^*|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,\mathcal{N}_i^*=\mathcal{J},Z_i}(s) \times f_{\mathcal{N}_i^*|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(\mathcal{J}) \\
&= \sum_{\mathcal{J}\in\Omega_{\mathcal{N}^*},\ s.t.\ |\mathcal{J}|=n^*} f_{S_i^*|\mathcal{F}_i^*=n^*,\mathcal{N}_i^*=\mathcal{J},Z_i}(s) \times f_{\mathcal{N}_i^*|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(\mathcal{J}) \\
&= f_{S_i^*|\mathcal{F}_i^*=n^*,Z_i}(s) \sum_{\mathcal{J}\in\Omega_{\mathcal{N}^*},\ s.t.\ |\mathcal{J}|=n^*} f_{\mathcal{N}_i^*|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(\mathcal{J}) \\
&= f_{S_i^*|\mathcal{F}_i^*=n^*,Z_i}(s), \qquad (4.9.59)
\end{aligned}
$$

where the second equality is due to $\mathcal{F}_i \perp S_i^*|Z_i,\mathcal{N}_i^*$, the third equality is because of $\mathcal{N}_i^* \perp S_i^*|Z_i,\mathcal{F}_i^*$ in proof (a). Hence, (4.9.59) permits that $\mathcal{F}_i \perp S_i^*|Z_i,\mathcal{F}_i^*$.

(b) Given $S_i = \sum_{j\in\mathcal{N}_i} D_j$, according to Assumptions 4.3.2 (a) and 4.3.4 (a), $\{D_i\}_{i\in\mathcal{P}}$ are i.i.d. and independent to $(Z_i,\mathcal{N}_i)$. Thus, applying the same arguments used to show (a), we can obtain for $s \le n$, $f_{S_i|\mathcal{F}_i=n,\mathcal{N}_i,Z_i}(s) = f_{S_i|\mathcal{F}_i=n,Z_i}(s) = C_n^s f_D^s(1) f_D(0)^{(n-s)}$, leading to $\mathcal{N}_i \perp S_i|Z_i,\mathcal{F}_i$.

Moreover, because $S_i = \sum_{j\in\mathcal{N}_i} D_j$ is a function of $(\mathcal{N}_i,\{D_j\}_{j\in\mathcal{N}_i})$, the randomness of S_i given $\mathcal{N}_i$ only comes from D_j for $j \in \mathcal{N}_i$. In addition, since $\{D_j\}_{j\in\mathcal{P}}$ are independent to $(Z_i,\mathcal{N}_i^*,\mathcal{N}_i)$ as in Assumption 4.3.4 (a), it implies that $\mathcal{F}_i^* \perp S_i|Z_i,\mathcal{N}_i$. Hence, for

$\forall(s,n,n^*) \in \Omega_{S,\mathcal{F},\mathcal{F}^*}$,

$$\begin{aligned} f_{S_i|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s) &= \sum_{\mathcal{J}\in\Omega_{\mathcal{N}},\ s.t.\ |\mathcal{J}|=n} f_{S_i|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,\mathcal{N}_i=\mathcal{J},Z_i}(s) \times f_{\mathcal{N}_i|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(\mathcal{J}) \\ &= \sum_{\mathcal{J}\in\Omega_{\mathcal{N}},\ s.t.\ |\mathcal{J}|=n} f_{S_i|\mathcal{F}_i=n,\mathcal{N}_i=\mathcal{J},Z_i}(s) \times f_{\mathcal{N}_i|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(\mathcal{J}) \\ &= f_{S_i|\mathcal{F}_i=n,Z_i}(s) \sum_{\mathcal{J}\in\Omega_{\mathcal{N}},\ s.t.\ |\mathcal{J}|=n} f_{\mathcal{N}_i|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(\mathcal{J}) \\ &= f_{S_i|\mathcal{F}_i=n,Z_i}(s), \end{aligned} \tag{4.9.60}$$

where the second equality is because $\mathcal{F}_i^* \perp S_i|Z_i,\mathcal{N}_i$, the third equality is due to $\mathcal{N}_i \perp S_i|Z_i,\mathcal{F}_i$ as shown at the beginning of this proof. Therefore, $S_i \perp \mathcal{F}_i^*|Z_i,\mathcal{F}_i$ from (4.9.60).

(c) The proof in this step follows directly from the proofs in (a) and (b). ■

Proof of Proposition 4.4.2. Recall that by Bayes' Theorem, we have

$$f_{S_i^*,\mathcal{F}_i^*|D_i,S_i,\mathcal{F}_i,Z_i} = \frac{f_{S_i,\mathcal{F}_i|D_i,S_i^*,\mathcal{F}_i^*,Z_i} \times f_{S_i^*,\mathcal{F}_i^*|D_i,Z_i}}{f_{S_i,\mathcal{F}_i|D_i,Z_i}}. \tag{4.9.61}$$

In what follows, we further rewrite the distributions in the numerator and the denominator to achieve the desired result. Based on Assumptions 4.3.2 and 4.3.4, we know that $\{D_i\}_{i\in\mathcal{P}}$ is i.i.d. and independent to $\{Z_i,\mathcal{N}_i^*,\mathcal{N}_i\}_{i\in\mathcal{P}}$. Thus, from the fact that $i \notin \mathcal{N}_i^*$ and $i \notin \mathcal{N}_i$, we can conclude that

$$D_i \perp (S_i^*, S_i, Z_i, \mathcal{N}_i^*, \mathcal{N}_i), \quad \text{for } \forall i \in \mathcal{P}. \tag{4.9.62}$$

It further yields that $D_i \perp S_i\big|(S_i^*, Z_i, \mathcal{F}_i^*, \mathcal{F}_i)$ and $D_i \perp \mathcal{F}_i\big|(S_i^*, Z_i, \mathcal{F}_i^*)$. Therefore, consider the first term in the numerator, for any $(s,n) \in \Omega_{S,\mathcal{F}}$

$$\begin{aligned} f_{S_i,\mathcal{F}_i|D_i,S_i^*,\mathcal{F}_i^*,Z_i}(s,n) &= f_{S_i|D_i,S_i^*,\mathcal{F}_i^*,\mathcal{F}_i=n,Z_i}(s) \times f_{\mathcal{F}_i|D_i,S_i^*,\mathcal{F}_i^*,Z_i}(n) \\ &= f_{S_i|S_i^*,\mathcal{F}_i^*,\mathcal{F}_i=n,Z_i}(s) \times f_{\mathcal{F}_i|S_i^*,\mathcal{F}_i^*,Z_i}(n) \\ &= f_{S_i|S_i^*,\mathcal{F}_i^*,\mathcal{F}_i=n,Z_i}(s) \times f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}(n), \end{aligned} \tag{4.9.63}$$

where the last equality is because of $\mathcal{F}_i \perp S_i^*\big|Z_i,\mathcal{F}_i^*$ in Lemma 4.4.1. Besides, again by

(4.9.62), we have $D_i \perp S_i^* \big| Z_i, \mathcal{F}_i^*$ and $D_i \perp \mathcal{F}_i^* \big| Z_i$. For the second term in the numerator,

$$\begin{aligned} f_{S_i^*,\mathcal{F}_i^*|D_i,Z_i}(s,n) =& f_{S_i^*|D_i,\mathcal{F}_i^*=n,Z_i}(s) \times f_{\mathcal{F}_i^*|D_i,Z_i}(n) \\ =& f_{S_i^*|\mathcal{F}_i^*=n,Z_i}(s) \times f_{\mathcal{F}_i^*|Z_i}(n). \end{aligned} \tag{4.9.64}$$

Similarly, by (4.9.62), we can rewrite the denominator

$$f_{S_i,\mathcal{F}_i|D_i,Z_i}(s,n) = f_{S_i|\mathcal{F}_i=n,Z_i}(s) \times f_{\mathcal{F}_i|Z_i}(n). \tag{4.9.65}$$

Now, substituting (4.9.63), (4.9.64) and (4.9.65) into (4.9.61) leads to the stated result.

■

Proof of Theorem 4.4.3. (a) Due to the Assumption 4.3.3, it is clear that

$$f_{Z_i},\ f_{\mathcal{F}_i^*|Z_i},\ f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*,Z_i},\ f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i},\ \text{are all identical across } i \in \mathcal{P}. \tag{4.9.66}$$

Now, according to $\tilde{\mathcal{F}}_i \perp \mathcal{F}_i | Z_i, \mathcal{F}_i^*$ in Assumption 4.4.1, we can obtain

$$f_{\mathcal{F}_i^*,\tilde{\mathcal{F}}_i,\mathcal{F}_i,Z_i} = f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i^*,Z_i} = f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i^*,Z_i}.$$

Because all the terms on the right hand side of the above equation are identical for all i, then $f_{\mathcal{F}_i^*,\tilde{\mathcal{F}}_i,\mathcal{F}_i,Z_i}$ is identical for all i, so as all the marginal and conditional distributions of $(\mathcal{F}_i^*, \tilde{\mathcal{F}}_i, \mathcal{F}_i, Z_i)$, which include $f_{\mathcal{F}_i|Z_i}$ and $f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|Z_i}$.

In addition, recall that $Y_i = r(D_i, S_i^*, \mathcal{F}_i^*, Z_i, \varepsilon_i)$ as in Assumption 4.3.1. By Lemma 4.9.13 and (4.9.62), we have $(\varepsilon_i, D_i) \perp (S_i^*, \tilde{\mathcal{F}}_i, \mathcal{F}_i) | Z_i, \mathcal{F}_i^*$. Moreover, from Lemma 4.4.1 we know that $S_i^* \perp (\tilde{\mathcal{F}}_i, \mathcal{F}_i) | Z_i, \mathcal{F}_i^*$. Therefore,

$$\begin{aligned} f_{\mathcal{F}_i^*,\tilde{\mathcal{F}}_i,\mathcal{F}_i,S_i^*,\varepsilon_i,D_i,Z_i} =& f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,S_i^*,\varepsilon_i,D_i|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i^*,Z_i} \\ =& f_D \times f_{\varepsilon_i|\mathcal{F}_i^*,Z_i} \times f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,S_i^*|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i^*,Z_i} \\ =& f_D \times f_{\varepsilon_i|\mathcal{F}_i^*,Z_i} \times f_{S_i^*|\mathcal{F}_i^*,Z_i} \times f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i^*,Z_i}. \end{aligned}$$

By the identical distribution of ε_i given $Z_i, \mathcal{F}_i^*$ in Assumption 4.3.3, and $f_{S_i^*|\mathcal{F}_i^*=n^*,Z_i}(s^*) = C_{n^*}^{s^*} f_D^{s^*}(1) f_D(0)^{(n^*-s^*)}$, together with (4.9.66), we can conclude that $(\mathcal{F}_i^*, \tilde{\mathcal{F}}_i, \mathcal{F}_i, S_i^*, \varepsilon_i, D_i, Z_i)$

is identically distributed and so as $(\mathcal{F}_i^*, \tilde{\mathcal{F}}_i, \mathcal{F}_i, Y_i, Z_i)$.

(b) In this proof, we first show that $f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*,Z_i}$ and $f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}$ are identified. We then verify the identification of $f_{\mathcal{F}_i^*|Z_i}$. By the law of total probability, for any $(\tilde{n}, n, y) \in \Omega_{\tilde{\mathcal{F}},\mathcal{F},Y}$

$$\begin{aligned}
&f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|Z_i}(\tilde{n}, n, y) \\
&= \sum_{n^* \in \Omega_{\mathcal{F}^*}} f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|\mathcal{F}_i^*=n^*,Z_i}(\tilde{n}, n, y) \times f_{\mathcal{F}_i^*|Z_i}(n^*) \\
&= \sum_{n^* \in \Omega_{\mathcal{F}^*}} f_{Y_i|\mathcal{F}_i^*=n^*,\tilde{\mathcal{F}}_i=\tilde{n},\mathcal{F}_i=n,Z_i}(y) \times f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(\tilde{n}, n) \times f_{\mathcal{F}_i^*|Z_i}(n^*) \\
&= \sum_{n^* \in \Omega_{\mathcal{F}^*}} f_{Y_i|\mathcal{F}_i^*=n^*,Z_i}(y) \times f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*=n^*,Z_i}(\tilde{n}) \times f_{\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(n) \times f_{\mathcal{F}_i^*|Z_i}(n^*), \qquad (4.9.67)
\end{aligned}$$

where the last equality is due to Assumption 4.4.1 and Lemma 4.9.14. Integrate both sides of (4.9.67)

$$\begin{aligned}
&\int_{y \in \Omega_Y} y f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|Z_i}(\tilde{n}, n, y) dy \\
&= \sum_{n^* \in \Omega_{\mathcal{F}^*}} \mathbb{E}[Y_i|\mathcal{F}_i^* = n^*, Z_i] \times f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*=n^*,Z_i}(\tilde{n}) \times f_{\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(n) \times f_{\mathcal{F}_i^*|Z_i}(n^*). \qquad (4.9.68)
\end{aligned}$$

Besides, for any $(\tilde{n}, n) \in \Omega_{\tilde{\mathcal{F}},\mathcal{F}}$, because of Assumption 4.4.1

$$\begin{aligned}
f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|Z_i}(\tilde{n}, n) &= \sum_{n^* \in \Omega_{\mathcal{F}^*}} f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(\tilde{n}, n) \times f_{\mathcal{F}_i^*|Z_i}(n^*) \\
&= \sum_{n^* \in \Omega_{\mathcal{F}^*}} f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*=n^*,Z_i}(\tilde{n}) \times f_{\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(n) \times f_{\mathcal{F}_i^*|Z_i}(n^*). \qquad (4.9.69)
\end{aligned}$$

Recall that the notations below from the main text: for $\forall y \in \Omega_Y$, the $K_\mathcal{F} \times K_\mathcal{F}$ matrices

$$E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} = \begin{bmatrix} \int_{y \in \Omega_Y} y f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|Z_i}(0,0,y)dy & \cdots & \int_{y \in \Omega_Y} y f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|Z_i}(0,K_\mathcal{F}-1,y)dy \\ \vdots & \ddots & \vdots \\ \int_{y \in \Omega_Y} y f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|Z_i}(K_\mathcal{F}-1,0,y)dy & \cdots & \int_{y \in \Omega_Y} y f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i,Y_i|Z_i}(K_\mathcal{F}-1,K_\mathcal{F}-1,y)dy \end{bmatrix},$$

$$F_{\tilde{\mathcal{F}},\mathcal{F}|Z} = \begin{bmatrix} f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|Z_i}(0,0) & \cdots & f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|Z_i}(0,K_\mathcal{F}-1) \\ \vdots & \ddots & \vdots \\ f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|Z_i}(K_\mathcal{F}-1,0) & \cdots & f_{\tilde{\mathcal{F}}_i,\mathcal{F}_i|Z_i}(K_\mathcal{F}-1,K_\mathcal{F}-1) \end{bmatrix}.$$

In addition, recall and denote two $K_{\mathcal{F}} \times K_{\mathcal{F}}$ diagonal matrices

$$\begin{aligned} T_{Y|\mathcal{F}^*,Z} =& diag\left(\mathbb{E}[Y_i|\mathcal{F}_i^* = 0, Z_i], \mathbb{E}[Y_i|\mathcal{F}_i^* = 1, Z_i], \cdots, \mathbb{E}[Y_i|\mathcal{F}_i^* = K_{\mathcal{F}} - 1, Z_i]\right), \\ T_{\mathcal{F}^*|Z} =& diag\left(f_{\mathcal{F}_i^*|Z_i}(0), f_{\mathcal{F}_i^*|Z_i}(1), \cdots, f_{\mathcal{F}_i^*|Z_i}(K_{\mathcal{F}} - 1)\right). \end{aligned}$$

Then, given the notations above, (4.9.67) and (4.9.69) can be rewritten in the following expressions:

$$E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} = F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z} \times T_{Y|\mathcal{F}^*,Z} \times T_{\mathcal{F}^*|Z} \times F'_{\mathcal{F}|\mathcal{F}^*,Z}, \tag{4.9.70}$$

$$F_{\tilde{\mathcal{F}},\mathcal{F}|Z} = F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z} \times T_{\mathcal{F}^*|Z} \times F'_{\mathcal{F}|\mathcal{F}^*,Z}, \tag{4.9.71}$$

where $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$ and $F_{\mathcal{F}|\mathcal{F}^*,Z}$ are defined in the main text. Based on Assumption 4.4.3, we know that $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$ and $F_{\mathcal{F}|\mathcal{F}^*,Z}$ are invertible. In addition, based on Assumption 4.4.4 (b), we have that $f_{\mathcal{F}_i^*|Z_i}(n) > \eta > 0$ for $\forall n \in \Omega_{\mathcal{F}^*}$ indicates the invertibility of $T_{\mathcal{F}^*|Z}$. Hence, (4.9.71) implies that $F_{\tilde{\mathcal{F}},\mathcal{F}|Z}$ is also invertible. It then yields from (4.9.70) and (4.9.71) that the square matrix $E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z}$ can be factorised as

$$E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z} = F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z} \times T_{Y|\mathcal{F}^*,Z} \times F^{-1}_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}, \tag{4.9.72}$$

where the matrix $E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z}$ on the left hand side of the above equation is identifiable from the observed data, and the right hand side corresponds to its eigen-decomposition, whose eigenvalues are the diagonal entries of $T_{Y|\mathcal{F}^*,Z}$.

By Assumption 4.4.4 (a), all the $K_{\mathcal{F}}$ eigenvalues in the diagonal matrix $T_{Y|\mathcal{F}^*,Z}$ are strictly positive and distinct. Thus, given the eigen-decomposition of matrix $E_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z}$ in (4.9.72), its $K_{\mathcal{F}}$ eigenvectors are linearly independent and are corresponding to the $K_{\mathcal{F}}$ columns of $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$. By simple algebra, we can solve the $K_{\mathcal{F}}$ eigenvectors, meaning that the columns of $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$ are identifiable. Moreover, Assumption 4.4.4 (b) ensures there is an unique maximum entry of each eigenvector, and its location reveals which eigenvalue it corresponds to. For example, if the largest value in some eigenvector appears in its first entry, then this eigenvector gives the latent probabilities $[f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*=0,Z_i}(0), f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*=0,Z_i}(1), ..., f_{\tilde{\mathcal{F}}_i|\mathcal{F}_i^*=0,Z_i}(K_{\mathcal{F}} - 1)]'$ and corresponds to the eigenvalue $\mathbb{E}[Y_i|\mathcal{F}_i^* = 0, Z_i]$. Because the summation of each column in the matrix $F_{\mathcal{F}|\mathcal{F}^*,Z}$ is natu-

rally normalised to be one, there is an unique solution for each eigenvector. The above discussions verify that $F_{\tilde{\mathcal{F}}|\mathcal{F}^*,Z}$ can be nonparametrically identified. Same arguments can be use to show the identification of $F_{\mathcal{F}|\mathcal{F}^*,Z}$.

Next, let us move on to $f_{\mathcal{F}_i^*|Z_i}$. Define two $K_\mathcal{F} \times 1$ vectors as

$$F_{\mathcal{F}^*|Z} = \left[f_{\mathcal{F}_i^*|Z_i}(0) \quad f_{\mathcal{F}_i^*|Z_i}(1) \quad \cdots \quad f_{\mathcal{F}_i^*|Z_i}(K_\mathcal{F}-1) \right]',$$
$$F_{\mathcal{F}|Z} = \left[f_{\mathcal{F}_i|Z_i}(0) \quad f_{\mathcal{F}_i|Z_i}(1) \quad \cdots \quad f_{\mathcal{F}_i|Z_i}(K_\mathcal{F}-1) \right]'.$$

Based on the law of total probability, it is easy to get $F_{\mathcal{F}|Z} = F_{\mathcal{F}|\mathcal{F}^*,Z} \times F_{\mathcal{F}^*|Z}$. Since $F_{\mathcal{F}|\mathcal{F}^*,Z}$ is invertible, multiplying both sides of the above equation by $F^{-1}_{\mathcal{F}|\mathcal{F}^*,Z}$ gives

$$F_{\mathcal{F}^*|Z} = F^{-1}_{\mathcal{F}|\mathcal{F}^*,Z} \times F_{\mathcal{F}|Z}, \tag{4.9.73}$$

which indicates the identifiability of $F_{\mathcal{F}^*|Z}$. ■

Proof of Lemma 4.4.4. Recall that $\Delta S_i := S_i - S_i^*$. (a) Consider the case $\mathcal{N}_i^* \subset \mathcal{N}_i$. For $\forall (s,n) \in \Omega_{S,\mathcal{F}}$ and $(s^*,n^*) \in \Omega_{S^*,\mathcal{F}^*}$ such that $n^* \leq n$

$$\begin{aligned} & f_{S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s) = f_{\Delta S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s-s^*) \\ = & \sum_{(\mathcal{J}^*,\mathcal{J}),\ s.t.\ \mathcal{J}^*\subset\mathcal{J},\ |\mathcal{J}^*|=n^*,|\mathcal{J}|=n} f_{\Delta S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,\mathcal{N}_i^*=\mathcal{J}^*,\mathcal{N}_i=\mathcal{J},Z_i}(s-s^*) \\ & \qquad \times f_{\mathcal{N}_i^*,\mathcal{N}_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(\mathcal{J}^*,\mathcal{J}), \end{aligned} \tag{4.9.74}$$

where the last line is based on the law of total probability. Because $\mathcal{N}_i^* \subset \mathcal{N}_i$, we have that $\mathcal{N}_i^*/\mathcal{N}_i$ is empty and $\Delta S_i = \sum_{j\in\mathcal{N}_i/\mathcal{N}_i^*} D_j$. In addition, $\mathcal{N}_i/\mathcal{N}_i^*$ and $\mathcal{N}_i^*$ are mutually exclusive, i.e. if $i \in \mathcal{N}_i/\mathcal{N}_i^*$ then $i \notin \mathcal{N}_i^*$. Due to the i.i.d. of $\{D_i\}_{i\in\mathcal{P}}$ (Assumption 4.3.2), and the independence between $\{D_i\}_{i\in\mathcal{P}}$ and $(Z_i,\mathcal{N}_i^*,\mathcal{N}_i)$ (Assumption 4.3.4), we have that $\Delta S_i \perp S_i^*|Z_i,\mathcal{N}_i^*,\mathcal{N}_i$. Therefore,

$$f_{\Delta S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,\mathcal{N}_i^*=\mathcal{J}^*,\mathcal{N}_i=\mathcal{J},Z_i}(s-s^*) = f_{\Delta S_i|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,\mathcal{N}_i^*=\mathcal{J}^*,\mathcal{N}_i=\mathcal{J},Z_i}(s-s^*). \tag{4.9.75}$$

Again by the independence of $\{D_i\}_{i\in\mathcal{P}}$, once conditional on $\mathcal{N}_i/\mathcal{N}_i^* = \mathcal{J}/\mathcal{J}^*$, we know

that $\Delta S_i = \sum_{j \in \mathcal{J}/\mathcal{J}^*} D_j$ follows a binomial distribution if $s^* \leq s$ and $\Delta s \leq \Delta n$, and is independent to the identity of network neighbours contained in $(\mathcal{N}_i^*, \mathcal{N}_i)$. Then (4.9.75) becomes to

$$\begin{aligned} & f_{\Delta S_i|\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,\mathcal{N}_i^*=\mathcal{J}^*,\mathcal{N}_i=\mathcal{J},Z_i}(s-s^*) \\ =& f_{\Delta S_i|\mathcal{F}_i^*=n^*,\mathcal{F}_i-\mathcal{F}_i^*=n-n^*,\mathcal{N}_i^*=\mathcal{J}^*,\mathcal{N}_i/\mathcal{N}_i^*=\mathcal{J}/\mathcal{J}^*,Z_i}(s-s^*) \\ =& f_{\Delta S_i|\mathcal{F}_i-\mathcal{F}_i^*=\Delta n,Z_i}(\Delta s), \end{aligned} \tag{4.9.76}$$

where the last equality follows the same arguments used to show Lemma 4.4.1 (a). Substituting (4.9.76) into (4.9.74) gives the desired result.

(b) Similar arguments used in proof for the case $\mathcal{N}_i^* \subset \mathcal{N}_i$ can be applied to obtain the result for the case $\mathcal{N}_i \subset \mathcal{N}_i^*$. Therefore, we omit the proof. ■

Proof of Theorem 4.4.5. (a) From Proposition 4.4.2, we know that

$$f_{S_i^*,\mathcal{F}_i^*|D_i,S_i,\mathcal{F}_i,Z_i} = \frac{f_{S_i|S_i^*,\mathcal{F}_i^*,\mathcal{F}_i,Z_i} \times f_{S_i^*|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i} \times f_{\mathcal{F}_i^*|Z_i}}{f_{S_i|\mathcal{F}_i,Z_i} \times f_{\mathcal{F}_i|Z_i}}. \tag{4.9.77}$$

Based on Lemma 4.4.1, we know that $f_{S_i^*|\mathcal{F}_i^*=n^*,Z_i}(s^*) = C_{n^*}^{s^*} f_D^{s^*}(1) f_D(0)^{(n^*-s^*)}$ and $f_{S_i|\mathcal{F}_i=n,Z_i}(s) = C_n^s f_D^s(1) f_D(0)^{(n-s)}$. Similarly, from Lemma 4.4.4,

$$f_{S_i|S_i^*=s^*,\mathcal{F}_i^*=n^*,\mathcal{F}_i=n,Z_i}(s) = C_{\Delta n}^{\Delta s} f_D^{\Delta s}(1) f_D(0)^{(\Delta n-\Delta s)}. \tag{4.9.78}$$

Because D_i is i.i.d., by Assumptions 4.3.3 and 4.3.4, $f_{\mathcal{F}_i^*|Z_i}$, $f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}$ and $f_{\mathcal{F}_i|Z_i}$ are identical for all i. Therefore, all distributions on the right hand side of (4.9.77) are identical across i, together with Theorem 4.4.3, $f_{S_i|S_i^*=s^*,\mathcal{F}_i=n,\mathcal{F}_i^*=n^*,Z_i}(s)$ can be nonparametrically identified. ■

Proof of Theorem 4.4.6. (a) From Theorem 4.4.5, we know that $f_{S_i^*,\mathcal{F}_i^*|D_i,S_i,\mathcal{F}_i,Z_i}$ is identical for all i, together with Proposition 4.3.3, we know m_i is also identical for all $i \in \mathcal{P}$.

(b) To ease the notation, denote $T_i = (S_i, \mathcal{F}_i)'$ and $T_i^* = (S_i^*, \mathcal{F}_i^*)'$. According to Assumption 4.4.2, the support of T_i and T_i^* are the same, and we denote it as $\Omega_T = \{t_1, t_2, ..., t_{K_T}\}$

with $t_k = (s_k, n_k) \in \Omega_{S,\mathcal{F}}$. Let us rank the possible values in Ω_T by the lexicographical ordering, according to the natural order of the integers in $\Omega_{S,\mathcal{F}}$, i.e.

$$\begin{aligned}
&t_1 = (0,0),\\
&t_2 = (0,1),\ t_3 = (1,1),\\
&t_4 = (0,2),\ t_5 = (1,2),\ t_6 = (2,2),\\
&\cdots\\
&t_{\frac{(K_{\mathcal{F}}-1)K_{\mathcal{F}}}{2}+1} = (0, K_{\mathcal{F}}-1), \cdots, t_{\frac{(K_{\mathcal{F}}-1)(K_{\mathcal{F}}+2)}{2}+1} = (K_{\mathcal{F}}-1, K_{\mathcal{F}}-1).
\end{aligned} \tag{4.9.79}$$

Because by result in (a), $m_i(\cdot)$ is identical for all i, thus we suppress the subscript i, i.e. $m(d,s,z,n) := \mathbb{E}[Y_i|D_i = d, S_i = s, Z_i = z, \mathcal{F}_i = n]$. By notation abuse, we ignore the arguments (d,z) in functions m and m^*, and introduce the following notations. For any $(d,z) \in \{0,1\} \times \Omega_Z$, denote $M_{Y|T,D=d,Z=z}$ and $M_{Y|T^*,D=d,Z=z}$ as two $K_T \times 1$ column vectors

$$M_{Y|T,D=d,Z=z}(m) = [m(t_1), m(t_2), \cdots, m(t_{K_T})]', \tag{4.9.80}$$

$$M_{Y|T^*,D=d,Z=z}(m^*) = [m^*(t_1), m^*(t_2), \cdots, m^*(t_{K_T})]', \tag{4.9.81}$$

where $m(t_k)$ represents the mean function $m(d, s_k, z, n_k) = \mathbb{E}[Y_i|D_i = d, S_i = s_k, Z_i = z, \mathcal{F}_i = n_k]$. Define the $K_T \times K_T$ matrix

$$F_{T^*|T,D=d,Z=z} = \begin{bmatrix} f_{T_i^*|T_i=t_1,D_i=d,Z_i=z}(t_1) & \cdots & f_{T_i^*|T_i=t_1,D_i=d,Z_i=z}(t_{K_T}) \\ \vdots & \ddots & \vdots \\ f_{T_i^*|T_i=t_{K_T},D_i=d,Z_i=z}(t_1) & \cdots & f_{T_i^*|T_i=t_{K_T},D_i=d,Z_i=z}(t_{K_T}) \end{bmatrix}. \tag{4.9.82}$$

From Proposition 4.3.3 and the notations in (4.9.80)-(4.9.82), we have for any $(d,z) \in \{0,1\} \times \Omega_Z$

$$M_{Y|T,D=d,Z=z}(m) = F_{T^*|T,D=d,Z=z} \times M_{Y|T^*,D=d,Z=z}(m^*). \tag{4.9.83}$$

Given Proposition 4.4.2, for $\forall (d,z) \in \{0,1\} \times \Omega_Z$, the elements in the main diagonal of

$F_{T^*|T,D=d,Z=z}$

$$
\begin{aligned}
&f_{S_i^*,\mathcal{F}_i^*|D_i=d,S_i=s,\mathcal{F}_i=n,Z_i=z}(s,n) \\
=&\frac{f_{S_i|S_i^*=s,\mathcal{F}_i^*=n,\mathcal{F}_i=n,Z_i=z}(s) \times f_{S_i^*|\mathcal{N}_i^*|=n,Z_i=z}(s) \times f_{\mathcal{F}_i^*|\mathcal{F}_i=n,Z_i=z}(n) \times f_{\mathcal{F}_i^*|Z_i=z}(n)}{f_{S_i|\mathcal{F}_i^*=n,Z_i=z}(s) \times f_{\mathcal{F}_i|Z_i=z}(n)} \\
=&\frac{f_{\mathcal{F}_i^*|\mathcal{F}_i=n,Z_i=z}(n) \times f_{\mathcal{F}_i^*|Z_i=z}(n)}{f_{\mathcal{F}_i|Z_i=z}(n)},
\end{aligned}
$$

where the second equality is because of Lemma 4.4.1 and Lemma 4.4.4. In addition, based on Assumption 4.4.4 (b), we know that $f_{\mathcal{F}_i^*|\mathcal{F}_i=n,Z_i=z}(n) > 0$, which also leads to $f_{\mathcal{F}_i^*|Z_i=z}(n) > 0$. Therefore, by the preassumption that $f_{\mathcal{F}_i|Z_i=z}(n) > 0$, we can conclude that

$$
f_{S_i^*,\mathcal{F}_i^*|D_i=d,S_i=s,\mathcal{F}_i=n,Z_i=z}(s,n) > 0 \quad \text{for } \forall (s,n) \in \Omega_{S^*,\mathcal{F}^*}. \tag{4.9.84}
$$

In what follows, we prove the desired result in two steps. Firstly, we show that the square matrix $F_{T^*|T,D=d,Z=z}$ is invertible. Secondly, we show that the CASF m^* is identifiable from (4.9.83).

Step 1. Consider any $t^* = (s^*, n^*)$ and $t = (s, n)$ such that $0 \leq s^* \leq n^*$ and $0 \leq s \leq n$. Under Assumption 4.4.5, we need to consider two cases.

Firstly, suppose $\mathcal{N}_i^* \subset \mathcal{N}_i$ holds. Then, we know that $S_i^* \leq S_i$ and $\mathcal{F}_i^* \leq \mathcal{F}_i$. Thus, $f_{T_i^*|T_i=t,D_i=d,Z_i=z}(t^*) = 0$ if at least one of the restrictions $s^* \leq s$ and $n^* \leq n$ is violated. Similarly, when $\mathcal{N}_i \subset \mathcal{N}_i^*$ holds, we have that $S_i \leq S_i^*$ and $\mathcal{F}_i \leq \mathcal{F}_i^*$. Then, $f_{T_i^*|T_i=t,D_i=d,Z_i=z}(t^*) = 0$ if at least one of the restrictions $s \leq s^*$ and $n \leq n^*$ is violated. Given the lexicographical ordering of the elements in Ω_T, it is easy to see that the matrix $F_{T^*|T,D=d,Z=z}$ is lower triangular if $\mathcal{N}_i^* \subset \mathcal{N}_i$, and is upper triangular if $\mathcal{N}_i \subset \mathcal{N}_i^*$. Moreover, (4.9.84) implies that all the elements on the main diagonal of the triangular matrix $F_{T^*|T,D=d,Z=z}$ are strictly positive. Since the eigenvalues of a triangular matrix are its diagonal entries, the matrix $F_{T^*|T,D=d,Z=z}$ is therefore invertible.

Step 2. Next, we show that the CASF m^* is identifiable. Suppose m^* is not identifiable, then there exists $\tilde{m}^* \neq m^*$ such that $\tilde{m}^*$ is observationally equivalent to m^*, in the sense

that (4.9.83) also holds for $\tilde{m}^*$:

$$M_{Y|T,D=d,Z=z}(m) = F_{T^*|T,D=d,Z=z} M_{Y|T^*,D=d,Z=z}(\tilde{m}^*). \tag{4.9.85}$$

It then yields from (4.9.83) and (4.9.85) that

$$0 = F_{T^*|T,D=d,Z=z} \left[M_{Y|T^*,D=d,Z=z}(m^*) - M_{Y|T^*,D=d,Z=z}(\tilde{m}^*) \right]. \tag{4.9.86}$$

Since $F_{T^*|T,D=d,Z=z}$ is invertible, it follows from (4.9.86) that

$$M_{Y|T^*,D=d,Z=z}(m^*) = M_{Y|T^*,D=d,Z=z}(\tilde{m}^*),$$

meaning that $\tilde{m}^*(t_k) = m^*(t_k)$ for all $k = 1, 2, ..., K_T$, which contradicts $\tilde{m}^* \neq m^*$. Therefore, we can conclude that m^* is identifiable. ■

Proofs of Section 4.5

Proof of Theorem 4.5.2. For illustration simplicity, by notation abuse, we denote W_i as any generic vector of observable variables of interest, where $W_i = (W_i^{c\prime}, W_i^{d\prime})' \in \Omega_{W^c} \times \Omega_{W^d}$, with the $Q \times 1$ vector $W_i^c := (W_{i1}^c, ..., W_{iQ}^c)'$ containing continuous variables and the vector W_i^d containing discrete variables. In this proof, we focus on the uniform convergence rate of the kernel estimation $\hat{f}_{W_i}(w)$. Then, replacing W_i by the observable variables of interest gives the stated results.

Denote $w = (w^{c\prime}, w^{d\prime})'$ with $w^c = (w_1^c, ..., w_Q^c)'$ and $\hat{f}_{W_i}(w) = 1/N \sum_{i=1}^N \hat{f}_i^{ker}(w)$, where

$$\hat{f}_i^{ker}(w) := K(W_i^c, w^c) 1\left[W_i^d = w^d\right], \tag{4.9.87}$$

with $K(W_i^c, w^c) = h^{-Q} \prod_{q=1}^Q \kappa\left((W_{iq}^c - w_q^c)/h\right)$. Let $f_{W_i}(w)$ be the true distribution of W_i. For any $w \in \Omega_W$,

$$\left|\hat{f}_{W_i}(w) - f_{W_i}(w)\right| \leq \left|\hat{f}_{W_i}(w) - \mathbb{E}\left[\hat{f}_{W_i}(w)\right]\right| + \left|\mathbb{E}\left[\hat{f}_{W_i}(w)\right] - f_{W_i}(w)\right|.$$

Given the inequality above, we prove the uniform convergence of $\left|\hat{f}_{W_i}(w) - f_{W_i}(w)\right|$ and its rate in two steps. In Step 1, we show that the bias of $\hat{f}_{W_i}(w)$, i.e. $|\mathbb{E}[\hat{f}_{W_i}(w)] - f_{W_i}(w)|$,

is $O(h^2)$ uniformly. In Step 2, we show the uniform convergence of $\hat{f}_{W_i}(w)$ to $\mathbb{E}[\hat{f}_{W_i}(w)]$ and establish its convergence rate.

Step 1. Firstly, let w^{d*} and $w^{c*} := (w_1^{c*}, ..., w_Q^{c*})'$ be any generic element in Ω_{W^d} and Ω_{W^c}, respectively. Then, for $w = (w^{c'}, w^{d'})'$

$$\mathbb{E}\left[\hat{f}_i^{ker}(w)\right] = h^{-Q} \sum_{w^{d*} \in \Omega_{W^d}} \left[1[w^{d*} = w^d] \int \prod_{q=1}^{Q} \kappa\left(\frac{w_q^{c*} - w_q^c}{h}\right) f_{W_i^c, W_i^d}\left(w^{c*}, w^{d*}\right) dw^{c*} \right],$$

by changing of variables using $v = (v_1, ..., v_Q)'$ with $v_q = (w_q^{c*} - w_q^c)/h$ and $q = 1, ..., Q$,

$$\begin{aligned}\mathbb{E}\left[\hat{f}_i^{ker}(w)\right] &= \sum_{w^{d*} \in \Omega_{W^d}} \left[1[w^{d*} = w^d] \int \prod_{q=1}^{Q} \kappa(v_q) f_{W_i^c, W_i^d}(w^c + hv, w^{d*}) dv \right] \\ &= \int f_{W_i^c, W_i^d}(w^c + hv, w^d) \prod_{q=1}^{Q} \kappa(v_q) dv, \end{aligned} \tag{4.9.88}$$

where we denote the shorthand notation $w^c + hv := (w_1^c + hv_1, ..., w_Q^c + hv_Q)$. Let the $Q \times 1$ vector $f_c^{(1)}(w) := \partial f_{W_i}(w)/\partial w^c$ represent the first order derivative of $f_{W_i}(w)$ with respect to w^c, and let the $Q \times Q$ matrix $f_c^{(2)}(w) := \partial^2 f_{W_i}(w)/\partial w^c \partial w^{c'}$ be the second order derivative of f_{W_i} with respect to w^c. Consider the Taylor series expansion of $f_{W_i^c, W_i^d}(w^c + hv, w^d)$ around w:

$$f_{W_i^c, W_i^d}(w^c + hv, w^d) - f_{W_i^c, W_i^d}(w^c, w^d) = h f_c^{(1)}(w)' v + h^2 v' f_c^{(2)}(\tilde{w}) v \tag{4.9.89}$$

where $\tilde{w}$ is between $(w^c + hv, w^d)$ and (w^c, w^d). Since W_i is identically distributed based on Theorems 4.4.3 and 4.4.5, we have $\mathbb{E}[\hat{f}_{W_i}(w)] = \mathbb{E}[\hat{f}_i^{ker}(w)]$. Plugging (4.9.89) into (4.9.88) gives

$$\begin{aligned}\mathbb{E}\left[\hat{f}_{W_i}(w)\right] - f_{W_i}(w) &= \int \left[h f_c^{(1)}(w)' v + h^2 v' f_c^{(2)}(\tilde{w}) v \right] \prod_{q=1}^{Q} \kappa(v_q) dv \\ &= h f_c^{(1)}(w)' \int v \prod_{q=1}^{Q} \kappa(v_q) dv + h^2 \int v' f_c^{(2)}(\tilde{w}) v \prod_{q=1}^{Q} \kappa(v_q) dv \end{aligned}$$

$$\begin{aligned}
\leq & Ch^2 \int v'v \prod_{q=1}^{Q} \kappa(v_q) dv \\
= & Ch^2 \sum_{q=1}^{Q} \int v_q^2 \kappa(v_q) dv_q, \qquad (4.9.90)
\end{aligned}$$

where the inequality is because that each element in $f_c^{(2)}$ is bounded uniformly in w^c, and the symmetric kernel function $\kappa(\cdot)$ in Assumption 4.5.2 (c) implies $\int \kappa(v_q) v_q dv_q = 0$, thus $\int v \prod_{q=1}^{Q} \kappa(v_q) dv = (\int v_1 \kappa(v_1) dv_1, ..., \int v_Q \kappa(v_Q) dv_Q)' = (0, ..., 0)'$. From (4.9.90), we get

$$\sup_{w \in \Omega_W} \left| \mathbb{E}\left[\hat{f}_{W_i}(w) \right] - f_{W_i}(w) \right| \leq \sup_{w \in \Omega_W} \left| Ch^2 \sum_{q=1}^{Q} \int \kappa(v_q) v_q^2 dv_q \right| \leq CK_1 Q h^2 = O(h^2). \qquad (4.9.91)$$

Step 2. Next, we show the uniform convergence of $|\hat{f}_{W_i}(w) - \mathbb{E}[\hat{f}_{W_i}(w)]|$. Since Ω_{W^c} is compact and Ω_{W^d} has finite dimension as in Assumption 4.5.2 (a), for some constant $C > 0$, Ω_W can be covered by less than $L_N = C l_N^{-Q}$ open balls of radius l_N, where for any $w = (w^{c'}, w^{d'})', \tilde{w} = (\tilde{w}^{c'}, \tilde{w}^{d'})'$ in the same ball, we let $w^d = \tilde{w}^d$. Denote the centers of these open balls as $\bar{w}_{j\epsilon}$ with $j = 1, 2, ..., J(\epsilon)$ and $J(\epsilon) \leq L_N$. For any $w, \tilde{w}$ in the same ball, the mean value theorem implies that

$$\begin{aligned}
\sup_{\|w - \tilde{w}\| < \epsilon} \left| \hat{f}_{W_i}(w) - \hat{f}_{W_i}(\tilde{w}) \right| \leq & \sup_{\|w - \tilde{w}\| < \epsilon} \frac{1}{N} \sum_{i=1}^{N} |K_W(W_i^c, w^c) - K_W(W_i^c, \tilde{w}^c)| \\
= & \sup_{\|w - \tilde{w}\| < \epsilon} \frac{1}{N h^Q} \sum_{i=1}^{N} \left| \prod_{q=1}^{Q} \kappa\left(\frac{W_{iq}^c - w_q^c}{h} \right) - \prod_{q=1}^{Q} \kappa\left(\frac{W_{iq}^c - \tilde{w}_q^c}{h} \right) \right| \\
\leq & \sup_{\|w - \tilde{w}\| < \epsilon} \frac{1}{N h^{Q+1}} \sum_{i=1}^{N} \left| \tilde{\kappa}'\left(w_h^{c*} \right) \right| \|w^c - \tilde{w}^c\| \\
\leq & C l_N h^{-(Q+1)}, \qquad (4.9.92)
\end{aligned}$$

where w_h^{c*} denotes some intermediate value between $(W_i^c - w^c)/h$ and $(W_i^c - \tilde{w}^c)/h$, and $\tilde{\kappa}'(v)$ represents the first order derivative of $\prod_{q=1}^{Q} \kappa(v_q)$ to $v = (v_1, ..., v_Q)'$. The last line of (4.9.92) is because of the boundedness of $\kappa(\cdot)$ and the uniform boundedness of its first order derivative (Assumption 4.5.2). Let $\bar{w}_{j\epsilon}$ denote the center of an open ball containing

w. Then,

$$\begin{aligned}\sup_{w\in\Omega_W}\left|\hat{f}_{W_i}(w)-\mathbb{E}[\hat{f}_{W_i}(w)]\right| \leq & \max_{1\leq j\leq L_N}\sup_{\|w-\bar{w}_{j\epsilon}\|<\epsilon}\left|\hat{f}_{W_i}(w)-\hat{f}_{W_i}(\bar{w}_{j\epsilon})\right| \\ & + \max_{1\leq j\leq L_N}\left|\hat{f}_{W_i}(\bar{w}_{j\epsilon})-\mathbb{E}[\hat{f}_{W_i}(\bar{w}_{j\epsilon})]\right| \\ & + \max_{1\leq j\leq L_N}\sup_{\|w-\bar{w}_{j\epsilon}\|<\epsilon}\left|\mathbb{E}[\hat{f}_{W_i}(w)]-\mathbb{E}[\hat{f}_{W_i}(\bar{w}_{j\epsilon})]\right| \\ := & R_1+R_2+R_3. \end{aligned} \tag{4.9.93}$$

By (4.9.92), we find immediately that R_1 and R_3 can be bounded as below

$$R_1 \leq C_1 l_N h^{-(Q+1)}, \text{ and } R_3 \leq C_3 l_N h^{-(Q+1)}, \tag{4.9.94}$$

for some constants C_1, C_3. The main task is then to find the convergence rate of R_2. Denote

$$Q_{N,i} := Q_{N,i}(w) = (\hat{f}_i^{ker}(w) - \mathbb{E}[\hat{f}_i^{ker}(w)])/N,$$

where to east the notation, we suppress the argument w in $Q_{N,i}(w)$. Then, $\hat{f}_{W_i}(w) - \mathbb{E}[\hat{f}_{W_i}(w)] = \sum_{i=1}^N Q_{N,i}$. Following the method of Masry (1996), which aims at approximating dependent random variables by independent ones, we further divide the proof for R_2 into two parts:

- Step 2.1 construct the approximation process;
- Step 2.2 shows that the independent random variable approximation converges uniformly and verifies the uniform convergence for the reminder term.

Step 2.1. Recall that $\mathbb{S}_1, ..., \mathbb{S}_{q_N}$ are the mutually exclusive partitions of index set $\{1, 2, ..., N\}$ with $\bigcup_{l=1,...,q_N} \mathbb{S}_l = \{1, 2, ..., N\}$. Define $V_N(k) = \sum_{i\in\mathbb{S}_k} Q_{N,i}$, for $k = 1, ..., q_N$ and

$$\begin{cases} W_N' = \sum_{k=1}^{q_N/2} V_N(2k-1), \quad W_N'' = \sum_{k=1}^{q_N/2} V_N(2k), & \text{if } q_N \text{ is even} \\ W_N' = \sum_{k=1}^{(q_N+1)/2} V_N(2k-1), \quad W_N'' = \sum_{k=1}^{(q_N-1)/2} V_N(2k), & \text{if } q_N \text{ is odd} \end{cases}$$

so that $\hat{f}_{W_i}(w) - \mathbb{E}[\hat{f}_{W_i}(w)] = W'_N + W''_N$ with W'_N and W''_N are the sums of $Q_{N,i}$ over the odd-numbered subsets $\{\mathbb{S}_{2k-1}\}$ and even-numbered subsets $\{\mathbb{S}_{2k}\}$, respectively. Then, for any $\eta > 0$,

$$\begin{aligned}\Pr(R_2 > \eta) \leq & \Pr\left(\max_{1\leq j\leq L_N} |W'_N(\bar{w}_{j\epsilon})| > \eta/2\right) + \Pr\left(\max_{1\leq j\leq L_N} \left|W''_N(\bar{w}_{j\epsilon})\right| > \eta/2\right) \\ \leq & 2L_N \sup_{w\in\Omega_W} \Pr\left(|W'_N(w)| > \eta/2\right). \end{aligned} \tag{4.9.95}$$

Next, we bound $\Pr\left(|W'_N(w)| > \eta/2\right)$ by applying Lemma 4.9.4 and approximating the odd-numbered $\{V_N(2k-1)\}$ series by independent random variables. Enlarging the probability space if necessary, let us introduce a random variable sequence $\{U_1, U_2, ...\}$ of mutually independent uniform $[0,1]$ random variables, which is also independent to the odd-numbered sequence $\{V_N(2k-1)\}$. Define $V^*_N(0) = 0$ and $V^*_N(1) = V_N(1)$. Then by Lemma 4.9.4, for each $k \geq 2$, there is a random variable $V^*_N(2k-1)$ that is a measurable function of $\{V_N(1), V_N(3), ...V_N(2k-1), U_k\}$ satisfying the three conditions below:

(a) $V^*_N(2k-1)$ is independent of $\{V_N(1), V_N(3), ..., V_N(2k-3)\}$;

(b) $V^*_N(2k-1)$ has the same distribution as $V_N(2k-1)$;

(c) for any μ such that $0 < \mu \leq \|V_N(2k-1)\|_\infty < \infty$,

$$\begin{aligned}&\Pr(|V^*_N(2k-1) - V_N(2k-1)| > \mu) \\ &\qquad \leq 18(\|V_N(2k-1)\|_\infty/\mu)^{1/2} \sup|\Pr(AB) - \Pr(A)\Pr(B)|, \end{aligned} \tag{4.9.96}$$

where the inequality follows by setting the γ in Lemma 4.9.4 as infinity, and the supremum is over all possible sets A and B, for A, B in the σ-field of events generated by $\{V_N(1), V_N(3), ..., V_N(2k-3)\}$ and by $V_N(2k-1)$, respectively. Most importantly, such construction of $V^*_N(2k-1)$ guarantees that $V^*_N(1), V^*_N(3), ..., V^*_N(2k-1)$ are mutually independent with each other based on condition (a) above. Up to here, we have established the approximation of the dependent random sequence $\{V_N(2k-1)\}$ by the independent one $\{V^*_N(2k-1)\}$.

Step 2.2. Without loss of generality, let q_N be an even number. Then,

$$\begin{aligned}
&\Pr\left(|W_N'(w)| > \eta/2\right) \\
=&\Pr\left(\left|\sum_{k=1}^{q_N/2}\left[V_N(2k-1) - V_N^*(2k-1)\right] + \sum_{k=1}^{q_N/2} V_N^*(2k-1)\right| > \eta/2\right) \\
\leq&\Pr\left(\left|\sum_{k=1}^{q_N/2} V_N^*(2k-1)\right| > \eta/4\right) + \Pr\left(\left|\sum_{k=1}^{q_N/2}\left[V_N(2k-1) - V_N^*(2k-1)\right]\right| > \eta/4\right) \\
:=&R_{21}(w) + R_{22}(w).
\end{aligned} \tag{4.9.97}$$

Firstly, we bound $R_{21}(w)$ as follows. Denote $r_i = |\Delta(i,N)|$, then $\bar{r}_N = \sup_{1\leq i\leq N} r_i$. Noting that $\kappa(\cdot)$ is bounded, let $\sup_{w^c\in\Omega_{W^c}}|\prod_{q=1}^{Q}\kappa(v_q)| = A_1$ for some constant $A_1 > 0$. Then, by construction,

$$|Q_{N,i}(w)| \leq 2A_1(Nh^Q)^{-1}, \text{ and } |V_N(k)| \leq 2r_kA_1(Nh^Q)^{-1} \leq 2\bar{r}_NA_1(Nh^Q)^{-1}. \tag{4.9.98}$$

Let $\lambda_N = C[Nh^Q\ln(N)]^{1/2}$ and we have that for N large enough, by choosing C properly,

$$\lambda_N|V_N(k)| = 2CA_1\bar{r}_N\left(\frac{\ln(N)}{Nh^Q}\right)^{1/2} \leq 1/2,$$

because of $\bar{r}_N[\ln(N)/(Nh^Q)]^{1/2} = O(1)$ in Assumption 4.5.2. By the inequality that $\exp(x) \leq 1 + x + x^2$ when $|x| \leq 1/2$, we can get

$$\exp\left(\pm\lambda_NV_N(2k-1)\right) \leq 1 \pm \lambda_NV_N(2k-1) + \lambda_N^2V_N^2(2k-1).$$

Thus, it yields from $\mathbb{E}[\lambda_NV_N(2k-1)] = 0$ and the same distribution of $V_N^*(2k-1)$ and $V_N(2k-1)$ that

$$\mathbb{E}\left[\exp(\pm\lambda_NV_N^*(2k-1))\right] = \mathbb{E}\left[\exp(\pm\lambda_NV_N(2k-1))\right] \leq 1 + \lambda_N^2\mathbb{E}[V_N^2(2k-1)]. \tag{4.9.99}$$

Moreover, because $1 + x \leq \exp(x)$ for $x \geq 0$, let $x = \mathbb{E}\left[\lambda_N^2V_N^2(2k-1)\right]$ we have

$$\mathbb{E}\left[\exp(\pm\lambda_NV_N^*(2k-1))\right] \leq \exp\left(\mathbb{E}\left[\lambda_N^2V_N^2(2k-1)\right]\right) = \exp\left(\mathbb{E}\left[\lambda_N^2V_N^{*2}(2k-1)\right]\right), \tag{4.9.100}$$

From the Markov inequality, for any generic random variable X, constants c and $a > 0$, we have $\Pr(X > c) \leq \frac{\mathbb{E}[\exp(aX)]}{\exp(ac)}$. Consequently, based on the independence of $\{V_N^*(2k-1)\}_{k=1}^{q_N/2}$ and (4.9.100),

$$\begin{aligned}
R_{21}(w) =& \Pr\left(\left|\sum_{k=1}^{q_N/2} V_N^*(2k-1)\right| > \eta/4\right) \\
=& \Pr\left(\sum_{k=1}^{q_N/2} V_N^*(2k-1) > \eta/4\right) + \Pr\left(-\sum_{k=1}^{q_N/2} V_N^*(2k-1) > \eta/4\right) \\
\leq& \left\{\mathbb{E}\left[\exp\left(\lambda_N \sum_{k=1}^{q_N/2} V_N^*(2k-1)\right)\right] + \mathbb{E}\left[\exp\left(-\lambda_N \sum_{k=1}^{q_N/2} V_N^*(2k-1)\right)\right]\right\} / \exp(\lambda_N \eta/4) \\
\leq& \left\{\prod_{k=1}^{q_N/2} \mathbb{E}\left[\exp\left(\lambda_N V_N^*(2k-1)\right)\right] + \prod_{k=1}^{q_N/2} \mathbb{E}\left[\exp\left(-\lambda_N V_N^*(2k-1)\right)\right]\right\} / \exp(\lambda_N \eta/4) \\
\leq& 2\prod_{k=1}^{q_N/2} \exp\left(\mathbb{E}\left[\lambda_N^2 V_N^{*2}(2k-1)\right]\right) / \exp(\lambda_N \eta/4) \\
\leq& 2\exp\left(-\lambda_N \eta/4 + \lambda_N^2 \sum_{k=1}^{q_N/2} \mathbb{E}\left[V_N^{*2}(2k-1)\right]\right)
\end{aligned} \tag{4.9.101}$$

where the first inequality is obtained by letting $a = \lambda_N$ and $c = \eta/4$ in the Markov inequality. Due that $\{V_N(2k-1)\}$ and $\{V_N^*(2k-1)\}$ have identical probability and $V_N(k) = \sum_{i \in \mathbb{S}_k} Q_{N,i}$,

$$\sum_{k=1}^{q_N/2} \mathbb{E}\left[V_N^{*2}(2k-1)\right] = \sum_{k=1}^{q_N/2} \mathbb{E}\left[V_N^2(2k-1)\right] = \sum_{k=1}^{q_N/2} \sum_{i,j \in \mathbb{S}_{2k-1}} \mathrm{Cov}\left(Q_{N,i}, Q_{N,j}\right).$$

Given that the density function $f_{W_i^c}$ is uniformly bounded (Assumption 4.5.2 (b)), there exists a constant A_2 such that $|f_{W_i^c}| < A_2$. Then, because

$$Q_{N,i} = \frac{1}{N}\left\{K(W_i^c, w^c)1[W_i^d = w^d] - \mathbb{E}\left[K(W_i^c, w^c)1[W_i^d = w^d]\right]\right\},$$

we have

$$\mathrm{Var}\left[Q_{N,i}\right] = \mathbb{E}\left[Q_{N,i}^2\right] \leq \frac{1}{N^2}\mathbb{E}\left[K^2(W_i^c, w^c)\right]$$

$$
\begin{aligned}
=&\frac{1}{(Nh^Q)^2}\int\prod_{q=1}^{Q}\kappa^2\left(\frac{w_q^{c*}-w_q^c}{h}\right)f_{W_i^c}(w_q^{c*})dw^{c*}\\
\leq&\frac{A_2}{N^2h^Q}\prod_{q=1}^{Q}\int\kappa^2\left(v_q\right)dv_q=\frac{A_3}{N^2h^Q},
\end{aligned}
\tag{4.9.102}
$$

with $A_3 = A_2K_2^Q$ and $A_3 < \infty$ due that $\int\kappa^2(v)\,dv = K_2 < \infty$. Recall that $r_i = |\Delta(i,N)|$. By the Cauchy–Schwarz inequality and (4.9.102)

$$
\begin{aligned}
\left|\sum_{k=1}^{q_N/2}\sum_{i,j\in\mathbb{S}_{2k-1}}\mathrm{Cov}\left(Q_{N,i},Q_{N,j}\right)\right|\leq&\sum_{k=1}^{q_N/2}\sum_{i,j\in\mathbb{S}_{2k-1}}\left|\mathrm{Cov}\left(Q_{N,i},Q_{N,j}\right)\right|\\
\leq&\sum_{k=1}^{q_N/2}\sum_{i,j\in\mathbb{S}_{2k-1}}\mathrm{Var}\left[Q_{N,i}\right]\\
\leq&\frac{A_3}{2N^2h^Q}\sum_{k=1}^{q_N/2}|\mathbb{S}_{2k-1}|(|\mathbb{S}_{2k-1}|-1),
\end{aligned}
$$

substituting $\sum_{k=1}^{q/2}|\mathbb{S}_{2k-1}| \leq N$ and $|\mathbb{S}_{2k-1}| \leq r_{i_{2k-1}}$ into the above inequality,

$$
\left|\sum_{k=1}^{q_N/2}\sum_{i,j\in\mathbb{S}_{2k-1}}\mathrm{Cov}\left(Q_{N,i},Q_{N,j}\right)\right|\leq\frac{A_3}{2N^2h^Q}\left(\sum_{k=1}^{q_N/2}r_{i_{2k-1}}^2+N\right)=\frac{A_4}{Nh^Q},
\tag{4.9.103}
$$

for some constant $A_4 > 0$, because $\sum_{k=1}^{q_N/2}r_{i_{2k-1}}^2 \leq \sum_{k=1}^{q_N}r_{i_k}^2 \leq \sum_{i=1}^{N}|\Delta(i,N)|^2 = O(N)$ (Assumption 4.5.2). Given (4.9.103), it is easy to see that (4.9.101) becomes to

$$
R_{21}(w)\leq 2\exp\left(-\frac{\lambda_N\eta}{4}+\lambda_N^2\frac{A_4}{Nh^Q}\right)=2\exp\left(-\frac{\lambda_N\eta}{4}+A_4\ln(N)\right).
\tag{4.9.104}
$$

Let $\eta = 4A_5[\ln(N)/(Nh^Q)]^{1/2}$ for some constant $A_5 > 0$. Then, we have $\lambda_N\eta = A_5\ln(N)$. We can bound $R_{21}(w)$ uniformly as

$$
\sup_{w\in\Omega_W}R_{21}(w)\leq 2\exp((A_4-A_5)\ln(N))=2N^{-\alpha},
\tag{4.9.105}
$$

and we choose A_5 large enough such that $\alpha > 0$ with $\alpha = A_5 - A_4$.

At last, we deal with $R_{22}(w)$. Let $B_{2k-3} \in \sigma\{V_N(1), V_N(3), ..., V_N(2k-3)\}$, $B'_{2k-1} \in$

$\sigma\{V_N(2k-1)\}$ and

$$\alpha_{2k-1} = \sup_{B_{2k-3}, B'_{2k-1}} \left|\Pr(B_{2k-3}, B'_{2k-1}) - \Pr(B_{2k-3})\Pr(B'_{2k-1})\right|.$$

Making use of (4.9.96), we can obtain that the reminder term

$$\begin{aligned} R_{22}(w) =& \Pr\left(\left|\sum_{k=1}^{q_N/2} [V_N(2k-1) - V_N^*(2k-1)]\right| > \eta/4\right) \\ \leq& \sum_{k=1}^{q_N/2} \Pr\left(|V_N(2k-1) - V_N^*(2k-1)| > \frac{\eta}{2q_N}\right) \\ \leq& 18 \sum_{k=1}^{q_N/2} \left(\frac{2q_N \|V_N(2k-1)\|_\infty}{\eta}\right)^{1/2} \alpha_{2k-1}. \end{aligned} \tag{4.9.106}$$

Furthermore, applying (4.9.98) and $\eta = 4A_5[\ln(N)/(Nh^Q)]^{1/2}$ to the above inequality,

$$\begin{aligned} R_{22}(w) \leq 18 \sum_{k=1}^{q_N/2} \left(\frac{2q_N A_1 r_{2k-1}}{\eta N h^Q}\right)^{1/2} \alpha_{2k-1} \leq& A_6 \left(\frac{q_N \bar{r}_N}{[\ln(N) N h^Q]^{1/2}}\right)^{1/2} \sum_{k=1}^{q_N/2} \alpha_{2k-1} \\ \leq& A_6 \left(\frac{N}{\ln(N)}\right)^{1/2} \sum_{k=1}^{q_N/2} \alpha_{2k-1} \end{aligned} \tag{4.9.107}$$

uniformly in w for some constant $A_6 > 0$, where the last line is due to $\bar{r}_N = O([Nh^Q/\ln(N)]^{1/2})$ and $q_N \leq N$. Now, substitute (4.9.105) and (4.9.107) into (4.9.97),

$$\sup_{w \in \Omega_W} \Pr\left(|W'_N(w)| > \eta/2\right) \leq 2N^{-\alpha} + A_6 \left(\frac{N}{\ln(N)}\right)^{1/2} \sum_{k=1}^{q_N/2} \alpha_{2k-1}$$

which, together with (4.9.95), further implies that

$$\Pr(R_2 > \eta) \leq 4L_N N^{-\alpha} + 2A_6 L_N \left(\frac{N}{\ln(N)}\right)^{1/2} \sum_{k=1}^{q_N/2} \alpha_{2k-1}. \tag{4.9.108}$$

Let $l_N = [\ln(N) h^{(Q+2)}/N]^{1/2} = \eta h^{Q+1} \to 0$, then $L_N = 1/l_N^Q = 1/[\eta h^{(Q+1)}]^Q \to \infty$ as $N \to \infty$. By properly choosing α, we can obtain the result that $L_n N^{-\alpha}$ is summable, i.e. $\sum_{N=1}^{\infty} L_n N^{-\alpha} < \infty$. In addition, by Assumption 4.5.3, we know that $L_N \left(\frac{N}{\ln(N)}\right)^{1/2} \sum_{k=1}^{q_N/2} \alpha_{2k-1}$ is also summable. It then follows from the Borel-Cantelli

lemma that

$$R_2 = O(\eta) = O\left(\left[\frac{\ln(N)}{Nh^Q}\right]^{1/2}\right) \text{ almost surely.} \tag{4.9.109}$$

Together with (4.9.91) and (4.9.94), we arrive the conclusion that

$$\sup_{w\in\Omega_W}\left|\hat{f}_{W_i}(w) - f_{W_i}(w)]\right| = O_p\left([\ln(N)/(Nh^Q)]^{1/2} + h^2\right). \tag{4.9.110}$$

■

Proof of Corollary 1. We prove the desired result in two steps. Step 1 aims at the uniform convergence of $\hat{f}_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}$. Step 2 fulfils the proof by establishing the uniform convergence of $\hat{f}_{S^*,\mathcal{F}_i^*|D_i,S_i,\mathcal{F}_i,Z_i}$.

Step 1. From (4.9.72) we know that $F_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z} = F_{\mathcal{F}|\mathcal{F}^*,Z} \times T_{Y|\mathcal{F}^*,Z} \times F^{-1}_{\mathcal{F}|\mathcal{F}^*,Z}$. Denote $\mathbf{B}(\gamma^0) := F_{\tilde{\mathcal{F}},\mathcal{F},Y|Z} \times F^{-1}_{\tilde{\mathcal{F}},\mathcal{F}|Z}$, and let $\lambda(\gamma^0)$ and $\psi(\gamma^0)$ represent the eigenvalues and eigenvectors of $\mathbf{B}(\gamma^0)$. Then, we have $\left(\mathbf{B}(\gamma^0) - \lambda(\gamma^0)\mathbf{I}_{K_T}\right)\psi(\gamma^0) = 0$.

Furthermore, recall that $T_{Y|\mathcal{F}^*,Z}$ is a diagonal matrix with all entries on its diagonal strictly positive. It then yields from the eigendecomposition that for the eigenvalue $\lambda(\gamma^0) = \mathbb{E}[Y_i|\mathcal{F}_i^* = n^*, Z_i]$, its eigenvector is $\psi(\gamma^0) = [f_{\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(0), ..., f_{\mathcal{F}_i|\mathcal{F}_i^*=n^*,Z_i}(K_{\mathcal{F}}-1)]'$. Andrew et al. (1993) shows the existence of a neighbourhood of γ^0 in the parameter space, denoted by $\mathcal{M}_0$, such that for any $\gamma \in \mathcal{M}_0$, there exist an eigenvalue function $\lambda(\gamma)$ and an eigenvector function $\psi(\gamma)$ that are both analytic functions of γ. Given the uniform convergence of $\hat{\gamma}_N$ to γ^0 proved in Theorem 4.5.2, we only need to consider the convergence of $\psi(\gamma)$ over a small neighbourhood of γ^0 such that $\|\gamma - \gamma^0\|_\infty \leq \eta$ with $\eta = o(1)$. The rest of the proof is exactly the same with the proof of Lemma 3 in Hu (2008), therefore ignored here due to space limitation. Let $\hat{\psi}_N := \psi(\hat{\gamma}_N)$ and $\psi_0 := \psi(\gamma^0)$, then we can show the uniform convergence

$$\sup_{\|\hat{\gamma}_N-\gamma^0\|_\infty\leq\eta}\left\|\hat{\psi}_N - \psi_0\right\|_\infty = O_p\left(\|\hat{\gamma}_N - \gamma^0\|_\infty\right),$$
$$\sup_{\|\hat{\gamma}_N-\gamma^0\|_\infty\leq\eta}\left\|\hat{\psi}_N - \psi_0 - \frac{\partial\psi(\gamma^0)}{\partial\gamma'}(\hat{\gamma}_N - \gamma^0)\right\|_\infty = O_p\left(\|\hat{\gamma}_N - \gamma^0\|_\infty^2\right).$$

Step 2. Again, because of the uniform convergence of $\hat{\gamma}_N$, in Step 2 we consider only a small neighbourhood of γ^0. Denote $\varphi = (\varphi_1, ..., \varphi_6)'$ where each of its elements represents one probability distribution on the right hand side of the equation (4.3):

$$\begin{aligned}\varphi_1 =& f_{S_i|S_i^*,\mathcal{F}_i^*,\mathcal{F}_i,Z_i}, \quad \varphi_2 = f_{S_i^*|\mathcal{F}_i^*,Z_i}, \quad \varphi_3 = f_{\mathcal{F}_i|\mathcal{F}_i^*,Z_i}, \quad \varphi_4 = f_{\mathcal{F}_i^*|Z_i},\\ \varphi_5 =& f_{S_i|\mathcal{F}_i,Z_i}, \quad \varphi_6 = f_{\mathcal{F}_i|Z_i}.\end{aligned}$$

where we actually have that $\varphi_3 = \psi$. Given Proposition 4.4.2, $\phi = \phi(\varphi) = \varphi_1\varphi_2\varphi_3\varphi_4/(\varphi_5\varphi_6)$, which is a twice continuously differentiable function of φ by Assumption 4.5.2. Beside, its estimator is constructed by $\hat{\phi}_N = \phi(\hat{\varphi}_N)$ with true value $\phi^0 = \phi(\varphi^0)$. Let the true value of φ be $\varphi^0 = \varphi(\gamma^0, \psi^0) = (\varphi_1^0, ..., \varphi_6^0)'$ and let its plug-in estimator be $\hat{\varphi}_N = \varphi(\hat{\gamma}_N, \hat{\psi}_N) = (\hat{\varphi}_{1,N}, ..., \hat{\varphi}_{6,N})'$. Then,

$$\frac{d\phi(\varphi)}{d\varphi'} = \left(\frac{\varphi_2\varphi_3\varphi_4}{\varphi_5\varphi_6}, \frac{\varphi_1\varphi_3\varphi_4}{\varphi_5\varphi_6}, \frac{\varphi_1\varphi_2\varphi_4}{\varphi_5\varphi_6}, \frac{\varphi_1\varphi_2\varphi_3}{\varphi_5\varphi_6}, -\frac{\varphi_1\varphi_2\varphi_3\varphi_4}{\varphi_5^2\varphi_6}, -\frac{\varphi_1\varphi_2\varphi_3\varphi_4}{\varphi_5\varphi_6^2}\right). \quad (4.9.111)$$

Recall that there exists a $\epsilon > 0$, such that φ_6^0 are uniformly bounded from below by ϵ based on the condition stated in Corollary 1. We also know that $\varphi_5^0 = C_n^s f_D^s(1) f_D(0)^{(n-s)} > \epsilon'$ uniformly over Ω_W for some constant ϵ'. Moreover, since φ_1 to φ_4 are all conditional probabilities of discrete random variables, their true values φ_1^0 to φ_4^0 all lie in $[0, 1]$. When we consider a uniform $o(1)$ neighbourhood of γ^0, by the uniform convergence of $\hat{\psi}_N$ in Step 1, we know that for large enough sample size, $\hat{\varphi}_{1,N}$ to $\hat{\varphi}_{4,N}$ are also uniformly bounded from above and $\hat{\varphi}_{5,N}$ and $\hat{\varphi}_{6,N}$ are uniformly bounded from below. Therefore, any intermediate value $\tilde{\varphi}$ between φ^0 and $\hat{\varphi}_N$ is uniformly bounded. Thus, for the derivative in (4.9.111) evaluated at $\tilde{\varphi}$, there exists some constant $C > 0$ such that $\|d\phi(\tilde{\varphi})/d\varphi'\| \leq C$ uniformly over Ω_W. By the mean value theorem, we then have that

$$\begin{aligned}\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \|\hat{\phi}_N - \phi^0\|_\infty =& \sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \|\phi(\hat{\varphi}_N) - \phi(\varphi^0)\|_\infty\\ \leq& \sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\|\frac{d\phi(\tilde{\varphi})}{d\varphi'}\right\|_\infty \|\hat{\varphi}_N - \varphi^0\|_\infty\\ \leq& C \sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \|\hat{\varphi}_N - \varphi^0\|_\infty, \quad (4.9.112)\end{aligned}$$

where $\tilde{\varphi}$ is an intermediate vector between φ^0 and $\hat{\varphi}_N$. Besides, because $\hat{\varphi}_N = \varphi(\hat{\gamma}_N, \hat{\psi}_N)$

and $\varphi^0 = \varphi(\gamma^0, \psi^0)$, together with the fact that $\varphi(\gamma, \psi)$ is continuously differentiable in (γ, ψ) with uniformly bounded first order derivative, we get that (4.9.112) can be further bounded by

$$\begin{aligned}\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \|\hat{\phi}_N - \phi^0\|_\infty \leq & C' \left(\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \|\hat{\psi}_N - \psi\|_\infty + \|\hat{\gamma}_N - \gamma^0\|_\infty \right) \\ = & O_p\left(\|\hat{\gamma}_N - \gamma^0\|_\infty \right),\end{aligned}$$

for some constant $C' > 0$, and the last line is from in Step 1. Furthermore, recall that $\phi = \phi(\psi)$, where $\psi = \psi(\gamma, \varphi)$ and $\varphi = \varphi(\gamma)$. Thus, ϕ can be regarded as a function of γ only. Applying similar arguments, we can also obtain that

$$\sup_{w \in \Omega_W} \left\| \hat{\phi}_N - \phi^0 - \frac{\partial \phi}{\partial \gamma}(\hat{\gamma}_N - \gamma^0) \right\|_\infty = O_p\left(\|\hat{\gamma}_N - \gamma^0\|_\infty^2 \right). \tag{4.9.113}$$

■

Proof of Theorem 4.5.3. Now, from $m(x; \theta, \phi) = \sum_{j=1}^{K_T} m^*(x_j; \theta) f_{T_i^*|X_i = x}(t_j)$ with $x_j = (d, s_j, z, n_j)$ and $t_j = (s_j, n_j)$, we can get

$$\begin{aligned}& \mathcal{L}_N(\theta, \hat{\phi}_N) - \mathcal{L}_N(\theta, \phi^0) \\ = & \frac{1}{N} \sum_{i=1}^N \tau_i \left\{ \left[Y_i - m\left(X_i; \theta, \hat{\phi}_N \right) \right]^2 - \left[Y_i - m\left(X_i; \theta, \phi^0 \right) \right]^2 \right\} \\ = & \frac{1}{N} \sum_{i=1}^N \tau_i \left[m\left(X_i; \theta, \hat{\phi}_N \right) - m\left(X_i; \theta, \phi^0 \right) \right]^2 \\ & - \frac{2}{N} \sum_{i=1}^N \tau_i \left[Y_i - m\left(X_i; \theta, \phi^0 \right) \right] \left[m\left(X_i; \theta, \hat{\phi}_N \right) - m\left(X_i; \theta, \phi^0 \right) \right] \\ = & \frac{1}{N} \sum_{i=1}^N \tau_i \left\{ \sum_{j=1}^{K_T} m^*(x_{i,j}; \theta) \left[\hat{f}_{T_i^*|X_i}(t_j) - f_{T_i^*|X_i}(t_j) \right] \right\}^2 \\ & - \frac{2}{N} \sum_{i=1}^N \tau_i \left[Y_i - m\left(X_i; \theta, \phi^0 \right) \right] \left\{ \sum_{j=1}^{K_T} m^*(x_{i,j}; \theta) \left[\hat{f}_{T_i^*|X_i}(t_j) - f_{T_i^*|X_i}(t_j) \right] \right\},\end{aligned} \tag{4.9.114}$$

where $x_{i,j} = (D_i, s_j, Z_i, n_j)$. Because of the uniform convergence of $\hat{\gamma}_N$, we only need to focus on a small neighbourhood of γ^0. Due to the boundedness of $\tau(x)$ and the

Cauchy–Schwarz inequality,

$$
\begin{aligned}
&\left|\mathcal{L}_N(\theta,\hat{\phi}_N)-\mathcal{L}_N(\theta,\phi^0)\right| \\
\leq&\frac{C}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_T} m^*(x_{i,j};\theta)^2\sum_{j=1}^{K_T}\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]^2 \\
&+\frac{2}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_T}\tau_i\left|Y_i-m\left(X_i;\theta,\phi^0\right)\right|\left|m^*(x_{i,j};\theta)\right|\left|\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right| \\
\leq& C\left(\sup_{\|\hat{\gamma}_N-\gamma^0\|_\infty\leq\eta}\left\|\hat{\phi}_N-\phi^0\right\|_\infty\right)^2\frac{1}{N}\sum_{j=1}^{K_T}\sum_{i=1}^{N} m^*(x_{i,j};\theta)^2 \\
&+2\sup_{\|\hat{\gamma}_N-\gamma^0\|_\infty\leq\eta}\left\|\hat{\phi}_N-\phi^0\right\|_\infty\sum_{j=1}^{K_T}\left[\frac{1}{N}\sum_{i=1}^{N}\tau_i\left|Y_i-m\left(X_i;\theta,\phi^0\right)\right|^2\right]^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\left|m^*(x_{i,j};\theta)\right|^2\right]^{1/2}.
\end{aligned}
\tag{4.9.115}
$$

Because (D_i, Z_i) is i.i.d., then $x_{i,j} = (D_i, s_j, Z_i, n_j)$ is also i.i.d. for any given $j = 1, ..., K_T$. Then, by Assumption 4.5.4 and the uniform convergence of i.i.d. samples (Lemma 2.4 of Newey and McFadden (1994))

$$
\begin{aligned}
\sup_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N} m^*(x_{i,j};\theta)^2 \leq&\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N} m^*(x_{i,j};\theta)^2-\mathbb{E}\left[m^*(x_{i,j};\theta)^2\right]\right|+\sup_{\theta\in\Theta}\left|\mathbb{E}\left[m^*(x_{i,j};\theta)^2\right]\right| \\
=&O_p(1),
\end{aligned}
\tag{4.9.116}
$$

because $\sup_{\theta\in\Theta}\mathbb{E}\left[m^*(x_{i,j};\theta)^2\right] \leq \mathbb{E}[h_1(x_{i,j})] < \infty$ by Assumption 4.5.4. Similarly, the uniform convergence of data with dependency neighbourhood structure in Lemma 4.9.3 leads to

$$
\sup_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\tau_i\left|Y_i-m\left(X_i;\theta,\phi^0\right)\right|^2=O_p(1),
\tag{4.9.117}
$$

because of Assumption 4.5.1 and Assumption 4.5.4 (e). Hence, we can conclude that

$$
\sup_{\theta\in\Theta}\left|\mathcal{L}_N(\theta,\hat{\phi}_N)-\mathcal{L}_N(\theta,\phi^0)\right|=O_p\left(\sup_{\|\hat{\gamma}_N-\gamma^0\|_\infty\leq\eta}\left\|\hat{\phi}_N-\phi^0\right\|_\infty\right)=O_p\left(\|\hat{\gamma}_N-\gamma^0\|_\infty\right).
\tag{4.9.118}
$$

Next, we show the uniform convergence of $\mathcal{L}_N(\theta, \phi^0)$ to $\mathcal{L}(\theta, \phi^0)$ by verifying the uniform

law of large number for dependent data as in Lemma 4.9.3. Firstly, condition (i), (ii), (iii) and (iv)-(c) of Lemma 4.9.3 are trivially sanctified by Assumption 4.5.4 (a), (c) and (e). Secondly, (iv) (a) of Lemma 4.9.3 holds because of Assumption 4.5.1. In addition, we have that $1/N\sum_{i=1}^{N}|\Delta(i,N)| \leq 1/N\sum_{i=1}^{N}|\Delta(i,N)|^2 = O(1)$ as in Assumption 4.5.2. Hence, we have verified that all required conditions of Lemma 4.9.3 are satisfied, implying

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\mathcal{L}_N(\theta,\phi^0)-\mathcal{L}(\theta,\phi^0)\right| && (4.9.119)\\
=&\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\tau_i\left[Y_i-m(X_i;\theta,\phi^0)\right]^2-\mathbb{E}\left[\tau_i\left[Y_i-m(X_i;\theta,\phi^0)\right]^2\right]\right|\\
=&o_p(1). && (4.9.120)
\end{aligned}$$

Then, making use of (4.9.118), (4.9.119) and Theorem 4.5.2, we can bound

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\mathcal{L}(\theta,\phi^0)-\mathcal{L}_N(\theta,\hat{\phi}_N)\right|\\
=&\sup_{\theta\in\Theta}\left|\mathcal{L}(\theta,\phi^0)-\mathcal{L}_N(\theta,\phi^0)+\mathcal{L}_N(\theta,\phi^0)-\mathcal{L}_N(\theta,\hat{\phi}_N)\right|\\
\leq&\sup_{\theta\in\Theta}\left|\mathcal{L}(\theta,\phi^0)-\mathcal{L}_N(\theta,\phi^0)\right|+\sup_{\theta\in\Theta}\left|\mathcal{L}_N(\theta,\phi^0)-\mathcal{L}_N(\theta,\hat{\phi}_N)\right|\\
=&\sup_{\theta\in\Theta}\left|\mathcal{L}(\theta,\phi^0)-\mathcal{L}_N(\theta,\phi^0)\right|+O_p\left(\|\hat{\gamma}_N-\gamma^0\|_\infty\right)\\
=&o_p(1). && (4.9.121)
\end{aligned}$$

As assumed in Assumption 4.5.4, θ^0 uniquely minimises the objective function $\mathcal{L}(\theta,\phi^0)$ over Θ. Then, for any $\delta>0$, there exists a $\epsilon>0$ such that $\|\hat{\theta}_N-\theta^0\|>\delta$ implies $\mathcal{L}(\hat{\theta}_N,\phi^0)-\mathcal{L}(\theta^0,\phi^0)>\epsilon$. Thus, by the definition of $\hat{\theta}_N$,

$$\begin{aligned}
&\Pr\left(\left\|\hat{\theta}_N-\theta^0\right\|>\delta\right)\\
\leq&\Pr\left(\mathcal{L}(\hat{\theta}_N,\phi^0)-\mathcal{L}(\theta^0,\phi^0)>\epsilon\right)\\
\leq&\Pr\left(\mathcal{L}(\hat{\theta}_N,\phi^0)-\mathcal{L}_N(\hat{\theta}_N,\hat{\phi}_N)+\mathcal{L}_N(\hat{\theta}_N,\hat{\phi}_N)-\mathcal{L}(\theta^0,\phi^0)>\epsilon\right)\\
\leq&\Pr\left(\mathcal{L}(\hat{\theta}_N,\phi^0)-\mathcal{L}_N(\hat{\theta}_N,\hat{\phi}_N)+\mathcal{L}_N(\theta^0,\hat{\phi}_N)-\mathcal{L}(\theta^0,\phi^0)>\epsilon\right)\\
\leq&\Pr\left(\sup_{\theta\in\Theta}\left|\mathcal{L}(\theta,\phi^0)-\mathcal{L}_N(\theta,\hat{\phi}_N)\right|>\epsilon\right)\\
\to&0, && (4.9.122)
\end{aligned}$$

where the last line is due to (4.9.121). It then follows from (4.9.122) that $\|\hat{\theta}_N - \theta^0\| = o_p(1)$. ■

Proof of Lemma 4.5.4. (a) Based on Theorems 4.5.2 and 1, we know that $\hat{\phi}_N = \phi(\hat{\gamma}_N)$ and $\hat{\gamma}_N \xrightarrow{p} \gamma^0$. Hence, in what follows, we can establish the consistency of $\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'}$ in a small neighbourhood of γ^0. For a small constant $\eta > 0$, by triangular inequality,

$$
\begin{aligned}
&\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'} - \mathbb{E}\left[\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'}\right] \right\| \\
\leq &\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'} - \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\phi^0)}{\partial \theta'} \right\| \\
&\quad + \left\| \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\phi^0)}{\partial \theta'} - \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'} \right\| \\
&\quad + \left\| \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'} - \mathbb{E}\left[\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'}\right] \right\| \\
:= &\mathcal{H}_1 + \mathcal{H}_2 + \mathcal{H}_3. \qquad (4.9.123)
\end{aligned}
$$

Given (4.9.123), it suffices to show that $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ are all $o_p(1)$. In what follows, we divide the rest of the proof into three steps.

Step 1. First, consider $\mathcal{H}_1$. By definition of $g(W_i;\theta,\phi)$, we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'} - \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\phi^0)}{\partial \theta'} \\
= &\frac{1}{N}\sum_{i=1}^N \tau_i \left\{ \left[Y_i - m(X_i;\tilde{\theta}_N,\hat{\phi}_N)\right] \frac{d^2 m(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta d\theta'} - \left[Y_i - m(X_i;\tilde{\theta}_N,\phi^0)\right] \frac{d^2 m(X_i;\tilde{\theta}_N,\phi^0)}{d\theta d\theta'} \right\} \\
&- \frac{1}{N}\sum_{i=1}^N \tau_i \left[\frac{dm(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta} \frac{dm(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta'} - \frac{dm(X_i;\tilde{\theta}_N,\phi^0)}{d\theta} \frac{dm(X_i;\tilde{\theta}_N,\phi^0)}{d\theta'} \right]. \qquad (4.9.124)
\end{aligned}
$$

Making use of the identity $\hat{a}\hat{b} - ab = (\hat{a}-a)b + a(\hat{b}-b) + (\hat{a}-a)(\hat{b}-b)$ and applying it

to both terms on the right hand side of (4.9.124) give us

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial\theta'}-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i;\tilde{\theta}_N,\phi^0)}{\partial\theta'}\\
=&-\frac{1}{N}\sum_{i=1}^{N}\tau_i\left[m(X_i;\tilde{\theta}_N,\hat{\phi}_N)-m(X_i;\tilde{\theta}_N,\phi^0)\right]\frac{d^2m(X_i;\tilde{\theta}_N,\phi^0)}{d\theta d\theta'}\\
&+\frac{1}{N}\sum_{i=1}^{N}\tau_i\left[Y_i-m(X_i;\tilde{\theta}_N,\phi^0)\right]\left[\frac{d^2m(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta d\theta'}-\frac{d^2m(X_i;\tilde{\theta}_N,\phi^0)}{d\theta d\theta'}\right]\\
&-\frac{1}{N}\sum_{i=1}^{N}\tau_i\left[m(X_i;\tilde{\theta}_N,\hat{\phi}_N)-m(X_i;\tilde{\theta}_N,\phi^0)\right]\left[\frac{d^2m(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta d\theta'}-\frac{d^2m(X_i;\tilde{\theta}_N,\phi^0)}{d\theta d\theta'}\right]\\
&-\frac{1}{N}\sum_{i=1}^{N}\tau_i\left[\frac{dm(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta}-\frac{dm(X_i;\tilde{\theta}_N,\phi^0)}{d\theta}\right]\frac{dm(X_i;\tilde{\theta}_N,\phi^0)}{d\theta'}\\
&-\frac{1}{N}\sum_{i=1}^{N}\tau_i\frac{dm(X_i;\tilde{\theta}_N,\phi^0)}{d\theta}\left[\frac{dm(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta'}-\frac{dm(X_i;\tilde{\theta}_N,\phi^0)}{d\theta'}\right]\\
&-\cdot\frac{1}{N}\sum_{i=1}^{N}\tau_i\left[\frac{dm(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta}-\frac{dm(X_i;\tilde{\theta}_N,\phi^0)}{d\theta}\right]\left[\frac{dm(X_i;\tilde{\theta}_N,\hat{\phi}_N)}{d\theta'}-\frac{dm(X_i;\tilde{\theta}_N,\phi^0)}{d\theta'}\right].
\end{aligned}
\tag{4.9.125}
$$

Recall that $m(X_i;\theta,\phi)=\sum_{j=1}^{K_T}m^*(x_{i,j};\theta)f_{T_i^*|X_i}(t_j)$ and $x_{i,j}=(D_i,s_j,Z_i,n_j)$. We can further rewrite (4.9.125) as

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial\theta'}-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i;\tilde{\theta}_N,\phi^0)}{\partial\theta'}\\
=&-\frac{1}{N}\sum_{i=1}^{N}\tau_i\left\{\sum_{j=1}^{K_T}m^*(x_{i,j};\tilde{\theta}_N)\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\right\}\sum_{j=1}^{K_T}\frac{d^2m^*(x_{i,j};\tilde{\theta}_N)}{d\theta d\theta'}f_{T_i^*|X_i}(t_j)\\
&+\frac{1}{N}\sum_{i=1}^{N}\tau_i\left[Y_i-m(X_i;\tilde{\theta}_N,\phi^0)\right]\left\{\sum_{j=1}^{K_T}\frac{d^2m^*(x_{i,j};\tilde{\theta}_N)}{d\theta d\theta'}\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\right\}\\
&-\frac{1}{N}\sum_{i=1}^{N}\tau_i\left\{\sum_{j=1}^{K_T}m^*(x_{i,j};\tilde{\theta}_N)\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\right\}\\
&\qquad\times\left\{\sum_{j=1}^{K_T}\frac{d^2m^*(x_{i,j};\tilde{\theta}_N)}{d\theta d\theta'}\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\right\}\\
&-\frac{1}{N}\sum_{i=1}^{N}\tau_i\left\{\sum_{j=1}^{K_T}\frac{\partial m^*(x_{i,j};\tilde{\theta}_N)}{\partial\theta}\left[\hat{f}_{T_i^*|X_i}(t_j)-f_{T_i^*|X_i}(t_j)\right]\right\}\sum_{j=1}^{K_T}\frac{dm^*(x_{i,j};\tilde{\theta}_N)}{d\theta'}f_{T_i^*|X_i}(t_j)
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{N}\sum_{i=1}^{N}\tau_i \sum_{j=1}^{K_T} \frac{dm^*(x_{i,j};\tilde{\theta}_N)}{d\theta} f_{T_i^*|X_i}(t_j) \left\{ \sum_{j=1}^{K_T} \frac{\partial m^*(x_{i,j};\tilde{\theta}_N)}{\partial \theta'} \left[\hat{f}_{T_i^*|X_i}(t_j) - f_{T_i^*|X_i}(t_j) \right] \right\} \\
&- \cdot \frac{1}{N}\sum_{i=1}^{N}\tau_i \left\{ \sum_{j=1}^{K_T} \frac{\partial m^*(x_{i,j};\tilde{\theta}_N)}{\partial \theta} \left[\hat{f}_{T_i^*|X_i}(t_j) - f_{T_i^*|X_i}(t_j) \right] \right\} \\
&\qquad \times \left\{ \sum_{j=1}^{K_T} \frac{\partial m^*(x_{i,j};\tilde{\theta}_N)}{\partial \theta'} \left[\hat{f}_{T_i^*|X_i}(t_j) - f_{T_i^*|X_i}(t_j) \right] \right\}. \qquad (4.9.126)
\end{aligned}
$$

Because that for a $k \times k$ matrix $A = ab'$ where $a, b \in \mathbb{R}^k$, then $\|A\| = \|a\|\|b\|$, the boundedness of $f_{T_i^*|X_i}$ and (4.9.126),

$$
\begin{aligned}
\mathcal{H}_1 \leq & C \sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \hat{\phi}_N - \phi^0 \right\|_\infty \frac{1}{N} \sum_{j,l=1}^{K_T} \sum_{i=1}^{N} \left| m^*(x_{i,j};\tilde{\theta}_N) \right| \left\| \frac{d^2 m^*(x_{i,l};\tilde{\theta}_N)}{d\theta d\theta'} \right\| \\
& + C \sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \hat{\phi}_N - \phi^0 \right\|_\infty \frac{1}{N} \sum_{j=1}^{K_T} \sum_{i=1}^{N} \tau_i \left| Y_i - m(X_i;\tilde{\theta}_N, \phi^0) \right| \left\| \frac{d^2 m^*(x_{i,j};\tilde{\theta}_N)}{d\theta d\theta'} \right\| \\
& + C \left(\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \hat{\phi}_N - \phi^0 \right\|_\infty \right)^2 \frac{1}{N} \sum_{j,l=1}^{K_T} \sum_{i=1}^{N} \left| m^*(x_{i,j};\tilde{\theta}_N) \right| \left\| \frac{d^2 m^*(x_{i,l};\tilde{\theta}_N)}{d\theta d\theta'} \right\| \\
& + 2C \sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \hat{\phi}_N - \phi^0 \right\|_\infty \frac{1}{N} \sum_{j,l=1}^{K_T} \sum_{i=1}^{N} \left\| \frac{\partial m^*(x_{i,j};\tilde{\theta}_N)}{\partial \theta} \right\| \left\| \frac{dm^*(x_{i,l};\tilde{\theta}_N)}{d\theta'} \right\| \\
& C \left(\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \hat{\phi}_N - \phi^0 \right\|_\infty \right)^2 \frac{1}{N} \sum_{j,l=1}^{K_T} \sum_{i=1}^{N} \left\| \frac{\partial m^*(x_{i,j};\tilde{\theta}_N)}{\partial \theta} \right\| \left\| \frac{dm^*(x_{i,l};\tilde{\theta}_N)}{d\theta'} \right\| \\
:= & \mathcal{H}_{11} + \mathcal{H}_{12} + \mathcal{H}_{13} + \mathcal{H}_{14} + \mathcal{H}_{15}. \qquad (4.9.127)
\end{aligned}
$$

By the Cauchy–Schwarz inequality, we can further bound $\mathcal{H}_{11}$ as

$$
\begin{aligned}
\mathcal{H}_{11} \leq & C \sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \hat{\phi}_N - \phi^0 \right\|_\infty \sum_{j,l=1}^{K_T} \left[\frac{1}{N} \sum_{i=1}^{N} \left| m^*(x_{i,j};\tilde{\theta}_N) \right|^2 \right]^{1/2} \left[\frac{1}{N} \sum_{i=1}^{N} \left\| \frac{d^2 m^*(x_{i,l};\tilde{\theta}_N)}{d\theta d\theta'} \right\|^2 \right]^{1/2} \\
\leq & O_p \left(\sup_{\|\hat{\gamma}_N - \gamma^0\|_\infty < \eta} \left\| \hat{\phi}_N - \phi^0 \right\|_\infty \right) \\
= & o_p(1), \qquad (4.9.128)
\end{aligned}
$$

where the second line is due to (4.9.116) and Lemma 4.9.8, and the last line is because of Corollary 1. For $\mathcal{H}_{12}$, it follows again from the Cauchy–Schwarz inequality and Corollary

1 that

$$\begin{aligned}\mathcal{H}_{12} \leq & o_p(1) \sum_{j=1}^{K_T} \left[\frac{1}{N}\sum_{i=1}^{N} \tau_i \left|Y_i - m(X_i;\theta^0,\phi^0)\right|^2\right]^{1/2} \left[\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{d^2 m^*(x_{i,j};\tilde{\theta}_N)}{d\theta d\theta'}\right\|^2\right]^{1/2} \\ = & o_p(1), \end{aligned} \tag{4.9.129}$$

where the last line is due to the uniform convergence in (4.9.119) and that proved in Lemma 4.9.8. Given $\mathcal{H}_{11} = o_p(1)$, it is apparent that $\mathcal{H}_{13}$ is also a $o_p(1)$. Similarly, if we know that $\mathcal{H}_{14} = o_p(1)$, then $\mathcal{H}_{15} = o_p(1)$. Again, by the Cauchy–Schwarz inequality and Lemma 4.9.8,

$$\begin{aligned}\mathcal{H}_{14} \leq & o_p(1) \sum_{j,l=1}^{K_T} \left[\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{dm^*(x_{i,j};\tilde{\theta}_N)}{d\theta}\right\|^2\right]^{1/2} \left[\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{dm^*(x_{i,l};\tilde{\theta}_N)}{d\theta'}\right\|^2\right]^{1/2} \\ = & o_p(1), \end{aligned} \tag{4.9.130}$$

Thus, based on (4.9.128), (4.9.129) and (4.9.130), we can conclude that $\mathcal{H}_1 = o_p(1)$.

Step 2. Consider the term inside the absolute value in $\mathcal{H}_2$

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i;\tilde{\theta}_N,\phi^0)}{\partial\theta'} - \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial\theta'} \\ = & \frac{1}{N}\sum_{i=1}^{N}\tau_i\left\{\left[Y_i - m(X_i;\tilde{\theta}_N,\phi^0)\right]\frac{\partial^2 m(X_i;\tilde{\theta}_N,\phi^0)}{\partial\theta\partial\theta'} - \left[Y_i - m(X_i;\theta^0,\phi^0)\right]\frac{\partial^2 m(X_i;\theta^0,\phi^0)}{\partial\theta\partial\theta'}\right\} \\ & + \frac{1}{N}\sum_{i=1}^{N}\tau_i\left[\frac{\partial m(X_i;\tilde{\theta}_N,\phi^0)}{\partial\theta}\frac{\partial m(X_i;\tilde{\theta}_N,\phi^0)}{\partial\theta'} - \frac{\partial m(X_i;\theta^0,\phi^0)}{\partial\theta}\frac{\partial m(X_i;\theta^0,\phi^0)}{\partial\theta'}\right]. \end{aligned} \tag{4.9.131}$$

Applying again the identity $\hat{a}\hat{b} - ab = (\hat{a}-a)b + a(\hat{b}-b) + (\hat{a}-a)(\hat{b}-b)$ to (4.9.131) and substituting $m(X_i;\theta,\phi) = \sum_{j=1}^{K_T} m^*(x_{i,j};\theta) f_{T_i^*|X_i}(t_j)$ give us

$$\begin{aligned}\mathcal{H}_2 \leq & \frac{C}{N}\sum_{i=1}^{N}\sum_{j,l=1}^{K_T}\left[\left|m^*(x_{i,j};\tilde{\theta}_N) - m^*(x_{i,j};\theta^0)\right|\left\|\frac{d^2 m^*(x_{i,l};\theta^0)}{d\theta d\theta'}\right\|\right] \\ & + \frac{C}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_T}\left[\tau_i\left|Y_i - m(X_i;\theta^0,\phi^0)\right|\left\|\frac{d^2 m^*(x_{i,j};\tilde{\theta}_N)}{d\theta d\theta'} - \frac{d^2 m^*(x_{i,j};\theta^0)}{d\theta d\theta'}\right\|\right] \end{aligned}$$

$$
\begin{aligned}
&+\frac{C}{N}\sum_{i=1}^{N}\sum_{j,l=1}^{K_T}\left[\left|m^*(x_{i,j};\tilde{\theta}_N)-m^*(x_{i,j};\theta^0)\right|\left\|\frac{d^2m^*(x_{i,l};\tilde{\theta}_N)}{d\theta d\theta'}-\frac{d^2m^*(x_{i,l};\theta^0)}{d\theta d\theta'}\right\|\right]\\
&+\frac{2C}{N}\sum_{i=1}^{N}\sum_{j,l=1}^{K_T}\left[\left\|\frac{dm^*(x_{i,j};\tilde{\theta}_N)}{d\theta}-\frac{dm^*(x_{i,j};\theta^0)}{d\theta}\right\|\left\|\frac{dm^*(x_{i,l};\theta^0)}{d\theta'}\right\|\right]\\
&+\frac{C}{N}\sum_{i=1}^{N}\sum_{j,l=1}^{K_T}\left[\left\|\frac{dm^*(x_{i,j};\tilde{\theta}_N)}{d\theta}-\frac{dm^*(x_{i,l};\theta^0)}{d\theta'}\right\|\left\|\frac{dm^*(x_{i,l};\tilde{\theta}_N)}{d\theta'}-\frac{dm^*(x_{i,l};\theta^0)}{d\theta'}\right\|\right]\\
:=&\mathcal{H}_{21}+\mathcal{H}_{22}+\mathcal{H}_{23}+\mathcal{H}_{24}+\mathcal{H}_{25}. \qquad (4.9.132)
\end{aligned}
$$

By the Cauchy–Schwarz inequality and Lemma 4.9.8, it is easy to show $\mathcal{H}_{21}$ to $\mathcal{H}_{25}$ are all $o_p(1)$. Consequently, we know that $\mathcal{H}_2 = o_p(1)$.

Step 3. Next, consider $\mathcal{H}_3$. Let $g_r(W_i;\theta,\phi)$ be the r-th element in the column vector $g(W_i;\theta,\phi)$. Then, we can rewrite $\mathcal{H}_3^2$ as

$$
\begin{aligned}
\mathcal{H}_3^2=&\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial\theta'}-\mathbb{E}\left[\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial\theta'}\right]\right\|^2\\
=&\sum_{r,q=1}^{d_\theta}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial g_r(W_i;\theta^0,\phi^0)}{\partial\theta_q}-\mathbb{E}\left[\frac{\partial g_r(W_i;\theta^0,\phi^0)}{\partial\theta_q}\right]\right)\right]^2. \qquad (4.9.133)
\end{aligned}
$$

Because $\mathbb{E}[|\partial g_r(W_i;\theta^0,\phi^0)/\partial\theta_q|^2]<\infty$ as in Assumption 4.5.5, the variance of $\partial g_r(W_i;\theta^0,\phi^0)/\partial\theta_q$ exists and is finite for all $r,q=1,...,d_\theta$. Then, the Chebyshev's inequality implies

$$
\begin{aligned}
&\Pr\left[\left|\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_r(W_i;\theta^0,\phi^0)}{\partial\theta_q}-\mathbb{E}\left[\frac{\partial g_r(W_i;\theta^0,\phi^0)}{\partial\theta_q}\right]\right|>\epsilon\right]\\
\leq&\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_r(W_i;\theta^0,\phi^0)}{\partial\theta_q}\right]/\epsilon^2\\
=&\frac{1}{\epsilon^2N^2}\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\mathrm{Cov}\left(\frac{\partial g_r(W_i;\theta^0,\phi^0)}{\partial\theta_q},\frac{\partial g_r(W_j;\theta^0,\phi^0)}{\partial\theta_q}\right)+s.o.\\
\leq&\frac{C}{\epsilon^2N^2}\sum_{i=1}^{N}|\Delta(i,N)|+s.o.\\
=&O\left(\frac{1}{\epsilon^2N}\right),
\end{aligned}
$$

where the second equality comes from Assumption 4.5.1, and the last line is because that $1/N\sum_{i=1}^{N}|\Delta(i,N)|=O(1)$ (Assumption 4.5.2), and set ϵ such that $\epsilon\to 0$ and $\epsilon^2N\to\infty$

as $N \to \infty$. Thus,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_r(X_i;\theta^0,\phi^0)}{\partial \theta_q} - \mathbb{E}\left[\frac{\partial g_r(X_i;\theta^0,\phi^0)}{\partial \theta_q}\right] \xrightarrow{p} 0, \text{ for all } r,q = 1,...,d_\theta, \tag{4.9.134}$$

leading to $\mathcal{H}_3 = o_p(1)$. Based on the results in the above three steps, we can make the conclusion that the stated result holds.

(b) This proof is analogue to the proof of Theorem 8.1 in Newey and McFadden (1994). All the sufficient conditions are verified in the Lemmas 4.9.9, 4.9.10 and 4.9.11. Recall that $\tilde{F}_W(w) = 1/N\sum_{i=1}^{N} 1[W_i \leq w]$ represents the empirical distribution and $\int \delta(w)d\tilde{F}_W(w) = 1/N\sum_{i=1}^{N}\delta(W_i)$. By triangular inequality, we have

$$\begin{aligned}
&\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[g(W_i;\theta^0,\hat{\phi}_N) - g(W_i;\theta^0,\phi^0) - \delta(W_i)\right]\right\| \\
\leq &\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[g(W_i;\theta^0,\hat{\phi}_N) - g(W_i;\theta^0,\phi^0) - G(W_i;\tilde{\gamma}_N - \gamma^0)\right]\right\| \\
&\quad + \left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[G(W_i;\tilde{\gamma}_N - \gamma^0) - \int G(w;\hat{\gamma}_N - \bar{\gamma})dF_W(w)\right]\right\| \\
&\quad + \left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\int G(w;\hat{\gamma}_N - \bar{\gamma})dF_W(w) - \int \delta(w)d\hat{F}_W(w)\right]\right\| \\
&\quad + \left\|\sqrt{N}\left[\int \delta(w)d\hat{F}_W(w) - \int \delta(w)d\tilde{F}_W(w)\right]\right\| \\
= &o_p(1),
\end{aligned} \tag{4.9.135}$$

where the last line follows from Lemmas 4.9.9, 4.9.10 and 4.9.11. ■

Proof of Theorem 4.5.5. By Assumption 4.5.4 and the construction of $\delta(w)$, we know that $\mathbb{E}[\tilde{g}_i]$=0. Since the dependency neighbourhood $\Delta(i,N)$ is symmetric as in Assumption 4.5.7, we know that $\Sigma_N^{\tilde{g}}$ is symmetric: because for $\forall r,q = 1,2,...,d_\theta$, its (r,q)-th entry

$$\sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\mathbb{E}[\tilde{g}_{i,r}\tilde{g}_{j,q}] = \sum_{j=1}^{N}\sum_{i\in\Delta(j,N)}\mathbb{E}[\tilde{g}_{j,r}\tilde{g}_{i,q}] = \sum_{i=1}^{N}\sum_{j\in\Delta(i,N)}\mathbb{E}[\tilde{g}_{i,q}\tilde{g}_{j,r}],$$

where the first equality follows from change of index and the second equality is due to

the symmetry of $\Delta(i,N)$. Under Assumption 4.5.7, the sufficient conditions for the CLT under neighbourhood dependent data required in Lemma 4.9.7 are satisfied. Thus, we can show that $\left[\Sigma_N^{\tilde{g}}\right]^{-1/2}\mathbf{S}_N^{\tilde{g}} \xrightarrow{d} \mathbb{N}(0,\mathbf{I}_{d_\theta})$. Next, we show the asymptotic normality for $\sqrt{N}(\hat{\theta}_N-\theta^0)$.

From (4.13) and Lemma 4.5.4 (b), we have

$$-\left[\frac{1}{\sqrt{N}}\sum_{i=1}^N \tilde{g}_i + o_p(1)\right] = \frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'}\sqrt{N}(\hat{\theta}_N-\theta^0).$$

Since from Lemma 4.5.4 (a), we have that $\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'} \xrightarrow{p} \mathbb{E}\left[\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'}\right]$, where by Assumption 4.5.5 $\mathbb{E}\left[\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'}\right]$ is invertible. Thus, $\left[\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'}\right]^{-1}$ exists for large enough N. Moreover, recall that Ω_N is symmetric and $\Omega_N \xrightarrow{p} \Omega$ with Ω being positive definite and nonsingular. It indicates that $\Omega_N^{-1/2}$ also exists for large enough N. Then, because $\|\Omega_N^{-1/2}\| = O(1)$ and $\Omega_N^{-1/2} = \sqrt{N}[\Sigma_N^{\tilde{g}}]^{-1/2}$, we can obtain

$$\begin{aligned}
\sqrt{N}(\hat{\theta}_N-\theta^0) &= -\left[\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'}\right]^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^N \tilde{g}_i + o_p(1)\right] \\
&= -\left[\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'}\right]^{-1}\Omega_N^{1/2}\left[\Omega_N^{-1/2}\frac{1}{\sqrt{N}}\sum_{i=1}^N \tilde{g}_i + \Omega_N^{-1/2}o_p(1)\right] \\
&= -\left[\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'}\right]^{-1}\Omega_N^{1/2}\left[[\Sigma_N^{\tilde{g}}]^{-1/2}\mathbf{S}_N^{\tilde{g}} + o_p(1)\right] \\
&\xrightarrow{d} \mathbb{N}(0, H^{-1}\Omega H^{-1}),
\end{aligned}$$

where the last line is because of

$$\left[\frac{1}{N}\sum_{i=1}^N \frac{\partial g(W_i;\tilde{\theta}_N,\hat{\phi}_N)}{\partial \theta'}\right] \xrightarrow{p} \mathbb{E}\left[\frac{\partial g(W_i;\theta^0,\phi^0)}{\partial \theta'}\right] \text{ and } [\Sigma_N^{\tilde{g}}]^{-1/2}\mathbf{S}_N^{\tilde{g}} \xrightarrow{d} \mathbb{N}(0,\mathbf{I}_{d_\theta}).$$

■

Proof of Corollary 2. To simplify notation, denote $\hat{\tilde{g}}_i = g(W_i;\hat{\theta}_N,\hat{\phi}_N) + \hat{\delta}(W_i)$. Then,

$$\left\|\hat{\Omega}_N - \Omega\right\| = \left\|\frac{1}{N}\sum_{i=1}^N \sum_{j\in\Delta(i,N)}\left(\hat{\tilde{g}}_i\hat{\tilde{g}}_j' - \mathbb{E}[\tilde{g}_i\tilde{g}_j']\right)\right\|$$

$$\leq \left\| \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \left(\hat{\tilde{g}}_i \hat{\tilde{g}}'_j - \tilde{g}_i \tilde{g}'_j \right) \right\| + \left\| \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \left(\tilde{g}_i \tilde{g}'_j - \mathbb{E}[\tilde{g}_i \tilde{g}'_j] \right) \right\|$$
$$:= \Delta\Omega_1 + \Delta\Omega_2. \tag{4.9.136}$$

Step 1. Consider $\Delta\Omega_1$ and by simple algebra

$$\begin{aligned} \Delta\Omega_1 \leq & \left\| \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \left[\left(\hat{\tilde{g}}_i - \tilde{g}_i \right) \left(\hat{\tilde{g}}'_j - \tilde{g}'_j \right) + \tilde{g}_i \left(\hat{\tilde{g}}'_j - \tilde{g}'_j \right) + \left(\hat{\tilde{g}}_i - \tilde{g}_i \right) \tilde{g}'_j \right] \right\| \\ \leq & \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \left[\left\| \hat{\tilde{g}}_i - \tilde{g}_i \right\| \left\| \hat{\tilde{g}}'_j - \tilde{g}'_j \right\| + \|\tilde{g}_i\| \left\| \hat{\tilde{g}}'_j - \tilde{g}'_j \right\| + \left\| \hat{\tilde{g}}_i - \tilde{g}_i \right\| \|\tilde{g}'_j\| \right] \end{aligned} \tag{4.9.137}$$

Given (4.9.137), it suffices to show $\Delta\Omega_1 = o_p(1)$ by verifying that (a) $\tilde{g}_i$ and $\hat{\tilde{g}}_i$ are bounded, and (b) $\frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \Delta(i,N)} \|\hat{\tilde{g}}_i - \tilde{g}_i\| = o_p(1)$.

Firstly, (a) is satisfied if $|g(w; \theta, \phi) + \delta(w; \theta, \phi)|$ is uniformly bounded over Ω_W and $\Theta \times [0, 1]$. We know that $m^*(x; \theta)$ is continuous differentiable in θ to order three (Assumption 4.5.5) and Θ is compact, implying for $\forall x \in \Omega_X$

$$|m^*(x;\theta)|, \quad \left| \frac{\partial m^*(x;\theta)}{\partial \theta} \right|, \quad \left| \frac{\partial^2 m^*(x;\theta)}{\partial \theta \partial \theta'} \right| \text{ are bounded uniformly over } \Theta. \tag{4.9.138}$$

Furthermore, since $\nu(w; \theta, \gamma)$ is almost everywhere (a.e.) continuously differentiable in w^c (Assumption 4.5.6), it implies (by definition of $\nu(w; \theta, \gamma)$) that $m^*(x; \theta)$ and $\frac{\partial m^*(x;\theta)}{\partial \theta}$ are also continuous in w^c a.e. within the compact Ω_{W^c}. Therefore, for $\forall \theta \in \Theta$,

$$|m^*(x;\theta)|, \quad \left| \frac{\partial m^*(x;\theta)}{\partial \theta} \right|, \quad \left| \frac{\partial^2 m^*(x;\theta)}{\partial \theta \partial \theta'} \right| \text{ are bounded uniformly over } \Omega_X. \tag{4.9.139}$$

Then, (4.9.138) and (4.9.139) together indicate the uniform boundedness of $|m^*(x; \theta)|$ and its first and second derivatives over Ω_X and Θ. Thus,

$$\begin{aligned} \sup_{w \in \Omega_W,\ (\theta,\phi) \in \Theta \times [0,1]} |g(w; \theta, \phi)| = & \sup_{w \in \Omega_W,\ (\theta,\phi) \in \Theta \times [0,1]} \left| \tau(x)(y - m(x; \theta, \phi)) \frac{\partial m^*(x; \theta)}{\partial \theta} \right| \\ \leq & C \sup_{w \in \Omega_W,\ (\theta,\phi) \in \Theta \times [0,1]} \left| \frac{\partial m^*(x; \theta)}{\partial \theta} \right| \leq C_1, \end{aligned}$$

where the first inequality is because the maximum of y and $m(x; \theta, \phi)$ are finite since

Ω_{W^c} is compact, and $\tau(\cdot)$ is bounded (Assumption 4.5.2).

For $\delta(W_i;\theta,\phi) = \nu(W_i;\theta,\phi) - \mathbb{E}[\nu(W_i;\theta,\phi)]$, with $\nu(W_i;\theta,\phi) = \tau(X_i)\frac{\partial \mathcal{R}(W_i;\theta,\phi)}{\partial\theta}\frac{\partial\phi(\mathbf{t};\gamma)}{\partial\gamma'}\mathbf{1}_{d_\gamma}$ and the $d_\theta \times K_T$ vector

$$\frac{\partial \mathcal{R}(W_i;\theta,\phi)}{\partial\theta} = \begin{bmatrix} -\frac{\partial m(X_i;\theta,\phi)}{\partial\theta} m^*(x_{i,1};\theta) + (Y_i - m(X_i;\theta,\phi))\frac{\partial m^*(x_{i,1};\theta)}{\partial\theta} \\ \vdots \\ -\frac{\partial m(X_i;\theta,\phi)}{\partial\theta} m^*(x_{i,K_T};\theta) + (Y_i - m(X_i;\theta,\phi))\frac{\partial m^*(x_{i,K_T};\theta)}{\partial\theta} \end{bmatrix}',$$

it is easy to see that $\delta(W_i;\theta,\phi)$ is a function of $m^*(x;\theta)$, $\frac{\partial m^*(x;\theta)}{\partial\theta}$ and $\frac{\partial\phi(\gamma)}{\partial\gamma}$, and it is linear in ϕ. Moreover, ϕ is the probability function of discrete random variables therefore strictly lies in $[0,1]$. Hence, the above dicussion together with the uniform boundedness of $\frac{\partial\phi(\gamma)}{\partial\gamma}$ provided in the proof of Corollary 1 leads to $\sup_{w\in\Omega_W,\ (\theta,\phi)\in\Theta\times[0,1]} |\delta(w;\theta,\phi)| \leq C_2$ for constant $C_2 > 0$. So far we have established that (a) holds.

Secondly, move on to (b). For θ_N^* between θ^0 and $\hat{\theta}_N$, the triangular inequality and the mean value theorem lead to

$$\begin{aligned}
\left\|\hat{\tilde{g}}_i - \tilde{g}_i\right\| \leq{}& \left\|g(W_i;\hat{\theta}_N,\hat{\phi}_N) - g(W_i;\theta^0,\hat{\phi}_N)\right\| + \left\|g(W_i;\theta^0,\hat{\phi}_N) - g(W_i;\theta^0,\phi^0)\right\| \\
&+ \left\|\delta(W_i;\hat{\theta}_N,\hat{\phi}_N) - \delta(W_i;\theta^0,\hat{\phi}_N)\right\| + \left\|\delta(W_i;\theta^0,\hat{\phi}_N) - \delta(W_i;\theta^0,\phi^0)\right\| \\
\leq{}& \left\|\frac{\partial g(W_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta'}\right\| \left\|\hat{\theta}_N - \theta^0\right\| + \left\|g(W_i;\theta^0,\hat{\phi}_N) - g(W_i;\theta^0,\phi^0)\right\| \\
&+ \left\|\frac{\partial \delta(W_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta'}\right\| \left\|\hat{\theta}_N - \theta^0\right\| + \left\|\delta(W_i;\theta^0,\hat{\phi}_N) - \delta(W_i;\theta^0,\phi^0)\right\|.
\end{aligned} \tag{4.9.140}$$

Start from the first term of (4.9.140), when sample size is large enough (i.e. $\hat{\phi}_N$ is close to ϕ^0),

$$\begin{aligned}
\left\|\frac{\partial g(W_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta'}\right\| = \Bigg\|\tau(X_i)\Bigg[& -\frac{\partial m(X_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta}\frac{\partial m(X_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta'} \\
& + \left[Y_i - m(X_i;\theta_N^*,\hat{\phi}_N)\right]\frac{\partial^2 m(X_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta\partial\theta'}\Bigg]\Bigg\|
\end{aligned}$$

$$\leq C\left(\sum_{j,l=1}^{K_T}\left\|\frac{\partial m^*(x_{i,j};\theta_N^*)}{\partial\theta}\right\|\left\|\frac{\partial m^*(x_{i,l};\theta_N^*)}{\partial\theta'}\right\|+\sum_{j=1}^{K_T}\left\|\frac{\partial^2 m^*(x_{i,j};\theta_N^*)}{\partial\theta\partial\theta'}\right\|\right)$$
$$\leq C_3, \tag{4.9.141}$$

where the last line is because of (4.9.138) and (4.9.139). For the second term of (4.9.140), it yields from the calculation in (4.9.35) that

$$\begin{aligned}\left\|g(W_i;\theta^0,\hat{\phi}_N)-g(W_i;\theta^0,\phi^0)\right\| \leq& \left\|\hat{\phi}_N-\phi^0\right\|_\infty\left[\sum_{j,l=1}^{K_T}\left|m^*(x_{i,j};\theta^0)\right|\left\|\frac{\partial m^*(x_{i,l};\theta^0)}{\partial\theta}\right\|\right.\\ &\left.+\tau_i\left|Y_i-m(X_i;\theta^0,\phi^0)\right|\sum_{j=1}^{K_T}\left\|\frac{\partial m^*(x_{i,j};\theta^0)}{\partial\theta}\right\|\right]+s.o.\\ \leq& C\left\|\hat{\phi}_N-\phi^0\right\|_\infty\sum_{j=1}^{K_T}\left\|\frac{\partial m^*(x_{i,l};\theta^0)}{\partial\theta}\right\|+s.o.\\ \leq& C_4\left\|\hat{\phi}_N-\phi^0\right\|_\infty.\end{aligned} \tag{4.9.142}$$

where the second inequality is because of the compactness of Ω_Y which implies both $m^*(x_{i,j};\theta^0)$ and $|Y_i-m(X_i;\theta^0,\phi^0)|$ are bounded, and the last inequality is due to (4.9.138) and (4.9.139). To bound the third term of (4.9.140), by the dominated convergence theorem, we have

$$\begin{aligned}\frac{\partial\delta(W_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta'} =&\tau(X_i)\frac{\partial}{\partial\theta'}\left(\frac{\partial\mathcal{R}(W_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta}\frac{\partial\phi(\mathbf{t};\hat{\gamma}_N)}{\partial\gamma'}\mathbf{1}_{d_\gamma}\right)\\ &-\mathbb{E}\left[\tau(X_i)\frac{\partial}{\partial\theta'}\left(\frac{\partial\mathcal{R}(W_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta}\frac{\partial\phi(\mathbf{t};\hat{\gamma}_N)}{\partial\gamma'}\mathbf{1}_{d_\gamma}\right)\right].\end{aligned}$$

Based on similar arguments used to obtain (4.9.141) and the uniform boundedness of $\frac{\partial\phi(\gamma)}{\partial\gamma'}$ over $\gamma\in[0,1]$ provided in the proof of Corollary 1, we can get $\left\|\frac{\partial\delta(W_i;\theta_N^*,\hat{\phi}_N)}{\partial\theta'}\right\|\leq C_5$ for some constant $C_5>0$. At last,

$$\begin{aligned}&\left\|\delta(W_i;\theta^0,\hat{\phi}_N)-\delta(W_i;\theta^0,\phi^0)\right\|\\ \leq&|\tau(X_i)|\left\|\frac{\partial\mathcal{R}(W_i;\theta^0,\hat{\phi}_N)}{\partial\theta}\frac{\partial\phi(\mathbf{t};\hat{\gamma}_N)}{\partial\gamma'}-\frac{\partial\mathcal{R}(W_i;\theta^0,\phi^0)}{\partial\theta}\frac{\partial\phi(\mathbf{t};\gamma^0)}{\partial\gamma'}\right\|\|\mathbf{1}_{d_\gamma}\|\\ \leq&C\left(\left\|\frac{\partial\mathcal{R}(W_i;\theta^0,\hat{\phi}_N)}{\partial\theta}-\frac{\partial\mathcal{R}(W_i;\theta^0,\phi^0)}{\partial\theta}\right\|+\left\|\frac{\partial\phi(\mathbf{t};\hat{\gamma}_N)}{\partial\gamma'}-\frac{\partial\phi(\mathbf{t};\gamma^0)}{\partial\gamma'}\right\|\right)\end{aligned}$$

$$\leq C_6 \left\|\hat{\gamma}_N - \gamma^0\right\|_\infty .$$

Given the results above, by Corollary 1, (4.9.137) can be bounded as

$$\Delta\Omega_1 \leq \frac{C}{N}\sum_{i=1}^{N} |\Delta(i,N)| \left(\|\hat{\theta}_N - \theta^0\| + \left\|\hat{\gamma}_N - \gamma^0\right\|_\infty\right) = o_p(1),$$

based on the consistency of $\hat{\theta}_N$ and $\hat{\gamma}_N$, and the fact that $1/N\sum_{i=1}^{N}|\Delta(i,N)| = O(1)$.

Step 2. Next, let us deal with $\Delta\Omega_2$. Based on (4.9.25) and Assumption 4.5.7,

$$\begin{aligned}\mathbb{E}[\|\Delta\Omega_2\|^2] \leq & \frac{d_\theta}{N^2}\left\|\sum_{i,k=1}^{N}\sum_{j\in\Delta(i,N)}\sum_{l\in\Delta(k,N)}\mathbb{E}\Big[\left(\tilde{g}_i\tilde{g}_j' - \mathbb{E}[\tilde{g}_i\tilde{g}_j']\right)'\left(\tilde{g}_k\tilde{g}_l' - \mathbb{E}[\tilde{g}_k\tilde{g}_l']\right)\Big]\right\|_\infty \\ \leq & \frac{d_\theta}{N^2} o\left(\left\|[\Sigma_N^{\tilde{g}}]^2\right\|_\infty\right) = o(1), \end{aligned} \tag{4.9.143}$$

where the last line comes from (4.9.28) that $o(\|[\Sigma_N^{\tilde{g}}]^2\|_\infty/N^2) = o(1)$. Hence, $\left\|\hat{\Omega}_N - \Omega\right\| = o_p(1)$. ■

Chapter 5

Conclusion

This thesis explores issues regarding the identification and estimation of causal treatment effects using the instrumental variable (IV) approaches.

Chapter 2 proposes a weak IV test for discrete outcome models employing a distorted version of the J-statistic, and generalises the notion of rule-of-thumb to accommodate the nonlinear nature of the models. We find that blindly applying the conventional weak IV tests for linear models to the discrete choice models would over reject the truth of weak IVs, which may further lead to misleading causal implications in certain circumstances. These findings demonstrate the importance of the test proposed in this chapter, as it provides the practitioners with a feasible weak IV test when working with discrete choice models.

One caveat of the proposed test should be noted. By construction, the distorted J-test is conservative in the sense that its rejecting rate under the null hypothesis (i.e. the size of the test) is lower than the nominal rate, resulting in a relatively low power for detecting those IVs with moderate strength (i.e. the IVs that are neither weak nor sufficiently strong). Therefore, one possible extension of this chapter is to improve the power of the distorted J-test, by deriving the asymptotic distribution of the test statistic and adopting critical values that make the test less conservative.

Chapter 3 studies an important topic related to IV strength. We use the reduction in the size of the ATE identified set as a measure for the identification power, and conduct the analysis of the identification gains achieved by the ATE bounds. We show that, when the ATE is only partially identified, it is via the extreme values of the conditional treatment propensity score that the instruments exert their influence. It indicates that two instrument sets producing the same propensity score range will actually make identical contributions to the ATE identification gains, regardless of their conventional measures of the IV strength, such as the F-statistics or the pseudo-R^2. In addition, we find that the endogeneity degree plays a key role in determining the IV identification power. Thus, the traditional notion of IV strength, measured by the explanatory power of IVs in the first-stage regression is no longer a suitable measure of the IV identification power in the context considered in this chapter.

We emphasise that the usefulness of the proposed IV identification power (IIP) for selecting irrelevant IVs is preliminary, and rigorous investigation of its finite sample performance and its theoretical asymptotic properties are left for future exploration. The study in this chapter also opens new questions, such as what instruments can be regarded as weak IV in the framework of ATE bounding analyses. Explorations along this direction are also left for future research.

Chapter 4 proposes a nonparametric point identification strategy for the treatment and spillover effects of a randomised intervention, when the network data suffers from measurement errors. The point identification is achieved if there exist two network proxies, one of which acts as an instrument for the latent network. In addition, it also requires that the two proxies contain relevant and distinct information of the true network structure. The identification result relies on two sufficient conditions, the exclusion restriction and the one type of measurement error restriction. Nevertheless, the simulation results have shown that the proposed method outperforms the naive approach ignoring the measurement errors, when either of the sufficient conditions is mildly violated.

There are several extensions worth exploring. They include more flexible estimation methods to mitigate the possibility of model misspecification; accommodating situations where there is imperfect compliance; and exploring the consequences of higher-order

network interference.

Exploration of these potential extensions is left for future research, and hopefully the current thesis will be useful for any future endeavours in related fields.

Bibliography

Advani, A. and B. Malde (2018): "Credibly identifying social effects: Accounting for network formation and measurement error," *Journal of Economic Surveys*, 32, 1016–1044.

Ahn, H. and J. L. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3–29.

Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

Andrew, A. L., K.-W. E. Chu, and P. Lancaster (1993): "Derivatives of eigenvalues and eigenvectors of matrix functions," *SIAM Journal on Matrix Analysis and Applications*, 14, 903–926.

Andrews, D. W. and X. Cheng (2012): "Estimation and inference with weak, semi-strong, and strong identification," *Econometrica*, 80, 2153–2211.

——— (2014): "GMM estimation and uniform subvector inference with possible identification failure," *Econometric Theory*, 30, 287–333.

Andrews, I. (2018): "Valid two-step identification-robust confidence sets for GMM," *Review of Economics and Statistics*, 100, 337–348.

Angelucci, M., G. De Giorgi, M. A. Rangel, and I. Rasul (2010): "Family networks and school enrolment: Evidence from a randomized social experiment," *Journal*

of Public Economics, 94, 197–221.

Angelucci, M. and V. Di Maro (2016): "Programme evaluation and spillover effects," *Journal of Development Effectiveness*, 8, 22–43.

Angrist, J. and W. Evans (1998): "Children and their parents' labor supply: Evidence from exogenous variation in family size," *American Economic Review*, 88, 450–77.

Angrist, J. D. and W. N. Evans (2009): "Replication data for: Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," https://doi.org/10.7910/DVN/4W9GW2, Harvard Dataverse, V1, UNF:3:gmuGDmy3Gcf/k1/lAJqw/A==.

Antoine, B. and E. Renault (2009): "Efficient GMM with nearly-weak instruments," *The Econometrics Journal*, 12.

——— (2012): "Efficient minimum distance estimation with multiple rates of convergence," *Journal of Econometrics*, 170, 350–367.

——— (2020): "Testing identification strength," *Journal of Econometrics*.

Aral, S. and D. Walker (2012): "Identifying influential and susceptible members of social networks," *Science*, 337, 337–341.

Arendt, J. N. (2005): "Does education cause better health? A panel data analysis using school reforms for identification," *Economics of Education Review*, 24, 149–160.

Aronow, P. M. and C. Samii (2017): "Estimating average causal effects under general interference, with application to a social network experiment," *The Annals of Applied Statistics*, 11, 1912–1947.

Athey, S., D. Eckles, and G. W. Imbens (2018): "Exact p-values for network interference," *Journal of the American Statistical Association*, 113, 230–240.

Athey, S. and G. W. Imbens (2017): "The econometrics of randomized experiments," in *Handbook of Economic Field Experiments*, Elsevier, vol. 1, 73–140.

Auerbach, E. (2019): "Identification and estimation of a partially linear regression model using network data," *arXiv preprint arXiv:1903.09679*.

Baird, S., J. A. Bohren, C. McIntosh, and B. Özler (2018): "Optimal design of experiments in the presence of interference," *Review of Economics and Statistics*, 100, 844–860.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The diffusion of microfinance," *Science*, 341, 1236498.

Barrera-Osorio, F., M. Bertrand, L. L. Linden, and F. Perez-Calle (2011): "Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia," *American Economic Journal: Applied Economics*, 3, 167–95.

Basse, G. and A. Feller (2018): "Analyzing two-stage experiments in the presence of interference," *Journal of the American Statistical Association*, 113, 41–55.

Battistin, E., M. De Nadai, and B. Sianesi (2014): "Misreported schooling, multiple measures and returns to educational qualifications," *Journal of Econometrics*, 181, 136–150.

Battistin, E. and B. Sianesi (2011): "Misclassified treatment status and treatment effects: An application to returns to education in the United Kingdom," *Review of Economics and Statistics*, 93, 495–509.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2012): "Treatment effect bounds: An application to Swan–Ganz catheterization," *Journal of Econometrics*, 168, 223–243.

Block, J. H., L. Hoogerheide, and R. Thurik (2013): "Education and entrepreneurial choice: An instrumental variables analysis," *International Small Busi-*

ness Journal, 31, 23–33.

Blundell, R. W. and J. L. Powell (2004): "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, 71, 655–679.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement error in survey data," in *Handbook of Econometrics*, Elsevier, vol. 5, 3705–3843.

Bradley, R. C. et al. (1983): "Approximation theorems for strongly mixing random variables." *The Michigan Mathematical Journal*, 30, 69–81.

Breusch, T., H. Qian, P. Schmidt, and D. Wyhowski (1999): "Redundancy of moment conditions," *Journal of Econometrics*, 91, 89–111.

Breza, E., A. G. Chandrasekhar, T. H. McCormick, and M. Pan (2020): "Using aggregated relational data to feasibly identify network structure without network data," *American Economic Review*, 110, 2454–2484.

Cai, J., A. De Janvry, and E. Sadoulet (2015a): "Replication data for: Social networks and the decision to insure," Nashville, TN: American Economic Association, Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-10-12. https://doi.org/10.3886/E113593V1.

——— (2015b): "Social networks and the decision to insure," *American Economic Journal: Applied Economics*, 7, 81–108.

Calvi, R., A. Lewbel, and D. Tommasi (2018): "Women's empowerment and family health: Estimating LATE with mismeasured treatment," *Available at SSRN 2980250*.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009): "Peer effects and social networks in education," *The Review of Economic Studies*, 76, 1239–1267.

Candelaria, L. E. and T. Ura (2020): "Identification and inference of network formation games with misclassified links," Tech. rep., University of Warwick, Department of Economics.

Caner, M. (2009): "Testing, estimation in GMM and CUE with nearly-weak identification," *Econometric Reviews*, 29, 330–363.

Cawley, J. and C. Meyerhoefer (2012): "The medical care costs of obesity: An instrumental variables approach," *Journal of Health Economics*, 31, 219–230.

Chandrasekhar, A. (2016): "Econometrics of network formation," *The Oxford Handbook of the Economics of Networks*, 303–357.

Chandrasekhar, A. and R. Lewis (2011): "Econometrics of sampled networks," *Working Paper*.

Chandrasekhar, A. G. and M. O. Jackson (2016): "A network formation model based on subgraphs," *Available at SSRN 2660381*.

Chaudhuri, S. and E. Renault (2020): "Score tests in GMM: Why use implied probabilities?" *Journal of Econometrics*.

Chen, L. H., L. Goldstein, and Q.-M. Shao (2010): *Normal approximation by Stein's method*, Springer Science & Business Media.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear models of measurement errors," *Journal of Economic Literature*, 49, 901–37.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models 1," *Econometrica*, 75, 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: Estimation and inference," *Econometrica*, 81, 667–737.

Chesher, A. (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73, 1525–1550.

——— (2010): "Instrumental variable models for discrete outcomes," *Econometrica*, 78, 575–601.

Chiburis, R. C. (2010): "Semiparametric bounds on treatment effects," *Journal of Econometrics*, 159, 267–275.

Chin, A. (2018): "Central limit theorems via Stein's method for randomized experiments under interference," *arXiv preprint arXiv:1804.03105*.

Comola, M. and M. Fafchamps (2017): "The missing transfers: Estimating misreporting in dyadic data," *Economic Development and Cultural Change*, 65, 549–582.

Conley, T. G. and C. R. Udry (2010): "Learning about a new technology: Pineapple in Ghana," *American Economic Review*, 100, 35–69.

Currarini, S., M. O. Jackson, and P. Pin (2009): "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 77, 1003–1045.

——— (2010): "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences*, 107, 4857–4861.

De Paula, Á., S. Richards-Shubik, and E. Tamer (2018): "Identifying preferences in networks with bounded degree," *Econometrica*, 86, 263–288.

Dufour, J.-M. and J. Wilde (2018): "Weak identification in probit models with endogenous covariates," *AStA Advances in Statistical Analysis*, 102, 611–631.

Dupas, P. (2014): "Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment," *Econometrica*, 82, 197–228.

Erdös, P. and A. Rényi (1959): "On random graphs," *Mathematicae Debrecen*, 6, 290–297.

Fafchamps, M. and S. Lund (2003): "Risk-sharing networks in rural Philippines," *Journal of Development Economics*, 71, 261–287.

Finlay, K. and L. M. Magnusson (2009): "Implementing weak-instrument robust

tests for a general class of instrumental-variables models," *The Stata Journal*, 9, 398–421.

Flores, C. A. and X. Chen (2018): *Average treatment effect bounds with an instrumental variable: Theory and practice*, Springer.

Frazier, D. T., E. Renault, L. Zhang, and X. Zhao (June 28, 2019): "Weak instruments test in discrete choice models," Paper presented at the 2019 North American Summer Meeting of the Econometric Society, Seattle, Washington.

Freedman, D. A. and J. S. Sekhon (2010): "Endogeneity in probit response models," *Political Analysis*, 18, 138–150.

Gao, W. and S. Li (2019): "Identification and estimation of peer effects in latent networks," *Working Paper*.

Goldsmith-Pinkham, P. and G. W. Imbens (2013): "Social networks and the identification of peer effects," *Journal of Business & Economic Statistics*, 31, 253–264.

Goldstein, L. and Y. Rinott (1996): "Multivariate normal approximations by Stein's method and size bias couplings," *Journal of Applied Probability*, 33, 1–17.

Goto, U. and T. Iizuka (2016): "Cartel sustainability in retail markets: Evidence from a health service sector," *International Journal of Industrial Organization*, 49, 36–58.

Graham, B. S. (2017): "An econometric model of network formation with degree heterogeneity," *Econometrica*, 85, 1033–1063.

Hahn, J. and G. Kuersteiner (2002): "Discontinuities of weak instrument limiting distributions," *Economics Letters*, 75, 325–331.

Hall, A. R. (2005): *Generalized method of moments*, Oxford University Press.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007): "Information in general-

ized method of moments estimation and entropy-based moment selection," *Journal of Econometrics*, 138, 488–512.

Hall, A. R. and F. P. Peixe (2003): "A consistent method for the selection of relevant instruments," *Econometric Reviews*, 22, 269–287.

Han, S. and S. Lee (2019): "Estimation in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Applied Econometrics*, 34, 994–1015.

Han, S. and E. J. Vytlacil (2017): "Identification in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Econometrics*, 199, 63–73.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hansen, L. P., J. Heaton, and E. G. Luttmer (1995): "Econometric evaluation of asset pricing models," *The Review of Financial Studies*, 8, 237–274.

Hardy, M., R. M. Heath, W. Lee, and T. H. McCormick (2019): "Estimating spillovers using imprecisely measured networks," *arXiv preprint arXiv:1904.00136*.

Hausman, J. A. (1978): "Specification tests in econometrics," *Econometrica: Journal of the Econometric Society*, 1251–1271.

He, X. and K. Song (2018): "Measuring diffusion over a large network," *arXiv preprint arXiv:1812.04195*.

Heckman, J. (1990): "Varieties of selection bias," *The American Economic Review*, 80, 313–318.

Heckman, J. J. (1978): "Dummy endogenous variables in a simultaneous equation system," *Econometrica*, 46, 931–959.

Heckman, J. J. and R. Robb (1985): "Alternative methods for evaluating the impact

of interventions: An overview," *Journal of Econometrics*, 30, 239–267.

——— (1986): "Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes," in *Drawing Inferences from Self-selected Samples*, Springer, 63–107.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding instrumental variables in models with essential heterogeneity," *The Review of Economics and Statistics*, 88, 389–432.

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation 1," *Econometrica*, 73, 669–738.

Heckman, J. J. and E. J. Vytlacil (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

——— (2001): "Instrumental variables, selection models, and tight bounds on the average treatment effect," in *Econometric Evaluation of Labour Market Policies*, Springer, 1–15.

Hu, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 144, 27–61.

Hu, Y. and S. M. Schennach (2008): "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 76, 195–216.

Hudgens, M. G. and M. E. Halloran (2008): "Toward causal inference with interference," *Journal of the American Statistical Association*, 103, 832–842.

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Imbens, G. W. and C. F. Manski (2004): "Confidence intervals for partially identified

parameters," *Econometrica*, 72, 1845–1857.

Imbens, G. W. and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512.

Joe, H. (1997): *Multivariate models and multivariate dependence concepts*, Chapman and Hall/CRC.

Johnsson, I. and H. R. Moon (2015): "Estimation of peer effects in endogenous social networks: Control function approach," *Review of Economics and Statistics*, 1–51.

Kawaguchi, D., Y. Matsushita, and H. Naito (2017): "Moment estimation of the probit model with an endogenous continuous regressor," *The Japanese Economic Review*, 68, 48–62.

Kinda, T. (2010): "Investment climate and FDI in developing countries: firm-level evidence," *World Development*, 38, 498–513.

Kitagawa, T. (2009): "Identification region of the potential outcome distributions under instrument independence," Tech. rep., CEMMAP Working Paper.

——— (2015): "A test for instrument validity," *Econometrica*, 83, 2043–2063.

Kleibergen, F. (2005): "Testing parameters in GMM without assuming that they are identified," *Econometrica*, 73, 1103–1123.

Kleibergen, F. and R. Paap (2006): "Generalized reduced rank tests using the singular value decomposition," *Journal of Econometrics*, 133, 97–126.

Kojevnikov, D., V. Marmer, and K. Song (2019): "Limit theorems for network dependent random variables," *arXiv preprint arXiv:1903.01059*.

Kossinets, G. (2006): "Effects of missing data in social networks," *Social Networks*, 28, 247–268.

Kremer, M. and E. Miguel (2007): "The illusion of sustainability," *The Quarterly*

Journal of Economics, 122, 1007–1065.

Kuersteiner, G. M. (2019): "Limit theorems for data with network structure," *arXiv preprint arXiv:1908.02375*.

Lee, Y. and E. L. Ogburn (2020): "Network dependence can lead to spurious associations and invalid inference," *Journal of the American Statistical Association*, 1–31.

Leung, M. P. (2019a): "Causal inference under approximate neighborhood interference," *Available at SSRN 3479902*.

——— (2019b): "A weak law for moments of pairwise stable networks," *Journal of Econometrics*, 210, 310–326.

——— (2020a): "Dependence-robust inference using resampled statistics," *arXiv preprint arXiv:2002.02097*.

——— (2020b): "Treatment and spillover effects under network interference," *Review of Economics and Statistics*, 102, 368–380.

Leung, M. P. and H. R. Moon (2019): "Normal approximation in large network models," *arXiv preprint arXiv:1904.11060*.

Lewbel, A. (2007): "Estimation of average treatment effects with misclassification," *Econometrica*, 75, 537–551.

——— (2019): "The identification zoo: Meanings of identification in econometrics," *Journal of Economic Literature*, 57, 835–903.

Lewbel, A., X. Qu, and X. Tang (2019): "Social networks with misclassified or unobserved links," Tech. rep., Boston College Department of Economics.

Li, C., D. S. Poskitt, and X. Zhao (2018): "Bounds for average treatment effect: A comparison of nonparametric and quasi maximum likelihood estimators," Tech. rep., Working Paper, Monash University.

——— (2019): "The bivariate probit model, maximum likelihood estimation, pseudo true parameters and partial identification," *Journal of Econometrics*, 209, 94–113.

Li, Q. and J. S. Racine (2007): *Nonparametric econometrics: Theory and practice*, Princeton University Press.

Liu, L. and M. G. Hudgens (2014): "Large sample randomization inference of causal effects in the presence of interference," *Journal of the American Statistical Association*, 109, 288–301.

Liu, X. (2013): "Estimation of a local-aggregate network model with sampled networks," *Economics Letters*, 118, 243–246.

Lochner, L. and E. Moretti (2004): "The effect of education on crime: Evidence from prison inmates, arrests, and self-reports," *American Economic Review*, 94, 155–189.

Ma, Y., Y. Wang, and V. Tresp (2020): "Causal inference under networked interference," *arXiv preprint arXiv:2002.08506*.

Machado, C., A. Shaikh, and E. Vytlacil (2013): "Instrumental variables and the sign of the average treatment effect," *Unpublished Manuscript, Getúlio Vargas Foundation, University of Chicago, and New York University.[2049]*.

Maddala, G. S. (1986): *Limited-dependent and qualitative variables in econometrics*, 3, Cambridge University Press.

Magnusson, L. M. (2007): "Weak instruments robust tests for limited dependent variable models," Tech. rep., Working Paper, Brown University (RI).

——— (2010): "Inference in limited dependent variable models robust to weak identification," *The Econometrics Journal*, 13.

Mahajan, A. (2006): "Identification and estimation of regression models with misclassification," *Econometrica*, 74, 631–665.

Manski, C. F. (1990): "Nonparametric bounds on treatment effects," *The American Economic Review*, 80, 319–323.

——— (2001): "Designing programs for heterogeneous populations: The value of covariate information," *American Economic Review*, 91, 103–106.

——— (2013): "Identification of treatment response with social interactions," *The Econometrics Journal*, 16, S1–S23.

Manski, C. F. and E. Tamer (2002): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70, 519–546.

Marra, G. and R. Radice (2011): "Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity," *Canadian Journal of Statistics*, 39, 259–279.

Masry, E. (1996): "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *Journal of Time Series Analysis*, 17, 571–599.

McKenzie, D. and H. Rapoport (2011): "Can migration reduce educational attainment? Evidence from Mexico," *Journal of Population Economics*, 24, 1331–1358.

Meckes, E. et al. (2009): "On Stein's method for multivariate normal approximation," in *High dimensional probability V: the Luminy volume*, Institute of Mathematical Statistics, 153–178.

Miguel, E. and M. Kremer (2004): "Worms: identifying impacts on education and health in the presence of treatment externalities," *Econometrica*, 72, 159–217.

Miguel, E., S. Satyanath, and E. Sergenti (2004): "Economic shocks and civil conflict: An instrumental variables approach," *Journal of Political Economy*, 112, 725–753.

Montiel Olea, J. L. and C. Pflueger (2013): "A robust test for weak instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

Mourifié, I. and R. Méango (2014): "A note on the identification in two equations probit model with dummy endogenous regressor," *Economics Letters*, 125, 360–363.

Mroz, T. A. (1987): "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," *Econometrica: Journal of the Econometric Society*, 765–799.

Newey, W. K. (1994): "Kernel estimation of partial means and a general variance estimator," *Econometric Theory*, 10, 1–21.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67, 565–603.

Newey, W. K. and F. Windmeijer (2009): "Generalized method of moments with many weak moment conditions," *Econometrica*, 77, 687–719.

Nunn, N. and N. Qian (2014): "US food aid and civil conflict," *American Economic Review*, 104, 1630–66.

Opper, I. M. (2019): "Does helping john help sue? Evidence of spillovers in education," *American Economic Review*, 109, 1080–1115.

Oster, E. and R. Thornton (2012): "Determinants of technology adoption: Peer effects in menstrual cup take-up," *Journal of the European Economic Association*, 10, 1263–1293.

Patacchini, E., E. Rainone, and Y. Zenou (2017): "Heterogeneous peer effects in education," *Journal of Economic Behavior & Organization*, 134, 190–227.

Pflueger, C. E. and S. Wang (2015): "A robust test for weak instruments in Stata," *The Stata Journal*, 15, 216–225.

Powell, L. M., J. A. Tauras, and H. Ross (2005): "The importance of peer effects, cigarette prices and tobacco control policies for youth smoking behavior," *Journal of Health Economics*, 24, 950–968.

Qu, X. and L.-f. Lee (2015): "Estimating a spatial autoregressive model with an endogenous spatial weight matrix," *Journal of Econometrics*, 184, 209–232.

Rivers, D. and Q. H. Vuong (1988): "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, 39, 347–366.

Romano, J. P. and A. M. Shaikh (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," *The Annals of Statistics*, 40, 2798–2822.

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Ross, N. (2011): "Fundamentals of Stein's method," *Probability Surveys*, 8, 210–293.

Rubin, D. B. (1990): "Formal mode of statistical inference for causal effects," *Journal of Statistical Planning and Inference*, 25, 279–292.

Ruseski, J. E., B. R. Humphreys, K. Hallman, P. Wicker, and C. Breuer (2014): "Sport participation and subjective well-being: Instrumental variable results from German survey data," *Journal of Physical Activity and Health*, 11, 396–403.

Sävje, F. (2019): "Causal inference with misspecified exposure mappings," Tech. rep., Yale University.

Sävje, F., P. M. Aronow, and M. G. Hudgens (2017): "Average treatment effects in the presence of unknown interference," *arXiv preprint arXiv:1711.06399*.

Shaikh, A. and E. Vytlacil (2005): "Threshold crossing models and bounds on treatment effects: A nonparametric analysis," Tech. rep., National Bureau of Economic Research.

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial identification in triangular systems of equations with binary dependent variables," *Econometrica*, 79, 949–955.

Sobel, M. E. (2006): "What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference," *Journal of the American Statistical Association*, 101, 1398–1407.

Song, K. (2018): "Measuring the graph concordance of locally dependent observations," *Review of Economics and Statistics*, 100, 535–549.

Staiger, D. and J. H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557–586.

Stein, C. (1986): "Approximate computation of expectations," *Lecture Notes-Monograph Series*, 7, i–164.

Stock, J. and D. Andrews (2005): "Inference with weak instruments," *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol III, Cambridge University Press, url http://www.economics.harvard.edu/faculty/stock/files/worldcongresspaper9.pdf.*

Stock, J. H. and J. H. Wright (2000): "GMM with weak identification," *Econometrica*, 68, 1055–1096.

Stock, J. H. and M. Yogo (2005): "Testing for weak instruments in linear IV regression. Chapter 5 in Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg, edited by DWK Andrews and JH Stock," .

Swanson, S. A., M. A. Hernán, M. Miller, J. M. Robins, and T. S. Richardson (2018): "Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes," *Journal of the American Statistical Association*, 113, 933–947.

Tamer, E. (2010): "Partial identification in econometrics," *Annual Review of Economics*, 2, 167–195.

Tauchen, G. (1985): "Diagnostic testing and evaluation of maximum likelihood models," *Journal of Econometrics*, 30, 415–443.

Tchetgen, E. J. T. and T. J. VanderWeele (2012): "On causal inference in the presence of interference," *Statistical Methods in Medical Research*, 21, 55–75.

Terza, J. V., A. Basu, and P. J. Rathouz (2008): "Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling," *Journal of Health Economics*, 27, 531–543.

Thirkettle, M. (2019): "Identification and estimation of network statistics with missing link data," Tech. rep., Working Paper.

van der Laan, M. J. (2014): "Causal inference for a population of causally connected units," *Journal of Causal Inference*, 2, 13–74.

van der Vaart, A. W. and J. A. Wellner (1996): "Weak convergence," in *Weak Convergence and Empirical Processes*, Springer, 16–28.

Vazquez-Bare, G. (2019): "Identification and estimation of spillover effects in randomized experiments," *arXiv preprint arXiv:1711.02745*.

——— (2020): "Causal spillover effects using instrumental variables," *arXiv preprint arXiv:2003.06023*.

Veall, M. R. and K. F. Zimmermann (1992): "Pseudo-R^2's in the ordinal probit model," *Journal of Mathematical Sociology*, 16, 333–342.

——— (1996): "Pseudo-R^2 measures for some common limited dependent variable models," *Journal of Economic Surveys*, 10, 241–259.

Viviano, D. (2019): "Policy targeting under network interference," *arXiv preprint arXiv:1906.10258*.

Vuong, Q. and H. Xu (2017): "Counterfactual mapping and individual treatment

effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 8, 589–610.

Vytlacil, E. and N. Yildiz (2007): "Dummy endogenous variables in weakly separable models," *Econometrica*, 75, 757–779.

White, H. and I. Domowitz (1984): "Nonlinear regression with dependent observations," *Econometrica: Journal of the Econometric Society*, 143–161.

Wilde, J. (2000): "Identification of multiple equation probit models with endogenous dummy regressors," *Economics Letters*, 69, 309–312.

Windmeijer, F. (2019): "Two-stage least squares as minimum distance," *The Econometrics Journal*, 22, 1–9.

Windmeijer, F. A. (1995): "Goodness-of-fit measures in binary choice models," *Econometric Reviews*, 14, 101–116.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT Press.

——— (2014): "Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables," *Journal of Econometrics*, 182, 226–234.

——— (2015): "Control function methods in applied econometrics," *Journal of Human Resources*, 50, 420–445.